Practical Approaches To Construct RS-HDMR Component Functions

Genyuan Li,[†] Sheng-Wei Wang,[‡] and Herschel Rabitz*,[†]

Department of Chemistry, Princeton University, Princeton, New Jersey 08544, and Environmental and Occupational Health Sciences Institute, 170 Frelinghuysen Road, Piscataway, New Jersey 08854

Received: December 18, 2001; In Final Form: June 4, 2002

A general set of quantitative model assessment and analysis tools, termed high-dimensional model representations (HDMR), has been introduced recently for improving the efficiency of deducing high-dimensional input−output system behavior. HDMR is a particular family of representations where each term in the representation reflects the independent and cooperative contributions of the inputs upon the output. When data are randomly sampled, a RS (random sampling)-HDMR can be constructed. To reduce the sampling effort, different analytical basis functions, such as orthonormal polynomials, cubic B splines, and polynomials may be employed to approximate the RS-HDMR component functions. Only one set of random input−output samples is necessary to determine all the RS-HDMR component functions, and a few hundred samples may give a satisfactory approximation, regardless of the dimension of the input variable space. It is shown in an example that judicious use of orthonormal polynomials can provide a sampling saving of $\sim 10^3$ in representing a system compared to employing a direct sampling technique. This paper discusses these practical approaches: their formulas and accuracy along with an illustration from atmospheric modeling.

1. Introduction

Many problems in science and engineering reduce to efficiently constructing a map of the relationship between high-dimensional system input and output variables. The system may be described by a mathematical model (e.g., typically a set of differential equations), where the input variables might be specified initial and/or boundary conditions, parameters, or functions residing in the model, and the output variable(s) would be the solution to the model or a functional of it. The input−output (IO) behavior may also be based on observations in the laboratory or field where a mathematical model cannot readily be constructed for the system. In this case the IO system is simply considered as a black box where the input consists of the measured laboratory or field (control) variables and the output(s) is the observed system response. Regardless of the circumstances, the input is often very high dimensional with many variables even if the output is only a single quantity. We refer to the input variables collectively as $\mathbf{x} = (x_1, x_2, ..., x_n)$, with n ranging up to $\sim 10^2$ to 10^3 or more, and the output as $f(\mathbf{x})$. For simplicity in the remainder of the paper and without loss of generality, we shall refer to the system as a model regardless of whether it involves modeling, laboratory experiments, or field studies.

High-dimensional model representation (HDMR) is a general set of quantitative model assessment and analysis tools for capturing high-dimensional IO system behavior.[1−5] As the impact of the multiple input variables on the output can be independent and cooperative, HDMR expresses the model output $f(\mathbf{x})$ as a finite hierarchical correlated function expansion in terms of the input variables:

$$f(\mathrm{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\le i<j\le n} f_{ij}(x_i,x_j) + ... + \sum_{1\le i_1<...<i_l\le n} f_{i_1 i_2 ... i_l}(x_{i_1},x_{i_2},...,x_{i_l}) + ... + f_{12...n}(x_1,x_2,...,x_n) \quad (1)$$

where the zeroth-order (i.e., $l = 0$) component function f_0 is a constant representing the mean response to $f(\mathbf{x})$, and the first-order (i.e., $l = 1$) component function $f_i(x_i)$ gives the independent contribution to $f(\mathbf{x})$ by the ith input variable acting alone, the second-order (i.e., $l = 2$) component function $f_{ij}(x_i,x_j)$ gives the pair correlated contribution to $f(\mathbf{x})$ by the input variables x_i and x_j, etc. The last term $f_{12...n}(x_1,x_2,...,x_n)$ contains any residual nth-order correlated contribution of all input variables.

A critical feature of the HDMR expansion is that its component functions are optimal choices tailored to a given $f(\mathbf{x})$ over the entire desired domain Ω of $\mathbf{x}$. Experience shows that the high-order terms in the expansion often are negligible[3] such that an HDMR expansion to second order

$$f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\le i<j\le n} f_{ij}(x_i,x_j) \quad (2)$$

can provide a satisfactory description of $f(\mathbf{x})$ for many high-dimensional systems when the input variables are properly chosen.

Distinct, but formally equivalent, HDMR expansions, all of the same structure as eq 1, may be constructed. When the input data are randomly sampled, RS (random sampling)-HDMR component functions can be constructed. For RS-HDMR, we first rescale the variables x_i by some suitable transformations such that $0 \le x_i \le 1$ for all i. The output function $f(\mathbf{x})$ is then defined in the unit hypercube $K^n = \{(x_1, x_2, ..., x_n) | \ 0 \le x_i \le 1,$

[†] Princeton University.
[‡] Environmental and Occupational Health Sciences Institute.

10.1021/jp014567t CCC: $22.00
Published on Web 08/24/2002

$i = 1, 2, ..., n\}$. The component functions of RS-HDMR possess the following forms:[3]

$$f_0 = \int_{K^n} f(\mathbf{x})\, d\mathbf{x} \tag{3}$$

$$f_i(x_i) = \int_{K^{n-1}} f(\mathbf{x})\, d\mathbf{x}^i - f_0 \tag{4}$$

$$f_{ij}(x_i,x_j) = \int_{K^{n-2}} f(\mathbf{x})\, d\mathbf{x}^{ij} - f_i(x_i) - f_j(x_j) - f_0 \tag{5}$$

...

where $d\mathbf{x}^i$ and $d\mathbf{x}^{ij}$ are just the product $dx_1dx_2\cdots dx_n$ without dx_i and dx_idx_j, respectively. Finally, the last term $f_{12...n}(x_1,x_2,...,x_n)$ is determined from the difference between $f(\mathbf{x})$ and all the other component functions in eq 1. The RS-HDMR component functions satisfy the following condition: the integral of a component function of RS-HDMR with respect to any of its own variables is zero, i.e.,

$$\int_0^1 f_{i_1i_2...i_l}(x_{i_1},x_{i_2},...,x_{i_l})\, dx_s = 0 \qquad s \in \{i_1, i_2, ..., i_l\} \tag{6}$$

which defines the orthogonality relation between two RS-HDMR component functions as

$$\int_{K^n} f_{i_1i_2...i_l}(x_{i_1},x_{i_2},...,x_{i_l}) f_{j_1j_2...j_k}(x_{j_1},x_{j_2},...,x_{j_k})\, d\mathbf{x} = 0 \tag{7}$$
$$\{i_1, i_2, ..., i_l\} \neq \{j_1, j_2, ..., j_k\}$$

The component functions $f_i(x_i)$, $f_{ij}(x_i,x_j)$, ... are typically provided numerically, at discrete values of the input variables x_i, x_j, ... produced from sampling the output function $f(\mathbf{x})$ for employment on the right-hand side (rhs) of eqs 3−5. Thus, numerical data tables can be constructed for these component functions, and the approximate value of $f(\mathbf{x})$ for an arbitrary point $\mathbf{x}$ can be determined from these tables by performing only low-dimensional interpolation over $f_i(x_i)$, $f_{ij}(x_i,x_j)$, ...

To construct the numerical data tables for the RS-HDMR component functions, one needs to evaluate the above integrals. Evaluation of the high-dimensional integrals in the RS-HDMR expansion may be carried out by Monte Carlo random sampling.[6] For instance, N samples of the n-dimensional vector $\mathbf{x}^{(s)} = (x_1^{(s)}, x_2^{(s)}, ..., x_n^{(s)})$ $(s = 1, 2, ..., N)$ are randomly generated uniformly in K^n, and then f_0 is approximated by the average value of $f(\mathbf{x})$ at all $\mathbf{x}^{(s)}$:

$$f_0 = \int_{K^n} f(\mathbf{x})\, d\mathbf{x} \approx \frac{1}{N}\sum_{s=1}^{N} f(\mathbf{x}^{(s)}) \tag{8}$$

When $N \to \infty$, an accurate value of f_0 can be obtained. Very often the integrals converge quite fast, and a modest value of N may give a very good result. Moreover, the approximation of an integral by Monte Carlo sampling often does not depend significantly on the dimension n. This property is extremely beneficial for high-dimension systems.

The direct determination of all RS-HDMR component functions at different values of x_i, x_j, ... by Monte Carlo integration requires a large number of random samples.[7] For example, to determine $f_i(x_i)$, different sets of random samples of $f(x_i,\mathbf{x}^i)$ at $(x_i,\mathbf{x}^i)^{(s)} = (x_1^{(s)}, x_2^{(s)}, ..., x_{i-1}^{(s)}, x_i, x_{i+1}^{(s)}, ..., x_n^{(s)})$ with distinct fixed values of x_i are needed, i.e.,

$$f_i(x_i) = \int_{K^{n-1}} f(\mathbf{x})\, d\mathbf{x}^i - f_0 \approx \frac{1}{N}\sum_{s=1}^{N} f((x_i,\mathbf{x}^i)^{(s)}) - \frac{1}{N}\sum_{s=1}^{N} f(\mathbf{x}^{(s)}) \tag{9}$$

If the table mesh for x_i takes m distinct values, then mN random samples are necessary to construct the $f_i(x_i)$ numerical table.

Similarly, to construct the $f_{ij}(x_i,x_j)$ numerical table, different sets of random samples of $f(x_i,x_j,\mathbf{x}^{ij})$ at $(x_i, x_j, \mathbf{x}^{ij})^{(s)} = (x_1^{(s)}, x_2^{(s)}, ..., x_{i-1}^{(s)}, x_i, x_{i+1}^{(s)}, ..., x_{j-1}^{(s)}, x_j, x_{j+1}^{(s)}, ..., x_n^{(s)})$ with distinct fixed values of (x_i, x_j) are needed, i.e.,

$$f_{ij}(x_i, x_j) = \int_{K^{n-2}} f(\mathbf{x})\, d\mathbf{x}^{ij} - f_i(x_i) - f_j(x_j) - f_0 \approx \frac{1}{N}\sum_{s=1}^{N} f((x_i,x_j,\mathbf{x}^{ij})^{(s)}) - \frac{1}{N}\sum_{s=1}^{N} f((x_i,\mathbf{x}^i)^{(s)}) - \frac{1}{N}\sum_{s=1}^{N} f((x_j,\mathbf{x}^j)^{(s)}) + \frac{1}{N}\sum_{s=1}^{N} f(\mathbf{x}^{(s)}) \tag{10}$$

If the table meshes for both x_i and x_j take m distinct values, then m^2N random samples are necessary to construct the $f_{ij}(x_i,x_j)$ table. The required number of random samples increases exponentially with the order of the required RS-HDMR component functions. Thus, the direct approach is prohibitively expensive for the construction of high-order RS-HDMR component function numerical tables.

To reduce the sampling effort, the RS-HDMR component functions may be approximated by expansions in terms of a suitable set of functions, such as orthonormal polynomials, spline functions, or even simple polynomial functions:[5]

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) \tag{11}$$

$$f_{ij}(x_i,x_j) \approx \sum_{p=1}^{l}\sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i,x_j) \tag{12}$$

...

where k, l, l' are integers, α_r^i, β_{pq}^{ij} are constant coefficients to be determined, and $\varphi_r(x_i)$, $\varphi_{pq}(x_i, x_j)$ are one- and two-variable basis functions. With these formulas, eq 1 can be expressed as

$$f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) + \sum_{1\le i<j\le n}\sum_{p=1}^{l}\sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i,x_j) + ... \tag{13}$$

Each coefficient $\xi \in \{\alpha_r^i, \beta_{pq}^{ij}, ...\}$ may be determined by minimization of the functional

$$\min_{\xi\in\{\alpha_r^i,\beta_{pq}^{ij},...\}} \int_{K^n} [f(\mathbf{x}) - f_0 - \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) - \sum_{1\le i<j\le n}\sum_{p=1}^{l}\sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i,x_j) - ...]^2\, d\mathbf{x} \tag{14}$$

When the basis functions with different numbers of variables are orthogonal, i.e.,

$$\int_{K^n} \varphi_{r_1r_2...r_p}(x_{i_1},x_{i_2},...,x_{i_p}) \varphi_{s_1s_2...s_q}(x_{j_1},x_{j_2},...,x_{j_q})\, d\mathbf{x} = 0 \tag{15}$$
$$(p \neq q)$$

the approximations for the RS-HDMR component functions given by eqs 11 and 12 will preserve the mutual orthogonality in eq 7, and eq 14 is equivalent to

$$\min_{\alpha_r^i} \int_0^1 [f_i(x_i) - \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i)]^2 \, dx_i \tag{16}$$

$$\min_{\beta_{pq}^{ij}} \int_0^1 \int_0^1 [f_{ij}(x_i,x_j) - \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i,x_j)]^2 \, dx_i \, dx_j \tag{17}$$

...

Then each set of coefficients for the basis functions with the same variables can be obtained by solving a linear equation

$$\mathbf{Ay} = \mathbf{b} \tag{18}$$

where **A** is a constant nonsingular matrix, **b** is a vector whose elements are integrals over a product of $f(\mathbf{x})$ times the basis functions, and **y** is the vector of coefficients for the basis functions associated with the same variables. For example, consider α_r^i. In this case, the (r, r')-entry of **A** is

$$A_{rr'} = \int_0^1 \varphi_r(x_i)\varphi_{r'}(x_i)dx_i \qquad r, r' = 1, 2, ..., k \tag{19}$$

and

$$\mathbf{y} = (\alpha_1^i \alpha_2^i ... \alpha_k^i)^T \tag{20}$$

$$\mathbf{b} = \begin{pmatrix} \int_0^1 f_i(x_i)\, \varphi_1(x_i)\, dx_i \\ \int_0^1 f_i(x_i)\, \varphi_2(x_i)\, dx_i \\ \vdots \\ \int_0^1 f_i(x_i)\, \varphi_k(x_i)\, dx_i \end{pmatrix} \tag{21}$$

Substituting eq 9 into eq 21 yields

$$\mathbf{b} = \begin{pmatrix} \int_{K^n} f(\mathbf{x})\, \varphi_1(x_i)\, d\mathbf{x} \\ \int_{K^n} f(\mathbf{x})\, \varphi_2(x_i)\, d\mathbf{x} \\ \vdots \\ \int_{K^n} f(\mathbf{x})\, \varphi_k(x_i)\, d\mathbf{x} \end{pmatrix} \approx \frac{1}{N} \sum_{s=1}^{N} \begin{pmatrix} f(\mathbf{x}^{(s)})\, \varphi_1(x_i^{(s)}) \\ f(\mathbf{x}^{(s)})\, \varphi_2(x_i^{(s)}) \\ \vdots \\ f(\mathbf{x}^{(s)})\, \varphi_k(x_i^{(s)}) \end{pmatrix} \tag{22}$$

As no restriction is imposed on the values of the elements of **x** for $f(\mathbf{x})$ in the above integrals, only one set of random samples for $f(\mathbf{x})$ is necessary to determine the elements of **b** by Monte Carlo integration. All the coefficients α_r^i are given by $\mathbf{A}^{-1}\mathbf{b}$, and then $f_i(x_i)$ is obtained. The linear equation for β_{pq}^{ij} can be constructed similarly, and $f_{ij}(x_i,x_j)$ will be obtained from the same set of random samples. The sampling effort is then dramatically reduced. This paper will discuss different analytical basis functions, including orthonormal polynomials, cubic B splines, and polynomials for approximating the RS-HDMR component functions.

The paper is organized as follows. In section 2 the direct determination of the RS-HDMR component functions by Monte Carlo integration is presented and the results will be used in a comparison with the following analytical basis function approximations. Sections 3−5 respectively present the approximations with orthonormal polynomials, cubic B spline functions and polynomial functions. Finally, section 6 contains conclusions and a discussion.

2. Direct Determination by Monte Carlo Integration

The direct determination of RS-HDMR component functions f_0, $f_i(x_i)$, ... by Monte Carlo integration at discrete values of x_i, x_j, ... are performed by eqs 8−10 and other similar formulas for higher order component functions. Because the error of Monte Carlo integration decreases as $\sim 1/\sqrt{N}$, the accuracy of the resultant RS-HDMR component functions depends on the sample size N.[6] Therefore, for a given application, we need first to find the sample size that will give the desired accuracy.

TABLE 1: Ranges of Input Variables

input	lower bound	upper bound
relative humidity, x_1 (%)	5	100
CO, x_2 (ppb)	10	200
NO_x, x_3 (ppt)	50	950
O_3, x_4 (ppb)	10	150

TABLE 2: Constant f_0 for P and D Obtained from Different Sample Sizes

sample size (N)	P (ppb/day)	D (ppb/day)
1000	18.4	24.7
3000	18.4	24.7
5000	18.4	24.8

For illustration consider the following example: a zero-dimensional photochemical box model designed to treat the ozone chemistry in the background troposphere for the study of 3-dimensional global chemical-transport.[8] This box model consists of 63 reactions and 28 chemical species. Using this box model the rates of ozone production P and destruction D are calculated and incorporated into the overall 3-dimensional model. The details of this process are not relevant here, but the box model provides a good testing ground for the construction of RS-HDMR component functions. The rates of ozone production P and destruction D are chosen as two output variables controlled by the four independent input variables $\mathbf{x} = (x_1, x_2, x_3, x_4)$ corresponding to the concentrations of four precursors: H_2O, CO, NO_x, and O_3. The ranges of the four inputs are shown in Table 1.

The data were generated for a set of 5000 random samples of $\mathbf{x}^{(s)}$, as well as for given distinct values x_i or (x_i, x_j) also with 5000 random samples for $(x_i, \mathbf{x}^i)^{(s)}$ or $(x_i, x_j, \mathbf{x}^{ij})^{(s)}$ $(s = 1-5000)$. These samples were chosen within the ranges in Table 1 generated by quasi-random sampling.[6] The corresponding outputs P and D at $\mathbf{x}^{(s)}$, $(x_i,\mathbf{x}^i)^{(s)}$ or $(x_i, x_j,\mathbf{x}^{ij})^{(s)}$ are obtained by solving the differential equations of the box model. Using these data, the RS-HDMR component functions up to second-order f_0, $f_i(x_i)$, $f_{ij}(x_i,x_j)$ for the outputs P and D are constructed by eqs 8−10 at different sample sizes (1000−5000). The results for f_0 are given in Table 2. Some results for $f_3(x_3)$ and $f_{13}(x_1,x_3)$ are shown in Figures 1 and 2. Table 2 and Figures 1 and 2 show that the resultant f_0, $f_3(x_3)$ and $f_{13}(x_1,x_3)$ coincide very well for different sample sizes. This implies that a data set of 1000 samples already gives a convergent result. The results for other component functions are similar.

To attain a quantitative estimate for the accuracy of the collective RS-HDMR component functions, the second-order RS-HDMR approximations for P and D given by eq 2 were compared to the exact solutions obtained from 53 312 box-modelruns that uniformly covered the full region of the 4-dimensional input variable space. The component functions of the second-order RS-HDMR were constructed from different sample sizes. The results in Table 3 show that there is no significant difference between the second-order RS-HDMR approximations whose component functions are obtained from different sample sizes. This implies that Monte Carlo integration converges quite fast and a few thousand random samples can give reliable results.

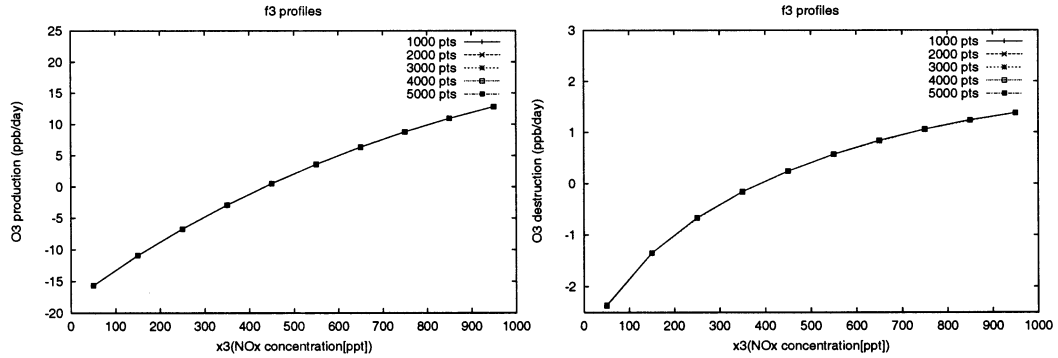

Figure 1. Function $f_3(x_3)$ for outputs P and D constructed from different sample sizes.

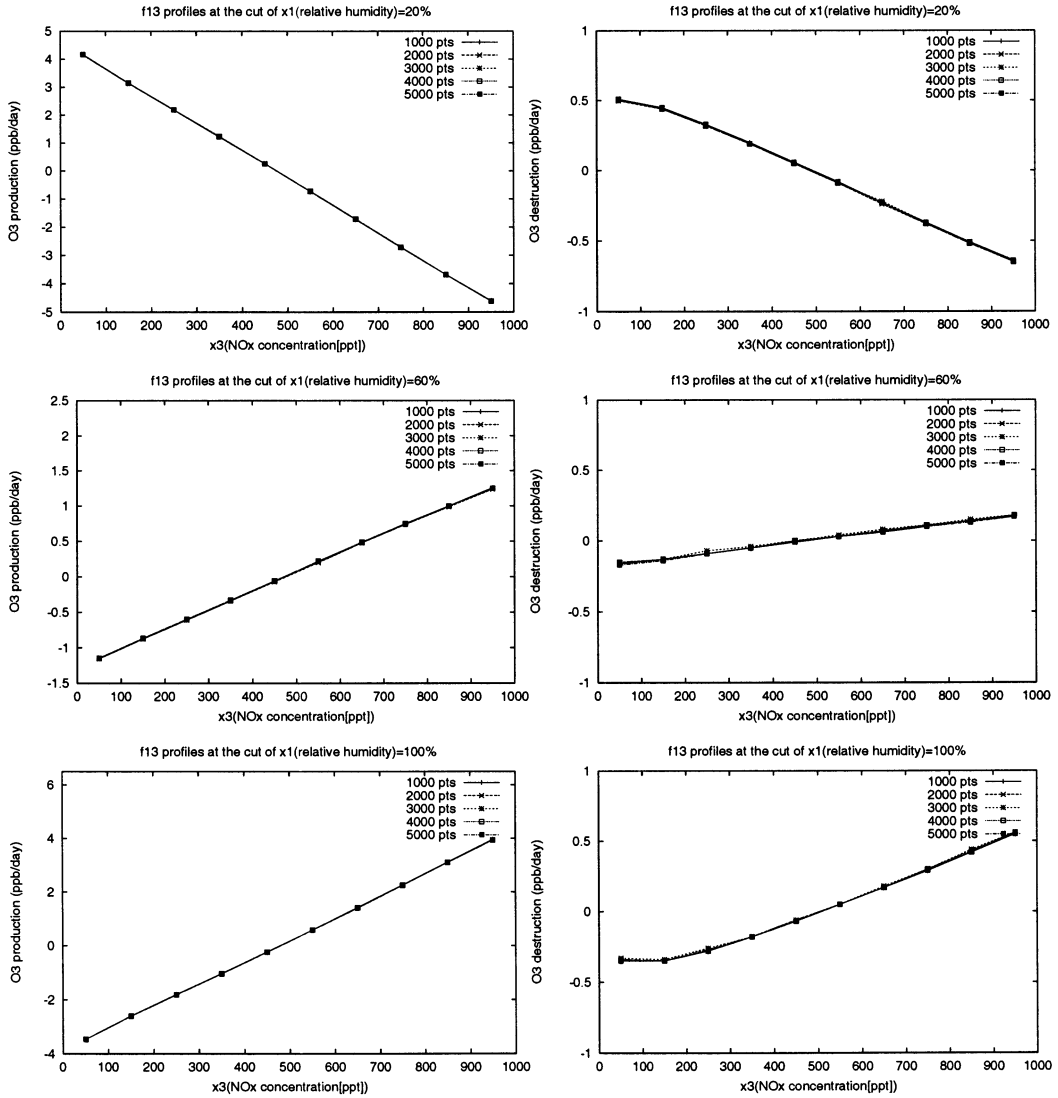

Figure 2. Function $f_{13}(x_1,x_3)$ for outputs P and D constructed from different sample sizes.

Compared to the 53 312 exact results, all the second-order RS-HDMR approximations, whose component functions were constructed from different sample sizes, have more than 88% and 97% of the tested points with relative errors less than 5% for P and D, respectively. The accuracy is quite satisfactory. As the values of f_0, $f_i(x_i)$, and $f_{ij}(x_i,x_j)$ have converged, in the following analytical basis function approximations, the RS-HDMR component functions up to the second order obtained by Monte Carlo integration with 5000 random samples will be used as a standard for comparison. Note that according to the meshes used for the four inputs, all together 1 454 000 or 7 270 000 random samples are needed to construct the RS-HDMR component functions up to second order when 1000 or 5000 points are respectively used in the Monte Carlo integration. If the third-order RS-HDMR component functions are also constructed in the same way, the required number of random samples are even bigger. Thus, the direct determination of RS-HDMR component functions by Monte Carlo integration is prohibitively expensive for use in many high-dimensional systems. The procedures introduced below dramatically reduce the necessary sample size while retaining excellent accuracy.

TABLE 3: Relative Errors of Second-Order RS-HDMR Approximations[a]

sample size (N)	relative error (%)	data portion (%)[b] P	D
1000	5	88.0	97.3
	10	96.7	99.4
	20	99.2	99.9
3000	5	88.5	97.4
	10	96.7	99.4
	20	99.2	99.9
5000	5	88.1	97.1
	10	96.6	99.4
	20	99.2	99.9

[a] The component functions are obtained from different sample sizes. [b] The percentage of 53 312 data with a relative error not larger than a given value.

3. Orthonormal Polynomial Approximation

Orthonormal polynomials were used as analytical basis functions to approximate the RS-HDMR component functions. The polynomials $\varphi_k(x)$ in the domain $[a, b]$ are referred to as orthonormal when they satisfy

$$\int_a^b \varphi_k(x)\, dx = 0 \qquad k = 1, 2, ... \tag{23}$$

$$\int_a^b \varphi_k^2(x)\, dx = 1 \qquad k = 1, 2, ... \tag{24}$$

$$\int_a^b \varphi_k(x)\, \varphi_l(x)\, dx = 0 \qquad k \neq l \tag{25}$$

i.e., they have a zero mean and unit norm and are mutually orthogonal. For the domain [0, 1], the orthonormal polynomials can be readily constructed from the above conditions:

$$\varphi_1(x) = \sqrt{3}(2x - 1) \tag{26}$$

$$\varphi_2(x) = 6\sqrt{5}\left(x^2 - x + \frac{1}{6}\right) \tag{27}$$

$$\varphi_3(x) = 20\sqrt{7}\left(x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}\right) \tag{28}$$

...

Using this set of basis functions, the RS-HDMR component functions $f_i(x_i)$, $f_{ij}(x_i,x_j)$, ... are represented as eqs 11 and 12 with

$$\varphi_{pq}(x_i,x_j) = \varphi_p(x_i)\, \varphi_q(x_j) \tag{29}$$

and Using the orthonormality property of the polynomials, the

$$f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) + \sum_{1 \le i < j \le n}\sum_{p=1}^{l}\sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i)\, \varphi_q(x_j) + ... \tag{30}$$

A matrices for α_r^i, β_{pq}^{ij}, ... in eq 18 are all identity matrices, and then

$$\alpha_r^i = \int_{K^n} f(\mathbf{x})\, \varphi_r(x_i)\, d\mathbf{x} \approx \frac{1}{N}\sum_{s=1}^{N} f(\mathbf{x}^{(s)})\, \varphi_r(x_i^{(s)}) \tag{31}$$

$$\beta_{pq}^{ij} = \int_{K^n} f(\mathbf{x})\, \varphi_p(x_i)\, \varphi_q(x_j)\, d\mathbf{x} \approx \frac{1}{N}\sum_{s=1}^{N} f(\mathbf{x}^{(s)})\, \varphi_p(x_i^{(s)})\, \varphi_q(x_j^{(s)}) \tag{32}$$

...

TABLE 4: Comparison between Second-Order RS-HDMR Approximations Whose Component Functions Were Obtained from Different Sample Sizes N and Different Orders of Orthonormal Polynomial Expansions

expansion order	sample size (N)	data portion (%)[a] 5%[b] P	5%[b] D	10%[b] P	10%[b] D	20%[b] P	20%[b] D
$k, l, l' = 3$	500	57.1	34.6	78.3	59.6	91.7	80.2
	1000	72.6	58.6	88.3	81.4	96.6	93.6
	3000	85.9	86.4	95.7	95.5	99.1	99.2
	5000	90.4	91.8	96.8	97.6	99.4	99.6
$k, l, l' = 2$	500	72.9	81.6	88.0	93.2	96.0	98.7
	1000	87.0	82.1	94.9	92.8	98.7	97.5
	3000	89.1	93.2	95.7	98.2	98.7	99.7
	5000	90.4	94.4	96.7	98.9	99.3	99.9
$k, l, l' = 1$	500	35.2	68.5	72.2	84.4	87.9	92.3
	1000	35.5	69.0	70.5	85.2	87.7	92.1
	3000	34.9	67.7	69.0	85.1	87.4	92.3
	5000	34.9	67.6	68.9	85.0	87.5	91.8

[a] The percentage of 53 312 data with a relative error not larger than a given value. [b] Relative error.

The accuracy of orthonormal polynomial approximation depends on the order of orthonormal polynomials used. In many cases, to achieve adequate accuracy using $\varphi_1(x)$, $\varphi_2(x)$, and $\varphi_3(x)$ is sufficient (i.e., $k, l, l' \le 3$). Because Monte Carlo integration is employed in eqs 31 and 32, the accuracy also depends on the sample size. Thus, different sample sizes and different orders of orthonormal polynomials were used to determine the coefficients α_r^i and β_{pq}^{ij}, and consequently the RS-HDMR component functions up to second order so that a comparison can be made with the results given by direct determination of Monte Carlo integration in section 2. Similarly, the accuracy of the resultant second-order RS-HDMR approximations whose component functions were approximated by orthonormal polynomials was determined by comparison with the previously mentioned 53 312 exact data. The results are shown in Table 4.

Table 4 shows that combining linear and quadratic (i.e., $k, l, l' = 2$) orthonormal polynomials gives the best results. When $N = 5000$, the accuracy is similar to that of direct determination by Monte Carlo integration (i.e., compare with Table 3). The accuracy is poor when only linear polynomials (i.e., $k, l, l' = 1$) are used. When the sample size N is small, Monte Carlo integration has large errors. This error may cause a poor approximation for $f_i(x_i)$ and especially for $f_{ij}(x_i, x_j)$. As the third-order (linear, quadratic, and cubic) polynomial expansion (i.e., $k, l, l' = 3$) has more terms ($4 \times 3 = 12$ for $f_i(x_i)$; $6 \times 3 \times 3 = 54$ for $f_{ij}(x_i, x_j)$) than the second-order (linear and quadratic) polynomial expansion ($4 \times 2 = 8$ for $f_i(x_i)$; $6 \times 2 \times 2 = 24$ for $f_{ij}(x_i,x_j)$), and each term has its own Monte Carlo integration error, the third-order polynomial expansion often has large errors when N is small. When the sample size becomes large and the Monte Carlo integration error becomes small, the accuracy given by the third-order polynomial expansion can be better than the second-order one. This behavior is observed in Figure 3, which gives the $f_3(x_3)$ for P and D obtained from different sample sizes and orders of orthonormal polynomials.

There are oscillations around the exact values for $f_3(x_3)$ of D when the third-order orthonormal polynomial expansion is used with small sample sizes. These oscillations introduce large errors. When the second-order orthonormal polynomial expansion is used, there is no such oscillation even if the sample size is smaller than 1000. This is the reason high-order polynomial expansions (i.e., $k, l, l' > 3$) may not be suitable for approxima-

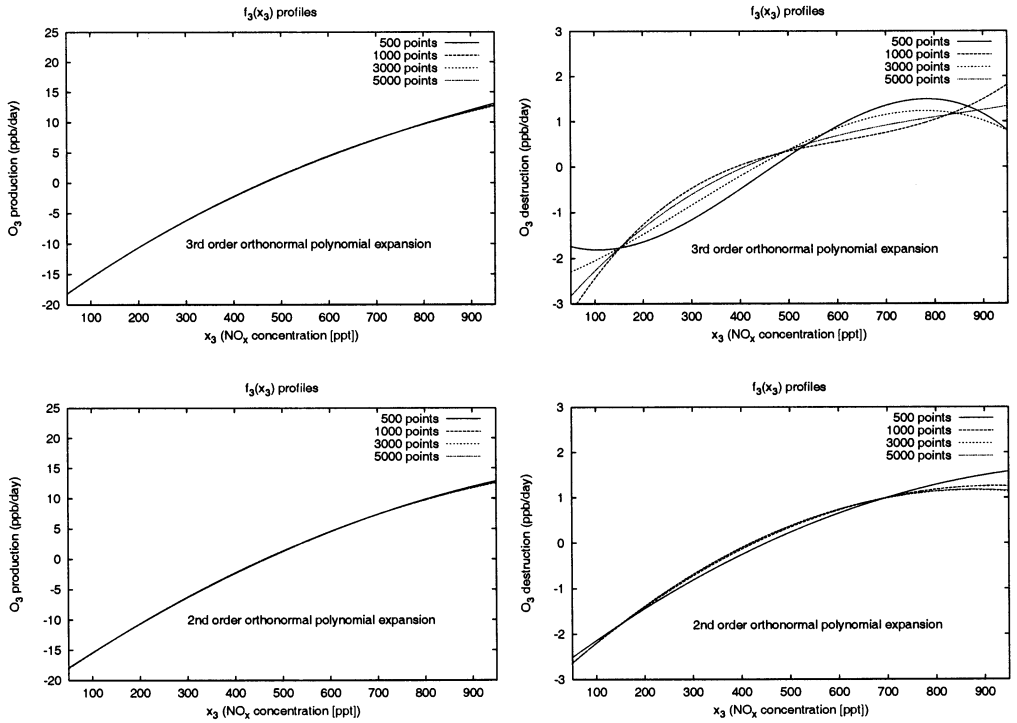

Figure 3. Function $f_3(x_3)$ for outputs P and D constructed from different sample sizes and different orders of orthonormal polynomials.

tion when Monte Carlo integration is involved. However, when the sample size is 5000, there is no significant difference between the accuracy of second- and third-order orthonormal polynomial expansions.

To diminish the oscillations produced by high-order orthonormal polynomial expansions, new objective functionals were introduced that also minimize the second-order derivatives of the approximation for a RS-HDMR component function. Writing the orthonormal polynomials in a general form,

$$\varphi_1(x) = a_1 x + a_0 \tag{33}$$

$$\varphi_2(x) = b_2 x^2 + b_1 x + b_0 \tag{34}$$

$$\varphi_3(x) = c_3 x^3 + c_2 x^2 + c_1 x + c_0 \tag{35}$$

...

where a_0, a_1, b_0, ..., c_3 are constant coefficients, the new functionals corresponding to eqs 16, 17, and 29 become

$$\min_{\alpha_r^i} \int_0^1 [f_i(x_i) - \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i)]^2 \, dx_i + \lambda_i \int_0^1 [\partial^2(\sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i))/\partial {x_i}^2]^2 \, dx_i \tag{36}$$

$$\min_{\beta_{pq}^{ij}} \int_0^1 \int_0^1 [f_{ij}(x_i,x_j) - \sum_{p=1}^{l}\sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_p(x_i)\, \varphi_q(x_j)]^2 \, dx_i \, dx_j + \lambda_{ij} \sum_{s,t\in\{i,j\}} \int_0^1 \int_0^1 [\partial^2(\sum_{p=1}^{l}\sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_p(x_i)\, \varphi_q(x_j))/\partial x_s \partial x_t]^2 \, dx_i \, dx_j \tag{37}$$

...

where λ_i and λ_{ij} are regularization weight parameters introduced to damp out the oscillations in representing the component functions.

For eq 36 and $k = 3$, the minimization yields

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 + 4{b_2}^2\lambda_i & 2b_2(2c_2 + 3c_3)\lambda_i \\ 0 & 2b_2(2c_2 + 3c_3)\lambda_i & 1 + 4\lambda_i({c_2}^2 + 3c_2c_3 + 3{c_3}^2) \end{pmatrix} \begin{pmatrix} \alpha_1^i \\ \alpha_2^i \\ \alpha_3^i \end{pmatrix} = \begin{pmatrix} \int_0^1 f_i(x_i)\, \varphi_1(x_i)\, dx_i \\ \int_0^1 f_i(x_i)\, \varphi_2(x_i)\, dx_i \\ \int_0^1 f_i(x_i)\, \varphi_3(x_i)\, dx_i \end{pmatrix} \tag{38}$$

Using the coefficients in eqs 26−28 the above equation is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 + 720\lambda_i & 0 \\ 0 & 0 & 1 + 8400\lambda_i \end{pmatrix} \begin{pmatrix} \alpha_1^i \\ \alpha_2^i \\ \alpha_3^i \end{pmatrix} = \begin{pmatrix} \int_0^1 f_i(x_i)\, \varphi_1(x_i)\, dx_i \\ \int_0^1 f_i(x_i)\, \varphi_2(x_i)\, dx_i \\ \int_0^1 f_i(x_i)\, \varphi_3(x_i)\, dx_i \end{pmatrix} \tag{39}$$

Equation 39 shows that one can choose a λ_i to reduce the contributions from α_2^i and especially α_3^i.

Figure 4 gives the results of $f_2(x_2)$ for D with $\lambda_2 = 0.9 \times 10^6/N^3$.

The oscillations around the exact value are diminished. As the oscillations decrease with the sample size N, the λ_i's were chosen to be proportional to $1/N^3$ such that for $N = 5000$ introducing λ_i has no significant influence on the results. Similar results were obtained for other $f_i(x_i)$'s.

Similarly, for eq 37 and l, $l' = 3$, the minimization yields

$$A\beta = \mathbf{b} \tag{40}$$

where

$$A = \begin{Bmatrix} 1+144\lambda_{ij} & 0 & 48\sqrt{21}\lambda_{ij} & 0 & 0 & 0 & 48\sqrt{21}\lambda_{ij} & 0 & 336\lambda_{ij} \\ 0 & 1+144\lambda_{ij} & 0 & 0 & 0 & 0 & 0 & 240\sqrt{21}\lambda_{ij} & 0 \\ 48\sqrt{21}\lambda_{ij} & 0 & 1+10416\lambda_{ij} & 0 & 0 & 0 & 336\lambda_{ij} & 0 & 672\sqrt{21}\lambda_{ij} \\ 0 & 0 & 0 & 1+144\lambda_{ij} & 0 & 240\sqrt{21}\lambda_{ij} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1+5040\lambda_{ij} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 240\sqrt{21}\lambda_{ij} & 0 & 1+19200\lambda_{ij} & 0 & 0 & 0 \\ 48\sqrt{21}\lambda_{ij} & 0 & 336\lambda_{ij} & 0 & 0 & 0 & 1+10416\lambda_{ij} & 0 & 672\sqrt{21}\lambda_{ij} \\ 0 & 240\sqrt{21}\lambda_{ij} & 0 & 0 & 0 & 0 & 0 & 1+19200\lambda_{ij} & 0 \\ 336\lambda_{ij} & 0 & 672\sqrt{21}\lambda_{ij} & 0 & 0 & 0 & 672\sqrt{21}\lambda_{ij} & 0 & 1+45024\lambda_{ij} \end{Bmatrix} \tag{41}$$

$$\beta = (\beta_{11}^{ij}\beta_{12}^{ij}\cdots\beta_{32}^{ij}\beta_{33}^{ij})^T \tag{42}$$

$$\mathbf{b} = \begin{Bmatrix} \int_0^1\int_0^1 f_{ij}(x_i,x_j)\,\varphi_1(x_i)\,\varphi_1(x_j)\,dx_i\,dx_j \\ \int_0^1\int_0^1 f_{ij}(x_i,x_j)\,\varphi_1(x_i)\,\varphi_2(x_j)\,dx_i\,dx_j \\ \vdots \\ \int_0^1\int_0^1 f_{ij}(x_i,x_j)\,\varphi_3(x_i)\,\varphi_2(x_j)\,dx_i\,dx_j \\ \int_0^1\int_0^1 f_{ij}(x_i,x_j)\,\varphi_3(x_i)\,\varphi_3(x_j)\,dx_i\,dx_j \end{Bmatrix} \tag{43}$$

Equation 41 shows that for a given λ_{ij} the oscillations related to high-order orthonormal polynomials $\varphi_k(x_i)$ and $\varphi_k(x_j)$ can be managed.

Table 5 gives the results of simultaneously minimizing the second-order derivatives of third-order orthonormal polynomial expansions for the second-order RS-HDMR approximation.

The regularization parameters λ_i and λ_{ij} were determined in two steps. First, a value for λ_i or λ_{ij} was chosen to damp out the oscillation for the corresponding functions $f_i(x_i)$ or $f_{ij}(x_i,x_j)$. Second, the resultant values of λ_i and λ_{ij} were adjusted to achieve the best accuracy of the resultant second-order RS-HDMR approximation for the N samples. Although λ_i and λ_{ij} were determined only from the N samples, Table 5 shows that the resultant second-order RS-HDMR approximation has excellent accuracy for the 53 312 data that uniformly cover the whole desired domain of **x**. This implies that choosing the values of λ_i and λ_{ij} from small samples can ensure the accuracy of the second-order RS-HDMR approximation in the whole domain of **x**. The results in Table 5 show that the accuracy of the orthonormal polynomial approximation obtained even from samples smaller than 1000 is better than that given by direct determination of Monte Carlo integration in Table 3 (the only exception is D with a relative error not larger than 5%. The reason is that some of the exact data of D are very small, and the orthonormal polynomial approximation can have relative errors larger than 5% for these data).

All the above results were obtained by using the orthonormality property of $\varphi_r(x_i)$; i.e., the matrix A is an identity matrix. If the integrals $\int\varphi_r(x_i)\varphi_r'(x_i)\,dx_i$, $\int\varphi_r(x_i)\,\varphi_r'(x_j)\,dx_i\,dx_j$, $\int\varphi_r(x_i)\,\varphi_p(x_i)\,\varphi_p(x_j)\,dx_i\,dx_j$, ... are also approximated by Monte Carlo integration, then the matrix A is no longer an identity matrix, but a symmetric one with diagonal elements close to unity and off-diagonal elements close to zero, and the coefficients $\{\alpha_r^i, \beta_{pq}^{ij}, ...\}$ will be determined by solving linear algebraic equations. Table 6 gives the accuracy of the resultant second-order RS-HDMR approximations whose coefficients were determined either by solving a single linear algebraic equation for all α_r^i, β_{pq}^{ij} corresponding to all $f_i(x_i)$ and $f_{ij}(x_i,x_j)$ simultaneously or by solving two linear algebraic equations for α_r^i corresponding to all $f_i(x_i)$ and β_{pq}^{ij} corresponding to all $f_{ij}(x_i,x_j)$ separately. The RS-HDMR component functions were approximated by third-order orthonormal polynomial expansions, and different sample sizes N were tested.

The results of Table 6 show that, compared to Table 4 without simultaneously minimizing the second-order derivatives, only considering the coupling of the coefficients within each order of RS-HDMR functions does not improve the accuracy. Considering the coupling of all α_r^i, β_{pq}^{ij} gives a better result, but it is generally worse than the result obtained by simultaneously minimizing the second-order derivatives (see the results of Table 5). Moreover, for a large n the matrix A will be very big, and solving high-dimensional algebraic equations is not computationally efficient.

As less than 1000 samples are necessary in the orthonormal polynomial approximation with regularization, compared to 1 454 000 or 7 270 000 samples in direct determination of Monte Carlo integration, the computational saving is very significant. Hence, the orthonormal polynomial approximation with regularization provides a practical way to construct RS-HDMR component functions.

4. Spline Function Approximation

Polynomial spline functions[9,10] can be used as another basis for approximating the RS-HDMR component functions $f_i(x_i)$, $f_{ij}(x_i,x_j)$, ... Cubic B splines $B_k(x)$ ($k = -1, 0, ..., m+1$) defined in interval $[a, b]$

$$B_k(x) = \frac{1}{h^3} \times \begin{cases} (y_{k+2}-x)^3 & y_{k+1} < x \le y_{k+2} \\ (y_{k+2}-x)^3 - 4(y_{k+1}-x)^3 & y_k < x \le y_{k+1} \\ (y_{k+2}-x)^3 - 4(y_{k+1}-x)^3 + 6(y_k-x)^3 & y_{k-1} < t \le y_k \\ (y_{k+2}-x)^3 - 4(y_{k+1}-x)^3 + 6(y_k-x)^3 - 4(y_{k-1}-x)^3 & y_{k-2} < x \le y_{k-1} \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

where

$$h = \frac{b-a}{m} \tag{45}$$

$$y_k = a + kh \tag{46}$$

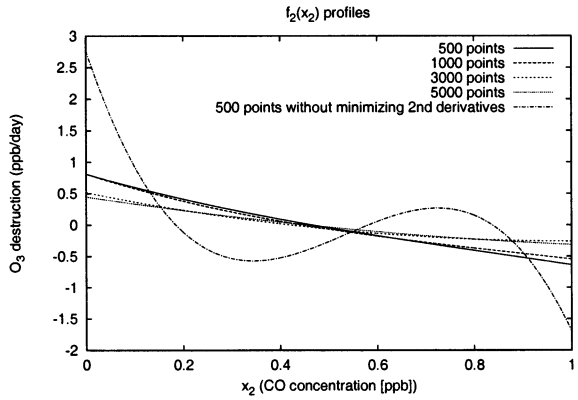

Figure 4. Function $f_2(x_2)$ for output D constructed from different sample sizes with a third-order orthonormal polynomial expansion using $\lambda_2 = 0.9 \times 10^6/N^3$.

TABLE 5: Comparison between Second-Order RS-HDMR Approximations Whose Component Functions Were Obtained from Different Sample Sizes *N* Using Third-Order Orthonormal Polynomial Expansions While Simultaneously Minimizing the Second-Order Derivatives

		data portion (%)[a]					
		5%[b]		10%[b]		20%[b]	
tested data	sample size (N)	P	D	P	D	P	D
N	500	79.6	85.0	92.2	94.4	97.4	98.4
	1000	85.1	89.3	95.2	97.4	97.8	99.9
	3000	87.1	93.3	95.8	98.4	97.9	99.8
	5000	88.0	94.6	95.8	98.6	98.2	99.5
53 312	500	92.8	93.0	99.5	98.3	100	99.5
	1000	93.9	94.3	99.5	99.6	100	100
	3000	95.0	96.7	99.7	99.8	100	100
	5000	96.8	97.1	100	99.8	100	100

[a] The percentage of tested data with a relative error not larger than a given value. λ_i/N^3: for P, $\lambda_1 = 0.9 \times 10^4$, $\lambda_i = 0.0$ ($i = 2, 3, 4$); for D, $\lambda_1 = 0.0$, $\lambda_2 = 0.7 \times 10^6$, $\lambda_3 = 0.4 \times 10^5$, $\lambda_4 = 0.1 \times 10^5$. λ_{ij}/N^3: for P, $\lambda_{12} = 0.9 \times 10^6$, $\lambda_{13} = 0.9 \times 10^4$, $\lambda_{14} = 0.9 \times 10^9$, $\lambda_{23} = 0.1 \times 10^6$, $\lambda_{24} = 0.2 \times 10^8$, $\lambda_{34} = 0.0$; for D, $\lambda_{12} = 0.1 \times 10^8$, $\lambda_{13} = 0.2 \times 10^7$, $\lambda_{14} = 0.2 \times 10^5$, $\lambda_{23} = 0.7 \times 10^7$, $\lambda_{24} = 0.7 \times 10^4$, $\lambda_{34} = 0.1 \times 10^5$. [b] Relative error.

were tested for this purpose. When the domain of $\mathbf{x}$ is a unit hypercube K^n,

$$h = \frac{1}{m} \tag{47}$$

$$y_k = \frac{k}{m} \tag{48}$$

The first and second-order RS-HDMR component functions $f_i(x_i)$ and $f_{ij}(x_i,x_j)$ can be approximately expanded as

$$f_i(x_i) \approx \sum_{r=-1}^{m+1} \alpha_r^i B_r(x_i) \tag{49}$$

$$f_{ij}(x_i,x_j) \approx \sum_{p=-1}^{m+1}\sum_{q=-1}^{m+1} \beta_{pq}^{ij} B_p(x_i)\, B_q(x_j) \tag{50}$$

...

where α_r^i and β_{pq}^{ij} are constant coefficients to be determined.

The cubic B splines with different variables are not mutually orthogonal. However, one cannot determine all the coefficients α_r^i and β_{pq}^{ij} simultaneously because the matrix A is singular.

TABLE 6: Comparison between Second-Order RS-HDMR Approximations Whose Component Functions Were Obtained from Different Sample Sizes *N* and Using Third-Order Orthonormal Polynomial Expansions Where the Integrals of the Orthonormal Polynomials Were Approximated by Monte Carlo Integration

		data portion (%)[a]					
		5%[b]		10%[b]		20%[b]	
determination of $\alpha_r^i,\beta_{pq}^{ij}$	sample size (N)	P	D	P	D	P	D
simultaneously	500	91.1	95.6	96.6	99.3	99.1	99.9
	1000	91.0	95.9	96.5	99.2	99.2	99.9
	3000	91.4	95.5	96.8	99.2	99.2	99.9
	5000	91.5	95.6	96.8	99.2	99.3	99.9
separately	500	50.5	40.4	75.5	63.5	91.6	82.3
	1000	66.8	57.9	85.5	80.7	94.8	93.5
	3000	86.2	80.2	95.5	92.8	99.2	98.7
	5000	90.5	90.2	96.7	97.0	99.3	99.5

[a] The percentage of 53 312 data with a relative error not larger than a given value. [b] Relative error.

The singularity arises because for different x_i the cubic B splines have the same form. Similar to the orthonormal polynomial approximation, only considering the coupling within each order of the RS-HDMR functions gives unsatisfactory results. Therefore, the α_r^i and β_{pq}^{ij} are still obtained by minimization of eqs 16 and 17. Then, α_r^i can be obtained by solving the linear equation $A\alpha = \mathbf{b}$:

$$A\begin{pmatrix} \alpha_{-1}^i \\ \alpha_0^i \\ \vdots \\ \alpha_{m+1}^i \end{pmatrix} = \begin{pmatrix} \int_0^1 f_i(x_i)\, B_{-1}(x_i)\, \mathrm{d}x_i \\ \int_0^1 f_i(x_i)\, B_0(x_i) \mathrm{d}x_i \\ \vdots \\ \int_0^1 f_i(x_i)\, B_{m+1}(x_i)\, \mathrm{d}x_i \end{pmatrix} \tag{51}$$

where A is an $(m + 3) \times (m + 3)$ symmetric and nonsingular matrix whose (k, l)-entry

$$A_{kl} = \int_0^1 B_k(x_i)\, B_l(x_i)\, \mathrm{d}x_i \qquad k, l = -1, 0, ..., m + 1 \tag{52}$$

Using the definition of $f_i(x_i)$ given in eq 4, eq 51 can be expressed as

$$A\begin{pmatrix} \alpha_{-1}^i \\ \alpha_0^i \\ \vdots \\ \alpha_{m+1}^i \end{pmatrix} = \begin{pmatrix} \int_{K^n} f(\mathbf{x})\, B_{-1}(x_i)\, \mathrm{d}\mathbf{x} - f_0 \int_0^1 B_{-1}(x_i)\, \mathrm{d}x_i \\ \int_{K^n} f(\mathbf{x})\, B_0(x_i)\, \mathrm{d}\mathbf{x} - f_0 \int_0^1 B_0(x_i)\, \mathrm{d}x_i \\ \vdots \\ \int_{K^n} f(\mathbf{x})\, B_{m+1}(x_i)\, \mathrm{d}\mathbf{x} - f_0 \int_0^1 B_{m+1}(x_i)\, \mathrm{d}x_i \end{pmatrix}$$

$$\approx \frac{1}{N}\sum_{s=1}^{N} \begin{pmatrix} f(\mathbf{x}^{(s)})\, B_{-1}(x_i^{(s)}) - f(\mathbf{x}^{(s)}) \int_0^1 B_{-1}(x_i)\, \mathrm{d}x_i \\ f(\mathbf{x}^{(s)})\, B_0(x_i^{(s)}) - f(\mathbf{x}^{(s)}) \int_0^1 B_0(x_i)\, \mathrm{d}x_i \\ \vdots \\ f(\mathbf{x}^{(s)})\, B_{m+1}(x_i^{(s)}) - f(\mathbf{x}^{(s)}) \int_0^1 B_{m+1}(x_i)\, \mathrm{d}x_i \end{pmatrix} \tag{53}$$

The integrals $\int B_k(x_i)\, \mathrm{d}x_i$ and $\int B_k(x_i)\, B_l(x_i)\, \mathrm{d}x_i$ contained in eqs 52 and 53 can be readily determined by using the definition eq 44. Then α is given by $A^{-1}\mathbf{b}$. As Monte Carlo integration has been used in eq 53, and the accuracy of cubic B splines approximation depends on the number of subintervals m, the accuracy of the resultant $f_i(x_i)$ is related to sample size N and

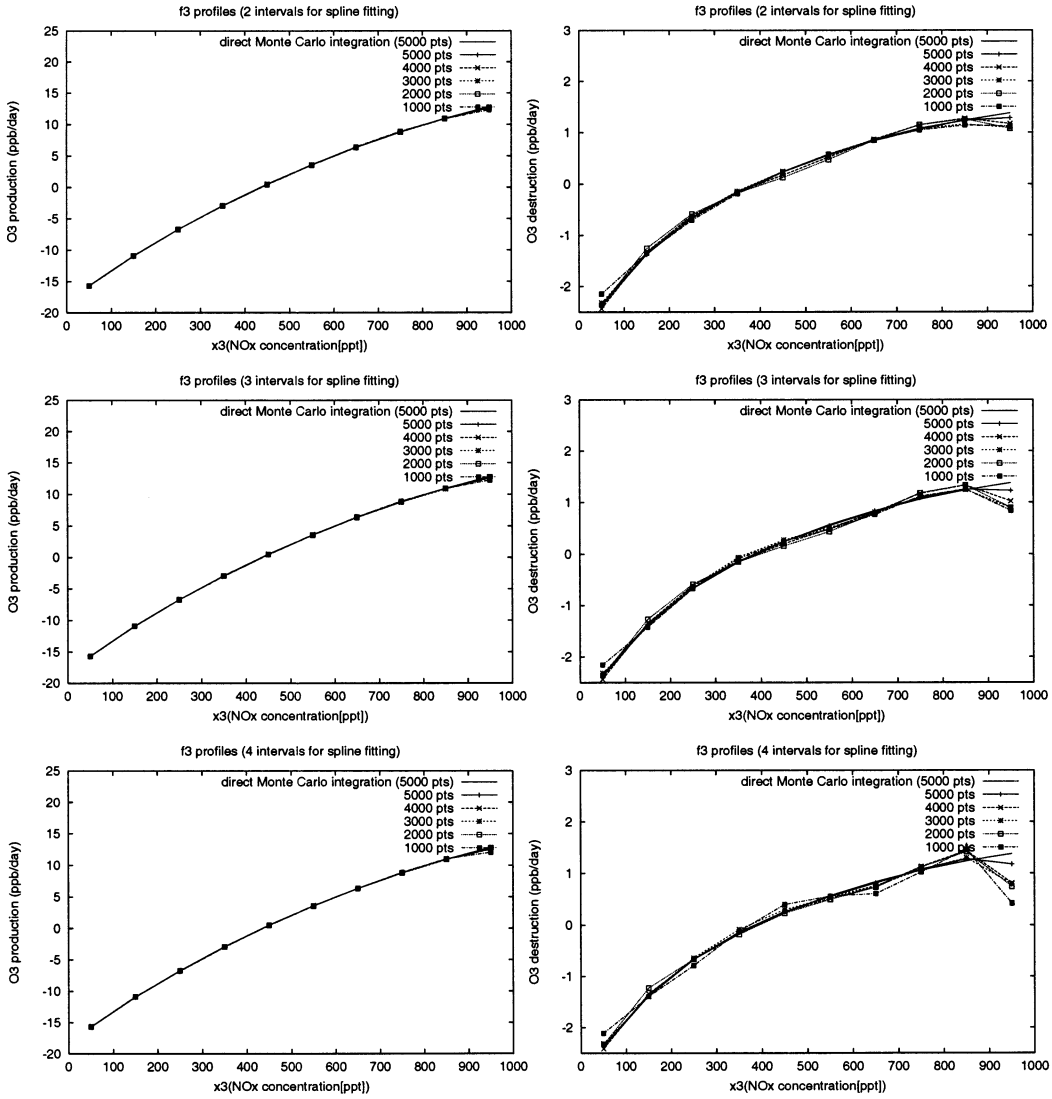

Figure 5. Function $f_3(x_3)$ for outputs *P* and *D* constructed from different sample sizes and cubic B splines with two to four subintervals.

the value of *m*. The coefficients β_{kl}^{ij} can be determined similarly. In this case, the elements of *A* and **b** are the integrals $\int\int B_p(x_i)\, B_q(x_j)\, B_r(x_i)\, B_s(x_j)\, dx_i\, dx_j$ and $\int\int f_{ij}(x_i,x_j)\, B_p(x_i)\, B_q(x_j)\, dx_i\, dx_j$, respectively. To save space, the formulas are not given here. The cubic B splines approximation was applied to the same model used for testing orthonormal polynomial approximations. $f_i(x_i)$ and $f_{ij}(x_i,x_j)$ were expanded by eqs 49 and 50. Different sample sizes (1000−5000) and numbers of subintervals (m = 2−4) were tested.

The results for $f_3(x_3)$ and $f_{13}(x_1,x_3)$ of *P* and *D* are given in Figures 5 and 6, respectively. The other component functions have similar figures. The figures show that the convergence is good and the resultant $f_3(x_3)$ are close to those given by direct Monte Carlo integration with 5000 points. For $f_{13}(x_1,x_3)$, oscillations around the exact values can be observed, especially when the sample size is small. Moreover, large errors occur at the end of the interval. The resultant second-order RS-HDMR approximations were compared to the table of exact solutions obtained from 53 312 box-model runs. The results are shown in Tables 7−9.

Notice that $B_k(x)$ may vanish on some subintervals. Thus, only part of the *N* data are actually used for $f(\mathbf{x}^{(s)})$ in Monte Carlo integration $\sum_{s=1}^{N} f(\mathbf{x}^{(s)})\, B_k(x^{(s)}{}_i)/N$, which is proportional to $1/m$. Therefore, for a given sample size *N*, large *m* may yield poor accuracy. The results in Tables 7−9 show that $m = 2$ gives the best approximation. Compared to the results of orthonormal polynomial approximation without regularization (see Table 4), its accuracy is a little worse for $N = 5000$, but much worse for small *N*. When *m* is larger, the accuracy becomes even worse. Like the orthonormal polynomial approximation, simultaneously

TABLE 7: Comparison between Second-Order RS-HDMR Approximations Whose Component Functions Were Obtained from Different Sample Sizes *N* and $m = 2$ Cubic B Splines Approximation

		data portion (%)[a]	
sample size (*N*)	relative error (%)	*P*	*D*
1000	5	45.7	33.8
	10	69.1	57.4
	20	86.3	79.5
3000	5	83.7	69.7
	10	93.7	87.3
	20	98.1	96.5
5000	5	88.6	84.5
	10	95.9	94.7
	20	98.8	98.8

[a] The percentage of 53 312 data with a relative error not larger than a given value.

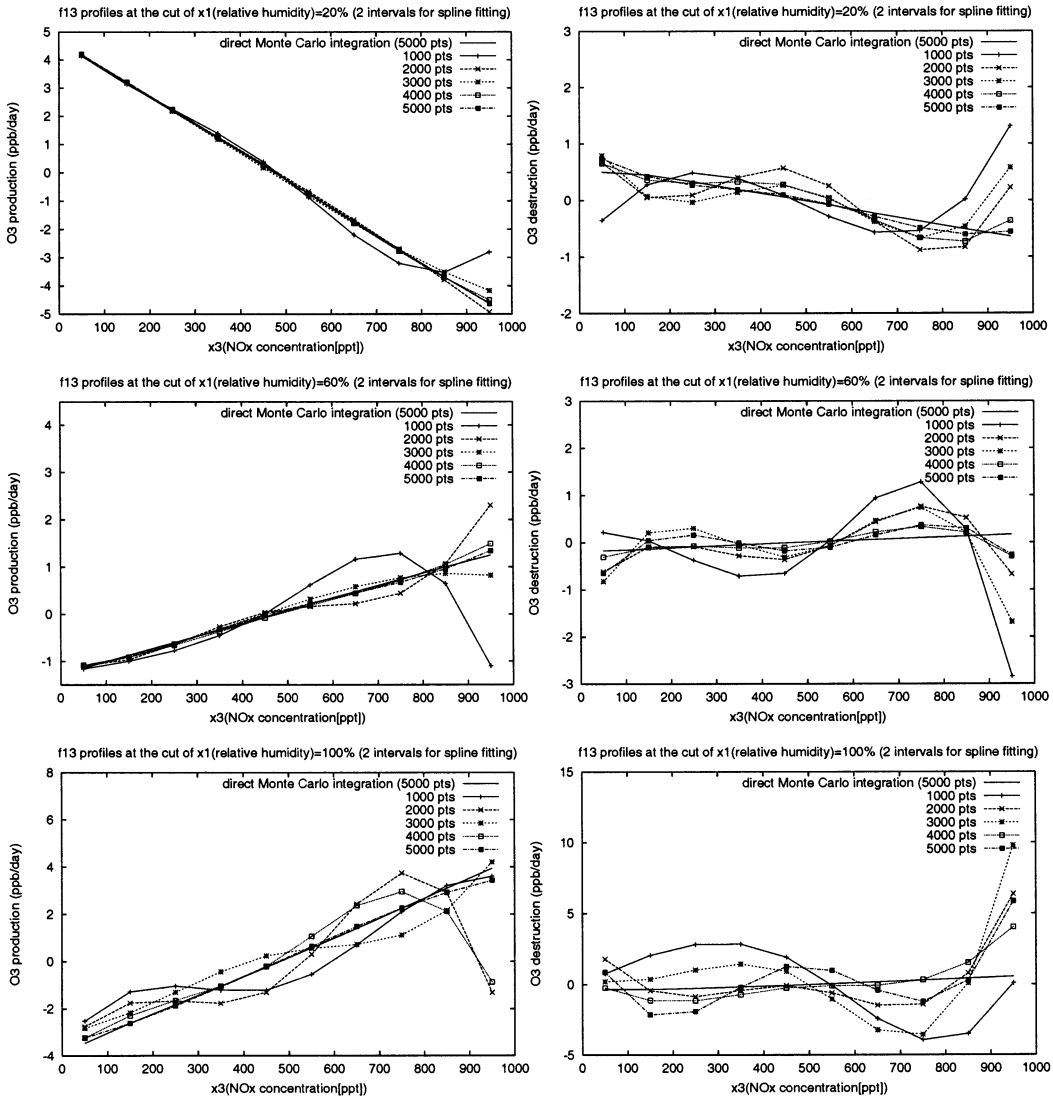

Figure 6. Function $f_{13}(x_1,x_3)$ for outputs P and D constructed from different sample sizes and cubic B splines with two subintervals.

TABLE 8: Comparison between Second-Order RS-HDMR Approximations Whose Component Functions Were Obtained from Different Sample Sizes N and $m = 3$ Cubic B Splines Approximation

		data portion (%)[a]	
sample size (N)	relative error (%)	P	D
1000	5	26.9	24.2
	10	49.2	45.0
	20	73.6	69.5
3000	5	68.4	57.0
	10	85.1	79.2
	20	94.0	92.3
5000	5	71.5	72.4
	10	86.8	88.2
	20	94.9	96.2

[a] The percentage of 53 312 data with a relative error not larger than a given value.

TABLE 9: Comparison between Second-Order RS-HDMR Approximations Whose Component Functions Were Obtained from Different Sample Sizes N and $m = 4$ Cubic B Splines Approximation

		data portion (%)[a]	
sample size (N)	relative error (%)	P	D
1000	5	27.4	23.1
	10	49.9	42.8
	20	75.1	66.4
3000	5	64.7	51.4
	10	83.0	74.3
	20	92.8	89.4
5000	5	72.0	68.0
	10	86.3	84.7
	20	94.8	94.8

[a] The percentage of 53 312 data with a relative error not larger than a given value.

minimizing the second-order derivatives of the cubic B splines approximation will increase the accuracy and damp out the large error at the end of the interval. However, considering that (1) the cubic B splines are not an orthogonal basis, (2) only part of N is used for the determination of the parameters α_r^i and β_{pq}^{ij}, and (3) the cubic B spline approximation has more terms ($m + 3$ for $f_i(x_i)$, $(m + 3)^2$ for $f_{ij}(x_i,x_j)$) than the orthonormal polynomial approximation (usually, 3 for $f_i(x_i)$, 9 for $f_{ij}(x_i,x_j)$), especially for large m and each term has the Monte Carlo integration error (more terms result in larger total error in the approximation), we do not expect that the regularized cubic B splines approximation can have better accuracy than the regularized orthonormal polynomial approximation.

5. Polynomial Approximation

The RS-HDMR component functions can be directly approximated by polynomial functions,

$$f_i(x_i) \approx \sum_{r=0}^{k} \alpha_r^i x_i^r \tag{54}$$

$$f_{ij}(x_i,x_j) \approx \sum_{p=0}^{l}\sum_{q=0}^{l'} \beta_{pq}^{ij} x_i^p x_j^q \tag{55}$$

...

Similar to the cubic B spline approximation, for the polynomial approximation the matrix **A** is singular, even if all α_r^i, β_{pq}^{ij} corresponding to all $f_i(x_i)$ and $f_{ij}(x_i,x_j)$ are determined simultaneously or if α_r^i corresponding to all $f_i(x_i)$ and β_{pq}^{ij} corresponding to all $f_{ij}(x_i,x_j)$ are determined separately. Therefore, the constant coefficients α_k^i and β_{kl}^{ij} were approximately determined by minimizing the integrals

$$\min_{\alpha_k^i} \int_0^1 [f_i(x_i) - \sum_{r=0}^{k} \alpha_i^i x_i^r]^2 \, dx_i \tag{56}$$

$$\min_{\beta_{kl}^{ij}} \int_0^1 \int_0^1 [f_{ij}(x_i,x_j) - \sum_{p=0}^{l}\sum_{q=0}^{l'} \beta_{pq}^{ij} x_i^p x_j^q]^2 \, dx_i \, dx_j \tag{57}$$

Linear equations for coefficients α_r^i and β_{pq}^{ij} can be obtained from these minimizations. For eq 56 the minimization gives

$$\mathbf{A}\alpha = \mathbf{b} \tag{58}$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & {}^1/_2 & \cdots & 1/(k+1) \\ {}^1/_2 & {}^1/_3 & \cdots & 1/(k+2) \\ & \cdots & \cdots & \\ 1/(k+1) & 1/(k+2) & \cdots & 1/(2k+2) \end{pmatrix} \tag{59}$$

$$\alpha = (\alpha_0^i \alpha_1^i \cdots \alpha_k^i)^T \tag{60}$$

and

$$\mathbf{b} = \begin{pmatrix} \int_0^1 f_i(x_i)\, dx_i \\ \int_0^1 f_i(x_i) x_i \, dx_i \\ \vdots \\ \int_0^1 f_i(x_i) x_i^k \, dx_i \end{pmatrix} = \begin{pmatrix} 0 \\ \int_{K^n} f(\mathbf{x})\, x_i \, d\mathbf{x} - f_0 \int_0^1 x_i \, dx_i \\ \vdots \\ \int_{K^n} f(\mathbf{x})\, x_i^k \, d\mathbf{x} - f_0 \int_0^1 x_i^k \, dx_i \end{pmatrix}$$

$$\approx \frac{1}{N}\sum_{s=1}^{N} \begin{pmatrix} 0 \\ f(\mathbf{x}^{(s)}) x_i^{(s)} - f(\mathbf{x}^{(s)})/2 \\ \vdots \\ f(\mathbf{x}^{(s)})(x_i^{(s)})^k - f(\mathbf{x}^{(s)})/(k+1) \end{pmatrix} \tag{61}$$

As **A** is symmetric and nonsingular, α is given by $\mathbf{A}^{-1}\mathbf{b}$. The calculation procedures are similar to cubic B splines approximations. The coefficients β_{pq}^{ij} can be determined similarly.

Different sample sizes and different orders of polynomials were used for the polynomial approximations. The best results for $f_i(x_i)$ and $f_{ij}(x_i,x_j)$ in the zero-dimensional photochemical box model are given by the following polynomials:

$$f_i(x_i) = \alpha_i^0 + \alpha_i^1 x_i + \alpha_i^2 x_i^2 \tag{62}$$

$$f_{ij}(x_i, x_j) = \beta_{ij}^{00} + \beta_{ij}^{10} x_i + \beta_{ij}^{01} x_j + \beta_{ij}^{20} x_i^2 + \beta_{ij}^{02} x_j^2 + \beta_{ij}^{11} x_i x_j + \beta_{ij}^{30} x_i^3 + \beta_{ij}^{03} x_j^3 + \beta_{ij}^{21} x_i^2 x_j + \beta_{ij}^{12} x_i x_j^2 + \beta_{ij}^{31} x_i^3 x_j + \beta_{ij}^{13} x_i x_j^3 + \beta_{ij}^{22} x_i^2 x_j^2 \tag{63}$$

The coefficients α_r^i and β_{pq}^{ij}, and consequently the RS-HDMR component functions up to second order, were determined so that the comparison with the results given by direct determination of Monte Carlo integration can be made. Results for $f_3(x_3)$ and $f_{13}(x_1,x_3)$ for *P* and *D* obtained from different sample sizes are given in Figures 7 and 8, respectively. Other component functions have similar behavior. The figures show that the convergence is good. When the sample size is 5000, the resultant $f_i(x_i)$ and $f_{ij}(x_i,x_j)$ are close to those given by direct Monte Carlo integration with 5000 points. In $f_{ij}(x_i,x_j)$, the error is larger for *D* on the boundary of the input hypercube.

The accuracy of the resultant second-order RS-HDMR approximations whose component functions were approximated by polynomials was determined by comparison with the 53 312 exact data set. The results are given in Table 10, which shows that the polynomial approximation has a better accuracy than the cubic B spline approximation with $m = 2$, especially when the sample size is smaller than 5000, but it is worse than the nonregularized orthonormal polynomial approximations (see Table 4). Similarly, regularization will improve the accuracy of the polynomial approximation. However, considering that the polynomials with different numbers of variables are not orthogonal, and that the nonregularized polynomial approximation is worse than nonregularized orthonormal polynomial approximation, we do not expect that its regularization can provide better accuracy than that for the regularized orthonormal polynomial approximation.

6. Conclusions and a Discussion

HDMR is a general set of quantitative model assessment and analysis tools for high-dimensional input−output systems. When data are randomly sampled, an RS-HDMR can be constructed. The RS-HDMR component functions involve high-dimensional integrals that may be approximately calculated by Monte Carlo integration. Because the direct determination of high-order RS-HDMR component functions by Monte Carlo integration is prohibitively expensive, analytical basis functions, including orthonormal polynomials, cubic B spline functions, and polynomials were employed to approximate RS-HDMR component functions. With such basis functions, only one set of random samples of the output is needed to determine all RS-HDMR component functions. Thus, the sampling effort is dramatically reduced. In a test example, the direct determination by Monte Carlo integration needed millions of samples, but employing the basis function approximations of the RS-HDMR component functions needed only thousands or even hundreds of samples with comparable accuracy. Therefore, analytical basis function approximations form a practical approach for RS-HDMR applications.

Three basis functions, orthonormal polynomials, cubic B splines, and polynomials, were used for the approximation of RS-HDMR component functions. The formulas to determine the expansion coefficients α_r^i, β_{pq}^{ij}, ... were constructed by using the orthogonality of $f_i(x_i)$, $f_{ij}(x_i,x_j)$, ..., and the determination of the coefficients involves Monte Carlo integration approxima-

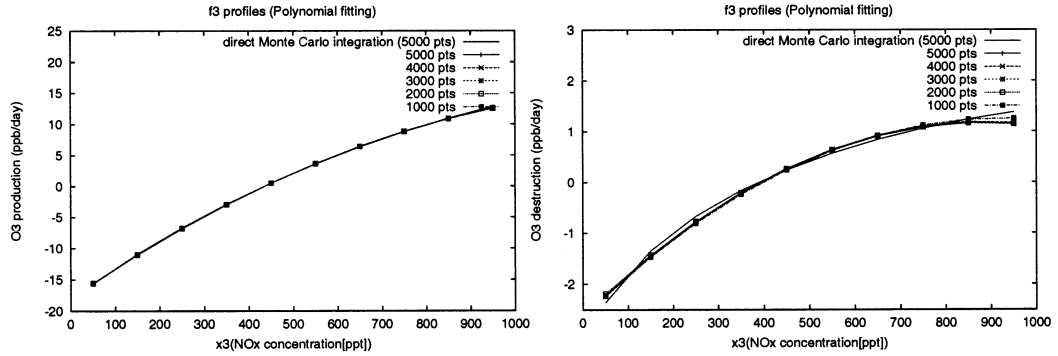

Figure 7. Function $f_3(x_3)$ for outputs P and D constructed from different sample sizes and polynomial approximation (eq 62).

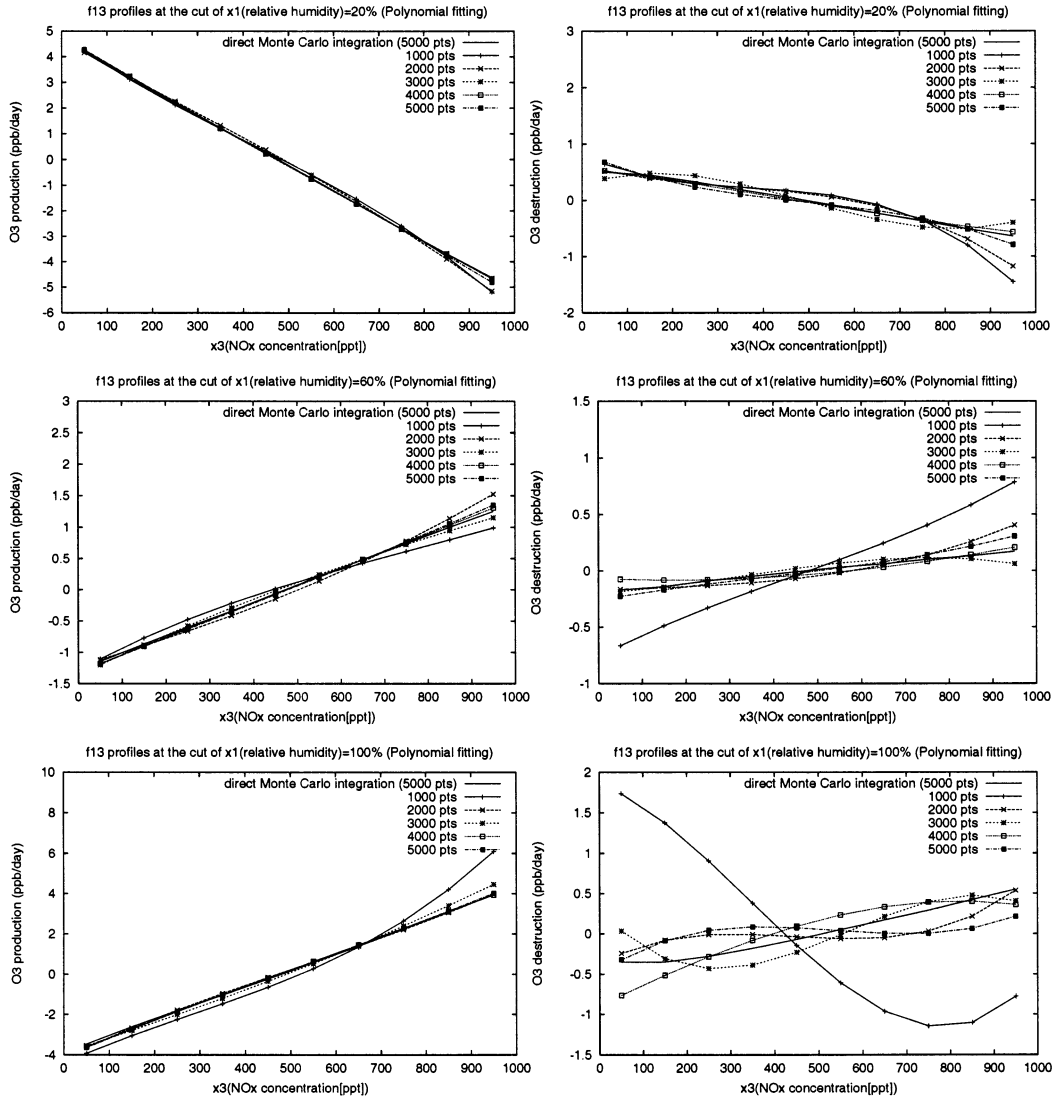

Figure 8. Function $f_{13}(x_1,x_3)$ for outputs P and D constructed from different sample sizes and polynomial approximation (eq 63).

tion whose error decreases as $\sim 1/\sqrt{N}$. Hence, the accuracy of analytical basis approximations for RS-HDMR component functions depends on the orthogonality of the basis functions and the sample size used in the approximation. Orthonormal polynomials provided the best accuracy. The cubic B spline function approximation has the worst accuracy of the three basis functions because it is not orthogonal and only uses a part of the N in Monte Carlo integration points. Increasing the number of subintervals m often improved the accuracy for cubic B splines in other problems. However, when Monte Carlo integration is involved, large m decreases the accuracy because the fraction of data used in N is proportional to $1/m$. Moreover, the cubic B splines approximation has more terms and each term has its own Monte Carlo integration error. Large terms cause a large total approximation error. All these factors make the cubic B splines approximation the worst one for the RS-HDMR component functions. Polynomial approximation has an accuracy in between. Simultaneous minimization of the second-order derivatives of the approximate functions for $f_i(x_i)$, $f_{ij}(x_i,x_j)$, ... dramatically improved the accuracy of the approximation, which can provide a sampling saving of $\sim 10^3$ in representing a system compared to employing a direct sampling technique.

TABLE 10: Comparison between Second-Order RS-HDMR Approximations Whose Component Functions Were Obtained from Different Sample Sizes *N* and Polynomial Approximations

		data portion (%)[a]	
sample size (*N*)	relative error (%)	*P*	*D*
1000	5	77.2	70.7
	10	90.0	85.9
	20	96.9	94.7
3000	5	85.3	87.9
	10	94.5	94.8
	20	98.5	98.6
5000	5	85.7	90.9
	10	94.5	97.2
	20	98.6	99.6

[a] The percentage of 53 312 data with a relative error not larger than a given value.

Thus, analytical basis function approximations with regularization for RS-HDMR component functions form a practical approach for the application of RS-HDMR.

Acknowledgment. We acknowledge support from the Air Force Office of Scientific Research, the Hercules Corporation, and the Petroleum Research Fund of American Chemical Society.

References and Notes

(1) Rabitz, H.; Alis, O. F.; Shorter J.; Shim, K. Efficient Input-Output Model Representations. *Comput. Phys. Commun.* **1998**, *115*, 1−10.

(2) Shim, K.; Rabitz, H. Independent and Correlated Composition Behavior of Material Properties: Application to Energy Band Gaps for the $Ga_\alpha In_{1-\alpha}P_\beta As_{1-\beta}$ and $Ga_\alpha In_{1-\alpha}P_\beta Sb_\gamma As_{1-\beta-\gamma}$ Alloys. *Phys. Rev. B* **1998**, *58*, 1940−1946.

(3) Alis, O. F.; Rabitz, H. General Foundations of High Dimensional Model Representations. *J. Math. Chem.* **1999**, *25*, 197−233.

(4) Shorter J.; Rabitz, H. An Efficient Chemical Kinetics Solver Using High Dimensional Model Representations. *J. Phys. Chem. A* **1999**, *103*, No. 36, 7192−7198.

(5) Alis, O. F.; Rabitz, H. Efficient Implementation of High Dimensional Model Representations. To appear in *Mathematical and Statistical Methods for Sensitivity Analysis*; Saltelli, A., Ed.; John Wiley and Sons: New York, 2000.

(6) Press: W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN*; Cambridge University Press: New York, 1992; p 299−319.

(7) Sobol, I. M. Sensitivity Estimates for Nonlinear Mathematical Models, *Math. Model. Computational Experiments* **1993**, *1*, 407−414.

(8) Wang, S. W.; Levy, H., II.; Li, G.; Rabitz, H. Fully Equivalent Operational Models for Atmospheric Chemical Kinetics Within Global Chemistry-Transport Models. *J. Geophys. Res.* **1999**, *104*, D23, 30417−30426.

(9) Prenter, P. M. *Splines and Variational Methods*; John Wiley & Sons: New York, 1989; p 77−115.

(10) Li, G.; Wang, S. W.; Rabitz, H.; Wang S. K.; Jaffé, P. Global Uncertainty Assessments by High Dimensional Model Representations (HDMR). *Chem. Eng. Sci.*, in press.