THE JOURNAL OF
PHYSICAL
CHEMISTRY A

VOLUME 105, NUMBER 33, AUGUST 23, 2001

FEATURE ARTICLE

High Dimensional Model Representations

Genyuan Li,[†] Carey Rosenthal,[‡] and Herschel Rabitz*[,†]

Department of Chemistry, Princeton University, Princeton, New Jersey 08544, and Department of Chemistry, Drexel University, Philadelphia, Pennsylvania 19104

Received: February 6, 2001; In Final Form: June 1, 2001

In the chemical sciences, many laboratory experiments, environmental and industrial processes, as well as modeling exercises, are characterized by large numbers of input variables. A general objective in such cases is an exploration of the high-dimensional input variable space as thoroughly as possible for its impact on observable system behavior, often with either optimization in mind or simply for achieving a better understanding of the phenomena involved. An important concern when undertaking these explorations is the number of experiments or modeling excursions necessary to effectively learn the system input → output behavior, which is typically a nonlinear relationship. Although simple logic suggests that the number of runs could grow exponentially with the number of input variables, broadscale evidence indicates that the required effort often scales far more comfortably. This paper considers an emerging family of high dimensional model representation concepts and techniques capable of dealing with such input → output problems in a practical fashion. A summary of the state of the subject is presented, along with several illustrations from various areas in the chemical sciences.

1. Background

This paper is concerned with a common situation associated with either the performance of experiments or the modeling of chemical/physical systems where there are large numbers of input variables accessible for alteration. The latter alteration of the input variables may be done with some design strategy in mind, or it may occur randomly due to natural uncontrolled variations in the input. In either circumstance, a frequent goal is to perform as many runs as possible, aiming at an exploration of the input variable space with respect to its impact on one or more system observables of interest. Such exercises may be performed either to gain a physical understanding of the role of the input variables, or often, ultimately for purposes of optimization to achieve one or more desired physical objectives by special choice of the input variables. This paper addresses the use of high dimensional model representation (HDMR) for making such efforts more feasible.

Before dealing with the technical issues involved with high dimensional (i.e., systems with large numbers of input variables) input → output (IO) mapping at the heart of HDMR, consideration of some typical examples is useful. Cases of high dimensional IO mappings abound, and the discussion here does not aim to be thorough; further consideration of HDMR applications will be returned to later in section 3. As a first example, many complex materials are specified by input variables that consist of the chemical components prescribing the substance of interest. In a mixture formulation, these input variables may be expressed as the mole fractions of the chemical species, where the observables are one or more properties of the mixture (e.g., the viscoelastic properties of a polymer blend). Such mixture problems are complex, as typically, the properties of the mixture are not just a linear combination of the properties of the input chemical components. A similar problem arises in chemical kinetics, where the input variables are the initial chemical concentrations, and the outputs are the concentrations at some latter time. Molecular materials (i.e., a sample consisting of a single type of molecule) encompass many applications, including mutated proteins and pharmaceuticals. For a molecular

[*] To whom correspondence should be addressed.
[†] Department of Chemistry, Princeton University.
[‡] Department of Chemistry, Drexel University.

10.1021/jp010450t CCC: $20.00 © 2001 American Chemical Society
Published on Web 07/26/2001

material, the *i*th variable may be associated with the *i*th site for chemical functionalization on a reference molecular structure. In the case of a protein subject to amino acid mutation, the total number of variables (i.e., sites for mutation) can be very large, and variable x_i associated with protein backbone site i may take on up to 20 values over the naturally occurring amino acids. In contrast, pharmaceutical molecules are typically of modest-size, generated by functionalization of a small number of sites on a reference chemical scaffold. For pharmaceuticals, the *i*th site variable could take on a large number of values, as a rather arbitrary set of chemical moieties may be considered for substitution on a suitable molecular scaffold. Molecular materials inherently differ from those of mixture formulations, as molecular moiety input variables are discrete (e.g., methyl-, ethyl-, chloro-, etc.), while the component mole fractions as input variables in mixtures can take on continuous values. Mixture materials drawn from a large set of possible molecular species have both discrete and continuous variables. All of these material problems, characterized by either large numbers of input variables or large numbers of variable values, has led to much interest in high throughput synthesis and screening techniques in an attempt to deal with the potentially exploding number of samples that may be considered.

Another class of problems with high dimensional input occurs in molecular modeling, where the inter- or intramolecular potential surface as an input *function* dictates all relevant dynamical evolution and properties, leading to one or more observables such as cross sections, rate constants, etc. In this case, the input potential function lies in a space that, in principle, is described by an infinite number of variables, or more practically, large numbers of discretized variables that define a realistic family of potential surfaces. A physically distinct, but mathematically analogous, input → output mapping problem involves solar radiation transport through the atmosphere, where the input consists of the column densities of various trace gases and the atmospheric temperature as a function of altitude, and the output is the atmospheric heating rate (i.e., as a function of the altitude) due to net solar radiation absorption. This application is of relevance to global warming, atmospheric kinetics, and general weather modeling. In this case, as for the previous example with potential surfaces, it would be natural to discretize the input functions to form a set of variables representing reasonable spatial resolution. In this context an input variable is, for example, the value of an input column density at a particular altitude.

A common characteristic of all the aforementioned illustrations, and many others, is the large number of variables that may naturally arise to describe the input. The notion of "large" in this context depends on the particular application, and especially the difficulty of either appropriately observing or calculating the system output corresponding to any single specification of all of the input variables. For example, in the case of semiconductor materials (e.g., the quaternary compound $Ga_xIn_{1-x}As_ySb_{1-y}$ which is of two dimensions, x and y), the system dimension is low, but the time and cost for synthesizing even a single sample can be quite high. The search for pharmaceuticals is generally of a similar nature involving low numbers of variables (i.e., the number of sites for functionalization on a molecular scaffold), but the number of moiety values for each of these variables can be very large, ranging up to 10^2 or more. In this case, making one potential pharmaceutical molecule may be easy, but making all relevant possibilities gets out of hand. In other problems, the number of variables involved can be inherently large, and one example occurs when the input is a function and good sampling resolution is required, thereby leading to hundreds or more discretized input variables.

Regardless of the circumstance, these IO problems become challenging, as the numbers of samples (i.e., whether experiments or modeling calculations) over the input variable space rise to reveal the essential structure of the IO mapping. The potential scope of such problems may be understood for a system with n variables $\mathbf{x} = (x_1, x_2, ..., x_n)$. For simplicity, here we may consider each variable x_i to run over the same number of discrete values $(x_{i_1}, x_{i_2}, ..., x_{is_i})$, such that $s_i = s$, for all i, and it is evident that the input variable space is covered by a grid of s^n points. Considering that s could be 10 or more in many cases, and that n could even be hundreds in some systems, reveals the very daunting task of thoroughly exploring typical realistic input variable spaces in chemical/physical systems. This apparent exponential growth in exploration effort is sometimes referred to as the "curse of dimensionality", and it produces a longstanding function mapping problem $\mathbf{x} \rightarrow f$ where $f(\mathbf{x}) = f(x_1, x_2, ..., x_n)$, $n \gg 1$, and $f(\mathbf{x})$ is a system observable output of interest. Problems of this nature occur in virtually all areas of science and engineering, as well as other disciplines where similar levels of possible exponential difficulty loom as roadblocks to the exploration of the full input variable space to an acceptable degree of resolution and quality.

The fundamental question is whether the effort of determining physically based IO mappings is expected to typically scale exponentially in difficulty, or possibly in a more attractive, less dramatic fashion. Fortunately, there is ample evidence suggesting that much more reasonable scaling should often exist as the number of variables n rises. This fortunate circumstance seems to occur ubiquitously for a number of reasons. In the limit of a totally irregular IO mapping $\mathbf{x} \rightarrow f$, then an exponentially growing number of samples s^n will be needed. The notion of irregular here means that every point $\mathbf{x}$ could produce distinctly unrelated output behavior to that of other even nearby points. Thus, it is crucial whether the system $f(\mathbf{x})$ contains identifiable regular structure with respect to the space of variables $\mathbf{x} = (x_1, x_2, ..., x_n)$. In practice, highly irregular IO maps often do not occur, and to appreciate this comment, it is most useful to view the breadth of IO behavior in terms of the degree of cooperativity (or correlations) among the input variables $\mathbf{x} = (x_1, x_2, ..., x_n)$ for their impact upon $f(\mathbf{x})$. In this sense, a hierarchy can be envisioned, starting with the variables acting independently (but still possibly nonlinearly) of each other for their impact on f, and then in all possible pairs for their impact, in all possible triples for their impact, etc., finally to the highest *n*th level of inseparable cooperativity among all of the input variables. The extent of high order variable cooperativity depends on the choice of input variables; there is much freedom in this choice for any system. In any application, there is a natural predilection to choosing the chemical/physical input variables so that they act as independently as possible, and this fact alone naturally leans toward the existence of a limited degree of input variable cooperativity upon f in the hierarchy discussed above.

Perhaps the strongest generic evidence for the typical lack of high order input variable cooperativity can be found in the overwhelming body of statistical analysis data of many systems, where it is rarely found that more than covariances (i.e., cooperativity order $l = 2$) are necessary to describe the input multivariate contributions to virtually any system output. In the extreme limit of the input variables acting totally independently (i.e., cooperativity order $l = 1$), although not necessarily linearly, the number of runs or experiments necessary to learn the IO map of the system will scale as $\sim ns$, and similarly, at $l = 2$, the scaling will be $\sim(ns)^2$. Thus, the degree of cooperativity among the input variables is crucial for determining the true scaling or complexity required to learn system IO response behavior. As a special case, chemical systems described at the atomic and molecular level are naturally dominated by few-body interactions, which is fully consistent with the same low order variable cooperativity behavior. Perhaps, most surprising is the apparently ubiquitous dominance of low order input variable cooperativity across science and engineering, at length scales beyond atomic dimensions.

Every chemical/physical problem will have its own characteristics in terms of the degree of variable cooperativity present, and the premise of HDMR is that realistic well-defined systems are dominated by low order behavior such that $l \ll n$. Accepting this premise, the key issue is how to exploit this behavior for translation into specific algorithms to guide the taking of laboratory or simulation data and the representation of the system IO in a physically transparent and quantitatively convenient fashion. In this manner, for $l \ll n$, the goal is to perform a modest number of experiments or model runs while still retaining full fidelity of the IO map throughout the input variable space. This problem, addressed by HDMR, can be viewed as interpolation of the system output throughout the input variable space, which is possibly of very high dimension n. A broad literature exists on function representations,[1-10] and the present paper will focus on HDMR, as ample reviews exist including a discussion on the background to HDMR. Section 2 below will discuss the mathematical and algorithmic IO representation problem, followed by a summary of various aspects of HDMR. Section 3 will present several chemical/physical illustrations of HDMR on problems ranging from dimension $n = 2$ to $n = 1000$ in order to give a sense of the scope and power of the concepts. The motivation in all applications of HDMR ultimately reduces to matters of efficiency and speed, both for learning IO mappings as well as exploiting them for subsequent optimization or other purposes. Section 4 will present some summarizing comments on this topic, which is still under active development.

2. HDMR Formulations

Many problems in science and engineering reduce to finding an efficiently constructed map of the relationship between sets of 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 variables would be the solutions to the model or a functional of them. The 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 $\sim 10^2 - 10^3$ 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.

2.1. Theoretical Foundations of HDMR.

High dimensional model representation (HDMR) is a general set of quantitative model assessment and analysis tools[11-15] for capturing high dimensional IO system behavior. As the impact of the multiple input variables on the output can be independent and cooperative, it is natural to express the model output $f(\mathbf{x})$ as a finite hierarchical correlated function expansion in terms of the input variables:

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

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. The above HDMR expansion has a finite number of terms and is always exact. Other expansions have been suggested,[16] but they commonly have an infinite number of terms with some specified functions (e.g., Hermite polynomials).

The basic conjecture underlying HDMR is that the component functions in eq 1 arising in typical real problems are likely to exhibit only low order l cooperativity among the input variables such that the significant terms in the HDMR expansion are expected to satisfy the relation: $l \ll n$ for $n \gg 1$. Experience shows that an HDMR expansion to second order

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

often provides a satisfactory description of $f(\mathbf{x})$ for many high dimensional systems when the input variables are properly chosen. HDMR attempts to exploit this observation to efficiently determine high-dimensional IO system mapping. The presence of only low order variable cooperativity does not necessarily imply a small set of significant variables nor does it limit the nonlinear nature of the IO relationship. Figure 1 gives an example of typical first and second-order HDMR component functions which reveal the nonlinear relationships between model inputs and outputs.

2.1.1. Optimization Procedures to Determine HDMR Component Functions. Exploiting the expected low order variable cooperativity in high dimensional systems can only be done if practical formulations of the HDMR component functions can be found. The HDMR expansion component functions f_0, $f_i(x_i)$, $f_{ij}(x_i, x_j)$, ... are optimally tailored to each particular $f(\mathbf{x})$ over the entire domain Ω of $\mathbf{x}$. A component function $f_{i_1 i_2 \cdots i_l}(x_{i_1}, x_{i_2}, ..., x_{i_l})$ ($l = 0, 1, ..., n - 1$ with f_0 corresponding to $l = 0$) is obtained by minimizing the functional

$$\min_{f_{i_1 i_2 ... i_l}} \int_\Omega w_{i_1 i_2 \cdots i_l}(\hat{\mathbf{x}}, \mathbf{u})[f(\mathbf{u}) - f_0 - \sum_{i=1}^{n} f_i(u_i) - \sum_{1\le i<j\le n} f_{ij}(u_i, u_j) - ... - \sum_{1\le i_1<...<i_l\le n} f_{i_1 i_2 \cdots i_l}(u_{i_1}, u_{i_2}, ..., u_{i_l})]^2 \,\mathrm{d}\mathbf{u} \quad (3)$$

under a suitable specified orthogonality condition which guar-

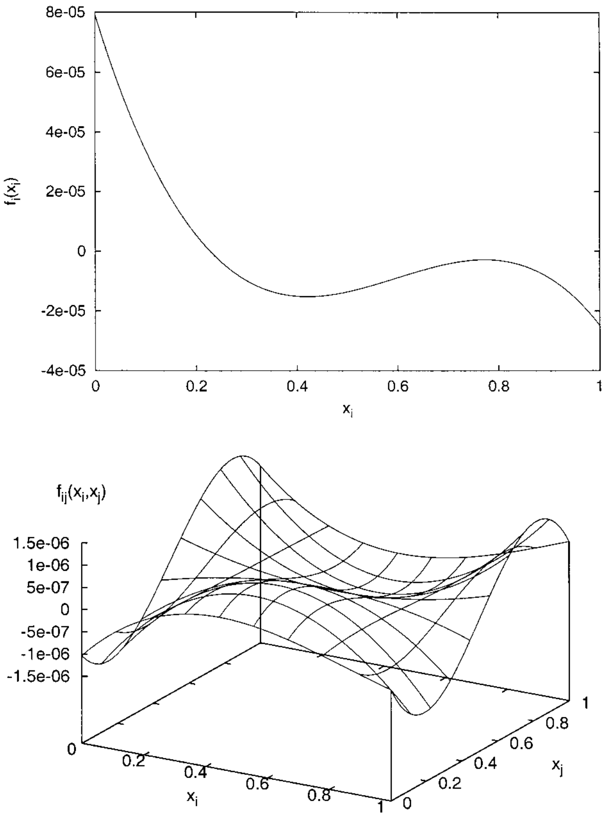

Figure 1. The functional behavior of typical (above) first- and (below) second-order HDMR component functions from a bioremediation model for a uranium soil contamination site.[17] Here the input variables x_i and x_j are rate constants in the model, and the output f is the accumulated flux of uranium U^{4+}passing through a given depth from the soil surface. This nonlinear behavior is typical of many chemical/physical models.

antees that all the component functions are determined step-by-step. Here, $\hat{\mathbf{x}} = (x_{i_1}, x_{i_2}, \ldots, x_{i_l})$, d$\mathbf{u}$ = du_1du_2...du_n, and $w_{i_1 i_2 \ldots i_l}(\hat{\mathbf{x}},\mathbf{u})$ is a weight function.

Different weight functions will produce distinct, but formally equivalent HDMR expansions, all of the same structure as eq 1. There are two commonly used HDMR expansions: Cut- and RS(Random Sampling)-HDMR. Cut-HDMR expresses $f(\mathbf{x})$ in reference to a specified cut point $\bar{\mathbf{x}}$ in Ω, while RS-HDMR depends on the average value of $f(\mathbf{x})$ over the whole domain Ω.

1. Cut-HDMR. For Cut-HDMR a reference point $\bar{\mathbf{x}}$ is first chosen in the n-dimensional input variable $\mathbf{x}$ space. When Cut-HDMR is taken to convergence, the representation of $f(\mathbf{x})$ is invariant to the choice of $\bar{\mathbf{x}}$. In practical circumstances it can be wise to choose $\bar{\mathbf{x}}$ within the neighborhood of interest in the input space.

The Cut-HDMR component functions with respect to reference point $\bar{\mathbf{x}}$ have the following forms:

$$f_0 = f(\bar{\mathbf{x}}) \tag{4}$$

$$f_i(x_i) = f(x_i, \bar{\mathbf{x}}^i) - f_0 \tag{5}$$

$$f_{ij}(x_i, x_j) = f(x_i, x_j, \bar{\mathbf{x}}^{ij}) - f_i(x_i) - f_j(x_j) - f_0 \tag{6}$$

$$\ldots\ldots$$

where

$$(x_i, \bar{\mathbf{x}}^i) = (\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_n) \tag{7}$$

$$(x_i, x_j, \bar{\mathbf{x}}^{ij}) = (\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_{j-1}, x_j, \bar{x}_{j+1}, \ldots, \bar{x}_n) \tag{8}$$

The last term $f_{12\ldots n}(x_1, x_2, \ldots, x_n)$ in eq 1 is determined by the difference between $f(\mathbf{x})$ and all other Cut-HDMR component functions.

The above formulas can be readily obtained from eq 3 or simply by substituting $(x_{i_1}, x_{i_2}, \ldots, x_{i_l}, \bar{\mathbf{x}}^{i_1 i_2 \ldots i_l})$ with different sets of $\{i_1, i_2, \ldots, i_l\} \subset \{1, 2, \ldots, n\}$ for $\mathbf{x}$ on the both sides of eq 1 and using the following specified condition: a component function of Cut-HDMR vanishes when any of its own variables takes the value of the corresponding element in $\bar{\mathbf{x}}$, i.e.,

$$f_{i_1 i_2 \ldots i_l}(x_{i_1}, x_{i_2}, \ldots, x_{i_l})|_{x_s = \bar{x}_s} = 0, \quad s \in \{i_1, i_2, \ldots, i_l\} \tag{9}$$

Equation 9 serves to define an orthogonal relation between two different component functions of Cut-HDMR as

$$f_{i_1 i_2 \ldots i_l}(x_{i_1}, x_{i_2}, \ldots, x_{i_l}) f_{j_1 j_2 \ldots j_k}(x_{j_1}, x_{j_2}, \ldots, x_{j_k})|_{x_s = \bar{x}_s} = 0$$
$$s \in \{i_1, i_2, \ldots, i_l\} \cup \{j_1, j_2, \ldots, j_k\} \tag{10}$$

The Cut-HDMR component functions $f_i(x_i)$, $f_{ij}(x_i, x_j)$, ... are typically attained 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 of eqs 4−6. Note that the Cut-HDMR component functions are defined along *cut* lines, planes, subvolumes, etc., through the reference point $\bar{\mathbf{x}}$ in Ω.

Since all the Cut-HDMR component functions satisfy a minimization condition in eq 3, they are optimal choices for a given output $f(\mathbf{x})$. Experience shows that often only low order terms of Cut-HDMR are needed to give a good approximation for $f(\mathbf{x})$. Numerical data tables can be constructed for these component functions, and the values of $f(\mathbf{x})$ for an arbitrary point $\mathbf{x}$ can be determined from these tables by performing only low dimensional interpolation over $f_i(x_i)$, $f_{ij}(x_i, x_j)$, ... If each input variable is sampled at s different values (with the cut center being one evaluation point), the required number of model runs to construct the $f_i(x_i)$, $f_{ij}(x_i, x_j)$ ... tables is

$$1 + n(s - 1) + \frac{n(n - 1)(s - 1)^2}{2} + \ldots$$

which grows only polynomically with n and s. The sample savings for large n are significant compared to traditional s^n sampling. Thus, Cut-HDMR renders the original exponential difficulty to a problem of only polynomic complexity.

2. RS-HDMR. For RS-HDMR, we first rescale variables x_i such that $0 \le x_i \le 1$ for all i. The output function $f(\mathbf{x})$ is then defined in the unit hypercube $K^n = \{(x_1, x_2, \ldots, x_n)| 0 \le x_i \le 1, i = 1, 2, \ldots, n\}$ by suitable transformations. The component functions of RS-HDMR possess the following forms:

$$f_0 = \int_{K^n} f(\mathbf{u})\, \mathrm{d}\mathbf{u} \tag{11}$$

$$f_i(x_i) = \int_{K^{n-1}} f(x_i, \mathbf{u}^i)\, \mathrm{d}\mathbf{u}^i - f_0 \tag{12}$$

$$f_{ij}(x_i, x_j) = \int_{K^{n-2}} f(x_i, x_j, \mathbf{u}^{ij})\, \mathrm{d}\mathbf{u}^{ij} - f_i(x_i) - f_j(x_j) - f_0 \tag{13}$$

$$\ldots\ldots$$

where d$\mathbf{u}^i$ and d$\mathbf{u}^{ij}$ are just the product du_1du_2...du_n without du_i and du_idu_j, respectively. Finally, the last term $f_{12\ldots n}(x_1, x_2, \ldots, x_n)$ is determined from the difference between $f(\mathbf{x})$ and all the other lower order component functions in eq 1.

Considering that the domain Ω is a unit hypercube, f_0 is the average value of $f(\mathbf{x})$ over the whole domain in contrast with f_0 of Cut-HDMR, which is the value of $f(\mathbf{x})$ at the specified single reference point $\bar{\mathbf{x}}$.

All the above formulas can be readily obtained from eq 3 or simply by integrating both sides of eq 1 with respect to different sets of input variables $\{x_{i_1}, x_{i_2}, ..., x_{i_l}\}$ ($l = n, n - 1, ..., 1$), and using the following specified condition: the integral of a component function of RS-HDMR with respect to any of its own variables is zero, i.e.,

$$\int_0^1 f_{i_1 i_2 \ldots i_l}(x_{i_1}, x_{i_2},..., x_{i_l})\, \mathrm{d}x_s = 0, \quad s \in \{i_1, i_2, ..., i_l\} \tag{14}$$

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

$$\int_{K^n} f_{i_1 i_2 \ldots i_l}(x_{i_1}, x_{i_2},..., x_{i_l}) f_{j_1 j_2 \ldots j_k}(x_{j_1}, x_{j_2},..., x_{j_k})\, \mathrm{d}\mathbf{x} = 0$$

$$\{i_1, i_2, ..., i_l\} \neq \{j_1, j_2, ..., j_k\} \tag{15}$$

Evaluation of the high dimensional integrals in the RS-HDMR expansion may be carried out by Monte Carlo random sampling,[18] and hence the name RS(random sampling)-HDMR.

According to the above formulas one can see that all the component functions of the Cut- and RS-HDMR expansions can be directly constructed from the values of output $f(\mathbf{x})$ either at ordered or randomly sampled points of $\mathbf{x}$, which makes the determination of f_0, $f_i(x_i)$, $f_{ij}(x_i, x_j)$, ... straightforward.

2.1.2. HDMR Component Functions Obtained from Orthogonal Projection Operators. To attain a better understanding of the HDMR expansions, we may view the concept from another perspective. The component functions of an HDMR can be obtained through application of a suitably defined set of linear operators $\mathcal{P}_0$, $\mathcal{P}_i$ ($i = 1, 2, ..., n$), $\mathcal{P}_{ij}$($1 \le i < j \le n$), ...

$$\mathcal{P}_0\, f(\mathbf{x}) = f_0 \tag{16}$$

$$\mathcal{P}_i\, f(\mathbf{x}) = f_i(x_i) \tag{17}$$

$$\mathcal{P}_{ij}\, f(\mathbf{x}) = f_{ij}(x_i, x_j) \tag{18}$$

Equations 4−6 and eqs 11−13 reveal the corresponding definitions of the operators for the Cut- and RS-HDMR component functions, respectively. It has been proven that all the operators for the Cut- and RS-HDMR expansions are commutative projection operators and they are mutually orthogonal to one another,[12] i.e., they obey 1. idempotency,

$$\mathcal{P}^2_{i_1 i_2 \ldots i_l} = \mathcal{P}_{i_1 i_2 \ldots i_l},\ \{i_1, i_2, ..., i_l\} \subset \{1, 2, ..., n\} \tag{19}$$

where $0 \le l \le n$, and $\mathcal{P}_0$ corresponds to $l = 0$; 2. orthogonality,

$$\mathcal{P}_{i_1 i_2 \ldots i_l}\mathcal{P}_{j_1 j_2 \ldots j_k} = 0.\ \{i_1, i_2, ..., i_l\} \neq \{j_1, j_2, ..., j_k\} \tag{20}$$

and 3. resolution of the identity,

$$\sum_{l=0}^{n} \sum_{1 \le i_1 < ... < i_l \le n} \mathcal{P}_{i_1 i_2 \ldots i_l} = \mathbf{1} \tag{21}$$

where $\mathbf{1}$ denotes the identity operator.

The projectors act on a linear space $\mathcal{F}$ composed of all n-variable functions $f(\mathbf{x})$. Each projector $\mathcal{P}_t$ provides an approximation $\mathcal{P}_t f(\mathbf{x})$ for $f(\mathbf{x})$, and has its range Φ_t which is a subspace of the linear space $\mathcal{F}$. Any function $f(\mathbf{x}) \in \Phi_t$ is invariant upon the action of $\mathcal{P}_t$, i.e.,

$$\mathcal{P}_t f(\mathbf{x}) = f(\mathbf{x}), \quad \forall f(\mathbf{x}) \in \Phi_t \tag{22}$$

This implies that upon the action of $\mathcal{P}_t$ there is no error for any function $f(\mathbf{x}) \in \Phi_t$. The larger the range Φ_t is, the better approximation $\mathcal{P}_t$ produces for $\mathcal{F}$.

Two projectors $\mathcal{P}_i$ and $\mathcal{P}_j$ are mutually orthogonal if

$$\mathcal{P}_i\mathcal{P}_j = \mathcal{P}_j\mathcal{P}_i = 0 \tag{23}$$

This is equivalent to

$$\Phi_i \cap \Phi_j = 0 \tag{24}$$

A sum of two mutually orthogonal projectors $\mathcal{P}_i + \mathcal{P}_j$ is also a projector whose range is $\Phi_i + \Phi_j$ which is larger than either Φ_i or Φ_j. Therefore, $\mathcal{P}_i + \mathcal{P}_j$ will produce an approximation for $\mathcal{F}$ with better accuracy than provided by either single operator $\mathcal{P}_i$ and $\mathcal{P}_j$.

Any set of commutative projectors generate a *distributive lattice* whose elements are obtained by all possible combinations (Boolean addition and multiplication) of the projectors in the set.[19] In particular, the lattice has a unique *maximal* projector $\mathcal{M}$ which provides the algebraically best approximation to $\mathcal{F}$. The range of the maximal projector $\mathcal{M}$ for the lattice generated by mutually commutative projectors $\{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_s\}$ is the union of all the ranges Φ_t, i.e.,

$$\Phi_{\mathcal{M}} = \Phi_1 \cup \Phi_2 \cup ... \cup \Phi_s \tag{25}$$

When the projectors are mutually orthogonal, the maximal projector is simply their sum

$$\mathcal{M} = \sum_{i=1}^{s} \mathcal{P}_s \tag{26}$$

and the range $\Phi_{\mathcal{M}}$ is $\sum_{i=1}^{s}\Phi_i$. When more orthogonal projectors are retained in the set, the resultant approximation to $\mathcal{F}$ obtained by its maximal projector $\mathcal{M}$ becomes better.

As an example, if we choose the subset $\mathcal{S}_1 = \{\mathcal{P}_0, \mathcal{P}_i(i = 1, 2, ..., n)\}$ of the above mutually orthogonal projectors to generate a lattice, its maximal projector is simply the sum of all these projectors:

$$\mathcal{M}_1 = \mathcal{P}_0 + \sum_{i=1}^{n} \mathcal{P}_i \tag{27}$$

and the best approximation of $f(\mathbf{x}) \in \mathcal{F}$ by the projectors in this lattice is

$$f(\mathbf{x}) \approx \mathcal{M}_1 f(\mathbf{x}) = \mathcal{P}_0 f(\mathbf{x}) + \sum_{i=1}^{n} \mathcal{P}_i f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) \tag{28}$$

which is the first-order HDMR approximation for $f(\mathbf{x})$. Similarly, for the subset $\mathcal{S}_2 = \{\mathcal{P}_0, \mathcal{P}_i(i = 1, 2, ..., n), \mathcal{P}_{ij}(1 \le i < j \le n)\}$, the best approximation of $f(\mathbf{x}) \in \mathcal{F}$ is given by

$$f(\mathbf{x}) \approx \mathcal{M}_2 f(\mathbf{x}) = \mathcal{P}_0 f(\mathbf{x}) + \sum_{i=1}^{n} \mathcal{P}_i f(\mathbf{x}) \sum_{1 \le i < j \le n} \mathcal{P}_{ij} f(\mathbf{x}) =$$

$$f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_i, x_j) \tag{29}$$

which is the second-order HDMR approximation for $f(\mathbf{x})$, and so on.

As $\mathcal{S}_1$ is a subset of $\mathcal{S}_2$, and $\mathcal{M}_2$ is the maximal projector in the lattice generated by $\mathcal{S}_2$, then $\Phi_{\mathcal{M}_1} \subset \Phi_{\mathcal{M}_2}$ and $\mathcal{M}_2$ is better than $\mathcal{M}_1$, i.e., the second-order approximation of HDMR is better than the first-order one. This implies that adding a new orthogonal projector into a sum of orthogonal projectors always produces a new projector with better accuracy. Finally, as the sum of all projectors of the HDMR expansion is the identity, the full HDMR expansion is exactly equal to $f(\mathbf{x})$.

It can be readily proven that the range for projector $\mathcal{P}_0 + \mathcal{P}_i$ is any constant and any function of variable x_i, and the range for $\mathcal{P}_0 + \sum_{i=1}^{n} \mathcal{P}_i$ is any constant and any linear combination of functions with one variable $x_i (i = 1, 2, ..., n)$. Similarly, the range for projector $\mathcal{P}_0 + \sum_{i=1}^{n} \mathcal{P}_i + \sum_{1 \le i < j \le n} \mathcal{P}_{ij}$ is any constant and any linear combination of functions with one or two variables x_i, x_j $(1 \le i < j \le n)$. This property can be employed to identify error free regions in $\mathbf{x}$ space for different order Cut-HDMR approximations.[12,13] As $f(x_i, \bar{\mathbf{x}}^i)$ is a one variable function, it is invariant to the projector $\mathcal{P}_0 + \sum_{i=1}^{n} \mathcal{P}_i$, i.e., there is no error for the first order Cut-HDMR approximation of $f(\mathbf{x})$ whenever the point $\mathbf{x}$ is located on a cut line through the reference point $\bar{\mathbf{x}}$ in Ω. Similarly, $f(x_i, x_j, \bar{\mathbf{x}}^{ij})$ is a two variable function, and thus there is no error for the second order Cut-HDMR approximation of $f(\mathbf{x})$ whenever the point $\mathbf{x}$ is located on any cut line or plane through the reference point $\bar{\mathbf{x}}$ in Ω. In summary, there is no error for the lth order Cut-HDMR approximation of $f(\mathbf{x})$ whenever the point $\mathbf{x}$ is located in any $k(k \le l)$-dimensional subvolume across the reference point $\bar{\mathbf{x}}$ in Ω.

2.2. Properties of HDMR Expansions. *2.2.1. Fast Convergence of HDMR Expansions.* As mentioned above, HDMR expansions have been observed to converge fast in realistic applications. The origin of this property can be understood from the following analysis. Suppose an output $f(\mathbf{x})$ defined in a unit hypercube of $\mathbf{x}$ can be expanded as a convergent Taylor series at reference point $\bar{\mathbf{x}}$, i.e.,

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \sum_{i=1}^{n} \frac{\partial f(\bar{\mathbf{x}})}{\partial x_i}(x_i - \bar{x}_i) + \sum_{i,j=1}^{n} \frac{1}{2!} \frac{\partial^2 \mathrm{f}(\bar{\mathbf{x}})}{\partial x_i \partial x_j}(x_i - \bar{x}_i)(x_j - \bar{x}_j) + ... \quad (30)$$

This Taylor expansion can be used to give clear mathematical meaning to the Cut-HDMR component functions. According to the definitions of f_0, $f_i(x_i)$, $f_{ij}(x_i, x_j)$, ... given in eqs 4−6, it is easy to prove that $f_0 = f(\bar{\mathbf{x}})$, i.e., the constant term of the Taylor series. Since $f_i(x_i) = f(x_i, \bar{\mathbf{x}}^i) - f(\bar{\mathbf{x}})$, substituting $(x_i, \bar{\mathbf{x}}^i)$ for $\mathbf{x}$ and subtracting $f(\bar{\mathbf{x}})$ from the both sides of eq 30 gives $f_i(x_i)$. As all the terms containing $x_j (j \ne i)$ vanish, the first-order component function $f_i(x_i)$ is the sum of *all* the Taylor series terms which only contain variable x_i. Similarly, the second-order component function $f_{ij}(x_i, x_j)$ is the sum of *all* the Taylor series terms which only contain both variables x_i and x_j, etc. Thus, the infinite number of terms in the Taylor series are partitioned into a finite number of distinct groups, and each group (still containing an infinite number of terms) corresponds to one Cut-HDMR component function, i.e., each component function of Cut-HDMR is composed of an infinite subclass of the full multidimensional Taylor series. Therefore, a truncated Cut-HDMR expansion is likely to give a better approximation of $f(\mathbf{x})$ than any truncated Taylor series because the latter only contains a finite number of terms. Furthermore, considering that $0 \le x_i \le 1$ $(i = 1, 2, ..., n)$ and $(x_i - \bar{x}_i) < 1$, the high order Cut-HDMR component functions are usually smaller than low order ones because the high order component functions involve the product $\prod_{s=1}^{l} (x_{i_s} - \bar{x}_{i_s})^{k_s}$ with larger l. Moreover, if the higher order derivatives have complex structure, especially with sign changes over their indices $\bar{x}_i$, $\bar{x}_j$, ..., then random phase arguments further suggest that the higher order HDMR component functions will tend to be small. Evidence supports this qualitative behavior, although there is no rigorous proof of the behavior.

Each of the subclasses of the Taylor series corresponding to different component functions of Cut-HDMR do not overlap one another, which is the basis for the orthogonal relation between two Cut-HDMR component functions. Other HDMR expansions possess the same property as Cut-HDMR because a one-to-one relationship between two different HDMR expansions can be established. Thus, if Cut-HDMR converges at certain order, so do the other HDMR expansions.[12]

2.2.2. Invariance of Conservation Laws for HDMR Approximations. If a set of physical outputs $\{f^{(1)}(\mathbf{x}), f^{(2)}(\mathbf{x}), ..., f^{(s)}(\mathbf{x})\}$ obey a set of linear-superposition conservation laws, their HDMR approximations at any order also obey these conservation laws,[12] i.e., if

$$\sum_{i=1}^{s} w_{ki} f^{(i)}(\mathbf{x}) = c_k, \quad k = 1, 2,..., m \quad (31)$$

where $\{w_{ki}\}$ and $\{c_k\}$ are two sets of constants, then

$$\sum_{i=1}^{s} w_{ki} [\mathcal{M}_l f^{(i)}(\mathbf{x})] = c_k, \quad k = 1, 2, ..., m; \; l = 0, 1, ..., n \quad (32)$$

here

$$\mathcal{M}_l = \mathcal{P}_0 + \sum_{i=1}^{n} \mathcal{P}_i + ... + \sum_{1 \le i_1 < ... < i_l \le n} \mathcal{P}_{i_1} \mathcal{P}_{i_2} ... \mathcal{P}_{i_l} \quad (33)$$

and $\mathcal{M}_l f^{(i)}(\mathbf{x})$ denotes the lth order HDMR approximation for $f^{(i)}(\mathbf{x})$. This property can be proven by applying operator $\mathcal{M}_l$ to the both sides of eq 31 and using the identity

$$\mathcal{M}_l c = c, \quad \text{c being a constant} \quad (34)$$

The invariance of conservation laws is very useful for the application of HDMR in physics, chemistry and other disciplines where conservation laws (e.g., mass, energy, momentum conservations, etc.) are important.

2.2.3. Decomposition of System Variance by RS-HDMR. Using the orthogonality property of the RS-HDMR component functions, it can be proven that the total variance σ_{f}^2 of $f(\mathbf{x})$ caused by all input variables sampled uniformly over their full range may be decomposed into distinct input contributions in the following manner.[12]

$$\begin{aligned}\sigma_{\mathrm{f}}^2 &= \int_{K^n} [f(\mathbf{x}) - \bar{f}]^2 \, \mathrm{d}\mathbf{x} = \int_{K^n} [f(\mathbf{x}) - f_0]^2 \, \mathrm{d}\mathbf{x} = \\ &\int_{K^n} [\sum_{i=1}^{\mathrm{n}} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_i, x_j) + ...]^2 \, \mathrm{d}\mathbf{x} \\ &= \sum_{i=1}^{n} \int_0^1 f_i^2(x_i) \, \mathrm{d}x_i + \sum_{1 \le i < j \le n} \int_0^1 \int_0^1 f_{ij}^2(x_i, x_j) \, \mathrm{d}x_i \, \mathrm{d}x_j + ... \\ &= \sum_{i=1}^{n} \sigma_i^2 + \sum_{1 \le i < j \le n} \sigma_{ij}^2 + ... \end{aligned} \quad (35)$$

where $\bar{f}$ is the mean value of $f(\mathbf{x})$ over the whole domain Ω. Thus, the total variance σ_f^2 is the sum of first-order variances

σ_i^2, second-order covariances σ_{ij}^2, etc. This property is useful for global uncertainty analysis because the above decomposition is valid over the whole domain. The magnitudes of the indices σ_i^2, σ_{ij}^2, etc., reveal how the output uncertainty is influenced by the input uncertainties and the nature of the cooperativities that exist. These multivariate indices are nonlinear analogues of the usual statistical moments for multivariance analysis. The fact that covariances are often adequate to describe multivariate system statistics is also supportive of the fast convergence of the HDMR expression.

2.3. Approximate and Extended HDMR. *2.3.1 Approximate Formulas for RS-HDMR Component Functions.* The direct determination of the component functions of RS-HDMR at different values of x_i, x_j, ... by Monte Carlo integration can require a large number of random samples. For instance, distinct Monte Carlo random samples for $f(x_i, \bar{\mathbf{x}}^i)$at different fixed values of x_i are needed to determine $f_i(x_i)$ in eq 12.[20] To reduce the sampling effort, the RS-HDMR component functions may be approximated to any desired level of accuracy by the following two means.

1. Analytical Approximation. The RS-HDMR component functions may be approximated by expansion in terms of a suitable set of functions, such as orthogonal polynomials, spline functions, or even simply polynomial functions,[15]

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r P_r(x_i) \tag{36}$$

$$f_{ij}(x_i, x_j) \approx \sum_{r=1}^{l} \sum_{s=1}^{l} \beta_{rs} P_{rs}(x_i, x_j) \tag{37}$$

where α_r, β_{rs} are constant coefficients, and $P_r(x_i)$, $P_{rs}(x_i, x_j)$ are one- and two-variable bases. Assuming that the functions are orthogonal, the coefficients are given by

$$\alpha_r = \frac{\int_{K^n} f(\mathbf{x}) p_r(x_i)\, \mathrm{d}\mathbf{x}}{\int_0^1 p_r^2(x_i)\, \mathrm{d}x_i} \tag{38}$$

$$\beta_{rs} = \frac{\int_{K^n} f(\mathbf{x}) p_{rs}(x_i, x_j)\, \mathrm{d}\mathbf{x}}{\int_0^1 \int_0^1 p_{rs}^2(x_i, x_j)\, \mathrm{d}x_i\, \mathrm{d}x_{\mathrm{j}}} \tag{39}$$

As no restriction is posed on the values of the elements of $\mathbf{x}$ for $f(\mathbf{x})$ in the above integrals, only one set of random samples for $f(\mathbf{x})$ are necessary to determine all the coefficients, and consequently all the component functions of RS-HDMR. The sampling effort is then dramatically reduced.

2. Numerical Approximation. The RS-HDMR component functions may be also approximated numerically by using reproducing kernels or filters. These kernels can be used to reduce the sampling burden as well as to act as a filter with noisy input data $f(\mathbf{x})$. For instance, the first- and second-order RS-HDMR component functions are given by

$$f_i(x_i) = \int_{K^n} f(\mathbf{u}) k(x_i; u_i)\, \mathrm{d}\mathbf{u} - f_0 \tag{40}$$

$$f_{ij}(x_i, x_j) = \int_{K^n} f(\mathbf{u}) k(x_i, x_j; u_i, u_j)\, \mathrm{d}\mathbf{u} - f_i(x_i) - f_j(x_j) - f_0 \tag{41}$$

where $k(x_i; u_i)$ and $k(x_i, x_j; u_i, u_j)$ are reproducing kernels.[21,22] Similarly, as no restriction is posed on the values of the elements of $\mathbf{u}$ for $f(\mathbf{u})$ in the above integrals, only one set of random samples for $f(\mathbf{x})$ are necessary to determine all the component functions of RS-HDMR at different values of the elements of $\mathbf{x}$.

2.3.2. Monomial Preconditioning Cut-HDMR. As argued earlier, very often the high order HDMR terms are small thereby making low (usually, first and second) order HDMR approximations satisfactory for practical purposes. However, in some cases the first- or second-order HDMR approximations may not provide the desired accuracy, and higher order HDMR approximations might have to be considered. For Cut-HDMR, the higher order terms demand a polynomically increasing number of data samples and possibly large computer storage. If the higher order component functions of Cut-HDMR can be approximately represented in a similar fashion as those for the zeroth-, first-, and second-order component functions, then higher order approximations of Cut-HDMR can be included without dramatically increasing the number of experiments or model runs as well as reducing computer storage requirements. One way to realize this concept is to represent a high order Cut-HDMR component function as products of low order Cut-HDMR component functions and some suitable functions of the remaining input variables. For instance, a third-order Cut-HDMR component function can be approximated as

$$\begin{aligned} f_{ijk}(x_i, x_j, x_k) \approx{}& \varphi_{ijk}(x_i, x_j, x_k)\, \bar{f}_0 + \varphi_{jk}(x_j, x_k)\, \bar{f}_i(x_i) + \\ & \varphi_{ik}(x_i, x_k)\, \bar{f}_j(x_j) + \varphi_{ij}(x_i, x_j)\, \bar{f}_k(x_k) + \varphi_k(x_k)\, \bar{f}_{ij}(x_i, x_j) + \\ & \varphi_j(x_j)\, \bar{f}_{ik}(x_i, x_k) + \varphi_i(x_i)\, \bar{f}_{jk}(x_j, x_k) \end{aligned} \tag{42}$$

where $\varphi_i(x_i)$, $\varphi_j(x_j)$, ..., $\varphi_{ijk}(x_i, x_j, x_k)$ are suitable known functions (e.g., the products of monomials $(x_i - b_i)$, $(x_j - b_j)$ and $(x_k - b_k)$ where the b's are constants), and $\bar{f}_0$, $\bar{f}_i(x_i)$, ..., $\bar{f}_{jk}(x_j, x_k)$ are Cut-HDMR component functions for some given function $\bar{f}(\mathbf{x})$-related to $f(\mathbf{x})$. Thus, the three-dimensional numerical table for $f_{ijk}(x_i, x_j, x_k)$ is replaced by some one- and two-dimensional numerical tables. The saving is large, especially for high order component functions. Using projector theory, an approach referred to as *monomial preconditioning* Cut-HDMR has been developed for this purpose.[23,24]

2.3.3. Multi-Cut-HDMR. The basic principles of HDMR may be extended to more general cases. Multi-Cut-HDMR is one extension where several lth order Cut-HDMR expansions at different reference points $\mathbf{a}(1)$, $\mathbf{a}(2)$, ..., $\mathbf{a}(m)$ are constructed, and $f(\mathbf{x})$ is approximately represented not by one but by all m Cut-HDMR expansions:

$$f(\mathbf{x}) = \sum_{k=1}^{m} w_k(\mathbf{x}) [f_0^{(k)} + \sum_{i=1}^{n} f_i^{(k)}(x_i) + ... + \sum_{1 \le i_1 < i_2 < ... < i_l \le n} f_{i_1 i_2 \ldots i_l}^{(k)}(x_{i_1}, ..., x_{i_l})] \tag{43}$$

The coefficients $w_k(\mathbf{x})$ possess the properties

$$w_k(\mathbf{x}) = \begin{bmatrix} 1 & \text{if } \mathbf{x} \text{ is in any cut subvol of the } k\text{th point expansions} \\ 0 & \text{if } \mathbf{x} \text{ is in any cut subvol of other point expansions} \end{bmatrix} \tag{44}$$

$$\sum_{k=1}^{m} w_k(\mathbf{x}) = 1 \tag{45}$$

The properties of the coefficients $w_k(\mathbf{x})$ imply that the contribution of all other Cut-HDMR expansions vanish except one when

x is located on any cut line, plane, or higher dimensional ($\leq l$) subvolumes through that reference point, and then the Multi-Cut-HDMR expansion reduces to single point Cut-HDMR expansion. As mentioned above, the lth order Cut-HDMR approximation does not have error when **x** is located on these subvolumes. When m Cut-HDMR expansions are used to construct a Multi-Cut-HDMR expansion, the error free region in input **x** space is m times that for a single reference point Cut-HDMR expansion. Therefore, the accuracy will be improved.

There are a variety of choices to define $w_k(\mathbf{x})$. For example, the metric distances $\rho_k^{i_1 i_2 \ldots i_l}$ from point **x** to an l-dimensional subvolume with variables $\{x_{i_1}, x_{i_2}, ..., x_{i_l}\}$ through reference point $\mathbf{a}(k)$ ($k = 1, 2, ..., m$)

$$\rho_k^{i_1 i_2 \ldots i_l}(\mathbf{x}) = [\sum_{\substack{i=1 \\ i \notin \{i_1, i_2, \ldots, i_l\}}}^{n} [x_i - a_i(k)]^2]^{1/2} \; \{i_1, i_2, ..., i_l\} \subset \{1, 2, ..., n\} \quad (46)$$

can be used to define

$$\bar{w}_k(\mathbf{x}) = \prod_{\substack{s=1 \\ s \neq k}}^{m} \prod_{1 \leq i_1 < \ldots < i_l \leq n} \rho_s^{i_1 i_2 \ldots i_l}(\mathbf{x}) \quad (47)$$

$$w_k(x) = \frac{\bar{w}_k(x)}{\sum_{s=1}^{m} \bar{w}_s(x)} \quad (48)$$

It can be readily proven that the defined $w_k(\mathbf{x})$ satisfies the required properties if different reference points $\mathbf{a}(k)$ do not share any coordinate. When the $\mathbf{a}(k)$'s have the same values for some elements, modified definitions for $w_k(\mathbf{x})$ may be used.

2.3.4. HDMR with Discrete Input Variables. HDMR can treat discrete as well as continuous input variables. The notion of inherently discrete variables refers to those that are naturally discrete (e.g., molecular moieties functionalized on a scaffold). A potentially serious difficulty in treating inherently discrete variables arises since often there is no a priori means to order the input data. There is a means for handling this problem within HDMR,[25] and the discrete input variable capabilities of HDMR have been successfully tested recently with protein mutations where the discrete variables are the amino acid residues (see section 3.2).

2.3.5. Functional HDMR. If inputs for a system consist of a set of functions, i.e., the input vector $\mathbf{x}(t) = (x_1(t), x_2(t), ..., x_n(t))$, then the system output becomes a functional. One approach to this functional mapping problem is to assume a discretization of the following form

$$x_i(t) = \sum_{k=1}^{N_i} c_{ik} \phi_k(t) \quad (49)$$

where $\{\phi_k(t)\}$ is a family of orthogonal functions. Then any "functional" becomes a "function" of the parameters c_{ik}, and the standard HDMR formulas are applicable.[12] This approach has been successfully implemented for a quantum scattering problem (see section 3.4) and an atmospheric radiative heating problem (see section 3.8).

3. Illustrations of High Dimensional Model Representation

At this stage of HDMR development, much of the activity is focused on testing the capabilities of the concepts and algorithms for realistic applications. This emphasis serves the dual purpose of exploiting HDMR for current problems of interest, as well identifying new algorithmic areas that need further development. This section will present a short synopsis of a variety of applications, spanning atomic and molecular phenomena up through macroscopic and environmental processes. Some of these applications have now been published while others are in press and some are awaiting documentation. Given the space limitations here, only a brief description of each application will be presented, and the interested reader is referred to the cited references and forthcoming works for complete details. The scope of the illustrations below will also indicate the breadth of applicability of HDMR.

3.1. Chemical Formulations. A common materials problem is the preparation of formulations (i.e., mixtures) with n component variables $\mathbf{x} = (x_1, x_2, ..., x_n)$ which are mole fractions implying the constraint $\sum_{i=1}^{n} x_i = 1$. A physical output property $f(\mathbf{x})$ may often have many input components $n \gg 1$, and the output can depend on the components **x** in a nonlinear fashion. Often, optimization of $f(\mathbf{x})$ over **x** is desired and this is an especially challenging task for large n, when each sample is expensive to make and/or test. Furthermore, additional constraints may also exist among some groupings of the chemical components, corresponding, for example, to miscibility criteria or material cost limitations, etc. Minimally, there will be the single total mass fraction constraint that defines a volume in the composition hypercube of dimension $n - 1$, such that $0 \leq \sum_{i=1}^{n-1} x_i \leq 1$. At first sight, this constraint would seem to cause considerable difficulty in exploring the composition input space, as the variables are mutually constrained. However, mitigating this difficulty is the fact that the physically accessible composition volume occupied in the unit hypercube goes down as $1/(n - 1)!$. Furthermore, the mean distance between any two arbitrary points in the volume also shrinks as $\sim 1/\sqrt{n}$. Thus, viewed from an interpolation point of view, this is an ideal circumstance, as the overall accessible space is shrinking primarily to a region around the origin, with thin narrow "fingers" shooting out the variable axes toward each of the limits $x_i \rightarrow 1$, $i = 1, ..., n - 1$. The fact that any two points in the space become increasingly close as n rises suggests that an HDMR of a chemical formulation, truncated to any order, should become more accurate as n rises, a result that first may appear to be surprising. This behavior was confirmed through simulations in up to $n = 20$ dimensions, using both Cut-HDMR sampling on the surface of the chemically reachable volume and with RS-HDMR operating on the interior of the volume. Random sampling techniques may also effectively treat additional constraints among the input chemical component variables. The ability of HDMR to readily capture the IO behavior of complex multicomponent mixtures in high dimensions may have significant implications for accelerating the search for successful materials formulations.

A simple mixture formulation illustration occurs for quaternary semiconductors $A_xB_{1-x}C_yD_{1-y}$ of overall dimension $n = 4$ (e.g., $Ga_xIn_{1-x}P_yAs_{1-y}$). A problem of this type actually corresponds to two coupled and constrained mixture problems (i.e, A_xB_{1-x} and C_yD_{1-y}), each of two dimensions, resulting in an overall formulation with dimension 2 when taking into account the dual mass fraction constraints. A number of semiconductor cases have been explored using laboratory one-dimensional ternary data with Cut-HDMR to estimate the quaternary compound electronic band gap as an output property. Tests of this type have proved to be quite successful, with

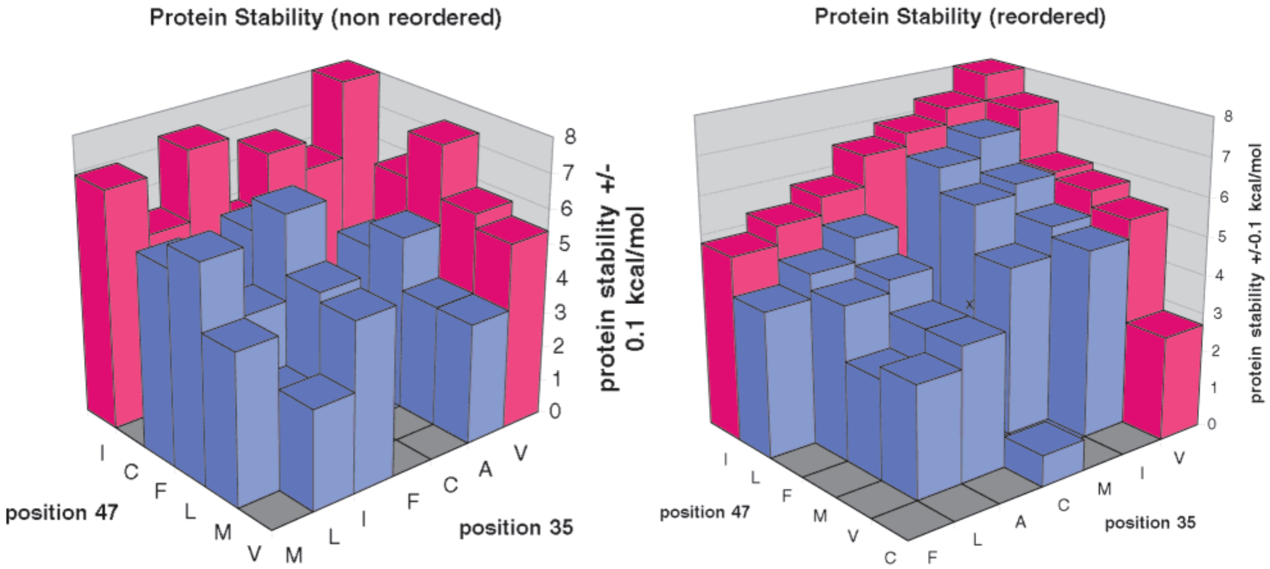

Figure 2. (Left) Histogram plot of stability data from mutations of the gene V protein at sites I47 and V35, organized as originally presented.[28,29] This work, as well as other such double and higher order protein mutation studies, did not reveal underlying regular patterns in the data. Identification of regular data patterns is essential to permit coarse sampling of mutations for an efficient description of the full space. The mutants colored red are single site variants referenced to the wild-type cut center. (Right) This figure is a rearrangement of the same data in the left panel. The measured value of the stability of each single-site mutant labeled in red was utilized to prescribe a monotonic ordering for the efficacy of the amino acids at each site, thereby revealing the underlying pattern of regular behavior amongst the double mutants.

quaternary band gap errors typically on the range of 1−3% over the accessible composition space.[13,26,27]

3.2. Protein Engineering Through Mutations. There is much interest in the artificial mutation of proteins, to both understand the role of the amino acid residues at individual backbone sites, and especially for engineering purposes, to create tailored proteins with enhanced or specialized functional properties for biomedical or industrial applications. The variables in this case are associated with the n backbone sites on the protein chosen for mutation, with each variable x_i taking on up to $s_i = 20$ residue values. At worst, the number of mutation experiments will grow as $\sim 20^n$, which is a frighteningly large number, as n could be 10 or more in some cases. Over the past few decades, many site-directed mutagenesis experiments have been performed on proteins. One general conclusion from these studies is that observed protein properties, often of a thermodynamic nature, are dominated by low order cooperativity among the residues at the various backbone sites. The residues at each site tend to act dominantly alone as contributors to the observed protein property along with some degree of pair cooperativity (i.e., residue−residue cooperative impact on the property), and perhaps a little residual triple cooperativity in some cases. This observed dominance of low order cooperativity clearly has its roots in the few-body nature of intramolecular forces, and it plays very attractively into the structure of HDMR.

An additional special feature of all molecular discovery problems is that the variables $\mathbf{x} = (x_1, x_2, ..., x_n)$ are inherently discrete, and in the present protein case, each variable takes on 20 values for the natural amino acids. The exploration of protein mutations for their observed functional response is a problem of judicious sampling and interpolation of the response $f(x)$ throughout the $\mathbf{x}$ space. As such, a potentially serious difficulty arises since there is no a priori means to order the input variable amino acid residues at each backbone site. Without some identified rational ordering, the response $f(\mathbf{x})$ will likely appear as random over the $\mathbf{x}$ space, and this behavior would prevent an efficient use of coarse sampling for interpolation. A good ordering of the residue variables at each site is defined as one that produces well-behaved “smooth” property variations $f(\mathbf{x})$-over the input $\mathbf{x}$ space. The ideal ordering at one site vs another will surely be different, and the ordering will generally depend on the particular measured protein property. It appears that a rational ordering of the variables can be found, based on the set of single mutations of first order HDMR at each backbone site. The observed response due to the single mutations at each site may be used to produce monotonic variation with respect to a suitable ordering of the residues. It is then natural to expect that the remaining behavior over the second-, or possibly third-, order HDMR mutation surfaces will be regular, if not monotonic. This variable ordering is crucial to make feasible coarse sampling among the mutations, and thereby efficiently employ HDMR to interpolate between the samples.[25] Figure 2 presents an illustration of the effect of protein mutation reordering with thermodynamic stability data from the gene V protein.[28,29] It is evident that a simple reordering of the amino acid variables reveals the regular structure in the mutation space. These data was taken in vitro, and similar behavior has been seen for in vivo experimental data in the same protein.

3.3. Pharmaceutical Discovery. The discovery of pharmaceuticals, from an HDMR perspective, is a sampling and interpolation problem analogous to the treatment of protein mutations in subsection 3.2 above. In this case, the input variables $x_1, x_2, ..., x_n$ label the sites for functionalization on some chosen molecular scaffold, and the variable values label the discrete moieties considered for functionalization at the sites. Unlike proteins, the number of input variables or sites is typically small $n \sim 2-4$, while the number of variable values (i.e., possible moieties at each site) can be arbitrarily large, often exceeding 100. In addition, the molecular scaffold itself could be treated as another variable. The objective is to coarsely sample this overall input space and interpolate over it so as to reveal its structure with respect to specific measures of pharmaceutical activity and other relevant properties $f^{(l)}(\mathbf{x})$, $l = 1, 2, ...$. Typically, the aim is to optimize pharmaceutical activity in balance with other goals such as minimizing the toxicity or other induced undesirable physiological processes associated with the drug. The search effort is also compounded by the fact that the ideal pharmaceutical may not lie within the original subspace of moieties considered, and this circumstance eventually calls for extrapolation. Presently, such extrapolations typically are carried out with only an incomplete understanding of pharmaceutical efficacy throughout the originally defined subspace of moieties. Without a rationalization and exploration of the original full subspace, extrapolation can be ineffective, and possibly lead syntheses down the wrong path, upon expansion of the subspace. An identification of the true regular behavior in the original subspace of moieties would provide the best information for an extrapolation beyond.

A critical component of pharmaceutical discovery with HDMR, as with protein mutations in subsection 3.2, is the ordering of the moiety variables in a rational fashion, based on first-order information on the observed pharmaceutical properties $f^{(l)}(\mathbf{x})$, $l = 1, 2, ...$. It may also occur that each observed property has its own best moiety orderings.[30] Although a full exploitation of this algorithm has not occurred at this time, the literature on pharmaceutical activity data does reveal the general dominance of the first order HDMR contributions. This behavior strongly suggests that suitable moiety ordering will produce regular behavior throughout the moiety space in analogy with what has been observed with protein mutations. Coarse sampling and interpolation should be key components of any molecular discovery effort guided by HDMR.

One conclusion from these molecular discovery algorithmic considerations is that it is essential to perform good quality functional observations $f^{(l)}(\mathbf{x})$, $l = 1, 2, ...$ for reliable interpolation. Although the current practice often involves qualitative high throughput screening observations, pharmaceutical discovery with HDMR offers the prospect of performing many fewer candidate syntheses, provided that appropriate good quality gray scale functional observations are performed to assess the effectiveness of the pharmaceutical candidates. These latter comments also apply to all molecular and material discovery efforts with HDMR.

3.4. Potential Surfaces and Dynamical Observables. In all applications of molecular dynamics, a general desire is to understand the influence of potential surface structure upon system observables, such as cross sections, rate constants, etc. As an IO mapping problem, the input potential $V(\mathbf{r})$ depending on p coordinates $\mathbf{r} = (r_1, r_2, ..., r_p)$ is a function and the observable output is a *functional* of the input. In practice, one may often discretize the potential at points $\mathbf{r}^{(l)} = (r_1^{(l)}, r_2^{(l)}, ..., r_p^{(l)})$, $l = 1, 2, ..., n$, in the configuration space to produce a large number of variables $\mathbf{x} = (x_1, x_2, ..., x_n)$, where the lth variable $x_l = V(\mathbf{r}^{(l)})$ is the value of the potential at the particular point $\mathbf{r}^{(l)}$. Considering x_l as an input variable is a meaningful perspective, as the precise value of the potential $x_l = V(\mathbf{r}^{(l)})$ at any point $\mathbf{r}^{(l)}$ is rarely known to high accuracy. In this fashion, an HDMR may be constructed for each observable, either through cut or random sampling techniques. Two issues of concern arise in generating such IO mappings. First, the natural desire for high resolution of the input potential over the coordinate space $\mathbf{r}$ implies that the number of variables n could be very large, ranging from hundreds to thousands even for problems of low configuration space dimension p. In addition, it is not at first immediately evident whether this definition of variables $x_l = V(\mathbf{r}^{(l)})$, $l = 1, 2, ..., n$, will inherently lead to low order cooperativity in the HDMR expansion. There is a special serendipitous circumstance with HDMR implying that higher potential spatial resolution (i.e., higher system dimension n) will in fact lead to simpler HDMR structure.

To appreciate the latter attractive feature, recall that an output observable will be a functional of the input potential, and the potential will appear in the IO map through some possibly complex layers of integrations. The presence of integrations over the potential is important, as it implies that the lth local region of the potential specified by $x_l = V(\mathbf{r}^{(l)})$ should have only a limited influence on the output, even over a reasonable dynamical range of its variability $V_{\min}^{(l)} \le x_l \le V_{\max}^{(l)}$. Furthermore, increasing the spatial resolution of the potential surface will just further contribute to this attractive behavior within HDMR, to likely produce only low order significant contributions to the output. Tests of this concept have been carried out to map atom−atom potentials upon total and differential elastic scattering cross sections as the output, with the potential discretized up to $n = 1000$ variables over a very broad dynamic range $V_{\min}^{(l)} \le x_l \le V_{\max}^{(l)}$, $l = 1, ..., n$. Figure 3 illustrates the attained dynamic range exhibiting good accuracy, based on first-order HDMR with respect to the potential variables and second-order HDMR coupling between the potential and the scattering energy, which is also treated as an input variable.[31,32]

The dynamic range is large, and most of the significant HDMR component functions are quite nonlinear with respect to their variables. Random test samples of potentials in the shaded region of the figure showed that the HDMR could estimate the cross section for virtually any potential in the domain, with errors less than 1%. Essentially the same quality results were obtained with either Cut- or RS-HDMR. The RS-

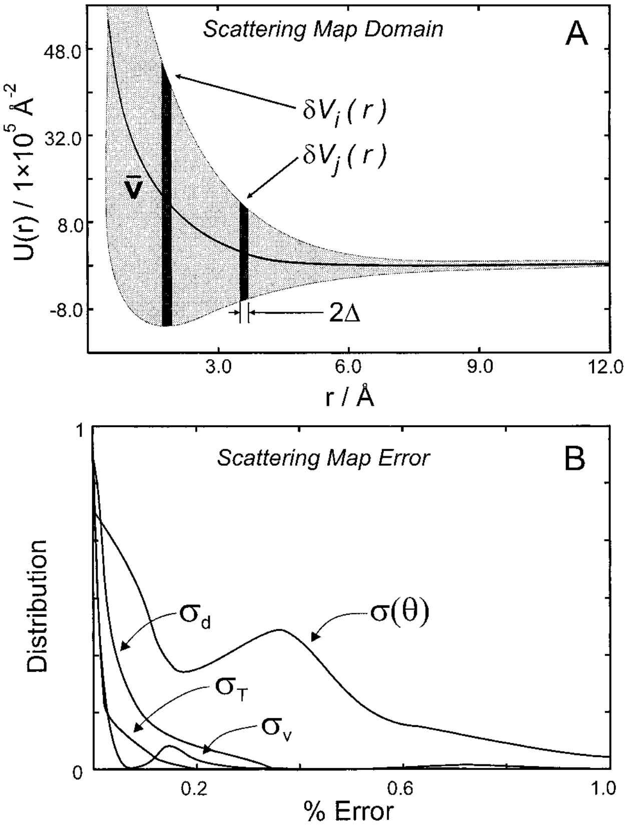

Figure 3. (A) An illustration of the dynamical range of a functional Cut-HDMR map for atom-atom scattering.[30] Here $\bar{\mathbf{v}}$ denotes the cut-center reference potential, $\delta V_i(r)$ and $\delta V_j(r)$ of width 2Δ indicate the full range of the ith and jth HDMR variables $x_i = \delta V_i$ and $x_j = \delta V_j$. The HDMR is first order in the potential variables and second order with respect to the potential and scattering energy. (B) The HDMR map over the domain of (A) produces scattering cross sections of high accuracy for all scattering potentials in the shaded region covering the main block elements of the periodic table. The error statistics from 100 000 random potentials is shown for the cross sections: elastic differential $\sigma(\theta)$, elastic integral σ_T, diffusion σ_d, and viscosity σ_v.

HDMR is more efficient at high values of n, and this issue should even be more important for similar multidimensional potential → observable HDMR maps.

3.5. Laboratory Data Inversion. Inverse problems are particularly challenging in virtually any area of science and engineering, and chemistry is no exception. Invariably, there is insufficient laboratory data or significant errors in the data to prevent a unique identification of the underlying model, and traditional techniques based on linearization procedures typically produce a single inverted model, giving essentially no indication of the breadth of possibilities. In practice, all inversion algorithms call for repeated solution of the system equations of motion (e.g., the Schrödinger equation in the case of quantum dynamics) in an iterative process. Input → output maps, such as generated by HDMR, can be employed in an inverse algorithm, to act as an equivalent stand-in for the original equations of motion. HDMRs, due to their use of simple low-dimensional interpolation, are typically extremely fast to evaluate. This fact, coupled with the ability of genetic-type algorithms to explore for families of solutions, provides a special capability for the inversion of chemical/physical data to identify an underlying family of models consistent with the data.[33−35] Although first generating an HDMR for IO mapping entails some computational overhead, that cost can be acceptable in many applications compared with the effort of attempting the same objective of identifying a family of consistent inverse solutions on the basis of calling up the original system dynamical equations many times upon each iteration. Considering Figure 3 again, as an example, the broad window of applicability of the IO HDMR map, when combined with laboratory data, can identify the family or distribution of potential surfaces consistent with the data and its quality. A successful simulation of this type of global inversion was carried out, using total elastic cross sections as data. The ability of HDMR to aid in the inversion by providing a family of models consistent with laboratory data is generic, and should be applicable to other inverse problems besides scattering.

3.6. Optimal Control of Molecular Motion. There is considerable interest in the design of optimal laser fields for implementation in the laboratory to control molecular dynamics phenomena. The process of optimal design produces a special type of inverse problem mathematically similar to those in subsection 3.5, but in this case, the multiplicity of solutions is an attractive feature for exploitation in laboratory control applications. Viewed as an inverse problem, the molecular control objectives are first prescribed, and these objectives are analogous to the observed data in subsection 3.5. As the objectives are often few and simple (e.g., breaking a particular chemical bond), finding some suitable control field $\epsilon(t)$ presents a problem that is generally highly underposed leading to the many possible good control solutions. The goal is to find at least one of these good solutions. Current laboratory procedures for this purpose have employed genetic algorithms, using the actual molecular sample in a closed-loop process. This procedure has proven successful, but it is also (a) subject to potential inefficiency as the number of significant input control variables (i.e., often taken as the control field frequency components) rise into the hundreds, and (b) no knowledge is gained about the independent, and especially the cooperative roles of the discretized spectral phases or amplitudes of the control field. An HDMR generated from laboratory data with the control field treated as input and the molecular target as the objective can address both of these concerns. The ability to treat item (b) is especially important, as there is basic interest in understanding the relationship between control field features and dynamical events. The HDMR decomposition should be capable of providing insights in this regard, as it inherently is based on variable cooperativity. Simulations of an HDMR-based control algorithm has been carried out[36−40] on model systems of up to 10 quantum levels; the algorithm presently awaits implementation in the laboratory.

3.7. Chemical Kinetics Mapping. Reactive flow is an important feature of many industrial and environmental processes. Typically, the modeling of such processes is broken into pieces, one of which is chemical kinetics (i.e., with the remaining processes often being mass, momentum, and energy transport). In many of these applications, the kinetics portion has grown to be a bottleneck due to the large number of species involved and the resultant excessive costs of solving the kinetic differential equations. In the latter circumstance, the chemical kinetics package in the overall modeling code may be called upon an enormous number of times during a long-term temporal calculation. For example, in the case of global stratospheric chemical modeling over a full year, the chemical kinetics package may be called upward of $\sim 10^8$ times, considering all of the spatial cells in the atmosphere and their generally distinct chemical processes. Currently, practical calculations of this magnitude are only made feasible by including oversimplified models of the chemical kinetics, perhaps also performed to less than the highest accuracy. HDMR offers a special capability as a IO chemical kinetics map, with the input being initial chemical concentrations and perhaps other variables (e.g., solar intensity for photochemical reactions), and the output similarly being chemical concentrations at a later time. Thus, by repeating this process for successive times, an HDMR effectively can act as an integrator, with perhaps a very large time step size. The potential increased efficiency of an HDMR due to the general speed of its evaluation, and the large time steps could in turn allow for the inclusion of enhanced chemical mechanisms. The generation of test HDMR's for this purpose has been carried out, both in terms of single box models in the atmosphere, as well as implementation into full three-dimensional reactive global circulation modeling. Figure 4 gives the comparison between the results provided by traditional look-up data table method and HDMR.[41]

HDMR map time steps of up to 24 h were successful, suggesting that this line of development should be fruitful for further applications. Analogous applications have also been carried out for optimal control of catalytic methanol conversion to formaldehyde[42] and reactive transport in soil media.[17,43] In the latter case, an HDMR was generated covering the aqueous chemistry, as well as much of the transport processes.

3.8. Atmospheric Solar Radiation Transport. In conjunction with the illustration in subsection 3.7, another component of realistic atmospheric modeling is solar radiation transport. Solar radiation can drive photochemical kinetics, and this process would be included in the chemical kinetics HDMR’s discussed in subsection 3.7. In addition, solar radiation may be absorbed by trace gases in the atmosphere, ultimately resulting in atmospheric heating when some of the absorbed radiation is transferred to molecular translation-rotation-vibration degrees of freedom through molecular collisions. This solar radiation energy transfer process is of special concern for issues of global warming, due to the presence of carbon dioxide, methane, and other atmospheric trace species that are strong radiative absorbers. The passage of radiation through the atmosphere, its reflection from the earth’s surface, and retransversal through the atmosphere into outer space, is traditionally modeled by

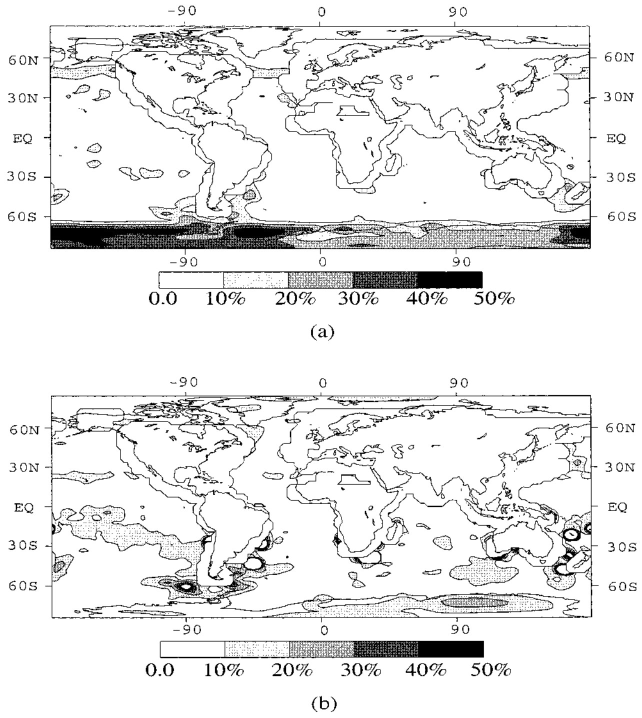

Figure 4. Percentage deviation from the exact solution of the ozone net chemical tendency (production−destruction) predicted by (a) the 4-way look-up table, (b) the HDMR approximation in the surface level during February. The HDMR results are overall more reliable.[41]

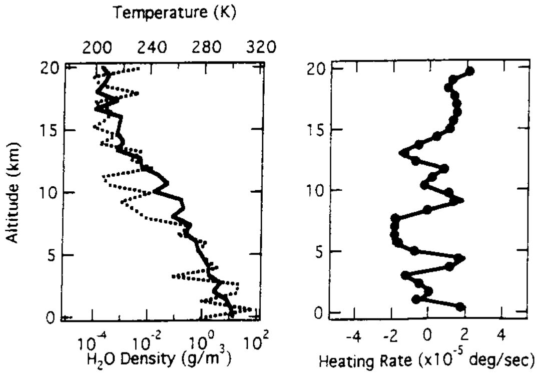

Figure 5. Results from the HDMR for atmospheric heating rates. The panel on the left is a random set of (dotted) atmospheric temperature and (solid) water profiles, with the corresponding heating rate in the panel to the right. The HDMR predictions (solid dot) are very accurate compared to the original heating rate code output. The HDMR output was determined at cost of ∼10^3 less than that of the original heating rate code.[44]

solving an appropriate set of differential equations describing the process. The input to such models is the column densities of the trace gases and the temperature profile as functions of altitude. As with chemical kinetics, such radiative transport calculations are performed an enormous number of times in the global treatment of atmospheric modeling over realistic time intervals. Enhanced efficiency of this component of atmospheric modeling could have a significant impact on overall model performance.

From a system IO perspective, atmospheric radiation transport is a functional mapping problem analogous to the potential surface mapping called for in subsection 3.4. The input consists of functions describing the trace species column densities and the temperature profile; in addition, the output heating rate is also a function of the altitude. In practice, the input functions would be discretized, possibly on a grid over the altitude, and the value of each of these discretized variables may be treated as an input variable to an HDMR. This problem has the same advantageous feature as quantum mechanical potential surface input in subsection 3.4, where better input spatial resolution also coincidentally tends to produce simplified and more accurate HDMR's. An illustration of the ability of an HDMR to capture radiation transport behavior has been carried out[44] with atmospheric temperature and water vapor concentration each discretized at 30 altitudes as input along with the earth's surface temperature and albedo, leading to a total of $n = 62$ variables. The output heating rate was also discretized into 30 atmospheric layers, and each of the local heating rates became an HDMR over the full space of 62 input variables. The resultant Cut-HDMR's, taken to second order, produced excellent results for predicting heating rates with errors no larger than ∼3%, with the HDMR simultaneously being nearly ∼10^3 times faster to evaluate than the original radiation transport code. An illustration of IO radiation transport behavior is shown in Figure 5 for an arbitrary water and temperature profile as input and the resultant heating profile from the HDMR, as well as that obtained by the original radiation transport package. This typical result illustrates the excellent quality of the HDMR. Much further development is needed to include additional atmospheric physical effects (e.g., clouds) and species to produce a viable HDMR radiative package for insertion into full atmospheric modeling.

4. Concluding Remarks

This article has focused on chemical/physical phenomena involving large numbers of input variables, and it should be evident that problems of this type are quite generic in many areas involving experimentation, plant operations, and modeling. Essentially the same methodology being developed for these chemical applications may be transferable to even broader classes of problems of equal significance in other domains. Exploration of the diverse applications of HDMR can, in turn, stimulate further development of the primary chemical/physical applications. As an example going beyond chemistry, economic systems of all types are frequently characterized as IO problems for understanding and estimation of future behavior. An illustration of this type was carried out considering derivatives (i.e., financial instruments whose value derives from the value of other commodities). An HDMR was generated with $n = 5$ input variables, based on real trading data. The quality of the results was excellent, with the HDMR showing predictive capability with errors typically no larger than ∼ 8%.[45]

Beyond this application, interestingly, there are others in economics which are mathematically analogous to those arising in chemistry. For example, industrial plant or economic system performance under conditions of constrained resources is a problem mathematically like that of chemical formulations theory in subsection 3.1 where a mass fraction constraint was present. Many analogies to other interesting problems also exist.

Acknowledgment. The authors acknowledge support from Department of Defense, the Environmental Protection Agency, and Hercules Incorporated.

References and Notes

(1) Diaconis, P.; Shahshahani, M. On Nonlinear Functions of Linear Combinations. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 175−191.

(2) Friedman, J.; Stuetzle, W. Projection Pursuit Regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817−823.

(3) Huber, P. Projection Persuit. *Ann. Statist.* **1985**, *13*, 435−525.

(4) Stone, C. J. Additive Regression and Other Nonparametric Models. *Ann. Stat.* **1985**, *13*, 689−705.

(5) Parker, D. Learning Logic. Working paper No. 47; Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology: Cambridge, MA, 1985.

(6) Piggio, T.; Girosi, F. Networks for Approximation and Learning. *Proc. IEEE* **1990**, *78*, 1481−1497.

(7) Girosi, F.; Poggio, T. Representation Properties of Networks: Kolmogorov's Theorem is Irrelevant. *Neural Comput.* **1989**, *1*, 465−469.

(8) Lorentz, G. G.; Golitschek, M. V.; Makovoz, Y. *Constructive Approximation*; Springer: New York, 1996.

(9) Frisch, H. L.; Borzi, C.; Ord, G.; Percus, J. K.; Williams, G. O. Approximate Representation of Functions of Several Variables in Terms of Functions of One Variable. *Phys. Rev. Lett.* **1989**, *63*, 927−929.

(10) Gorban, A. N. Approximation of Continuous Functions of Several Variables by an Arbitrary Nonlinear Continuous Function of One Variable, Linear Functions, and Their Superpositions. *Appl. Math. Lett.* **1998**, *11*, 45−49.

(11) Rabitz, H.; Alis, O. F.; Shorter, J.; Shim, K. Efficient Input-output Model Representations. *Comput. Phys. Commun.* **1999**, *117*, 11−20.

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

(13) Shim, K.; Rabitz, H. Independent and Correlated Composition Behavior of Material Properties: Application to Energy Band Gaps for the Ga$_\alpha$In$_{1-\alpha}$PbAs$_{1-\beta}$ and Ga$_\alpha$In$_{1-\alpha}$PbSbgAs$_{1-\beta-\gamma}$ Alloys. *Phys. Rev. B.* **1998**, *58*, 1940−1946.

(14) Shorter, J.; Precila, C. Ip.; Rabitz, H. An Efficient Chemical Kinetics Solver using High Dimensional Model Representations. *J. Phys. Chem. A.* **1999**, *103*, 7192−7198.

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

(16) Ghanem, R. G.; Spanos, P. D. *Stochastic Finite Elements: A Spectral Approach*; Springer-Verlag: New York, 1991.

(17) Wang, S.; Jaffe, P. R.; Li, G.; Wang, S. W.; Rabitz, H. Simulating Bioremediation of Uranium-Contaminated Aquifers; Uncertainty Assessments of Model Parameters. *J. Contam. Hydrol.* **2001**. Submitted for publication.

(18) 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.

(19) Gordon, W. J. Distributive Lattices and the Approximation of Multivariate Functions. In *Proceedings of the Symposium of the Approximation with Special Emphasis on Spline Functions*; Schoenberg, I. J., Ed.; Academic Press: New York, 1969; pp 223−277.

(20) Sobol, I. M. Sensitivity Estimates for Nonlinear Mathematical Models. *Mathematical Modelling and Computational Experiments* **1993**, *1*, 407−414.

(21) Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models*; Chapman and Hall: London, 1990; p 112.

(22) Hollebeek, T.; Ho, T. S.; Rabitz, H. *J. Chem. Phys.* **1997**, *106*, 7223−7227.

(23) Li, G.; Wang, S. W.; Rabitz, H.; Rosenthal, C. High Dimensional Model Representations Generated from Low Dimensional Data Samples I: mp-Cut-HDMR. *J. Math. Chem.* **2001**. In press.

(24) Wang, S. W.; Levy, H., II; Li, G.; Rabitz, H. Fully Equivalent Operational Models generated by a monomial preconditioning method for Atmospheric Chemical Kinetics within Global Chemistry-transport Models. *J. Geophys. Res.* **2000**. Submitted for publication.

(25) Rabitz, H.; Pierce, L.; Li, B.; Carey, J. Mapping Protein Functionality: Revealing Patterns in High-Dimensional Sequence Space. *Science* **2000**, Submitted for publication.

(26) Shim, K.; Rabitz, H. Electronic and Structure Properties of the Pentanary Alloy Ga$_x$In$_{1-x}$P$_y$Sb$_z$As$_{1-y-z}$. *J. Appl. Phys.* **1999**, *85*, 7705−7715.

(27) Rabitz, H.; Shim, K. Multicomponent Semiconductor Material Discovery Guided by a Generalized Correlated Function Expansion. *J. Chem. Phys.* **1999**, *111*, 10640−10651.

(28) Sandberg, W. S.; Terwilliger, T. C. Engineering Multiple Properties of a Protein by Combinatorial Mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 8367−8371.

(29) Sandberg, W. S.; Schlunk, P. M.; Zabin, H. B.; Terwilliger, T. C. Relationship between in Vivo Activity and in Vitro Measures of Function and Stability of a Protein. *Biochem.* **1995**, *34*, 11970−11978.

(30) Shenvi, N.; Geremia, J. M.; Rabitz, H. A Genetic Algorithm Solution to the Substituent Ordering Problem in Library Optimization. Manuscript in preparation.

(31) Geremia, J. M.; Rabitz, H.; Rosenthal, C. Constructing Global Functional Maps between Molecular Potentials and Quantum Observables. *J. Chem. Phys.* **2001**. In press.

(32) Geremia, J. M.; Rabitz, H. A Global, Nonlinear Method for Extracting Potentials from Spectral Data: The Singlet and Triplet States of Na2. *J. Chem. Phys.* **2001**. Submitted for publication.

(33) Shenvi, N.; Geremia, J. M.; Rabitz, H. Nonlinear Kinetics Parameter Identification by HDMR Map Inversion. Manuscript in preparation.

(34) Geremia, J. M.; Rabitz, H. A Global, Nonlinear Algorithm for Inverting Quantum Mechanical Observations. *Phys. Rev. A* **2001**. Submitted for publication.

(35) Geremia, J. M.; Rabitz, H. The Ar−HCl Potential Energy Surface From a Global Map-Facilitated Inversion of State-to-State Rotationally Resolved Differential Scattering Cross Sections and Rovibrational Spectral Data. *J. Chem. Phys.* **2001**. Submitted for publication.

(36) Geremia, J. M.; Zhu, W.; Rabitz, H. Incorporating Physical Implementation Concerns into Closed Loop Quantum Control Experiments. *J. Chem. Phys.* **2000**, *113*, 10841−10848.

(37) Geremia, J. M.; Weiss, E.; Rabitz, H. Achieving the Laboratory Control of Quantum Dynamics Phenomena Using Nonlinear Functional Maps. *Chem. Phys.* **2001**. In press.

(38) Geremia, J. M.; Rabitz, H. An Efficient Optimal Identification Algorithm: The Synthesis of Quantum Optimal Control and Map-Facilitated Inversion. *Chem. Phys.* **2001**. To be submitted.

(39) Biteen, J.; Geremia, J. M.; Rabitz, H. Closed-Loop Quantum Control Utilizing Time Domain Maps. Manuscript in preparation.

(40) Biteen, J.; Geremia, J. M.; Rabitz, H. Quantum Optimal Quantum Control Field Design Using Logarithmic Maps. Manuscript in preparation.

(41) 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.

(42) Faliks, A.; Yetter, R. A.; Floudas, C. A.; Bernasek, S. L.; Fransson, M.; Rabitz, H. Optimal Control of Catalytic Methanol Conversion to Formaldehyde. *J. Phys. Chem. A* **2000**. In press.

(43) Li, G.; Wang, S. W.; Rabitz, H.; Wang, S. K.; Jaffe, P. Global Uncertainty Assessments by High Dimensional Model Representations (HDMR). *Chem. Eng. Sci.* **2001**. To be submitted.

(44) Shorter, J.; Rabitz, H. Radiation Transport Simulation by Means of a Fully Equivalent Operational Model, *Geophys. Res. Lett.* **2000**, *27*, 3485−3488.

(45) Faliks, A.; Alis, O.; Rabitz, H. A Nonparametric Approach to Pricing Derivative Securities via High Dimensional Model Representations. Manuscript in preparation.