Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs

ARTICLE *in* THE JOURNAL OF PHYSICAL CHEMISTRY A · MAY 2010

Impact Factor: 2.69 · DOI: 10.1021/jp9096919 · Source: PubMed

7 AUTHORS, INCLUDING:

Genyuan Li
Princeton University
54 PUBLICATIONS **1,554** CITATIONS
SEE PROFILE

Herschel Rabitz
Princeton University
947 PUBLICATIONS **23,742** CITATIONS
SEE PROFILE

Oluwayemisi O. Oluwole
ANSYS
22 PUBLICATIONS **211** CITATIONS
SEE PROFILE

Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs

Genyuan Li and Herschel Rabitz*

Department of Chemistry, Princeton University, Princeton, New Jersey 08544

Paul E. Yelvington,† Oluwayemisi O. Oluwole, Fred Bacon, and Charles E. Kolb

Aerodyne Research, Inc., 45 Manning Road, Billerica, Massachusetts 01821

Jacqueline Schoendorf

SPARTA, Inc., 39 Simon Street, Suite 15, Nashua, New Hampshire 03060

Received: October 9, 2009; Revised Manuscript Received: February 24, 2010

The objective of a global sensitivity analysis is to rank the importance of the system inputs considering their uncertainty and the influence they have upon the uncertainty of the system output, typically over a large region of input space. This paper introduces a new unified framework of global sensitivity analysis for systems whose input probability distributions are independent and/or correlated. The new treatment is based on covariance decomposition of the unconditional variance of the output. The treatment can be applied to mathematical models, as well as to measured laboratory and field data. When the input probability distribution is correlated, three sensitivity indices give a full description, respectively, of the total, structural (reflecting the system structure) and correlative (reflecting the correlated input probability distribution) contributions for an input or a subset of inputs. The magnitudes of all three indices need to be considered in order to quantitatively determine the relative importance of the inputs acting either independently or collectively. For independent inputs, these indices reduce to a single index consistent with previous variance-based methods. The estimation of the sensitivity indices is based on a meta-modeling approach, specifically on the random sampling-high dimensional model representation (RS-HDMR). This approach is especially useful for the treatment of laboratory and field data where the input sampling is often uncontrolled.

1. Introduction

Suppose that an input−output system structure is described by a deterministic relation

$$y = f(\mathbf{x}) = f(x_1, x_2, ..., x_n) \tag{1}$$

where x_i denotes the ith input and y is the output. We will use upper-case letters, i.e., X_i, Y, when referring to the generic aspects of variables. Lower-case letters, i.e., x_i, y, represent their observed values. Boldface, as $\mathbf{X}$ or $\mathbf{x}$, is used to designate vectors. Sensitivity analysis is concerned with understanding how the system input variations influence the changes of the output. This is often motivated by the fact that there is uncertainty about the true values of the inputs used in a particular application. Thus, in sensitivity analysis, the X_i's are formally treated as random variables with specified distributions, and, consequently, Y is also a random variable with a probability distribution. The characterization of the empirical output distribution, given the input probability distribution, is the goal of uncertainty analysis. The assessment of the relative importance of the inputs in the above relation is the objective of sensitivity analysis.[1−3]

Variance-based methods are commonly used[4−7] in global sensitivity analysis for quantifying the sensitivity of the output Y to the inputs $\mathbf{X}$ in terms of a reduction in the variance of Y

$$S_i = V_i/V(Y) = \mathrm{Var}[E(Y|X_i)]/\mathrm{Var}(Y) \tag{2}$$

$$S_{ij} = V_{ij}/V(Y) = (\mathrm{Var}[E(Y|X_i, X_j)] - V_i - V_j)/\mathrm{Var}(Y)$$
$$\cdots \tag{3}$$

where $E(\cdot)$ and $\mathrm{Var}(\cdot)$ represent the expected value and variance; S_i and S_{ij} are referred to as the main and first-order interaction effects for X_i and X_i, X_j, respectively. These measures reflect the reduced portions of the output uncertainty caused by the inputs and their interactions when the true values of a subset of inputs $\mathbf{X}_p$ (where p is a subset of $\{1, 2, \cdots, n\}$) are known. Moreover, the total effect of X_i is defined as

$$S_{Ti} = V_{Ti}/V(Y) = (\mathrm{Var}(Y) - \mathrm{Var}[E(Y|\mathbf{X}_{-i})])/\mathrm{Var}(Y) \tag{4}$$

where $\mathbf{X}_{-i}$ indicates all inputs except X_i. S_{Ti} is the ratio of the remaining uncertainty of the output to the unconditional output uncertainty $V(Y)$ when the true values of all inputs except X_i are known.

The variance-based methods are closely related to the decomposition of $f(\mathbf{x})$ itself:[5]

* Corresponding author. Tel: +1-609-258-3917. Fax: +1-609-258-0967. E-mail: hrabitz@princeton.edu.

† Current address: Mainstream Engineering Corporation, 200 Yellow Place, Rockledge, FL 32955.

10.1021/jp9096919
Published on Web 04/26/2010

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\le i<j\le n} f_{ij}(x_i, x_j) + \cdots + f_{12\ldots n}(x_1, \ldots, x_n) = f_0 + \sum_{j=1}^{2^n-1} f_{p_j}(\mathbf{x}_{p_j}) \tag{5}$$

where

$$f_0 = E(Y) \tag{6}$$

$$f_i(x_i) = E(Y|x_i) - f_0 \tag{7}$$

$$f_{ij}(x_i, x_j) = E(Y|x_i, x_j) - f_i - f_j - f_0 \quad \cdots \tag{8}$$

with $E(f_{p_j}(\mathbf{X}_{p_j})) = 0$ for all the nonconstant component functions above. The last term is determined by the difference of y and all other terms on the right, thus $f(\mathbf{x})$ is exactly equal to y. The component functions in the above decomposition provide their best approximation to $f(\mathbf{x})$ in a least-squares sense. For independent inputs, all the component functions are mutually orthogonal, and the decomposition is unique. The determination of a component function, for example $f_i(X_i)$, individually (by minimizing $L = E[(f(\mathbf{X}) - f_i(X_i))^2]$) or simultaneously with other component functions by least-squares regression will give the same answer within the data error. For independent inputs, a unique decomposition of the unconditional variance $V(Y)$ for Y, parallel to the above decomposition of $f(\mathbf{x})$ can be obtained:[5]

$$V(Y) = \sum_{i=1}^{n} V_i + \sum_{1\le i<j\le n} V_{ij} + \cdots + V_{12\ldots n} = \sum_{j=1}^{2^n-1} V_{p_j} \tag{9}$$

with

$$V_{p_j} = \mathrm{Var}(f_{p_j}(\mathbf{X}_{p_j})) \tag{10}$$

and

$$1 = \sum_{j=1}^{2^n-1} V_{p_j}/V(Y) = \sum_{j=1}^{2^n-1} S_{p_j} \tag{11}$$

The importance rank of the inputs or subsets of inputs can be simply determined by comparing the magnitudes of the sensitivity indices.

Using this parallel relation between the $V(Y)$ and $f(\mathbf{x})$ decompositions, two methodological approaches for estimating sensitivity indices, classical and meta-modeling, have been developed. The classical approach[5,9−11] directly calculates the conditional variances. Specific Monte Carlo samples for the inputs $\mathbf{x}$ (e.g., FAST samples,[11] Sobol samples,[5] and replicated Latin Hypercube samples[12]) are generated, and the corresponding output values $y = f(\mathbf{x})$ are calculated to estimate the sensitivity indices. In the meta-modeling approach,[3,13−15] an effective model of the original system is constructed first, and all the sensitivity indices and the mapping of Y are based on this meta-model. The procedure calculates either the conditional variances or the variances of $f_{p_j}(\mathbf{X}_{p_j})$. The meta-modeling approach is often more efficient than the classical alternative that requires a large number of specific samples, and cannot analyze laboratory or field data where sampling is often uncontrolled.

When the inputs are correlated, some ambiguities arise in the definitions of sensitivity indices given by the variance-based methods. The conditional variances will generally depend on the existence of correlations in the input variables. Adopting the same definition of sensitivity indices given by the variance-based methods for a given subset of inputs can lead to contributions from other correlated inputs contaminating the result.[9] This problem was observed by Oakley and O'Hagan[1] who demonstrated that $V(Y)$ cannot be decomposed into a sum of squares as given in eq 9 and that the V_{p_j}'s do not partition $V(Y)$ for systems possessing a correlated input probability distribution. The sum of all sensitivity indices adopting the definitions given by the variance-based methods may not equal unity. Therefore, the resultant relative importance of the inputs is questionable based on comparing the magnitudes of the sensitivity indices. For a system with a correlated input probability distribution, a single sensitivity index cannot fully describe the input contributions.

Here we introduce a new unified global sensitivity analysis framework for systems whose input probability distribution has independent and/or correlated variables. The new treatment is based on covariance decomposition of the unconditional variance of the output. This analysis technique can be applied to mathematical models, as well as measured laboratory and field data. The definition of sensitivity indices given by the variance-based methods for systems with independent inputs is a special case of the new unified treatment. When the input probability distribution is correlated, three sensitivity indices S_{p_j}, $S^a_{p_j}$ and $S^b_{p_j}$ are defined to respectively give a full description of the total, structural (reflecting the system structure $Y = f(\mathbf{X})$), and correlative (reflecting the correlated input probability distribution) contributions for an input or a subset of inputs $\mathbf{X}_{p_j}$. We refer to this technique as the structural and correlative sensitivity analysis (SCSA) method. When the inputs are independent, the SCSA indices reduce to the single index S_{p_j}, consistent with the variance-based methods. In this paper, the estimation of sensitivity indices is based on the meta-modeling approach, specifically on the random sampling-high dimensional model representation (RS-HDMR) expansion,[17−19] whose component functions are approximated by cubic B splines.[20] The expansion coefficients are extracted from a given set of input−output data by a backfitting procedure utilizing the statistical F-test for identifying the significant component functions.[21]

This paper is organized as follows. Section 2 discusses methodology including the covariance decomposition and definition of the SCSA sensitivity indices along with the estimation of the sensitivity indices utilizing RS-HDMR. Section 3 illustrates three applications of the SCSA method: (1) a linear model whose input probability density function (pdf) is a joint normal distribution with or without correlation, (2) a nonlinear model used recently by Storlie et al.[15] to provide a comparison of the RS-HDMR method with other existing meta-modeling methods, and (3) ionospheric electron density characteristics assessed by measured ionosonde data. Section 4 presents conclusions.

2. Methodology

2.1. Covariance Decomposition of the Unconditional Variance $V(Y)$. Suppose that y can be approximated by n_p ($\ll 2^n - 1$) nonconstant component functions in eq 5:

$$y = f_0 + \sum_{j=1}^{n_p} f_{p_j}(\mathbf{x}_{p_j}) + \varepsilon \tag{12}$$

where $\varepsilon \sim N(0, \sigma^2)$ is random error. When all the f_{p_j}'s are determined from a set of input−output data by an unbiased method (e.g., least-squares regression), the difference between y and its approximation $f_0 + \sum_{j=1}^{n_p} f_{p_j}$ is orthogonal to the subspace spanned by all of the f_{p_j}'s ($j = 1, 2, ..., n_p$) in the Hilbert space,[22] i.e.,

$$(y - f_0 - \sum_{j=1}^{n_p} f_{p_j}, f_{p_k}) = (\varepsilon, f_{p_k}) = 0, (k = 1, 2, ..., n_p) \tag{13}$$

where $(\cdot,\cdot)$ denotes the inner product defined as

$$(h(\mathbf{x}), g(\mathbf{x})) = \int_{\Omega_n} w(\mathbf{x}) h(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} \tag{14}$$

with $w(\mathbf{x})$ being the pdf of $\mathbf{X}$, and Ω_n being the input domain.

Using eq 13 and $E(f_{p_j}(\mathbf{X}_{p_j})) = 0$, the unconditional variance of the output $V(Y)$ can be decomposed as the sum of all the covariances, $\mathrm{Cov}(f_{p_j}, Y)$, and the averaged square error $\overline{\varepsilon^2}$

$$\begin{aligned} V(Y) &= E[(Y - E(Y))^2] = \int_{\Omega_n} w(\mathbf{x})(y - f_0)^2\, d\mathbf{x} \\ &= (y - f_0, y - f_0) = (\sum_{j=1}^{n_p} f_{p_j} + \varepsilon, y - f_0) \\ &= \sum_{j=1}^{n_p} \mathrm{Cov}(f_{p_j}, Y) + (\varepsilon, \varepsilon) \\ &= \sum_{j=1}^{n_p} [\mathrm{Var}(f_{p_j}) + \mathrm{Cov}(f_{p_j}, \sum_{k=1, k\neq j}^{n_p} f_{p_k})] + \overline{\varepsilon^2} \end{aligned} \tag{15}$$

If $\overline{\varepsilon^2}$ is sufficiently small compared to $V(Y)$ (i.e., $f_0 + \sum_{j=1}^{n_p} f_{p_j}$ is a good approximation for $f(\mathbf{x})$), the sum of the covariances forms a good decomposition of $V(Y)$. When the f_{p_j}'s are all of the component functions in eq 5, then $\varepsilon = 0$, and consequently $\overline{\varepsilon^2} = 0$. In this case $V(Y)$ is exactly partitioned by all the $\mathrm{Cov}(f_{p_j}, Y)$'s.

A key difference between the covariance decomposition of $V(Y)$ in eq 15 from the variance decomposition of $V(Y)$ in eq 9 is that the terms $\mathrm{Cov}(f_{p_j}, Y)$ can be negative. The covariance $\mathrm{Cov}(f_{p_j}, Y)$ is the total contribution of f_{p_j} composed of its structure piece $\mathrm{Var}(f_{p_j})$ (which is always positive, reflecting f_{p_j}'s contribution in the system structure $Y = f(\mathbf{X})$) and a correlation piece $\mathrm{Cov}(f_{p_j}, \sum_{k=1, k\neq j}^{n_p} f_{p_k})$ (which can be positive or negative, reflecting the influence of the interaction between f_{p_j} and other component functions through the correlated input probability distribution). Fixing some inputs may influence the distributions of other inputs, and the total effect can decrease or increase the variance of the output, which yields a positive or negative $\mathrm{Cov}(f_{p_j}, Y)$.

The covariance $\mathrm{Cov}(h, Y)$ for a function $h(\neq 0)$ will be positive if h is a model approximating Y, i.e.,

$$y = h(\mathbf{x}) + \varepsilon \tag{16}$$

and $h(\mathbf{x})$ is determined by a unbiased method (e.g., least-squares regression). Then

$$\mathrm{Cov}(h, Y) = \mathrm{Cov}(h, h + \varepsilon) = \mathrm{Var}(h) > 0 \tag{17}$$

For example, $h(\mathbf{X}) = f_i(X_i)$. In this case, we have

$$\mathrm{Cov}[f_i(X_i), Y] = \mathrm{Var}[f_i(X_i)] = V_i \tag{18}$$

which is simply the conditional variance used in eq 2. Note that, for a probability distribution with correlated inputs, the resultant $f_i(X_i)$ will generally depend on whether it is determined individually (using eq 16) or simultaneously (using eq 12), and the corresponding $\mathrm{Cov}[f_i(X_i), Y]$ will also be different. The former covariance is always positive and is equal to V_i, but the latter covariance can be negative and represents the total contribution of $f_i(X_i)$, separate from the other f_{p_j}'s.

The covariance decomposition of $V(Y)$ given in eq 15 is general, while the variance decomposition given in eq 9 can be considered as a special case for systems with independent inputs and $n_p = 2^n - 1$. For independent inputs all of the f_{p_j}'s are mutually orthogonal, i.e., $\mathrm{Cov}(f_{p_j}, \sum_{k=1, k\neq j}^{n_p} f_{p_k}) = 0$, and $\mathrm{Cov}(f_{p_j}, Y) = \mathrm{Var}(f_{p_j})$. When the f_{p_j}'s include all of the component functions in eq 5, $\overline{\varepsilon^2} = 0$. As mentioned above, for independent inputs, either individual or simultaneous determination of the component functions by least-squares regression gives the same results, which implies that $\mathrm{Var}(f_{p_j}) = V_{p_j}$. Then eq 15 reduces to eq 9.

For systems with correlated inputs, a single sensitivity index cannot unambiguously describe the contributions of a single or a subset of inputs $\mathbf{X}_{p_j}$. On the basis of eq 15, three sensitivity indices are defined:

$$S_{p_j} = \mathrm{Cov}(f_{p_j}, Y)/V(Y) \tag{19}$$

$$S^a_{p_j} = \mathrm{Var}(f_{p_j})/V(Y) \tag{20}$$

$$S^b_{p_j} = \mathrm{Cov}(f_{p_j}, \sum_{k=1, k\neq j}^{n_p} f_{p_k})/V(Y) \tag{21}$$

which respectively represent the *total*, *structural*, and *correlative* contributions for $\mathbf{X}_{p_j} (j = 1, 2, ..., n_p)$ with

$$S_{p_j} = S^a_{p_j} + S^b_{p_j} \tag{22}$$

A similar treatment has been considered recently where the conditional variance V_i was decomposed as[23]

$$V_i = V_i^U + V_i^C \tag{23}$$

with V_i^U and V_i^C referring to the uncorrelated and correlated variations for X_i. Xu and Gertner calculated these quantities for a linear model $y = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + \varepsilon$. The SCSA treatment is distinct in that the decomposition is for all the covariances, $\mathrm{Cov}(f_i, Y)$, $\mathrm{Cov}(f_{ij}, Y)$, and so forth, and is not based on a linear model.

When $\overline{\varepsilon^2}$ is small, we have

$$\sum_{j=1}^{n_p} S_{p_j} = \sum_{j=1}^{n_p} \mathrm{Cov}(f_{p_j}, Y)/V(Y) \approx V(Y)/V(Y) = 1 \tag{24}$$

The magnitudes of S_{p_j}, $S^a_{p_j}$, and $S^b_{p_j}$ ($j = 1, 2, ..., n_p$) all need to be considered in order to quantitatively determine the relative importance of the inputs acting either independently or collectively. The deviation of the sum over all S_{p_j} from unity can be used to evaluate the quality of the sensitivity analysis. When all inputs are independent and the f_{p_j}'s are mutually orthogonal, then $S^b_{p_j} = 0$, and $S_{p_j} = S^a_{p_j}$, which is the sensitivity index given by the variance-based methods.

2.2. Estimation of Sensitivity Indices. As the sensitivity indices are related to the covariance of the component functions, f_{p_j}'s, with the output Y, the first step is the determination of the f_{p_j}'s from a set of input−output data by a suitable regression method. A large body of techniques for carrying out regression analysis has been pursued.[21,24,25] The advanced development of regression methods continues to be an area of active research, and new techniques have been considered for robust regression. Storlie and Helton review some of the traditional nonparametric regression procedures and other methods.[15,16] The present work does not compare the regression techniques, and any proper regression method can be used for the determination of the f_{p_j}'s. In this paper, RS-HDMR[17−19] combined with an F-test is employed. As illustrated in section 3, RS-HDMR combined with an F-test proved to be quite adequate to apply the new SCSA tools.

2.2.1. RS-HDMR. HDMR uses a general approach to optimally construct the component functions in eq 5 sequentially from lower to higher order, such that the lower order contributions are maximized and the higher order contributions are minimized. In this process, the high order component functions (if they exist) are decomposed and portions are included in the low order ones. Often utilizing only the first few low-order component functions gives a satisfactory approximation for $f(\mathbf{x})$ in practical applications. Moreover, distinct, but formally equivalent, HDMR expansions (e.g., Cut-HDMR, RS-HDMR) with the same structure as eq 5 can be constructed to meet various practical requirements.[8]

Cut-HDMR is useful in cases where the sampling may be controlled in an ordered fashion with the associated component functions constructed from numerical data tables along lines, planes, and other higher dimensional subvolumes with respect to a reference point in the input space. RS-HDMR results in the same form as that given in eqs 6−8 with the sampling of $\mathbf{x}$ following any given pdf to determine the component functions. Smoothing spline ANOVA models[26,27] and Generalized additive models[28] have similar formulas and treatments. RS-HDMR is especially useful for handling laboratory or field data where the sampling is often uncontrolled.

To reduce the sampling effort, the RS-HDMR component functions are approximated by expansions in terms of some suitable basis functions (e.g., polynomials, splines, etc.).[14,19] In this paper, cubic B spline functions $B_k(x)$ are used.[20] The first-, second-, and third-order RS-HDMR component functions, $f_i(x_i)$, $f_{ij}(x_i, x_j)$, and $f_{ijk}(x_i, x_j, x_k)$, respectively, can be approximately expanded as

$$f_i(x_i) \approx \sum_{r=-1}^{m+1} \alpha^i_r B_r(x_i) \tag{25}$$

$$f_{ij}(x_i, x_j) \approx \sum_{p=-1}^{m+1} \sum_{q=-1}^{m+1} \beta^{ij}_{pq} B_p(x_i) B_q(x_j) \tag{26}$$

$$f_{ijk}(x_i, x_j, x_k) \approx \sum_{p=-1}^{m+1} \sum_{q=-1}^{m+1} \sum_{r=-1}^{m+1} \gamma^{ijk}_{pqr} B_p(x_i) B_q(x_j) B_r(x_k) \tag{27}$$

where m is the number of knots, with α^i_r, β^{ij}_{pq}, and γ^{ijk}_{pqr} being constant coefficients that need to be determined.

2.2.2. Determination of the RS-HDMR Component Functions by Backfitting. After the RS-HDMR component functions are approximated by suitable basis functions, the resultant expression in eq 5 is an additive model.[21] In this case, the additive model builds on the ability to generate approximations from a set of low-dimensional functions combined with a statistical F-test, to provide confidence bands for the predicted functions, and so forth, and thereby determine which RS-HDMR component functions are retained.

The RS-HDMR component functions can be determined sequentially or simultaneously by least-squares regression. Backfitting may be used for high-dimensional systems. Forward or backward stepwise selection may be used to search for the significant component functions. The choice of algorithm depends on the particular system. For example, if the system has independent inputs, then sequential determination will be most efficient because the process does not depend on the order of the inputs and in each step only one component function needs to be determined. For correlated inputs, sequential determination is improper because the process depends on the order of the inputs. In this case, either simultaneous or backfitting determinations can be used, and they give the same result. However, backfitting is preferred for high-dimension systems because each iteration of backfitting only treats one component function whose dimension is always low.

In this work the backfitting procedure was used to determine the component functions, whereby a new estimate of f_{p_k} is obtained by solving the following equations using least-squares regression:

$$y^{(s)} - f_0 - \sum_{j=1, j\neq k}^{r} f_{p_j}(\mathbf{x}^{(s)}_{p_j}) = f_{p_k}(\mathbf{x}^{(s)}_{p_k}), \quad (s = 1, 2, ..., N) \tag{28}$$

Here, s denotes the sth sample, and N is the total number of samples. This procedure is performed for each component function f_{p_k} in turn, using the current estimates of the f_{p_j}'s to calculate the left-hand side of eq 28. This process continues until the estimated f_{p_j}'s converge.[21]

Suppose that RSS_1 represents the residual sum-of-squares for the least-squares regression of a large model $f_0 + \Sigma^r_{j=1} f_{p_j}$ with p_1 unknown parameters (e.g., the coefficients α^i_r, β^{ij}_{pq}, γ^{ijk}_{pqr} in the spline function expansions of eqs 25−27), and RSS_0 is the same quantity for a small model $f_0 + \Sigma^r_{j=1, j\neq k} f_{p_j}$ nested in the large model, but with p_0 unknown parameters. The F statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1)} \tag{29}$$

has an F distribution with $(p_1 - p_0)$ and $(N - p_1)$ degrees of freedom. If the observed F given by eq 29 is larger than the

tabulated value of the F distribution with $(p_1 - p_0)$ and $(N - p_1)$ degrees of freedom at the 99% confidence level (or other desired confidence level), then f_{p_k} is significant and should be included in the approximation. Otherwise, f_{p_k} can be excluded. Other means (e.g., Ridge regression, Lasso, MARS, ACOSSO, etc.)[15,21] for data fitting can be used as well.

The component functions are determined starting from first order. We may start by comparing the two models,

$$y = f_0 + \sum_{i=1}^{r-1} f_i(x_i), \quad (r = 1, 2, ..., n) \tag{30}$$

$$y = f_0 + \sum_{i=1}^{r} f_i(x_i) \tag{31}$$

to identify the significance of $f_r(x_r)$. In eq 30, $r = 1$ corresponds to $y = f_0$, and remaining functions in eqs 30 and 31 are determined by backfitting. Either forward selection (starting for f_0) or backward selection (starting for $f_0 + \Sigma_{i=1}^{n} f_i(x_i)$) for determining the significant first-order component functions can be used. The determination of the second-order component functions follows the same procedure. The only difference is that, for convenience, the identified significant first-order component functions are included in the small and large models without updating through backfitting. Similarly, the third-order component functions are determined with the identified significant first- and second-order component functions included in the small and large models without updating through backfitting.

2.2.3. Determination of the Sensitivity Indices. After all n_p significant component functions are identified, i.e.,

$$y \approx f_0 + \sum_{l=1}^{n_p} f_{p_j}(\mathbf{x}_{p_j}) \tag{32}$$

then the estimation of S_{p_j}, $S^a_{p_j}$, and $S^b_{p_j}$ is straightforward:

$$S_{p_j} = \mathrm{Cov}(f_{p_j}, Y)/V(Y) \approx \sum_{s=1}^{N} f_{p_j}(\mathbf{x}^{(s)}_{p_j})(y^{(s)} - \bar{y})/\sum_{s=1}^{N} (y^{(s)} - \bar{y})^2 \tag{33}$$

$$S^a_{p_j} = \mathrm{Var}(f_{p_j})/V(Y) \approx \sum_{s=1}^{N} (f_{p_j}(\mathbf{x}^{(s)}_{p_j}))^2/\sum_{s=1}^{N} (y^{(s)} - \bar{y})^2 \tag{34}$$

$$S^b_{p_j} = S_{p_j} - S^a_{p_j} \tag{35}$$

where $\bar{y}$ is the average value of the $y^{(s)}$'s. The total sensitivity indices S_{Ti}, S^a_{Ti}, and S^b_{Ti} also can be calculated by adding together all the sensitivity indices containing X_i. When $\Sigma_{j=1}^{n_p} S_{p_j} \approx 1$, the resultant total sensitivity indices can be considered as reliable.

3. Examples

In this section,the new SCSA treatment of global sensitivity analysis is illustrated by linear and nonlinear simulated models as well as a nonlinear ionospheric system based on field data. The simulated linear model has five input variables, which have either an independent or correlated multivariate normal distribution. Since the linear function and its input pdf are known, it is possible to explicitly establish the relationship between the sensitivity indices S_i, S^a_i, and S^b_i and the system structure as well as the parameters of the input pdf. This example is helpful for understanding the meaning of the new defined sensitivity indices and their advantages compared to that given by the variance-based methods. The nonlinear simulation model has three inputs and was used by Storlie, et al.[15] This model enables comparing RS-HDMR with other existing meta-models. The third illustration is for the analysis of measure field data involving characterization of the ionospheric electron density. Since the inputs were determined from ground-based ionosonde measurements, they represent neither controlled nor independent sampling. The pdf is correlated, and it is not explicitly known. Nevertheless, the SCSA method provides a clear identification of the important inputs.

3.1. Simulated Model: A Linear Function with a Multivariate Normal Distribution for Inputs. A simple linear mathematical model is used to examine how the sensitivity indices S^a_i and S^b_i reflect the structural and correlative contributions of the inputs to $V(Y)$. The model has five inputs: $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)^T$. The pdf for $\mathbf{X}$ is a multivariate normal distribution

$$w(\mathbf{x}) = \frac{1}{(2\pi)^{5/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T\mathbf{\Sigma}^{-1}(x - \mu)\right) \tag{36}$$

where $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$ is the expected value of $\mathbf{X}$, $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$, i.e., the (i, j)th entry of $\mathbf{\Sigma}$ is

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) \tag{37}$$

which quantifies the sampling correlations between X_i and X_j, and $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$. Since $\mathbf{\Sigma}$ is given, the independent or correlated sampling of the inputs is known. The relationship between S^a_i, S^b_i and the system structure as well as $\mathbf{\Sigma}$ can be established. Three cases are considered below.

3.1.1. Case 1: Equal Structural Contributions and Independent Sampling of the Inputs. First consider the simple case where the output y is a sum of all x_i's with equal contributions

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + \varepsilon \tag{38}$$

In this case, $\mu_i = 0.5$ $(i = 1, 2, ..., 5)$, and $\mathbf{\Sigma}$ is the identity matrix $\mathbf{I}_5$ (i.e., the inputs are sampled independently), and each input has an equal structural contribution to the output. The error $\varepsilon \sim N(0, \sigma^2)$ is a random variable with signal-to-noise ratio $\mathrm{SNR} = \mathrm{Var}(f(\mathbf{X}))/\sigma^2 = 100$. The confidence intervals (CI) of the sensitivity indices are determined by the bootstrap method.[15,21] Following the work of Storlie et al.,[15] 100 sets of random samples for $\mathbf{x}$ (each has the size $N = 300$) were generated according to the pdf (the pdf $w(\mathbf{x})$ has an infinite domain for x_i, but there is little possibility for obtaining a large value of x_i and the generated data are distributed within a small range around μ_i) and the corresponding values of y are calculated. Then, random error ε is added. In the construction of the RS-HDMR component functions, all of the input x_i's are normalized $(0 \le x_i \le 1)$. The expansion coefficients α^i_r, β^{ij}_{pq},... of the RS-HDMR component functions are determined by the backfitting procedure described above. Only the first-order RS-HDMR expansion is constructed because the function is linear. The

TABLE 1: Sensitivity Indices S_i, S_i^a, S_i^b for the Linear System Case 1

	eq 39			eq 40
input	S_i^a	S_i^b	S_i	$V_i/V(Y)$
X_1	0.19 ± 0.04	0.01 ± 0.04	0.21 ± 0.04	0.23 ± 0.08
X_2	0.19 ± 0.04	0.01 ± 0.04	0.20 ± 0.04	0.22 ± 0.07
X_3	0.19 ± 0.04	0.00 ± 0.05	0.19 ± 0.05	0.21 ± 0.08
X_4	0.20 ± 0.04	0.01 ± 0.04	0.20 ± 0.04	0.22 ± 0.07
X_5	0.19 ± 0.04	0.00 ± 0.05	0.19 ± 0.05	0.20 ± 0.08
sum	0.97 ± 0.13	0.02 ± 0.13	0.99 ± 0.00	1.08 ± 0.13

component functions $f_i(x_i)$'s are determined either simultaneously or individually:

$$y = f_0 + \sum_{i=1}^{5} f_i(x_i) \tag{39}$$

$$y = f_0 + f_i(x_i), (i = 1, 2, ..., 5) \tag{40}$$

The sensitivity indexes S_i, S_i^a, and S_i^b are calculated using the $f_i(x_i)$'s obtained by eq 39. The S_i obtained from eq 40 corresponds to $V_i/V(Y)$ given by the variance-based methods (see eq 18). The average value $\bar{S}$ and standard error se(S) of the sensitivity indices for the 100 data sets were calculated. Under the assumption that the error is normally distributed, the 95% CI for S is represented by $\bar{S} \pm 1.96\text{se}(S)$,[15,21] and given in Table 1.

Since the inputs are independent and have an equal structural contribution, we should have $S_i = S_i^a = 0.2$ and $S_i^b = 0$ for all i. For independent inputs, the SCSA method and the variance-based methods should give the same result. The values of S_i, S_i^a, S_i^b, and $V_i/V(Y)$ in Table 1 are very close to the expected outcomes for both the SCSA method and the variance-based methods, but the results given by the variance-based methods show somewhat more error.

The estimated bounds of CI (i.e., 1.96se(S)) for the individual sensitivity indices S_i^a, S_i^b, and S_i range from ±0.04 to ±0.05, but the estimated bounds of CI for $\sum_{i=1}^5 S_i^a$, $\sum_{i=1}^5 S_i^b$, and $\sum_{i=1}^5 S_i$ are ±0.13, ± 0.13, and ±0.00, respectively. This behavior can be understood as the sensitivity indices, for example the S_i^a's, are random variables (determined from 100 randomly sampled sets) and their sum ($\sum_{i=1}^5 S_i^a$) is a new random variable with[29]

$$\text{Var}(\sum_{i=1}^{5} S_i^a) = \sum_{i=1}^{5} \text{Var}(S_i^a) + 2 \sum_{1 \le i < j \le 5} \text{Cov}(S_i^a, S_j^a) \tag{41}$$

Since the S_i^a's are independent, i.e., $\text{Cov}(S_i^a, S_j^a) = 0$, then

$$\text{Var}(\sum_{i=1}^{5} S_i^a) = \sum_{i=1}^{5} \text{Var}(S_i^a) \tag{42}$$

The standard error se(S) is an estimate of $[\text{Var}(S)]^{1/2}$ from the 100 data sets. Hence, $1.96\text{se}(\sum_{i=1}^5 S_i^a) > 1.96\text{se}(S_i^a)$. Similar results can be obtained for S_i^b's. However, the S_i's are not independent and satisfy the restriction

$$\sum_{i=1}^{5} S_i = 1 \tag{43}$$

TABLE 2: Sensitivity Indices S_i, S_i^a, and S_i^b for the Linear System Case 2

	eq 39			eq 40
input	S_i^a	S_i^b	S_i	$V_i/V(Y)$
X_1	0.13 ± 0.02	0.11 ± 0.02	0.24 ± 0.03	0.46 ± 0.08
X_2	0.13 ± 0.02	0.11 ± 0.02	0.25 ± 0.03	0.47 ± 0.08
X_3	0.13 ± 0.03	0.06 ± 0.03	0.19 ± 0.03	0.29 ± 0.08
X_4	0.13 ± 0.02	0.03 ± 0.04	0.16 ± 0.04	0.21 ± 0.09
X_5	0.13 ± 0.02	0.03 ± 0.04	0.15 ± 0.04	0.20 ± 0.09
sum	0.65 ± 0.07	0.35 ± 0.08	0.99 ± 0.00	1.64 ± 0.17

Then the second set of terms $\text{Cov}(S_i, S_j)$'s in eq 41 have contributions. Suppose S_i has a positive error, then there must be another S_j having a negative error because the sum of S_i's should be 1. Thus the error for $\sum_{i=1}^5 S_i$ is expected to be smaller than the error of either S_i or S_j because the positive and negative errors cancel each other. This makes the estimated bound of CI $1.96\text{se}(\sum_{i=1}^5 S_i) < 0.01$. A similar discussion applies to the $V_i/V(Y)$'s and other analogous results below.

3.1.2. Case 2: Equal Structural Contributions and Correlated Sampling of the Inputs. All conditions are the same as in Case 1 except that the covariance matrix is now

$$\sum = \begin{bmatrix} 1.0 & 0.6 & 0.2 & 0.0 & 0.0 \\ 0.6 & 1.0 & 0.2 & 0.0 & 0.0 \\ 0.2 & 0.2 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.2 & 1.0 \end{bmatrix} \tag{44}$$

In this case, the inputs X_i's are correlated because some off-diagonal elements, σ_{ij}, are nonzero. For example, X_1 is correlated with X_2 and X_3 ($\sigma_{12} = 0.6$ and $\sigma_{13} = 0.2$). The sum of the nonzero off-diagonal elements for each X_i (e.g., $\sigma_{12} + \sigma_{13} = 0.8$) reflects the correlation of X_i with the other inputs X_j's. The corresponding sensitivity indexes S_i, S_i^a, and S_i^b, and $V_i/V(Y)$ are given in Table 2.

A significant difference between the SCSA method and the variance-based methods occurs for this case with a correlated input pdf. The S_i^a's obtained by the SCSA method are equal, which is consistent with the contribution arising from the model structure. Similarly, the S_i^b's obtained by the SCSA method show that the inputs can be divided into three groups: (X_1, X_2), (X_3), and (X_4, X_5). The ratios of S_i^b for the three groups are ~4:2:1, which reflects the ratios of the values of their respective sums $\sum_{j(j \neq i)}(\sigma_{ij})$, 0.8, 0.4, and 0.2. The correlative contribution of the inputs caused by the correlated pdf is correctly identified. The values $\sum_i S_i^a = 0.65 \pm 0.07$ and $\sum_i S_i^b = 0.35 \pm 0.08$ reveals that the total structural and correlative contributions are both significant. This information cannot be obtained by variance-based methods. Moreover, $\sum_i S_i$ is equal to 0.99, which implies that the sensitivity analysis provided by the SCSA method is reliable for this simple linear model. In contrast, the results given by the variance-based methods mix together the structural and correlative contributions of the inputs, and the information given by $V_i/V(Y)$ can be misleading. According to the values of $V_i/V(Y)$, the inputs X_1, X_2 are the most important, and the inputs X_4, X_5 are the least important. This conclusion is misleading because $V_1/V(Y)$ contains a large ($\sigma_{12} = 0.6$) contribution from X_2 and a small ($\sigma_{13} = 0.2$) contribution from X_3. Similarly, $V_2/V(Y)$ contains a large contribution from X_1 and a small contribution from X_3. For X_4 and X_5, each $V_i/V(Y)(i = 4, 5)$ contains a small ($\sigma_{45} = \sigma_{54} = 0.2$) contribution from the other.

TABLE 3: Sensitivity Indices S_i, S_i^a, and S_i^b for the Linear System Case 3

	eq 39			eq 40
input	S_i^a	S_i^b	S_i	$V_i/V(Y)$
X_1	0.28 ± 0.04	0.16 ± 0.02	0.44 ± 0.04	0.70 ± 0.06
X_2	0.17 ± 0.03	0.16 ± 0.02	0.33 ± 0.03	0.64 ± 0.06
X_3	0.10 ± 0.02	0.06 ± 0.02	0.16 ± 0.03	0.27 ± 0.08
X_4	0.04 ± 0.03	0.00 ± 0.03	0.04 ± 0.03	0.06 ± 0.05
X_5	0.02 ± 0.03	0.00 ± 0.01	0.02 ± 0.02	0.05 ± 0.04
sum	0.61 ± 0.05	0.38 ± 0.05	0.99 ± 0.00	1.71 ± 0.13

This makes the sum of all $V_i/V(Y)$ significantly larger than unity. Actually, all of the X_i's have the same structural contribution upon the output. It is impossible to judge whether $V_i/V(Y)$ contains contributions from other correlated inputs X_j's. Therefore, it is difficult to reliably rank the input importance order by simply comparing the magnitudes of $V_i/V(Y)$. Even though the particular approaches to the variance-based methods are different,[1−3] they should give S_i values close to $V_i/V(Y)$; all such methods cannot discern the structural and correlative contributions.

3.1.3. Case 3: Distinct Structural Contributions and Correlated Sampling of the Inputs. All of the Case 3 conditions are the same as for Case 2 except that

$$y = 5x_1 + 4x_2 + 3x_3 + 2x_4 + x_5 + \varepsilon \qquad (45)$$

Now the inputs have different structural contributions to Y and the sampling is correlated. The calculated sensitivity indexes S_i, S_i^a, and S_i^b and $V_i/V(Y)$ are given in Table 3.

The error in the sensitivity indices is related in a complicated fashion to the error in the estimating the component functions $f_i(x_i)$ and the presence of input covariances σ_{ij}. The coefficients of each x_i in the function are different along with finite correlated input samples, resulting in somewhat larger errors in the sensitivity indices than for the prior cases. However, the magnitudes of the S_i^a's qualitatively reflect the influence of the coefficient of each x_i in the model structure even though the ratios of S_i^a's do not exactly correspond to 5:4:3:2:1. The magnitudes of S_i^b's can still be separated into the same three groups as in Case 2 reflecting the structure in Σ. Compared to the $V_i/V(Y)$'s obtained by the variance-based methods, the SCSA method yields much more information and a clear view of the roles played by the inputs and their correlations.

As this model is a linear function and its pdf is a multivariate normal distribution, the relationship between S_i, S_i^a, S_i^b and the function $f(\mathbf{x})$ structure along with the pdf can be readily examined. For more complex systems, this relation will likely not be so simple, but the indices S_i, S_i^a, S_i^b should still correctly represent the structural and correlative contributions of the inputs to the variance of the output.

3.2. Simulated Model: A Nonlinear Function with Uniform and Multivariate Normal Distributions of Inputs. To compare the RS-HDMR method with other existing meta-models, a nonlinear function with three inputs

$$f(\mathbf{x}) = \sin(2\pi x_1 - \pi) + 7\sin^2(2\pi x_2 - \pi) + 0.1(2\pi x_3 - \pi)^4 \sin(2\pi x_1 - \pi) + \varepsilon \qquad (46)$$

is used for illustration. Storlie, et al.[15] calculated the total sensitivity indices S_{Ti} for this model with independent sampling using 10 different meta-models. Five meta-models, Adaptive COmponent Selection and Smoothing Operator (ACOSSO), Gaussian Process with Maximum Likelihood Estimation (MLE GP), Gaussian Process with MLE and Bayes estimates and Bayesian credible sets for S_{Ti} (MLE BGP), Recursive Partitioning (RPART), and Multivariate Adaptive Regression Splines (MARS), gave better results than the other five methods. This nonlinear system was also treated by RS-HDMR with independent and correlated sampling. For comparison, we used the same signal-to-noise ratio SNR = $\mathrm{Var}(f(\mathbf{x}))/\sigma^2 = 55$ used by Storlie.

TABLE 4: Sensitivity Indices for the Nonlinear System with a Uniform Distribution

input	S_i^a or S_{ij}^a	S_i^b or S_{ij}^b	S_i or S_{ij}	S_{Ti}
X_1	0.31 ± 0.06	0.01 ± 0.05	0.31 ± 0.06	0.51 ± 0.07
X_2	0.44 ± 0.09	0.00 ± 0.05	0.44 ± 0.07	0.44 ± 0.07
X_3	0.04 ± 0.03	0.01 ± 0.01	0.04 ± 0.04	0.24 ± 0.05
(X_1, X_3)	0.21 ± 0.05	− 0.01 ± 0.05	0.20 ± 0.05	

3.2.1. Case 1: Independent Sampling with a Uniform Distribution. We first sampled the data with a uniform distribution, and 100 sets of random samples for $\mathbf{x}$, each of the size $N = 300$, were generated following the procedure of Storlie et al.[15] RS-HDMR was used with these data sets. The functions $f_i(x_i)(i = 1, 2, 3)$ and $f_{13}(x_1, x_3)$ were identified to be significant. For independent sampling, the SCSA method and the variance-based methods should produce the same results, i.e., our results for S_{Ti} should coincide with Storlie's results. The outcome of the SCSA method is given in Table 4.

Since the inputs are independent, we should have $S_i^b = S_{ij}^b = 0$ and $S_i = S_i^a$, $S_{ij} = S_{ij}^a$. The true values of the total sensitivity indices are known[15] to be $S_{T1} = 0.55$, $S_{T2} = 0.45$, $S_{T3} = 0.24$. Since $S_1 + S_2 + S_3 + S_{13} = 0.99 \pm 0.03$ is close to unity, we can calculate the total sensitivity indices as

$$S_{T1} = S_1 + S_{13}, \quad S_{T2} = S_2, \quad S_{T3} = S_3 + S_{13} \qquad (47)$$

which are also given in Table 4. The results in Table 4 are quite satisfactory.

Storlie et al.[15] calculated the root mean squared error (RMSE) R_i for the resultant S_{Ti} from the 100 data sets as

$$R_i = \sqrt{\frac{1}{100}\sum_{r=1}^{100}(S_{Ti}^{(r)} - S_{Ti})^2} \qquad (48)$$

where $S_{Ti}^{(r)}$ denotes the value for the rth data set, and S_{Ti} is the known true value. The standard deviation s_{R_i} for R_i was also calculated by

$$s_{R_i} = \frac{1}{2R_i}\sqrt{\frac{1}{99}\sum_{r=1}^{100}[(S_{Ti}^{(r)} - S_{Ti})^2 - R_i^2]^2} \qquad (49)$$

R_i and s_{R_i} were used to compare the accuracy of the 10 methods in Storlie's work. Here we give the values of R_i and s_{R_i} for RS-HDMR and compare them with the five best meta-models used by Storlie et al. Table 5 shows that RS-HDMR gives results comparable to these methods.

3.2.2. Case 2: Correlated Sampling with a Multivariate Normal Distribution. The above results show the reliability of the SCSA method. The advantage of the SCSA method is that it can treat correlated sampling. Here, a joint multinormal distribution of inputs

$$\sum = \begin{bmatrix} 1.0 & 0.6 & 0.0 \\ 0.6 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \tag{50}$$

is used with the model in eq 46. Figure 1 gives scatter plots of output Y against the inputs X_i for (a) multinormal and (b) uniform distributions.

For the multinormal distribution, the values of the X_i's are concentrated in the center over the range (0.2, 0.8). In this region, the relation between Y and X_1, X_3 is quite flat, and the corresponding sensitivity indices should be small. In contrast, the contribution of the X_i's shows considerable variation in the uniform distribution. The resultant sensitivity indices for the multinormal distribution are given in Table 6, consistent with the scatter plots. The sensitivity indices of X_1, X_3 and (X_1, X_3) for the multinormal distribution are much smaller than those for the uniform distribution, and X_2 has a dominant contribution to the variance of the output.

TABLE 6: Sensitivity Indices for the Nonlinear System with a Multinormal Distribution

input	S_i^a or S_{ij}^a	S_i^b or S_{ij}^b	S_i or S_{ij}	S_{Ti}
X_1	0.10 ± 0.06	-0.03 ± 0.06	0.07 ± 0.04	0.08 ± 0.06
X_2	0.91 ± 0.11	-0.04 ± 0.07	0.87 ± 0.07	0.87 ± 0.07
X_3	0.01 ± 0.02	0.00 ± 0.02	0.01 ± 0.02	0.03 ± 0.03
(X_1, X_3)	0.03 ± 0.03	-0.02 ± 0.04	0.02 ± 0.03	

Table 6 also correctly shows that $S_1^b \approx S_2^b$ and $S_3^b \approx 0$. Similarly, since $S_1 + S_2 + S_3 + S_{13} = 0.96 \pm 0.04$ is close to unity, then the S_{Ti}'s should be reliable. Therefore, the resultant sensitivity indices correctly identify the important contributions of the inputs for the multinormal distribution data.

3.3. Application to Measured Ionosonde Field Data. The SCSA method has also been utilized for treating ionosonde data

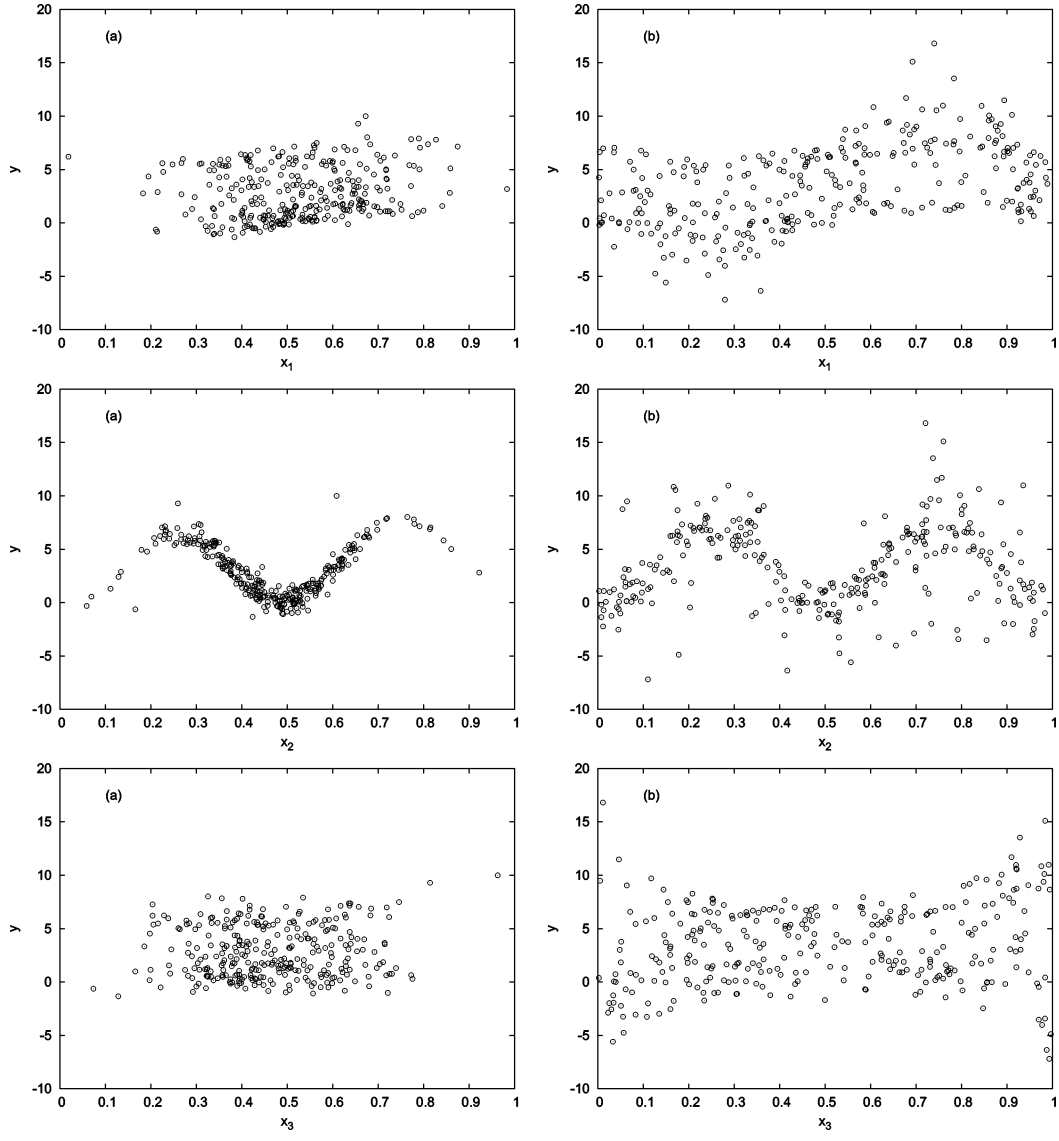

Figure 1. Scatter plots of output y against inputs x_i for (a) multinormal and (b) uniform distribution data.

TABLE 5: Comparison between RS-HDMR and the Five Meta-Models Used by Storlie et al.

	RS-HDMR	ACOSSO	MLE GP	MLE BGP	RPART	MARS
$R_1(s_{R_1})$	0.06(0.03)	0.06(0.00)	0.08(0.02)	0.05(0.01)	0.07(0.01)	0.10(0.04)
$R_2(s_{R_2})$	0.04(0.02)	0.05(0.00)	0.08(0.01)	0.12(0.02)	0.08(0.01)	0.09(0.04)
$R_3(s_{R_3})$	0.02(0.02)	0.07(0.00)	0.09(0.02)	0.04(0.01)	0.13(0.01)	0.11(0.05)

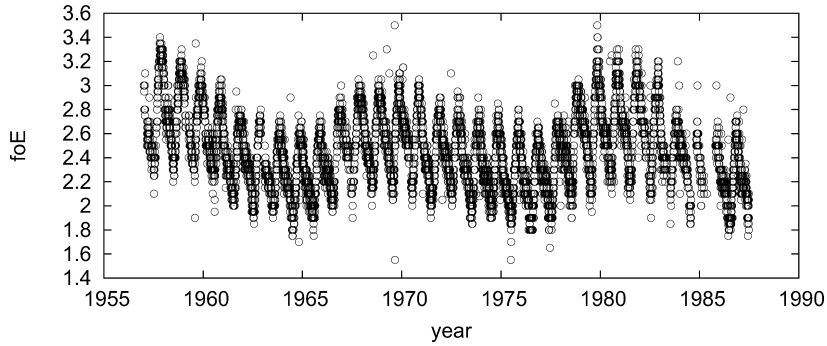

Figure 2. The measured *fo*E data at 12:00 UT from years 1957−1987. The yearly variations are superimposed on the 11 year solar cycle.

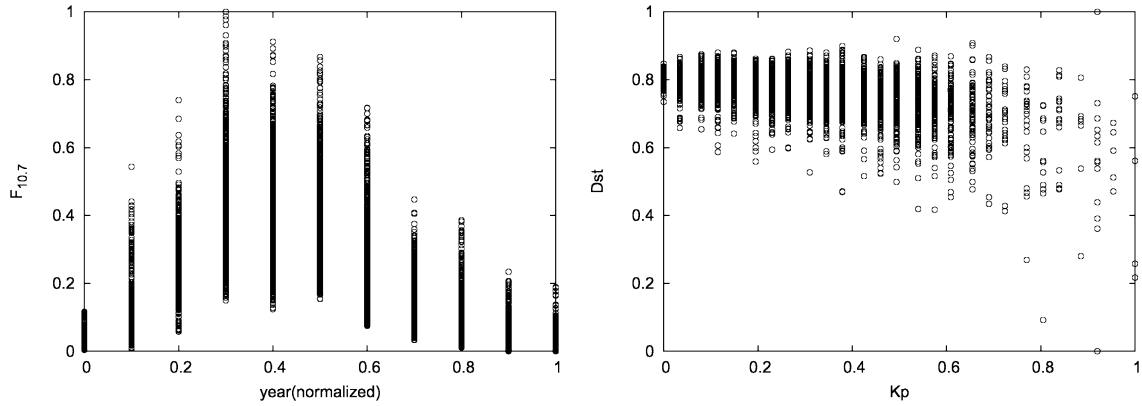

Figure 3. The correlation of the measured ionosonde field data for some of the input variables.

in Huancayo, Peru. A brief description of the physics is given here, and more details can be found elsewhere.[30] The ionosonde transmits radio wave signals that are reflected when the transmitted frequency is equal to the local plasma frequency in the ionosphere. Electron densities as a function of altitude and for a given time are calculated from these returned frequencies. The ionospheric electron density is characterized by the critical frequencies returned from the peak density in the E-region (*fo*E) and the peak density in the F-region (*fo*F2) of the ionosphere. *fo*E and *fo*F2 vary periodically in time. Figure 2 gives the measured data for *fo*E at 12:00 universal time (UT) for the years 1957−1987.

The critical frequencies *fo*E and *fo*F2 follow regular yearly variations superimposed over the 11 year solar cycle. The ionosphere exhibits much greater day-to-night variations within a 24 h period than it does at the same hour from day-to-day. The following analysis is for the specific hour 12 UT. The critical frequencies *fo*E and *fo*F2 may also be dependent on the measured geophysical parameters $F_{10.7}$, *Kp*, and *Dst*. Here $F_{10.7}$ represents the 10.7 cm solar flux index which is a surrogate for solar output: high values of $F_{10.7}$ occur during a solar maximum, and low values occur during a solar minimum. *Kp* is a 3-hourly index of the solar particle radiation derived from geomagnetic field variations measured at 13 subauroral locations. *Dst* is also an index based on the geomagnetic field, which is derived from mid- and low-latitude sites and reflects occurrences of magnetic storms. Therefore, the five input variables are "year" (considering a 11 year period, which is transformed to year = (year − 1957 mod 11)), "day", $F_{10.7}$, *Kp*, and *Dst*. Since the inputs were determined from ground-based ionosonde measurements, their sampling is not controlled, the variables are likely not independent, and the input pdf is not explicitly known. Figure 3 shows the correlation among some of the inputs.

The following analysis presents illustrative results for the output *fo*E. An RS-HDMR meta-model was constructed from the first 4000 points of 8694 measured data samples (training data), and another set of 4694 points are used for testing. To treat all input variables in a common fashion, they are normalized, i.e., $0 \le x_i \le 1 (i = 1, 2, ..., 5)$. The RS-HDMR component functions are approximated by cubic B splines. The backfitting algorithm was used to determine the expansion coefficients (i.e., α_r^i, β_{pq}^{ij}, ...) for the component functions, and the *F*-test was used to determine which component functions should be included. With a 99% confidence level, four $f_i(x_i)$ and two $f_{ij}(x_i, x_j)$ were identified as significant. A second-order RS-HDMR expansion with these component functions was constructed. The R^2 of the meta-model prediction is 0.93. The average relative errors for the training and testing data are 3.38% and 4.22%, respectively. The data portion with relative error less than a given value is shown in Table 7, and a comparison is shown in Figure 4 between the measured yearly variation of *fo*E and the second-order RS-HDMR approximation.

TABLE 7: The Relative Errors of Different Order RS-HDMR Approximations for Training and Testing Data

	data portion			
	training data		testing data	
relative error (%)	first order	second order	first order	second order
1	0.3090	0.3177	0.2793	0.2874
5	0.9125	0.9243	0.8617	0.8611
10	0.9925	0.9928	0.9808	0.9800
20	0.9972	0.9972	0.9962	0.9964

The results in Table 7 show that, for the second-order RS-HDMR approximation, more than 90% and 99% of the training data have relative errors of less than 5% and 10%, respectively. The accuracy for testing data is similar to training data, i.e., more than 86% and 98% of the testing data have relative errors of less than 5% and 10%, respectively. In Figure 4 the second-order RS-HDMR approximation constructed from the first ∼14

TABLE 8: The First-Order Sensitivity Indices Obtained from the SCSA Method and the Variance-Based Method for Measured Ionosonde Data (*fo*E)

	the SCSA method				
input	relative import.	S_i^a	S_i^b	S_i	$V_i/V(Y)$
day (X_2)	1	0.57	0.00	0.57	0.58
$F_{10.7}$ (X_3)	2	0.26	0.04	0.30	0.36
year (X_1)	3	0.01	0.04	0.06	0.29
Dst (X_5)	4	0.00	−0.01	−0.01	0.06
Kp (X_4)	[a]				0.02
sum		0.84	0.08	0.92	1.31

[a] Insignificant.

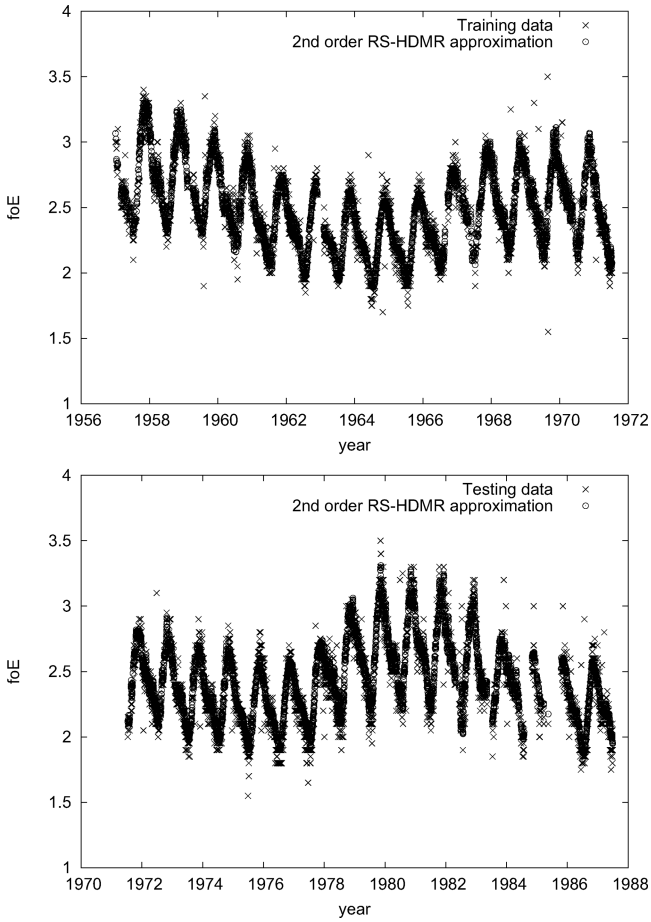

Figure 4. The comparison of the measured yearly variation of *fo*E and the RS-HDMR approximation.

years of measured data satisfactorily predicts *fo*E values for the following ∼16 years.

The sensitivity indices S_i, S_i^a, S_i^b obtained by the SCSA method from the resultant RS-HDMR component functions along with $V_i/V(Y)$ obtained by the variance-based methods are given in Table 8. Since the probability distribution is correlated, the coefficients S_i^b are not negligible for the measured ionosonde field data. The largest values of S_i^b are for $F_{10.4}(X_3)$ and "year" (X_1), which is consistent with Figure 3 already showing their correlation. All of the $f_i(X_i)$'s are identified to be significant except for $f_4(X_4)$ (*Kp*). From S_i and S_i^a, the most influential input is "day" (X_2), followed by $F_{10.7}(X_3)$, then "year" (X_1), and the least influential input is $Dst(X_5)$. $V_i/V(Y)$ gives the same order, but it does not distinguish the structural and correlative contributions of the inputs.

The significant second-order sensitivity indices are given in Table 9, and they are small. Since the sum $\sum S_i + \sum S_{ij} = 0.93$ is close to unity, the sensitivity analysis given by the SCSA method is reliable. The total sensitivity indices are given in Table 10. The relative importance order given by S_{Ti} or S_{Ti}^a is the same as that given by S_i or S_i^a. The similarity of ordering is reasonable because $\sum S_i = 0.92$, i.e., the first-order sensitivity indices dominate the analysis in this case.

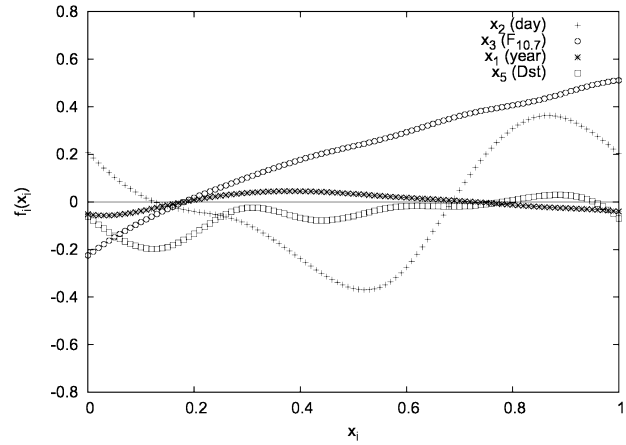

Figure 5. The four significant $f_i(x_i)$.

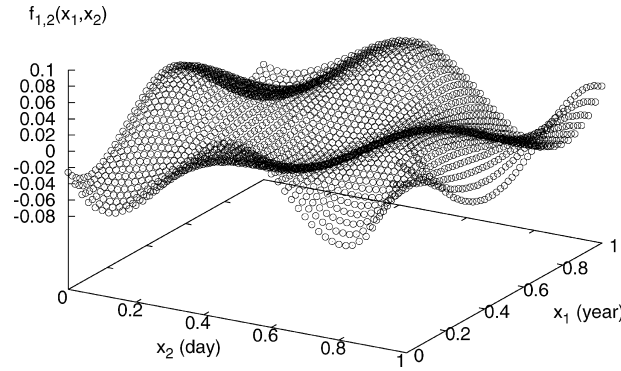

Figure 6. The significant function $f_{1,2}(x_1, x_2)$.

TABLE 9: The Second-Order Sensitivity Indices for the Measured Ionosonde Data (*fo*E)

inputs	relative import.[a]	S_{ij}^a	S_{ij}^b	S_{ij}
(X_1, X_2)	1	0.01	0.01	0.02
(X_2, X_3)	2	0.00	−0.02	−0.01
$\sum S_i$				0.92
$\sum S_i + \sum S_{ij}$				0.93

[a] Determined from S_{ij}^a.

TABLE 10: The Totals Ensitivity Indices S_{Ti}^a, S_{Ti}^b, S_{Ti} for the Measured Ionosonde Data (*fo*E)

input	relative import.	S_{Ti}^a	S_{Ti}^b	S_{Ti}
day (X_2)	1	0.58	−0.00	0.58
$F_{10.7}$ (X_3)	2	0.26	0.03	0.29
year (X_1)	3	0.02	0.06	0.08
Dst (X_5)	4	0.00	−0.01	−0.01

The component functions evaluated from RS-HDMR not only determine the magnitudes of the sensitivity indices, but the functions $f_i(x_i)$, $f_{ij}(x_i, x_j)$,... also provide qualitative descriptions of the influence patterns. Figure 5 gives all four significant functions $f_i(x_i)$, and Figure 6 shows $f_{1,2}(x_1, x_2)$. Figure 5 shows that $f_2(x_2)$ (day) and $f_3(x_3)$ ($F_{10.4}$) are large, while $f_1(x_1)$ (year) and $f_5(x_5)$ (*Dst*) are small. The function $f_3(x_3)$ monotonically increases with respect to $x_3(F_{10.4})$, while $f_2(x_2)$ changes over different seasons. The function $f_1(x_1)$ changes smoothly over

the 11 year period, while $f_5(x_5)$ has a more complex pattern with respect to $x_5(Dst)$. Figure 6 shows a very structured pattern for the cooperative influence of x_1 (year) and x_2 (day). The quantitative and qualitative information given by the RS-HDMR component functions and global sensitivity analysis are physically rich for this application.

4. Conclusions

The new global sensitivity analysis method introduced in this paper provides a unified framework for the treatment of systems whose input probability distribution may be independent and/or correlated. The analysis can be applied to mathematical models as well as to measured laboratory or field data. This paper established that the unconditional variance of the output can be decomposed into the covariances of the RS-HDMR component functions with the output. The covariance for each component function comprises two terms consisting of a structural contribution reflecting the nature of the system and a correlative contribution related to the input probability distribution. Three sensitivity indices S_{p_j}, $S^a_{p_j}$, and $S^b_{p_j}$ are defined to represent the total, structural, and correlative contributions of a single or a subset of inputs $\mathbf{X}_{p_j}$. The magnitudes of S_{p_j}, $S^a_{p_j}$, and $S^b_{p_j}$ all need to be considered in order to quantitatively determine the relative importance of the inputs acting either independently or collectively. When the inputs are independent, the three indices reduce to a single index, S_{p_j}, equivalent to that specified by the previous variance-based methods. The global sensitivity analysis given by the variance-based methods with independent inputs is a special case of the general treatment of the SCSA method for systems with arbitrary input probability distributions.

The estimation of the sensitivity indices is based on a meta-modeling approach in this paper, specifically, the RS-HDMR expansion. The RS-HDMR component functions are approximated by cubic B splines whose coefficients are determined by a backfitting procedure combining a statistical *F*-test for the identification of the significant component functions. This approach is especially useful for laboratory or field data where the sampling is often uncontrolled. Other meta-models could be employed as well. After the component functions $f_{p_j}(\mathbf{x}_{p_j})$'s in eq 12 are obtained, the sensitivity indices, S_{p_j}, $S^a_{p_j}$, and $S^b_{p_j}$, can be calculated using eqs 33−35.

A simple linear function with five input variables, which had either an independent or correlated multinormal distribution, was used for illustration. For the case of independent inputs, the results given by the SCSA method and the variance-based methods are almost the same, but for correlated inputs, the information obtained by the SCSA method provides further insights and is more reliable. In the nonlinear simulation model with three inputs, the results obtained from RS-HDMR correspond very well with the results given by other existing meta-models for independent sampling. The SCSA method was also applied to correlated inputs for this nonlinear simulation model. The characterization of ionospheric electron density from measured field data was used to successfully test the SCSA method under realistic conditions, including with no knowledge of the input pdf.

In summary, the correlation of inputs is very common in realistic applications. The variance-based methods are unable to properly treat such cases, which makes the sensitivity analysis of measured field and laboratory data a challenging task. The SCSA method provides a practical means to meet this need based on a covariance decomposition of the unconditional variance of the output.

Acknowledgment. This work was supported by the U.S. Army Small Business Technology Transfer (STTR) program under contract number W911NF-06-C-0181 to Aerodyne Research, Inc. and Princeton University. Support for this work has also been provided partially by the USEPA through the Center for Exposure and Risk Modeling (CERM - EPAR827033) and the Environmental Bioinformatics and Computational Toxicology Center (ebCTC - GAD R 832721-010).

References and Notes

(1) Oakley, J. E.; O'Hagan, A. *J. R. Stat. Soc. B* **2004**, *66*, 751.
(2) Saltelli, A.; Ratto, M.; Tarantola, S.; Campolongo, F. *Chem. Rev.* **2005**, *105*, 2811.
(3) Ratto, M.; Pagano, A.; Young, P. *Comput. Phys. Commun.* **2007**, *177*, 863.
(4) Cox, D. C. *IEEE Trans. Reliab.* **1982**, *31*, 265.
(5) Sobol, I. M. *Math. Model. Comput. Exp.* **1993**, *1*, 407.
(6) Homma, T.; Saltelli, A. *Reliab. Eng. Syst. Saf.* **1996**, *52*, 1.
(7) Saltelli, A.; Chan, K.; Scott, E. M., Eds.; *Sensitivity Analysis*; John Wiley & Sons, Ltd: New York, 2000.
(8) Rabitz, H.; Alis, Ö. F. *J. Math. Chem.* **1999**, *25*, 197.
(9) Saltelli, A.; Tarantola, S. *J. Am. Stat. Assoc.* **2002**, *97*, 702.
(10) McKay, M. D. *Evaluating Prediction Uncertainty, Report NUREG/CR-6311*, US Nuclear Regulation Commission and Los Alamos National Laboratory, **1995**.
(11) Saltelli, A.; Tarantola, S.; Chan, P. S. *Technometrics* **1999**, *41*, 39.
(12) McKay, M. D.; Beckman, R. J.; Conover, W. J. *Technometrics* **1979**, *21*, 239.
(13) Ziehn, T.; Hughes, K. J.; Griffiths, J. F.; Porter, R.; Tomlin, A. S. *Combust. Theory Modell.* **2009**, *13*, 589.
(14) Li, G.; Wang, S. W.; Rabitz, H.; Wang, S. K.; Jáffe, P. *Chem. Eng. Sci.* **2002**, *57*, 4445.
(15) Storlie, C. B.; Swiler, L. P.; Helton, J. C.; Sallaberry, C. J. *Reliab. Eng. Syst. Saf.* **2009**, *94*, 1735.
(16) Storlie, C. B.; Helton, J. C. *Reliab. Eng. Syst Saf.* **2008**, *93*, 28.
(17) Li, G.; Rosenthal, C.; Rabitz, H. *J. Phys. Chem. A* **2001**, *105*, 7765.
(18) Li, G.; Wang, S. W.; Rabitz, H. *J. Phys. Chem. A* **2002**, *106*, 8721.
(19) Li, G.; Hu, J. S.; Wang, S. W.; Georgopoulos, P. G.; Schoendorf, J.; Rabitz, H. *J. Phys. Chem. A* **2006**, *110*, 2474. Note that eqs 25 and 26 in this paper are valid only for systems with independent sampling, i.e., $w(x) = \prod w_i(x_i)$. For correlated sampling, eqs 25 and 26 can be used as approximations.
(20) Prenter, P. M. *Splines and Variational Methods*; John Wiley & Sons: New York, 1989; pp 7−115.
(21) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2001.
(22) Deutsch, F. *Best Approximation in Inner Product Space*; Springer-Verlag, Inc: New York, 2001; Chapter 4.
(23) Xu, C.; Gertner, G. Z. *Reliab. Eng. Syst. Saf.* **2008**, *93*, 1563.
(24) Berk, R. A. *Regression Analysis: A Constructive Critique*; Sage Publications: Thousand Oaks, CA, 2004.
(25) Freedman, D. A. *Statistical Models: Theory and Practice*; Cambridge University Press: Cambridge, NY, 2005.
(26) Wahba, G.; Wang, Y.; Gu, C; Klein, R.; Klein, B. *Ann. Stat.* **1995**, *23*, 1865.
(27) Gu, C. *Smoothing Spline ANOVA Models*; Springer: New York, 2002.
(28) Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models*; Chapman and Hall; New York, 1990.
(29) Evans, M. J.; Rosenthal, J. S. *Probability and Statistics: The Science of Uncertainty*; W. H. Freeman and Company; New York, 2004.
(30) Space Physics Interactive Data Resource (SPIDR) web page, National Geophysical Data Center (NGDC), NOAA Satellite and Information Service, 2009. (http://spidr.ngdc.noaa.gov/spidr).

JP9096919