Reliability Engineering and System Safety 112 (2013) 187–199

Contents lists available at SciVerse ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Separating the contributions of variability and parameter uncertainty in probability distributions

S. Sankararaman, S. Mahadevan *

Vanderbilt University, Department of Civil and Environmental Engineering, Nashville, TN 37235, USA

ARTICLE INFO

Article history:
Received 7 April 2011
Received in revised form
4 November 2012
Accepted 7 November 2012
Available online 11 December 2012

Keywords:
Variability
Family of distributions
Distribution parameter uncertainty
Sparse data
Interval data
Aleatory uncertainty
Epistemic uncertainty
Sensitivity analysis

ABSTRACT

This paper proposes a computational methodology to quantify the individual contributions of variability and distribution parameter uncertainty to the overall uncertainty in a random variable. Even if the distribution type is assumed to be known, sparse or imprecise data leads to uncertainty about the distribution parameters. If uncertain distribution parameters are represented using probability distributions, then the random variable can be represented using a family of probability distributions. The family of distributions concept has been used to obtain qualitative, graphical inference of the contributions of natural variability and distribution parameter uncertainty. The proposed methodology provides quantitative estimates of the contributions of the two types of uncertainty. Using variance-based global sensitivity analysis, the contributions of variability and distribution parameter uncertainty to the overall uncertainty are computed. The proposed method is developed at two different levels; first, at the level of a variable whose distribution parameters are uncertain, and second, at the level of a model output whose inputs have uncertain distribution parameters.

1. Introduction

In engineering design analysis, it is often required to calculate the uncertainty in a system response $Y = g(\boldsymbol{X})$ given the uncertainty in the input quantities $\boldsymbol{X}$. While uncertainty propagation methods such as Monte Carlo simulation (MCS), first-order reliability method (FORM), second-order reliability method (SORM), etc. have received much attention in the reliability literature [1], practical engineers still find it difficult to use these methods when limited data about input variables is available. In the context of a probabilistic approach, two difficulties are encountered. The first is the choice of the distribution type. Even if the distribution type is known (from previous experience or expert opinion), the second difficulty is lack of adequate data to estimate the distribution parameters with a high degree of confidence. *This paper focuses on the second difficulty, i.e., distribution parameter uncertainty and assumes that the distribution type is known.*

Consider a quantity with natural or physical variability (i.e., aleatory uncertainty) being represented using a probability distribution. A distribution type (e.g., normal, lognormal, etc.) is assumed for the quantity of interest, and the parameters of this probability distribution (e.g., mean and standard deviation in the case of a normal distribution) are usually estimated using techniques of statistical inference using observed data. When sufficient data is available, it may be reasonable to compute deterministic estimates of the parameters using techniques such as the method of moments [1], the method of maximum likelihood [2], etc. It can be proved that, under some conditions, these deterministic estimates approach the true estimates as the number of data approaches infinity [3]. When only finite data is available, there is uncertainty associated with these estimates; the importance of this uncertainty in the distribution parameters increases especially when the data is sparse and/or imprecise. While a deterministic parameter estimate completely characterizes the variability in the quantity of interest, the uncertainty in the parameter estimate adds to the uncertainty regarding the quantity of interest (assuming the distribution type is known).

The uncertainty in the distribution parameters is one example of epistemic uncertainty (which can be reduced with additional observational data). The topic of distribution parameter uncertainty has been studied by several researchers in the past [4–6] and this has also been referred to as statistical uncertainty [1,7] or second-order uncertainty [8,9]. Therefore, the total uncertainty in the quantity of interest is composed of two parts: variability and parameter uncertainty, assuming the distribution type is known. This paper proposes a computational methodology to compute the individual contributions of variability and parameter uncertainty to the overall uncertainty in the quantity of interest.

However, it is first essential to quantify the uncertainty in the distribution parameters based on the available data. There are two issues that need to be addressed: (1) The uncertainty in the

* Corresponding author. Tel.: +1 615 322 3040; fax: +1 615 322 3365.
E-mail address: sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

0951-8320/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ress.2012.11.024

parameter estimates must be computed in such a way that it is not only meaningful and easily interpreted, but also suitable for further probabilistic calculations such as uncertainty propagation, reliability analysis, etc. (2) In addition, previous studies have dealt with parameter estimation only in the presence of point data. However, the available data may also be in the form of intervals; in this case, the estimation of distribution parameters may not be straightforward.

Classical statistics-based frequentist methodology addresses the uncertainty in the distribution parameters by estimating statistical confidence intervals on the distribution parameters [3]. These intervals are calculated at some chosen significance level α. The interpretation of this interval is as follows: For a 95% confidence interval, the probability that the resultant confidence interval will contain the true value of the distribution parameter is equal to 0.95. In other words, if one were to collect 100 different sets of data and a confidence interval was calculated for each data set, then 95 of the 100 intervals would contain the true value of the distribution parameter. This is significantly different from the statement "the probability that the distribution parameter lies between the bounds of the interval is equal to 0.95". In the context of classical statistics, the distribution parameter is deterministic but unknown, and hence it is not meaningful to talk about the probability or the probability distribution of the distribution parameter [10]. Hence, statistical confidence intervals cannot be used further in uncertainty propagation, reliability analysis, etc.

In contrast, Bayesian statistics treats probability distributions not as relative frequencies of occurrence but as the state of belief or knowledge of the quantity of interest [10–14]. Therefore, the Bayesian approach can represent the uncertainty about the distribution parameters (which may be deterministic but unknown) through probability distributions, which are subjective estimates of probability [15,16], and facilitates further probabilistic calculations (in contrast with the confidence intervals approach). Let X denote the quantity of interest whose probability density function (PDF) is given by $(f_X(x|\boldsymbol{p})$ where p refers to the distribution parameters of the random variable X, and x is a realization of X. Suppose that the Bayesian approach yielded the PDF of the distribution parameters, denoted by $f_{\boldsymbol{P}}(\boldsymbol{p})$. Each realization $\boldsymbol{p}$ of the distribution parameters $\boldsymbol{P}$ results in a corresponding probability distribution (whose density is $f_X(x|\boldsymbol{P}=\boldsymbol{p})$) for the quantity of interest X. Thus, multiple realizations of distribution parameters result in a family of distributions for the quantity X. This family of distributions is shown in terms of the probability density function (PDF) and the cumulative distribution function (CDF) in Fig. 1.

Hence, the uncertainty estimates of distribution parameters obtained through the Bayesian approach can be easily used in further probabilistic calculations. This solves the first issue. The second issue is the presence of interval data for which parameter estimation may not be straightforward. Several researchers have used non-probabilistic methods [17] to tackle interval data; these methods include evidence theory [18–21], convex models of uncertainty [22], Zadeh's extension principle [21], fuzzy set theory [18], etc. Non-probabilistic methods may not use distribution parameters for uncertainty representation, and do not aid in the objectives of the present paper. Earlier, the authors have proposed [23–25] a likelihood-based approach for constructing a PDF for the random variable, based on available sparse point data and interval data. This likelihood-based approach calculates the probability distribution of the distribution parameters, and therefore, solves the second issue of parameter estimation using interval data, and is summarized later in Section 2.

Thus, having addressed both the issues of (1) estimating the uncertainty in distribution parameters, and (2) extending parameter estimation from point data to include interval data, the rest of this paper will focus on the overall goal of quantitatively distinguishing the contributions of parameter uncertainty and natural variability in the probability distribution of a particular quantity.

Consider the probability distributions in Fig. 1. Each PDF represents the natural variability in the quantity of interest, and the spread of the family of PDFs represents the uncertainty in the distribution parameters. This is only a qualitative (graphical) representation. The objective of the current paper is to quantify the individual contributions of these two types of uncertainty to the overall uncertainty in the variable.

The motivation for the proposed methodology is developed in two steps. First, the family of distributions-based approach is computationally expensive in the case of uncertainty propagation problems. This approach involves a nested (double loop or second-order) Monte Carlo sampling procedure, where the distribution parameters are sampled in the outer loop and the values of the variable are sampled in the inner loop. Therefore, in this paper, the family of distributions is replaced by an equivalent single probability distribution using the concepts of conditional probability and total probability; this single probability distribution includes both the variability and the uncertainty in the distribution parameters. The resultant distribution has referred to as a Bayesian predictive posterior distribution [4,5,7]. However, it may be argued that using a single distribution confounds the contribution of natural variability and distribution parameter uncertainty, and appears to lose the ability to see their individual contributions as in Fig. 1. Separating the individual contributions of variability and distribution parameter uncertainty will help in resource allocation for data collection, since natural variability is irreducible whereas parameter uncertainty is reducible.

The above argument leads to the second step and the main contribution of this paper, i.e., a computational method that explicitly calculates the individual contributions of natural variability and distribution parameter uncertainty. The tools of global sensitivity analysis [26,27] are used for this purpose.

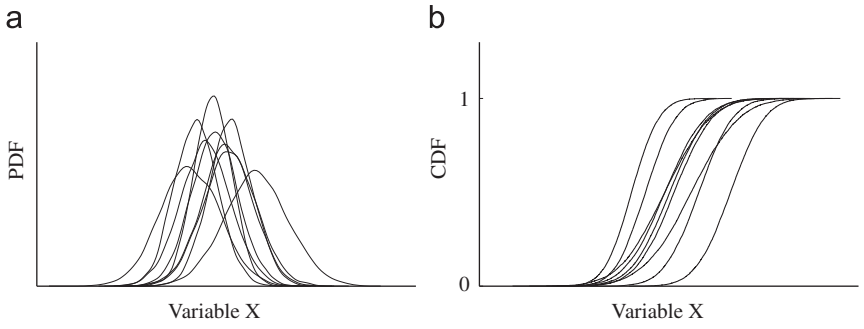

Fig. 1. Family of distributions: (a) probability density function and (b) cumulative distribution function.

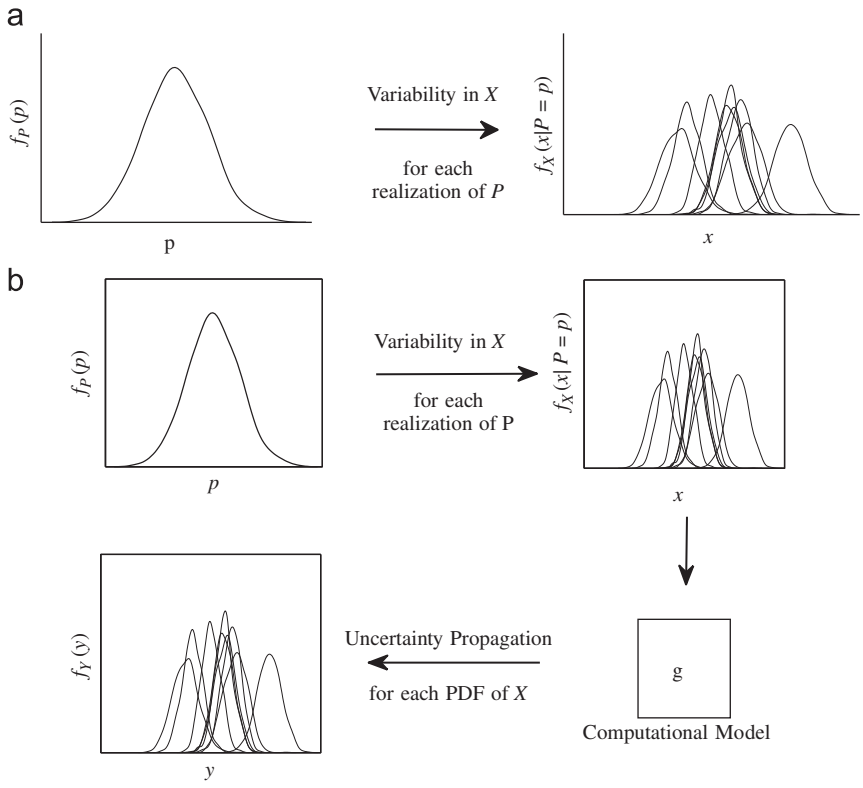

Fig. 2. Two types of problems: (a) problem type P1 and (b) problem type P2.

Two types of problems are considered in this paper (see Fig. 2):

1. Problem P1: Analysis of contributions of variability and distribution parameter uncertainty within a single random variable X. There may be more than one distribution parameter pertaining to this random variable; however, for the sake of illustration, only one distribution parameter (P instead of $\boldsymbol{P}$) and its PDF is indicated in Fig. 2(a).
2. Problem P2: Analysis of contributions of variability and distribution parameter uncertainty in multiple input random variables $\boldsymbol{X}$ to the output Y of a response function $g(\boldsymbol{X})$, i.e., $Y=g(\boldsymbol{X})$. For the sake of illustration, only one input quantity (X instead of $\boldsymbol{X}$) is shown in Fig. 2(b).

Note that the formulations of these two problems are considerably different from the problems formulated in the global sensitivity analysis literature. The method of global sensitivity analysis considers a deterministic one-to-one mapping from $\boldsymbol{X}$ to Y in the form of $Y=g(\boldsymbol{X})$, i.e., for a given realization $\boldsymbol{x}$ of the inputs $\boldsymbol{X}$ (the input is a vector, i.e., $\boldsymbol{X}=\{X^1,X^2,X^3\ldots X^n\}$) and a computational model G, there exists a corresponding realization y of the output Y. When the inputs $\boldsymbol{X}$ are random, the output Y is also random and needs to be expressed using a probability distribution. The method of global sensitivity analysis is based on the variance decomposition equation [27] which is true in the presence of a deterministic transfer function $Y=g(\boldsymbol{X})$, and apportions the variance of Y to the variance of each of the inputs X^i ($i=1$ to n). Sometimes, the output may be a vector ($\boldsymbol{Y}$ instead of Y), and it may be necessary to repeat the global sensitivity analysis procedure for each of the Y's. (In practical systems, there may be model uncertainty in G; the focus of the present paper is on quantifying the individual contributions of variability and distribution parameter uncertainty in probability distributions, and therefore, model uncertainty is not addressed here.)

It is not straightforward to apply the method of global sensitivity analysis to problems P1 and P2. In problem P1, there is no computational model at all; there are distribution parameters which in turn have their own distributions and therefore, a particular realization of the distribution parameters yields an entire distribution for X. In problem P2, though there is a computational model, each model input has uncertain distribution parameters. For a given sample of distribution parameters, there is a probability distribution for each input and hence a probability distribution for the output. Therefore, the output is represented using a family of distributions (see Ref. [28] for several practical examples of this kind). Even if each input were replaced by a single equivalent probability distribution, one can only calculate the relative contributions of the multiple input quantities and not separately quantify the contributions of variability and distribution parameter uncertainty in each input quantity, using currently available methods. Hence, the extension of global sensitivity analysis to solve this problem is not trivial and is the most important aspect of this paper.

The following sections describe the proposed methodology in detail. Section 2 describes the treatment of distribution parameter uncertainty, which is based on the authors' earlier work [23]. Section 3 develops the new global sensitivity analysis-based approach to quantify the individual contributions of variability and distribution parameter uncertainty to the overall uncertainty. Section 4 illustrates the proposed methodology using numerical examples.

2. Treatment of distribution parameter uncertainty

Consider a random variable X, which is defined in terms of a probability space triplet [29], given by $(\mathcal{X}, \mathbb{X}, \mathcal{P}_X)$. All further discussion regarding probability space triplets are presented for continuous random variables, since this is the primary focus of

the present paper. Therefore, $\mathcal{X} \subseteq \mathbb{R}$ (where $\mathbb{R}$ denotes the set of real numbers) is the sample space of the random variable X. The σ-algebra on $\mathcal{X}$ is denoted by $\mathbb{X}$ and consists of all intervals and satisfies the assumptions detailed in [30]. The probability measure $\mathcal{P}_X$ is a mapping from every element $\mathcal{E}_X \in \mathbb{X}$ to a number between 0 and 1. Conventionally, this mapping is specified in terms of the probability density function $f_X(x)$ (or the cumulative distribution function $F_X(x)$). When both the distribution type and distribution parameters are known, then the probability space triplet and the probability distribution function are unambiguously known. However, when either of them is uncertain, then the probability space triplet is not unambiguously known.

As mentioned earlier in Section 1, the distribution type is assumed to be known in this paper, and the focus is only on the uncertainty in the distribution parameters. These distribution parameters may be deterministic in reality; however, since they are unknown, they can be treated as random variables with probability distributions following the Bayesian interpretation of probability, according to which "lack of information" can be perceived to be equivalent to randomness [31]. Let $\boldsymbol{P}$ denote the random variables corresponding to the distribution parameters, and let $\boldsymbol{p}$ denote a particular realization of $\boldsymbol{P}$. If there are q distribution parameters, then $\boldsymbol{P} = \{P_i; i = 1 \text{ to } q\}$. The joint PDF of the distribution parameters is written as $f_{\boldsymbol{P}}(\boldsymbol{p})$, and the PDFs for each of the individual distribution parameters are written as $f_{P_i}(p_i)$ (i=1 to q); it is important to continue using the joint density function $f_{\boldsymbol{P}}(\boldsymbol{p})$ because it includes the dependency/correlation amongst distribution parameters. Further, the joint density function $f_{\boldsymbol{P}}(\boldsymbol{p})$ also uniquely defines the probability space triplets corresponding to the q different distribution parameters, as (P_i, $\mathbb{P}_i$, $\mathcal{P}_{P_i}$), where i=1 to q.

Since the distribution parameters are uncertain, the probability space triplet corresponding to X is uncertain in an epistemic sense. In fact, the probability space triplet, and therefore the probability density function of X, are dependent on the choice of distribution parameters, and can be expressed as ($\mathcal{X}(\boldsymbol{p})$, $\mathbb{X}(\boldsymbol{p})$, $\mathcal{P}_X(\boldsymbol{p})$) and $f_X(x|\boldsymbol{p})$, respectively. Note that this density is conditioned on a particular realization of $\boldsymbol{P}$, and therefore represents only the natural variability in X corresponding to a particular realization of the distribution parameter $\boldsymbol{P}$. A realization of the random variable $\boldsymbol{P}$ does not refer to an actual value the variable can assume in the real world, but is reflective of the analysts' uncertainty regarding what the actual value in the real world is. Lindley and Phillips [32] discuss in detail the interpretation of the probability distribution of the distribution parameter. In the rest of the paper, the uncertainty in $\boldsymbol{P}$ is simply referred to as parameter uncertainty, and the uncertainty in X for a given realization $\boldsymbol{p}$ of $\boldsymbol{P}$ is referred to as variability. While the former is represented using the PDF $f_{\boldsymbol{P}}(\boldsymbol{p})$, the latter is represented using the PDF $f_X(x|\boldsymbol{P}=\boldsymbol{p})$. Aven [33] refers to such an approach as a combined classical Bayesian approach, since there is a true underlying probability distribution (for X), and there are true unknown values for the distribution parameters ($\boldsymbol{P}$).

This section discusses the Bayesian approach for the estimation of the probability distribution ($f_{\boldsymbol{P}}(\boldsymbol{p})$) of the distribution parameters ($\boldsymbol{P}$), using data available on the random variable X. This data may be in the form of point and/or interval data. While point data may be available from experiments or real-world observations, interval data may be available through several sources [34,35]. In some cases, the only information available might come from physical and theoretical constraints that impose bounds on the quantities of interest. Data collected based on temporally spaced inspections may lead to intervals. Uncertainty and errors associated with calibrated instruments may result in experimental observations that are described using intervals. In some cases, subject matter experts may describe uncertain quantities using a range of values.

The point and/or interval data may be available on either the distribution parameters $\boldsymbol{P}$ (for e.g. see Ref. [34]) or the random variable X. The proposed uncertainty quantification and sensitivity analysis methods can address both these situations.

2.1. Data available on the random variable X

Suppose that the distribution parameters ($\boldsymbol{P}$) need to be statistically inferred from the given combination of point data (m data points, x_i, i=1 to m) and interval data (n intervals, $[a_i, b_i]$, i=1 to n). The likelihood of the distribution parameters can be calculated as the probability of observing the given data conditioned on the parameters [2,12].

First consider a point datum x_i. Then the likelihood function can be calculated by constructing an infinitesimally small interval around x_i as follows:

$$L(\boldsymbol{p}) \propto P\left(X \in \left(x_i - \frac{\epsilon}{2}, x_i + \frac{\epsilon}{2}\right) \middle| \boldsymbol{p}\right) = \int_{x_i-\epsilon/2}^{x_i+\epsilon/2} f_X(x|\boldsymbol{P}=\boldsymbol{p})\, dx \propto f_X(x_i|\boldsymbol{P}=\boldsymbol{p}) \tag{1}$$

By definition, the likelihood function is meaningful only up to a proportionality constant [12]. For each interval $[a_i, b_i]$, the likelihood function can be constructed similarly as

$$L(\boldsymbol{p}) \propto Prob(X \in [a_i, b_i]|\boldsymbol{P}=\boldsymbol{p}) = P(a_i \le X \le b_i|\boldsymbol{P}=\boldsymbol{p}) = \int_{a_i}^{b_i} f_X(x|\boldsymbol{P}=\boldsymbol{p})\, dx \tag{2}$$

If the multiple sources of point data and interval data are assumed to be independent of each other, then the overall likelihood of the distribution parameters can be derived as [23]

$$L(\boldsymbol{p}) \propto \left[\prod_{i=1}^{m} f_X(x = x_i|\boldsymbol{P}=\boldsymbol{p})\right]\left[\prod_{j=1}^{n} \int_{a_j}^{b_j} f_X(x|\boldsymbol{P}=\boldsymbol{p})\, dx\right] \tag{3}$$

(In Eq. (3), the independence implies that the sources of these data, i.e., different experiments, or different experts from which the data originate, are considered to be statistically independent.)

In Eq. (3), the likelihood of the parameters can be calculated as being proportional to the (1) PDF $f_X(x|\boldsymbol{P}=\boldsymbol{p})$ in the case of point data, and (2) difference in CDFs in the case of interval data. Note that the density function $f_X(x|\boldsymbol{P}=\boldsymbol{p})$ is actually conditioned on specific values of parameters $\boldsymbol{P}$, thereby representing only variability in X and not uncertainty in $\boldsymbol{P}$.

The maximum likelihood estimates of the parameters $\boldsymbol{P}$ can be calculated by maximizing the expression in Eq. (3) when point data and/or interval data are available. The maximum likelihood estimate yields only a deterministic estimate of the parameters, which may not be accurate, especially in the presence of sparse data. (The maximum likelihood estimate tends to the true value as the number of data points tends to infinity.) Hence, the likelihood function is used below to calculate the entire probability distribution ($f_{\boldsymbol{P}}(\boldsymbol{p})$) of the parameters ($\boldsymbol{P}$) through Bayes theorem. If $f'_{\boldsymbol{P}}(\boldsymbol{p})$ denotes the prior PDF, then Bayes theorem can be written as

$$f_{\boldsymbol{P}}(\boldsymbol{p}) = \frac{f'_{\boldsymbol{P}}(\boldsymbol{p})L(\boldsymbol{p})}{\int_{\mathfrak{D}'} f'_{\boldsymbol{P}}(\boldsymbol{p})L(\boldsymbol{p})\, d\boldsymbol{p}} \tag{4}$$

where the domain of integration $\mathfrak{D}'$ is the region where the prior PDF is non-zero, i.e., $\boldsymbol{p} \in \mathfrak{D}$ if and only if $f'_{\boldsymbol{P}}(\boldsymbol{p}) > 0$. While prior distributions are considered to add additional information, the use of non-informative prior distributions [36] overcomes this challenge. For example, the non-informative prior for the mean of a normal distribution is a uniform distribution on the interval $(-\infty, +\infty)$, while the non-informative prior distribution for the standard deviation is given by $f'(\sigma) \propto 1/\sigma$ where $\sigma \in (0,\infty)$. (One could also choose the prior for the standard deviation as

uniformly distributed on (0, ∞), but this is not non-informative.) All the aforementioned distributions are known as "improper" prior distributions (i.e., the integral of the prior PDF is non-finite), which are commonly used in Bayesian analysis [37].

Once the joint density function $f_{\boldsymbol{P}}(\boldsymbol{p})$ is calculated using Eq. (4), the marginal density function for each individual distribution parameter, i.e., $f_{P_i}(p_i)$ ($i=1$ to q), can be obtained by integrating the joint density function over all the other parameters. It is important to continue using the joint density because it properly accounts for the statistical dependence among the different distribution parameters, which is reflected in Eq. (4).

2.2. Data available on the distribution parameter P

Occasionally data – in the form of point values and/or intervals – may be directly available on a particular distribution parameter P. Since this section considers only one distribution parameter, that parameter is simply represented as P, implying that it is a scalar, and the subscript i is dropped for the sake of simplicity. Two such problems were considered in the Sandia epistemic uncertainty workshop [34], where interval data is available on the distribution parameters of a lognormally distributed quantity and a triangularly distributed quantity. One possible approach would be to assume a distribution type for P and estimate the parameters of the assumed distribution (i.e., the distribution parameters of the distribution parameter P); however, such an approach (estimating distribution parameters of a distribution parameter) would merely be a statistical exercise without much physical meaning, since the distribution parameters of P do not have a physical meaning. Therefore, the distribution of P needs to be constructed without an underlying assumption regarding the distribution type of P, thereby eliminating the need for distribution parameters of P. The non-parametric method developed in [23] facilitates such an approach; in this method, the PDF that best explains the available data is chosen using the method of maximum likelihood without any assumption about the distribution type or distribution parameters. While the details of this method can be found in [23], a brief summary is presented below.

Suppose that data on P is available in the form of point data (m data points, x_i, $i=1$ to m) and interval data (n intervals, $[a_i, b_i]$, $i=1$ to n), and this information needs to be used to construct the probability density function $f_P(p)$. A likelihood-based optimization procedure for this purpose is summarized below.

Since data is directly available on the distribution parameters, the range of values of the parameters is first obtained by computing the minimum and the maximum of the available point and interval data. Let Ω^P denote the resultant interval. Discretize Ω^P into a finite number of points, say θ_i, $i=1$ to Q. Assume that the probability density function values at each of these Q points are given by $f_P(p=\theta_i)=\alpha_i$ for $i=1$ to Q. Let $\boldsymbol{\alpha}=\{\alpha_1,\alpha_2,\alpha_3,\ldots,\alpha_{Q-1},\alpha_Q\}$. Using an interpolation technique, the entire probability density function $f_P(p)$ can be calculated for all $\theta\in\Omega^P$. Then the probability of observing the given data (both point data and interval data), i.e., the likelihood, can be calculated as

$$L(\boldsymbol{\alpha})=\left[\prod_{i=1}^{m} f_P(p=x_i|\boldsymbol{\alpha})\right]\left[\prod_{j=1}^{n}\int_{a_j}^{b_j} f_P(p|\boldsymbol{\alpha})\,dp\right] \tag{5}$$

Note that the above likelihood depends on (1) the discretization points selected, i.e., θ_i, $i=1$ to Q; (2) the corresponding probability density function values α_i; and (3) the type of interpolation technique used. In this paper, the discretization is fixed, i.e., uniformly spaced θ_i values ($i=1$ to Q) over the domain of P are chosen in advance. The value of Q (number of discretization points) is chosen based on computational power; higher the value of Q, better the results. The Gaussian process interpolation [38] technique is used because it does not assume any explicit functional form for the probability density function $f_P(p)$.

Since the discretization points and the interpolation technique are selected in advance, the likelihood function in Eq. (5) is expressed only as a function of $\boldsymbol{\alpha}$. The goal of the non-parametric approach is to calculate the values of α_i ($i=1$ to Q) that maximize the likelihood function in Eq. (5). A global optimization algorithm (e.g., genetic algorithm) may be used for this purpose, and the non-parametric PDF can be constructed. Details of this methodology can be found in [23].

If there are several distribution parameters and data is available individually on each distribution parameter, the same procedure can be repeated for all distribution parameters, and the marginal distributions of all the individual distribution parameters can be obtained. If any correlation information were available, then it is important to use the joint distribution of the distribution parameters in further analysis in Section 2.3.

Note that the non-parametric approach is used in this paper to construct the PDF of the distribution parameter P. However, the same approach non-parametric method can be used to construct the PDF of the random variable X as well (when data is available on X). In that case, X would not have any distribution parameters and hence the issue of distribution parameter uncertainty would not be relevant.

2.3. Unconditional probability distribution of X

Having constructed the probability density functions $f_{\boldsymbol{P}}(\boldsymbol{p})$ of the distribution parameters $\boldsymbol{P}$, the quantity X can be easily represented using a family of distributions. First, a sample realization of the distribution parameters is selected from $f_{\boldsymbol{P}}(\boldsymbol{p})$, and for this sampled value of $\boldsymbol{P}$, several samples of the quantity X are drawn and a probability distribution of X corresponding to the sampled value of $\boldsymbol{P}$ is constructed. This procedure is then repeated for multiple samples of $\boldsymbol{P}$ and the quantity of interest X can be represented using a family of distributions.

As mentioned earlier in Section 1, in the context of uncertainty propagation, the quantity X is an input to a computational model which is used to compute the response Y. If there is a family of PDFs for X, then each PDF of X can be used to estimate a corresponding PDF for Y, thus resulting in a family of estimated distributions for the model output Y. This requires a double loop Monte Carlo analysis, making the family of distributions approach computationally intensive and not affordable, in many cases.

An alternate approach is to use to construct a single unconditional PDF for X, which includes both the variability in X and the uncertainty in the distribution parameters as

$$f_X(x)=\int_{\mathfrak{D}} f_X(x|\boldsymbol{P}=\boldsymbol{p})f_{\boldsymbol{P}}(\boldsymbol{p})\,d\boldsymbol{p} \tag{6}$$

where the domain of integration $\mathfrak{D}$ is the region where the posterior PDF of $\boldsymbol{P}$ is non-zero, i.e., $\boldsymbol{p}\in\mathfrak{D}$ if and only if $f_{\boldsymbol{P}}(\boldsymbol{p})>0$. Some researchers refer to this PDF $f_X(x)$ as the predictive PDF [4,7] of X. Eq. (6) can be interpreted in different ways [23]: (1) calculating the expectation of the conditional PDF; (2) computing the mixture of conditional probability distributions; and (3) calculating the unconditional distribution of X by eliminating conditioning on the distribution parameters.

Note that Eq. (6) needs to be evaluated numerically and the resultant PDF is not analytical. Further, the resultant PDF is not the same distribution type as the initial $f_X(x|\boldsymbol{P})$; the parametric property is lost. However, it includes both the types of uncertainty—variability and distribution parameter uncertainty. A comparison of the family of distributions (conditioned on the parameters) and the unconditional distribution is shown—through PDFs in Fig. 3(a) and through CDFs in Fig. 3(b).

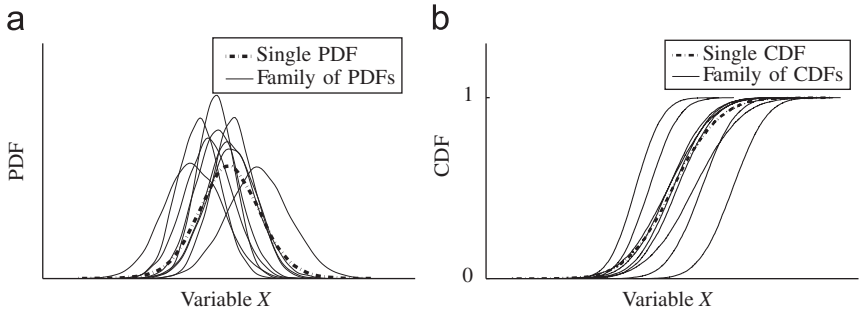

Fig. 3. Family of distributions vs. single distribution: (a) probability density function and (b) cumulative distribution function.

It is important to note that the integration in Eq. (6) does not involve the calculation of any performance function or a computational model (which is usually the more expensive calculation in practical problems of uncertainty propagation or reliability analysis). The integration in Eq. (6) characterizes input uncertainty (combining both variability and distribution parameter uncertainty) before the "uncertainty propagation" stage, thereby leading to a single PDF for X. The advantage of this approach is that Eq. (6) allows the two loops of sampling to be collapsed into a single loop for the sake of faster computation.

In some applications, it may be necessary to retain the difference between (1) uncertainty in X due to natural variability; and (2) uncertainty in X due to uncertainty in the distribution parameters [39]. As the proposed unconditional distribution combines these two uncertainties into a single distribution, it may appear that information regarding the individual contributions of the two types of quantities is lost.

However, note that the family of distributions in Fig. 3 only gives a graphical, qualitative measure of the relative contributions of the two types of uncertainty. The following section shows that the single distribution approach does not lose the information, and that in fact it is possible to develop a computational approach to quantify the individual contributions of variability and distribution parameter uncertainty.

3. Quantifying the contributions of variability and distribution parameter uncertainty

This section develops the computational methodology to distinguish and quantify the individual contributions of variability and distribution parameter uncertainty to the overall uncertainty. This section is organized into four subsections. Variance-based global sensitivity analysis (GSA) is used in the proposed methodology and hence, GSA methods are briefly reviewed in Section 3.1. Section 3.2 extends GSA and develops a methodology to quantify the individual contributions of variability and distribution parameter uncertainty. Section 3.3 calculates the contributions of these two types of uncertainty for a single random variable X. Section 3.4 calculates the contributions of these two types of uncertainty to the output Y of a response function $Y=g(\boldsymbol{X})$.

(When the discussion involves a single input random variable (scalar), the symbol X is used without any subscript. However, when there are several (say n) input random variables, the symbol $\boldsymbol{X}$ is used to represent the vector of inputs, and each input is represented by X^i, where i varies from 1 to n.)

3.1. Variance-based global sensitivity analysis

Consider a deterministic one-to-one transfer function $Y=h(\boldsymbol{X})$, where $\boldsymbol{X}=\{X^1,X^2,X^3\ldots X^n\}$ refers to the vector of inputs, and Y refers to the output (h is different from the computational model $Y=g(\boldsymbol{X})$ and the reason for using h will become clear in Sections 3.3 and 3.4). Let each of the model inputs be described using a precise probability distribution and the probability distribution of the output Y can be calculated using uncertainty propagation methods such as Monte Carlo simulation or reliability analysis-based methods such as first-order reliability method (FORM), second-order reliability method (SORM), etc. [1]. Global sensitivity analysis [27] focuses on apportioning the uncertainty in the model output Y to the uncertainty in the various model inputs $\boldsymbol{X}$. The term "global" refers to computing the sensitivity metric considering the entire probability distribution of the inputs.

Consider a particular input quantity X^i. In the global sensitivity analysis (GSA) approach, this input quantity is first fixed at a particular deterministic value and the expectation of the model output is calculated by considering the variation in other output quantities (denoted by X^{-i}). Thereby, the effect of the uncertainty of all other input quantities is averaged. Then, different deterministic values of the input quantity X^i are considered based on their probability distributions and the variance of the expectation is calculated. This metric is known as the first-order effect index of the input variable X^i on the variance of the output Y:

$$S_1^i=\frac{V_{X^i}(E_{X^{-i}}(Y|X^i))}{V(Y)} \tag{7}$$

The first-order effect measures the contribution of the variable X^i by itself. The sum of first order indices of all variables is always less than or equal to unity. The difference between this sum and unity is representative of the interaction between the input variables. Further, higher the first-order effect, more important the variable is.

Similarly, the interaction or combined effect of two variables X^i and X^j can also be calculated using the second-order effects index as

$$S_1^i+S_1^j+S_2^{i,j}=\frac{V_{X^{i,j}}(E_{X^{-i,j}}(Y|X^{i,j}))}{V(Y)} \tag{8}$$

The expression in Eq. (8) accounts not only for the individual effects of X^i and X^j but also for the interaction between X^i and X^j. The term $S_2^{i,j}$ is called as the second-order index, which explains only the interaction between X^i and X^j. Similarly the third-order effects, fourth-order effects, etc. can also be calculated. These quantities are collectively known as variance-based sensitivity indices or Sobol's indices [26,27]. Instead of calculating all the sensitivity indices of various orders, researchers often compute only the first-order index and the so-called total effects index for each input quantity.

The total effects index S_T^i of a particular input quantity X^i is defined as the sum of the first-effects index of X^i and the sum of all interactions of all orders of X^i with other input variables. However, to explicitly calculate the total-effects index, it is not necessary to compute all indices corresponding to all interactions of X^i. Consider the expression $V_{X^{-i}}(E_{X^i}(Y|X^{-i}))/V(Y)$. In analogy

with the above discussion regarding Eq. (8), this expression includes all interaction terms of all orders concerning all variables X^{-i}; any term involving X^i (both individual and any interaction with others) would not be included. As the sum of all the sensitivity indices must be equal to unity, the total effects (the sum of individual effects of X^i and all interactions with other quantities) can be calculated as

$$S_T^i = 1 - \frac{V_{X^{-i}}(E_{X^i}(Y|X^{-i}))}{V(Y)} \tag{9}$$

The sum of the total effects indices of all variables is always greater than or equal to unity; equality holds when there is no interaction between the input quantities. (In this case, the first-order effects indices are equal to the total effects indices.) If the total effects index is low, then it means that the input quantity is not important. In general, it is important to calculate both the first-order effects and the total effects indices of all input quantities for the purpose of sensitivity analysis.

The method of global sensitivity analysis has previously been applied only to assess the effects of input random variables on the output of a computational model, thus focusing only on natural variability. In the presence of epistemic uncertainty, global sensitivity analysis has been directly applied with respect to the epistemic parameters [40]; but, this approach does not distinguish the individual contributions of variability and epistemic uncertainty within a single quantity. In this paper, a global sensitivity analysis-based methodology is developed to quantify the individual contributions of variability and distribution parameter uncertainty to the overall uncertainty in a variable. The following subsection develops a methodology that facilitates the use of sensitivity indices to solve this problem.

3.2. Extension of GSA to separate variability and distribution parameter uncertainty

If a deterministic computational model $Y = g(\boldsymbol{X})$ has random inputs $\boldsymbol{X}$, then it is possible to use GSA in order to calculate the contributions of each of the inputs to the uncertainty in Y. Global sensitivity analysis can be applied when each input random variable (whose contribution is desired) can be represented using a probability distribution. If the distribution parameters $\boldsymbol{P}$ of all inputs X^i ($i=1$ to n) are precisely known, then all X^i ($i=1$ to n) have a precise PDF and unambiguously defined probability space triplets. Therefore, it is straightforward to calculate the contributions of each of the inputs to the uncertainty in the output; however, this is not the case in the two problems (P1 and P2) considered in the current paper due to the presence of distribution parameter uncertainty.

3.2.1. Challenges in problem P1

In problem P1, only one input variable X is considered, and it is desired to quantify the individual contributions of variability and distribution parameter uncertainty to the overall uncertainty in that variable X. GSA cannot be applied directly for this problem because of two challenges. First, there is no explicit deterministic transfer function (h in Section 3.1) in order to facilitate GSA. Note that this deterministic transfer function must calculate one realization of X, given one realization of the distribution parameters and one realization of natural variability. This leads to the second challenge, i.e., while distribution parameter uncertainty is represented by the uncertainty in $\boldsymbol{P}$, whereas there is no variable that explicitly represents the natural variability in X.

3.2.2. Challenges in problem P2

In problem P2, a vector of input variables $\boldsymbol{X}$ is considered, and it is desired to quantify the individual contributions of variability and distribution parameter uncertainty in $\boldsymbol{X}$ to the overall uncertainty in the output Y of a given computational model $Y = g(\boldsymbol{X})$. Each input variable (X^i; $i=1$ to n) has its own distribution parameters, which in turn have their own probability distributions. In order to delve deeper into the complexity of this problem, consider the probability spaces of the input variables $\boldsymbol{X}$.

First, consider a particular input X^i. Suppose that this input has q^i distribution parameters. Let the vector $\boldsymbol{P^i} = \{P_j^i; j=1 \text{ to } q^i\}$ denote the distribution parameters of X^i. Similar to the discussion in Section 2, the random variable X^i can then be represented using the probability space triplet ($\mathcal{X}^i(\boldsymbol{P^i})$, $\mathbb{X}^i(\boldsymbol{P^i})$, $\mathcal{P}_{X^i}(\boldsymbol{P^i})$). Since the distribution parameters are themselves uncertain, each P_j^i can be represented using its own probability space triplet, i.e., (P_j^i, $\mathbb{P}_j^i$, $\mathcal{P}_{P_j^i}$), where $j=1$ to q^i and $i=1$ to n. The PDF of each individual distribution parameter can be represented as $f_{P_j^i}(p_j^i)$. It is important to continue using the joint PDF of all distribution parameters (corresponding to one input variable), i.e., $f_{\boldsymbol{P^i}}(\boldsymbol{p^i})$, in order to account for any statistical dependence amongst them.

Further, the list of all distribution parameters for all the inputs $\boldsymbol{X}$ can be compactly represented as $\boldsymbol{P} = \{P_j^i; j=1 \text{ to } q^i \text{ and } i=1 \text{ to } n\}$, and their joint PDF can be expressed as $f_{\boldsymbol{P}}(\boldsymbol{P})$.

Similar to problem P1, GSA cannot be directly applied to this problem because (1) there is no variable that separately represents the natural variability in each input variable X^i; and (2) though the distribution parameter uncertainty can be represented by the uncertainty in $\boldsymbol{P}$, this will lead to a probabilistic output Y for a given realization of $\boldsymbol{P}$; GSA needs a deterministic input–output transfer function from $\boldsymbol{P}$ (and the variable that needs to represent variability) to Y.

3.2.3. The auxiliary variable

It is evident that two issues need to be overcome before GSA can be applied—explicit representation of natural variability and formulation of a deterministic function. Now, a new "auxiliary variable" is introduced in order to overcome these challenges. An auxiliary variable is introduced for each input variable, and therefore, it is sufficient to present the auxiliary variable considering only one input variable X.

This auxiliary variable is defined in such a way that it explicitly represents the variability in X so that Eqs. (7)–(9) can be used for sensitivity analysis. Recall that the variability in X is conditioned on the choice of the distribution parameters ($\boldsymbol{P}$), and the PDF corresponding to variability is denoted by $f_X(x|\boldsymbol{P}=\boldsymbol{p})$. Consider a random sample x drawn from this probability distribution. Then compute the CDF value $F_X(x|\boldsymbol{P}=\boldsymbol{p})$, as ($w$ is simply a dummy integration variable)

$$u = F_X(x|\boldsymbol{P}=\boldsymbol{p}) = \int_{-\infty}^{x} f_X(w|\boldsymbol{p})\, dw \tag{10}$$

Repeat the above process several times by considering different random numbers according to the variability of X, and obtain several corresponding CDF values. Then the resultant CDF values constitute another random variable, denoted by U. Every sample x of the random variable X has a one-to-one correspondence with a sample u of the random variable U. According to the probability integral transform theorem [41], the random variable U follows a uniform distribution on [0, 1]. Therefore the variability in X can be equivalently represented by another random variable U which is uniformly distributed on [0, 1]. In fact, this is concept behind inverse transform sampling or the golden rule [42], which is prominently used in Monte Carlo sampling; in order to draw a sample of X, a sample is first drawn from U(0, 1), and the CDF of X is inverted (inverse of Eq. (10)) to obtain the corresponding sample of X. This variable U is now defined to be the auxiliary variable in the rest of the paper, and since, it is corresponds to the variability in X, it is denoted by U_X hereafter. There is another

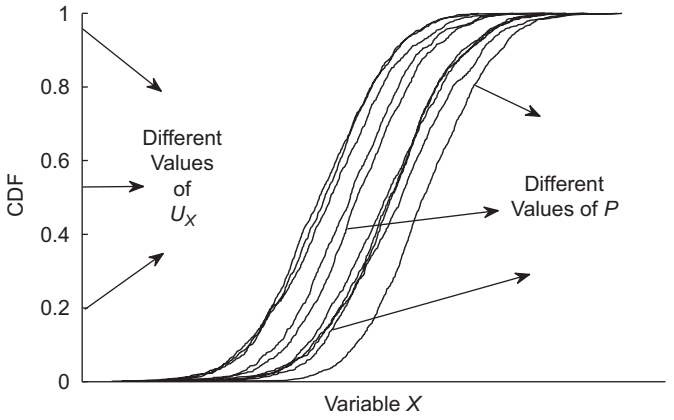

Fig. 4. Variability (U_X) and distribution parameter uncertainty ($\boldsymbol{P}$).

advantage of representing the variability in X using U: note that the distribution of U is statistically independent of the realization of $\boldsymbol{P}$; in other words, immaterial of what realization of $\boldsymbol{P}$ is considered, U is uniformly distributed on the interval [0, 1].

The introduction of the auxiliary variable not only overcomes the first difficulty (presence of a variable that explicitly represents the variability in X), but also addresses the second issue (presence of a deterministic transfer function, i.e., for a realization of U, and a realization of $\boldsymbol{P}$, there exists a one-to-one mapping to a realization of X, based on Eq. (10)). Further, it is now possible to delineate the overall uncertainty in X into two quantities—variability represented by U_X (uniformly distributed on [0, 1]) and distribution parameter uncertainty represented by $\boldsymbol{P}$ (whose distribution was calculated in Section 2), as seen in Fig. 4. The "spread" of X due to U_X corresponds to the variability whereas the "spread" of X due to $\boldsymbol{P}$ corresponds to distribution parameter uncertainty.

These two quantities U_X and $\boldsymbol{P}$ can be used in the context of Monte Carlo simulation, to generate both the family of distributions for X and the single unconditional (predictive) distribution of X. To generate the family of distributions, first sample $\boldsymbol{P}$ and then generate several samples of U_X and correspondingly several samples of X (using Eq. (10)) to generate one member of the family. Then repeat this procedure for multiple samples of $\boldsymbol{P}$. The single predictive distribution can be constructed by simultaneously generating a sample of U_X, and an independent sample of $\boldsymbol{P}$, and a corresponding sample of X (using Eq. (10)), and then repeating the entire procedure to generate multiple samples of X and hence a single distribution for X. The fundamental difference is that in the former case, multiple samples of U_X are drawn for each sample of $\boldsymbol{P}$ (double-loop sampling, i.e., sampling U_X within $\boldsymbol{P}$) whereas, in the latter case, one sample of U_X is drawn for each sample of $\boldsymbol{P}$ (single loop sampling, i.e., U_X and $\boldsymbol{P}$ are sampled together, simultaneously and independently. This is statistically justified because the distribution of U_X is independent of $\boldsymbol{P}$).

The following two subsections explain how the introduction of the auxiliary variable facilitates the application of global sensitivity analysis to problems P1 and P2, shown earlier in Fig. 2(a) and (b).

3.3. Variability and distribution parameter uncertainty within a variable (P1)

To summarize the above development, a variable X with known distribution type but uncertain distribution parameters ($\boldsymbol{P}$) is considered. The variability in X is given by the PDF $f_X(x|\boldsymbol{p})$. The uncertainty in the distribution parameters (reflective of the analyst's subjectivity) is represented through the PDF $f_{\boldsymbol{P}}(\boldsymbol{p})$, which was calculated earlier in Section 2. As explained above, now there are two variables that explicitly represent variability (U_X) and parameter uncertainty ($\boldsymbol{P}$); each has its own PDF and the aim is to calculate the contributions of variability (U_X) and distribution parameter uncertainty ($\boldsymbol{P}$) to the overall uncertainty in X.

Note that there is no computational model here; U_X and $\boldsymbol{P}$ are simply components of X. Therefore, the deterministic transfer function "h" required for GSA is now defined, as

$$x = h(u_X, \boldsymbol{p}) = F_X^{-1}(x = u_X|\boldsymbol{p}) \tag{11}$$

In Eq. (11), x, u_X, and $\boldsymbol{p}$ are realizations of the random variables X, U_X, and $\boldsymbol{P}$, respectively. Observe that u_X and $\boldsymbol{p}$ are the "inputs" to the deterministic transfer function "h", and x is the "output". This establishes deterministic transfer function needed for GSA, to calculate the contributions of variability (U_X) and distribution parameter uncertainty ($\boldsymbol{P}$) to the overall uncertainty in X. Hence, using the methods in Section 3.1, it is possible to calculate the first-order and total effects of both U_X and $\boldsymbol{P}$. In fact, it is not technically accurate to use the terms "first-order effects" and "total effects" any more. It was explained earlier that these terms are used to assess the effects of a single quantity (X^i in Section 3.1). However, $\boldsymbol{P}$ may be a vector; for example, $\boldsymbol{P}$ consists of two terms, mean and standard deviation, in the case of a normal distribution. Thus in order to calculate the contribution of distribution parameter uncertainty, one must consider the contribution of all variables in $\boldsymbol{P}$ and hence the term "first-order effects" is no more applicable. (In the above example of a normal distribution, calculating the effect of $\boldsymbol{P}$ would in fact require the calculation of second-order effect as in Eq. (8).) In order to avoid this confusion, this paper uses the terms "individual effects" and "overall effects" instead of "first-order effects" and "total effects". Note that the individual effects and total effects are not the same as first-order effects and total effects. It is necessary to calculate both individual and overall effects to quantify the individual contributions of variability and distribution parameter uncertainty.

The individual effects index is a higher order index, whose order to equal to the dimension of $\boldsymbol{P}$, and the total effects index is equal to one minus another higher order index, as defined below. The individual (I) and overall (O) contributions of variability to the overall uncertainty in X can be calculated as

$$S_U^I = \frac{V_{U_X}(E_{\boldsymbol{P}}(X|u_X))}{V(X)}$$
$$S_U^O = 1 - \frac{V_{\boldsymbol{P}}(E_{U_X}(X|\boldsymbol{p}))}{V(X)} \tag{12}$$

The individual (I) and overall (O) contributions of distribution parameter uncertainty to the overall uncertainty in X can be calculated as

$$S_{\boldsymbol{P}}^I = \frac{V_{\boldsymbol{P}}(E_{U_X}(X|\boldsymbol{p}))}{V(X)}$$
$$S_{\boldsymbol{P}}^O = 1 - \frac{V_{U_X}(E_{\boldsymbol{P}}(X|u_X))}{V(X)} \tag{13}$$

In the next subsection, this concept is extended where the contributions of variability and distribution parameter uncertainty to a response function are assessed.

3.4. Calculating the contributions to a response function (P2)

Now consider a computational model $Y = g(\boldsymbol{X})$ where $\boldsymbol{X}$ refers to the vector of inputs, and Y refers to the model output. Let X^i denote a particular input variable. Each input X^i now has a PDF $f_{X^i}(x^i|\boldsymbol{P^i})$ (type known), where $\boldsymbol{P^i}$ refers to the vector of distribution parameters of X^i, with PDF $f_{\boldsymbol{P}}^{\boldsymbol{i}}(\boldsymbol{p^i})$. Further, recall that all the distribution parameters of all the input variables are compactly

represented as $\boldsymbol{P}=\{P_j^i;\ j=1 \text{ to } q^i \text{ and } i=1 \text{ to } n\}$, and their joint PDF is expressed as $f_{\boldsymbol{P}}(\boldsymbol{P})$. Note that Y is not deterministic due to the uncertainty in P, and hence GSA cannot be applied directly.

However, in Section 3.1, the aim was to calculate the contribution of each X^i to Y. In problem P2, the aim is different; it is to calculate the contribution of variability and distribution parameter uncertainty in each X^i to the overall uncertainty in Y. Hence, the uncertainty in each X^i needs to be decomposed into two parts—variability (U_{X^i}) and distribution parameter uncertainty ($\boldsymbol{P^i}$)—similar to that in Section 3.3 where only one input variable was considered.

The deterministic transfer function "h" needed for GSA is constructed with two inputs — U_{X^i} and $\boldsymbol{P^i}$ — for each uncertain X^i. Hence, in $Y=g(\boldsymbol{X})$, each X^i is replaced with Eq. (11) using the corresponding U_{X^i} and $\boldsymbol{P^i}$ as inputs. Similar to the compact representation of all distribution parameters of all input variables, the vector of auxiliary variables can also be defined to be $\boldsymbol{U_X}=\{U_{X^i};i=1 \text{ to } n\}$. The output of h is now deterministic; i.e., a single Y for a choice of $\boldsymbol{U_X}$ and $\boldsymbol{P}$ can be calculated. It is now possible to compute several sensitivity indices regarding the contribution of the following to the variance of Y:

1. Individual and overall effects of the overall uncertainty of any X^i (by considering the corresponding U_{X^i} and $\boldsymbol{P^i}$ together).
2. Individual and overall effects of variability of any X^i (by considering the corresponding U_{X^i} alone).
3. Individual and overall effects of distribution parameter uncertainty of any X^i (by considering the corresponding $\boldsymbol{P^i}$ alone).
4. Individual and overall effects of variability of combinations of multiple X's, i.e., X^i, X^j, X^k, etc. (by considering the appropriate U_{X^i}, U_{X^j}, U_{X^k}, etc. together).
5. Individual and overall effects of distribution parameter uncertainty of combinations of X's, i.e., X^i, X^j, X^k, etc. (by considering the appropriate $\boldsymbol{P^i}$, $\boldsymbol{P^j}$, $\boldsymbol{P^k}$, etc. together).

Thus, an easy and efficient methodology is developed for computing the sensitivity indices that quantify the individual contributions of variability and distribution parameter uncertainty to the overall uncertainty in the model output.

Eqs. (12) and (13) involve the computation of "variance of expectation", which may intuitively require nested loop Monte Carlo sampling; i.e., an inner loop to calculate the expectation and an outer loop to calculate the variance of expectation. It may be argued that this requires the same computational expense as a family of distributions approach. However, there exist single loop sampling approaches to compute these sensitivity indices, as explained in Saltelli et al. [27]. Further, while the family of distributions approach provides only a graphical representation of relative contributions of variability and distribution parameter uncertainty, the proposed approach provides quantitative metrics based on the actual contribution to variance. While generating samples in order to evaluate Eqs. (12) and (13), the statistical dependence (or correlation) amongst multiple distribution parameters, and the statistical dependence (or correlation) between multiple input quantities can be easily included, by drawing samples from the respective joint density functions. There are several statistical approaches (for example, Markov Chain Monte Carlo sampling [43]) established in the literature for this purpose.

4. Numerical examples

This section illustrates the proposed methodology using three different examples. The first example considers a single random variable with uncertain distribution parameters (problem P1) whose probability distributions are known. The second example considers an uncertain propagation problem (problem P2) with three random variables, whose distribution parameters are uncertain and in turn are described using probability distributions. The third example considers another uncertainty propagation problem (problem P2) from the Epistemic Uncertainty Workshop organized by the Sandia National Laboratories [34], where there are two random variables and interval data is available on the distribution parameters.

4.1. Random variable with uncertain distribution parameters (P1)

Consider a variable X that is normally distributed with parameters mean (μ) and standard deviation (σ). Both these distribution parameters are assumed to be normally distributed for the sake of illustration. Three different cases are considered, as tabulated in Table 1.

For the sake of visualization, the family of PDFs and the single PDF of X are shown for the three cases in Figs. 5–7.

In Eqs. (12) and (13), $\boldsymbol{P}=[\mu,\sigma]$ is a vector of length two and to calculate the individual effects of $\boldsymbol{P}$, it would actually be necessary

Table 1
Three cases for problem P1.

Quantity →	Mean (μ)		Standard deviation (σ)	
Cases	Mean	Standard deviation	Mean	Standard deviation
Case 1	10	1	1	0.1
Case 2	10	1	2	0.1
Case 3	10	2	1	0.1

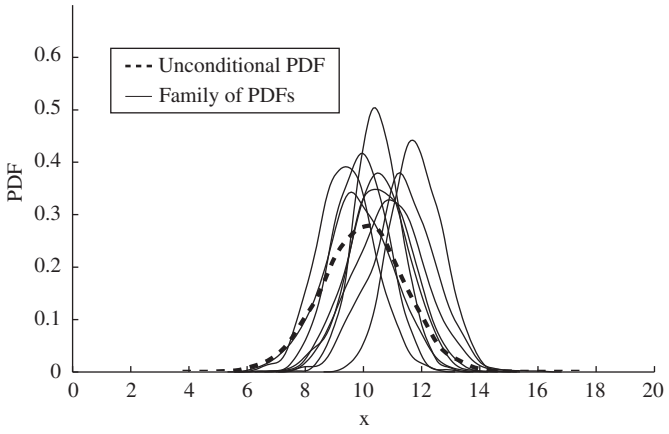

Fig. 5. Case 1: Family of PDFs and unconditional PDF.

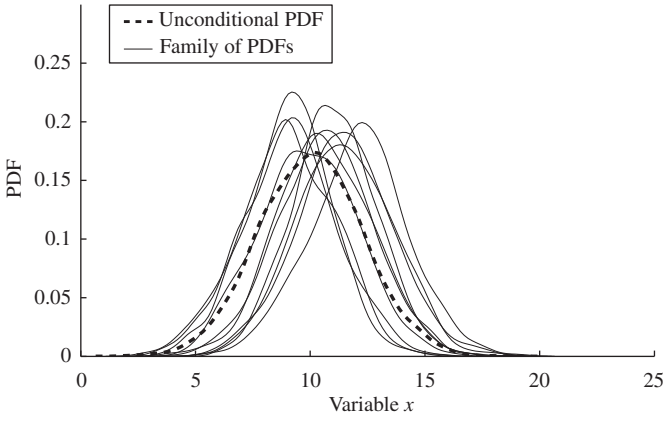

Fig. 6. Case 2: Family of PDFs and unconditional PDF.

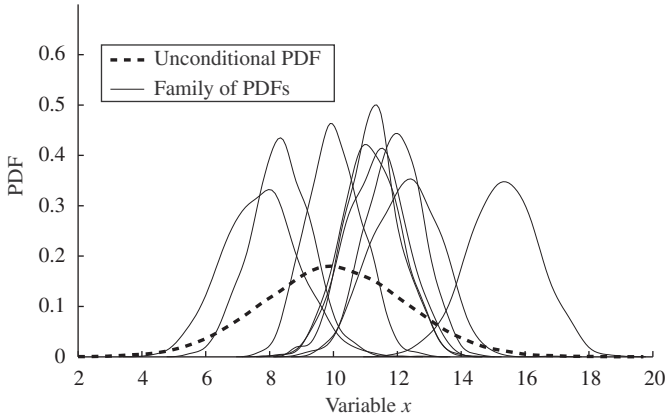

Fig. 7. Case 3: Family of PDFs and unconditional PDF.

Table 2
Three cases for problem P1: contributions.

Cases	Uncertainty	Individual effects (%)	Overall effects (%)
Case 1	Variability	0.50	0.50
	Parameter uncertainty	0.50	0.50
Case 2	Variability	0.80	0.80
	Parameter uncertainty	0.20	0.20
Case 3	Variability	0.20	0.20
	Parameter uncertainty	0.80	0.80

to calculate the second-order effects indices, as in Eq. (8). The deterministic function is constructed with inputs U_X, μ, and σ. The decomposition of variance is shown in the following equation:

$$S_{U_X}+S_\mu+S_\sigma+S_{U_X,\mu}+S_{U_X,\sigma}+S_{\mu,\sigma}+S_{U_X,\mu,\sigma}=100\% \quad (14)$$

In general, there are 2^k-1 terms in the decomposition of variance, where k represents the total number of inputs to the deterministic function "h". The individual and overall effects of variability and distribution parameter uncertainty are calculated based on Eqs. (12) and (13). The individual effect of variability is given by S_{U_X}; the individual effect of parameter uncertainty is given by $S_\mu+S_\sigma++S_{\mu,\sigma}$; the overall effect of variability is given by $S_{U_X}+S_{U_X,\mu}+S_{U_X,\sigma}+S_{U_X,\mu,\sigma}$; and the overall effect of parameter uncertainty is given by $S_\mu+S_\sigma+S_{U_X,\mu}+S_{U_X,\sigma}+S_{\mu,\sigma}+S_{U_X,\mu,\sigma}$. These sensitivities are tabulated in Table 2, in terms of fractions of the total variance.

The following observations are made from Table 2.

1. As seen in Eq. (14), the sum of individual effects of variability and the total effects of parameter uncertainty is equal to one. Similarly, the sum of individual effects of parameter uncertainty and the total effects of variability is equal to one.
2. The contributions of variability and distribution parameter uncertainty are almost equal in Case 1. In Case 2, the mean of the standard deviation is twice as in Case 1 and hence, this increases the contribution of variability and decreases the contribution of distribution parameter uncertainty. In Case 3, the standard deviation of the mean is twice as in Case 1, thereby increasing the distribution parameter uncertainty, and decreasing the contribution of variability.
3. Due to numerical errors that arise due to sampling, all percentage sensitivities are reported only to one decimal place. Though the overall effects indices were greater than the individual effects indices, this is not reflected in Table 2 due to the difference being less than 0.1 %. Thus, the overall effects indices are only marginally higher (in fact equal up to 1st decimal place) than the individual effects indices; hence, there is little interaction between the terms corresponding to variability and distribution parameter uncertainty).
4. This analysis suggests that by reducing the contribution of distribution parameter uncertainty, it is possible to reduce the uncertainty in the variable X, for e.g., in Case 1, by approximately 50%. The contribution due to variability is irreducible by collecting more data.

4.2. Uncertainty propagation problem (P2)

This section considers a simple uncertainty propagation problem, $y=g(\boldsymbol{X})$, where $\boldsymbol{X}=\{X^1,X^2,X^3\}$, and $g(\boldsymbol{X})\equiv X_1+X_2+X_3$ for the sake of illustration. Though a simple, additive function is chosen, this closed-form relationship is assumed to be known, "g" is still treated as a black-box function representative of complex computational models. All the inputs are random variables whose distribution parameters are themselves expressed using probability distributions, the details of which are provided in Table 3. The goal of this numerical example is to:

1. Consider the case where each input is of a different type of probability distribution, and therefore, has different types of distribution parameters.
2. Consider the case where the number of distribution parameters differ from one input to another.
3. Illustrate all the entities and concepts involved; the proceedings are complicated mainly due to the presence of more than one input.
4. Demonstrate the use of multiple auxiliary variables for multiple inputs.
5. Perform sensitivity analysis to quantify the contributions of variability and parameter uncertainty in each input to the overall uncertainty in y.

Consider the discussion regarding probability spaces in Section 3.2.2, where the distribution parameters are $\boldsymbol{P}=\{P_j^i;\ j=1 \text{ to } q^i \text{ and } i=1 \text{ to } n\}$, n is the total number of inputs, and q^i refers to the number of distribution parameters for the ith input variable. In this numerical example, $n=3$, $q^1=2$, $q^2=3$, and $q^3=3$, and the values of the distribution parameters are tabulated in Table 3.

Table 3
Distribution parameters for inputs.

Input symbol	Distribution type	Parameter name	Parameter symbol	Lower bound of parameter	Upper bound of parameter
X_1	Uniform	Lower bound	p_1^1	14	20
		Upper bound	p_2^1	23	28
X_2	Triangular	Lower bound	p_1^2	34	38
		Upper bound	p_2^2	41	44
		Mode	p_3^2	39	40
X_3	Truncated normal	Mean	p_1^3	37	43
		Std. deviation	p_2^3	1.8	2.2
		Truncation probability	p_3^3	0.03	0.06

Also, distribution parameters of the third input variable X_3 refer to the statistics before truncation; an α-truncation probability implies that the truncation points correspond to "α" and "$1-\alpha$" probability-levels of the normal distribution before truncation. For the purpose of illustration, all the distribution parameters are chosen to have uniform probability distributions, whose lower bounds and upper bounds are also indicated in Table 3.

Three auxiliary variables — U_{X^1}, U_{X^2}, and U_{X^3}—are defined as explained in Section 3.4 in order to facilitate sensitivity analysis. Each of the inputs X^i can be represented using a family of distributions or using a single, predictive distribution. In order to generate the family, sample several values of U_{X^i} are generated for one sample of $\boldsymbol{P^i}$ to obtain one member of the family, and then the entire procedure is repeated for several samples of $\boldsymbol{P^i}$ to obtain the entire family. Alternatively, the single predictive distribution for X^i can be obtained by generating samples of U_{X^i} and $\boldsymbol{P^i}$ simultaneously and independently. The family of distributions of $\boldsymbol{X}$ or the single predictive distribution of $\boldsymbol{X}$ can be propagated through the model $Y=g(\boldsymbol{X})$ to obtain the family of distributions or the single, predictive distribution of Y. Fig. 8 shows the comparison between the family of distributions and the single, predictive distribution for the output Y.

Sensitivity analysis is performed in order to quantify the contribution of variability and parameter uncertainty of each input to the overall uncertainty in Y. The deterministic transfer function for sensitivity results is constructed using three auxiliary variables and eight distribution parameters as inputs, and calculates y as the output. Therefore, there are $2^{11}-1$ terms in the decomposition of variance. Similar to the previous example, only the individual effects and total effects indices are calculated and reported in Table 4.

Table 4 lists the individual effects and total effects indices for the variability and the parameter uncertainty in each quantity. In addition, the indices are also reported for the overall variability and the overall parameter uncertainty. In Table 4, the term "quantity" refers to all the variables which are conditioned (fixed) in the inner loop in Eq. (7) (to calculate the individual effect) and Eq. (9) (to calculate the overall effect). Similar to the previous example, it is seen that the interaction between the variability and distribution parameter uncertainty is negligible and that the distribution parameter uncertainty contributes to about one-third of the overall uncertainty in the output Y. Such an analysis clearly helps to identify the important contributors of uncertainty and identify what proportion of the uncertainty can be decreased by collecting more data.

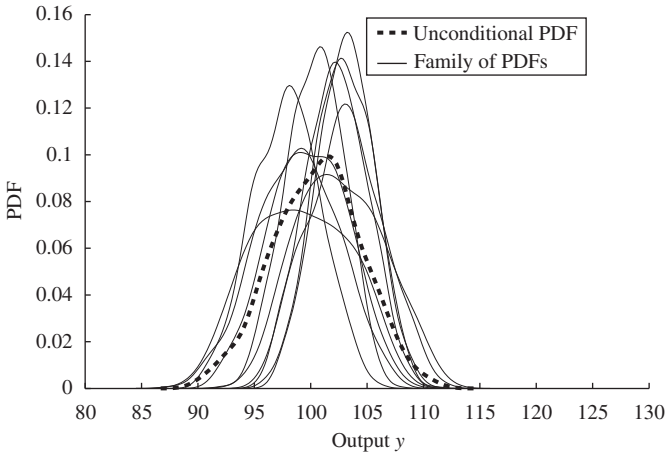

Fig. 8. Output Y: Family of PDFs and unconditional PDF.

Table 4
Contributions of variability and distribution parameter uncertainty.

Quantity	Meaning	Individual effects	Overall effects
U_{X^1}	Variability in X_1	0.39	0.42
U_{X^2}	Variability in X_2	0.12	0.12
U_{X^3}	Variability in X_3	0.00	0.00
$\boldsymbol{U_X}=\{U_{X^1},U_{X^2},U_{X^3}\}$	Total variability in $\boldsymbol{X}$	0.68	0.71
$\boldsymbol{P^1}=\{P_1^1,P_2^1\}$	Parameter uncertainty in X_1	0.08	0.11
$\boldsymbol{P^2}=\{P_1^2,P_2^2,P_3^2\}$	Parameter uncertainty in X_2	0.02	0.02
$\boldsymbol{P^3}=\{P_1^3,P_2^3,P_3^3\}$	Parameter uncertainty in X_3	0.17	0.18
$\boldsymbol{P}=\{\boldsymbol{P^1},\boldsymbol{P^2},\boldsymbol{P^3}\}$	Total parameter uncertainty in $\boldsymbol{X}$	0.29	0.32

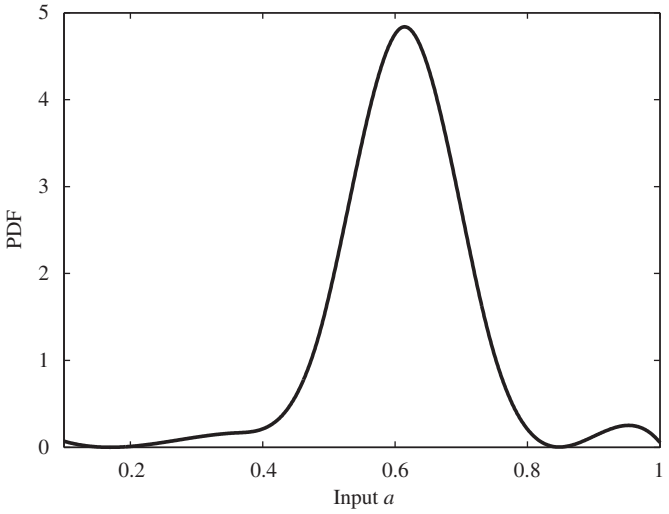

Fig. 9. Non-parametric PDF of input a.

4.3. Uncertainty propagation problem (P2) with interval data on distribution parameters

Sandia National Laboratories conducted an epistemic uncertainty workshop in 2003 [34] where uncertainty propagation problems were discussed using a function $y=(a+b)^a$. Multiple solution approaches using different uncertainty representation theories were reported by various researchers [44]. This uncertainty propagation problem helps to demonstrate the full potential of the proposed sensitivity analysis methodology. Information for the input variable a is in the form of three intervals [0.5, 0.7], [0.3, 0.8] and [0.1, 1.0]. Input variable b is specified as a lognormal PDF, but with imprecise parameters. The parameters (λ and ξ) of this lognormal distribution are described by multiple intervals ([0.6, 0.8], [0.2, 0.9], [0.0, 1.0]) and ([0.3, 0.4], [0.2, 0.45], [0.1, 0.5]), respectively.

Since the distribution type of the input variable a is not specified, it is not appropriate to assume any distribution type and estimate the parameters in the presence of interval data; instead a is represented using a non-parametric probability distribution. The PDF of a (denoted by $f_a(a)$) corresponding to the available interval data is shown in Fig. 9. Similarly, non-parametric PDFs are constructed for λ and ξ (denoted by $f_\lambda(\lambda)$ and $f_\xi(\xi)$, respectively) and shown in Fig. 10. For a non-parametric distribution, there is no issue of distinction between variability and distribution parameter uncertainty, since there are no distribution parameters. The construction of non-parametric probability distributions was discussed in Section 2.2.

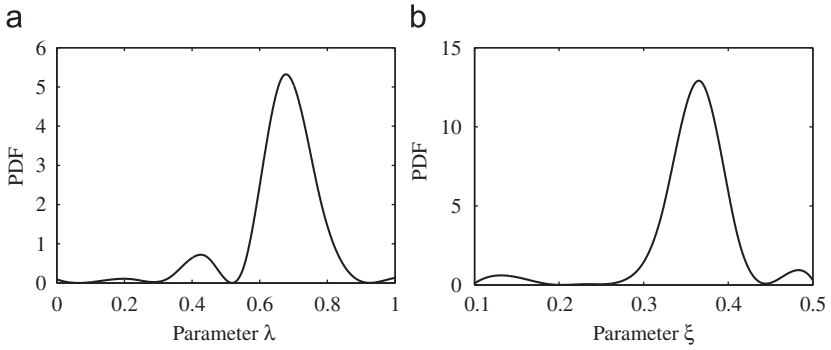

Fig. 10. Non-parametric PDFs of distribution parameters of input b: (a) PDF of λ; (b) PDF of ξ.

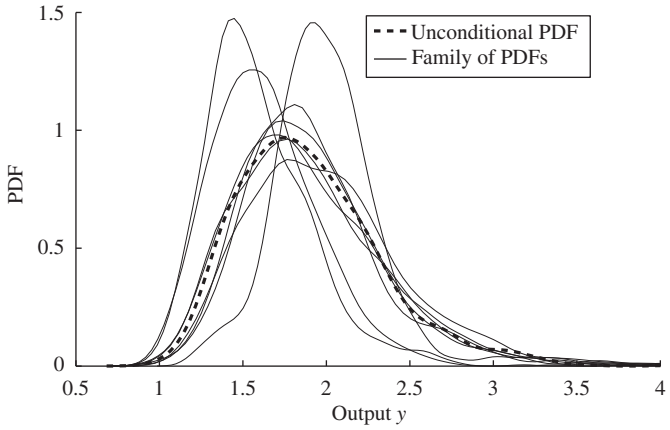

Fig. 11. Output Y: Family of PDFs and unconditional PDF.

Table 5
Contributions of variability and distribution parameter uncertainty.

Quantity	Meaning	Individual effects (%)	Overall effects (%)
(λ, ξ)	Parameter uncertainty in b	0.08	0.10
U_b	Variability in b	0.57	0.63
(λ, ξ, U_b)	Total uncertainty in b	0.67	0.71
a	Uncertainty in a	0.29	0.33

The family of distributions for y and the single unconditional distribution are shown in Fig. 11.

Sensitivity analysis is conducted next. The uncertainty in y can be attributed to: (1) uncertainty in a, shown in Fig. 9 (since this is a non-parametric PDF, the variability and distribution parameter uncertainty of a cannot be distinguished); (2) variability in b, which is a lognormally distributed quantity; as per the proposed sensitivity analysis method, an auxiliary variable U_b is introduced to represent this variability; and (3) uncertainty in distribution parameters (λ and ξ), shown in Fig. 10. There are 15 terms in the decomposition of variance.

Table 5 reports the individual and overall sensitivity indices (to one decimal place) of the following quantities: (1) parameter uncertainty in b; (2) variability in b; (3) total uncertainty in b; and (4) uncertainty in a. It is seen that there is little interaction between variability and parameter uncertainty of b. The contribution of distribution parameter uncertainty is about 10%, and the variability in b contributes to about 60% of the overall variance; while the former can be reduced, the latter is irreducible uncertainty.

5. Summary

This paper developed a three-step computational methodology to calculate the individual contributions of variability and distribution parameter uncertainty in random quantities. First, a Bayesian approach was used to construct the probability distributions of the distribution parameters. Second, a comparison of "family of distributions" approach vs. "single distribution" approach was made. The family of distributions approach required a double loop sampling approach which may be computationally expensive and hence, this approach was replaced by "the single unconditional distribution" approach. However, the use of a single distribution combined both variability and distribution parameter uncertainty into a single probability distribution. Therefore, third, global sensitivity analysis was extended to explicitly quantify the contributions of variability and distribution parameter uncertainty.

For each quantity, a uniformly distributed auxiliary variable (based on the probability integral transform theorem) was introduced to represent variability. The introduction of the auxiliary variable facilitated the extension of variance-based global sensitivity analysis to quantify the contributions of variability and distribution parameter uncertainty. The proposed methodology was implemented at two levels: (1) at the level of a single random variable; and (2) at the level of the output of a response function whose random inputs have both variability and distribution parameter uncertainty.

This paper quantified the sensitivity of distribution parameter uncertainty and variability to the overall uncertainty; however, the underlying distribution type was assumed to be known. This may not be the case in many practical applications, and it may be necessary to quantify the sensitivity of the distribution type as well. Further work may address this issue and extend the sensitivity analysis approach to quantify the effect of distribution type uncertainty.

Acknowledgment

The study in this paper was supported by funds from Sandia National Laboratories through Contract no. BG-7732 (Technical Monitor: Dr. Angel Urbina). The support is gratefully acknowledged.

References

[1] Haldar A, Mahadevan S. Probability, reliability, and statistical methods in engineering design. John Wiley & Sons Inc.; 2000.
[2] Pawitan Y. In all likelihood: statistical modelling and inference using likelihood. USA: Oxford University Press; 2001.
[3] Seber GAF, Wild CJ. Nonlinear regression. New York: John Wiley; 1989.
[4] Der Kiureghian A. Measures of structural safety under imperfect states of knowledge. Journal of Structural Engineering 1989;115(5):1119–40.

[5] Der Kiureghian A. Analysis of structural reliability under parameter uncertainties. Probabilistic Engineering Mechanics 2008;23(4):351–8.
[6] Zhang R, Mahadevan S. Integration of computation and testing for reliability estimation. Reliability Engineering & System Safety 2001;74(1):13–21.
[7] Ditlevsen O, Madsen HO. Structural reliability methods. New York: John Wiley; 1996.
[8] Hajagos JG. Interval Monte Carlo as an alternative to second-order sampling for estimating ecological risk. Reliable Computing 2007;13(1):71–81.
[9] Halpern EF, Weinstein MC, Hunink MGM, Gazelle GS. Representing both first- and second-order uncertainties by Monte Carlo simulation for groups of patients. Medical Decision Making 2000;20(3):314–22.
[10] Lee PM. Bayesian statistics. UK: Arnold London; 2004.
[11] Leonard T, Hsu JSJ. Bayesian methods. Cambridge Books; 2001.
[12] Edwards AWF. Likelihood. Cambridge University Press; 1984.
[13] Aven T. On how to define, understand and describe risk. Reliability Engineering & System Safety 2010;95(6):623–31.
[14] Aven T. The risk concept — historical and recent development trends. Reliability Engineering & System Safety 2011.
[15] North DW. Probability theory and consistent reasoning. Risk Analysis 2010;30(3):377–80.
[16] Warner North D. Uncertainties, precaution, and science: focus on the state of knowledge and how it may change. Risk Analysis 2011;31(10):1526–9.
[17] Helton JC, Johnson JD, Oberkampf WL. An exploration of alternative approaches to the representation of uncertainty in model predictions. Reliability Engineering & System Safety 2004;85(1):39–71.
[18] Rao SS, Annamdas KK. Dempster–Shafer theory in the analysis and design of uncertain engineering systems. Product Research 2009:135–60.
[19] Bae HR, Grandhi RV, Canfield RA. An approximation approach for uncertainty quantification using evidence theory. Reliability Engineering & System Safety 2004;86(3):215–25.
[20] Agarwal H, Renaud JE, Preston EL, Padmanabhan D. Uncertainty quantification using evidence theory in multidisciplinary design optimization. Reliability Engineering & System Safety 2004;85(1):281–94.
[21] Dubois D, Prade H, Harding EF. Possibility theory: an approach to computerized processing of uncertainty, vol. 2. New York: Plenum Press; 1988.
[22] Ben-Haim Y, Elishakoff I. Convex models of uncertainty in applied mechanics. Amsterdam: Elsevier; 1990.
[23] Sankararaman S, Mahadevan S. Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data. Reliability Engineering & System Safety 2011;96(7):814–24.
[24] Sankararaman S, Mahadevan S. Model validation under epistemic uncertainty. Reliability Engineering & System Safety 2011;96(9):1232–41 (Quantification of Margins and Uncertainties).
[25] Sankararaman S, Mahadevan S. Model parameter estimation with imprecise and unpaired data. Inverse Problems in Science and Engineering. Vol. 20, No. 7, 2012, http://dx.doi.org/10.1080/17415977.2012.675505, part of the Special Issue containing selected papers from the 3rd Inverse Problems, Design, and Optimization (IPDO) symposium conducted in João Pessoa, Brazil during 25–27 August 2010.
[26] Sobol IM. Sensitivity analysis for non-linear mathematical models. Mathematical Modeling and Computational Experiment 1993;1(1):407–14.
[27] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global sensitivity analysis: the primer. John Wiley & Sons; 2008.
[28] Helton JC, Johnson JD, Sallaberry CJ. Quantification of margins and uncertainties: example analyses from reactor safety and radioactive waste disposal involving the separation of aleatory and epistemic uncertainty. Reliability Engineering & System Safety 2011;96(9):1014–33.
[29] Kolmogorov AN. Foundations of the theory of probability. Chelsea Publishing Co.; 1950.
[30] Möller B, Beer M. Fuzzy randomness: uncertainty in civil engineering and computational mechanics. Springer Verlag; 2004.
[31] Calvetti D, Somersalo E. Introduction to Bayesian scientific computing. New York: Springer Verlag; 2007.
[32] Lindley DV, Phillips LD. Inference for a Bernoulli process (a Bayesian view). The American Statistician 1976;30(3):112–9.
[33] Aven T, Kvaløy JT. Implementing the Bayesian paradigm in risk analysis. Reliability Engineering & System Safety 2002;78(2):195–201.
[34] Oberkampf WL, Helton JC, Joslyn CA, Wojtkiewicz SF, Ferson S. Challenge problems: uncertainty in system response given uncertain parameters. Reliability Engineering & System Safety 2004;85(1):11–9.
[35] Du X, Sudjianto A, Huang B. Reliability-based design with the mixture of random and interval variables. Journal of Mechanical Design 2005;127(6):1068–76.
[36] Jeffreys H. Theory of probability. USA: Oxford University Press; 1998.
[37] O'Hagan A, Forster J. Kendall's advanced theory of statistics. Bayesian inference, vol. 2B. Halsted Press; 1994.
[38] McFarland JM. Uncertainty analysis for computer simulations through validation and calibration. PhD thesis, Vanderbilt University; 2008.
[39] Ferson S, Ginzburg LR. Different methods are needed to propagate ignorance and variability. Reliability Engineering & System Safety 1996;54(2):133–44.
[40] Borgonovo E, Apostolakis GE, Tarantola S, Saltelli A. Comparison of global sensitivity analysis techniques and importance measures in psa. Reliability Engineering & System Safety 2003;79(2):175–85.
[41] Angus JE. The probability integral transform and related results. SIAM Review 1994:652–4.
[42] Devroye L, Devroye L. Non-uniform random variate generation. New York: Springer Verlag; 1986.
[43] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970;57(1):97–109.
[44] Ferson S, Joslyn CA, Helton JC, Oberkampf WL, Sentz K. Summary from the epistemic uncertainty workshop: consensus amid diversity. Reliability Engineering & System Safety 2004;85(1):355–69.