SAMPLING VARIANCE AND BIAS OF WILKS' CONSERVATIVE ESTIMATE OF CONFIDENCE INTERVALS

J.P. Hessling
SP Technical Research Institute of Sweden, Measurement Technology,
Box 857, SE-50115 Boras, Sweden
peter.hessling@sp.se

P. Hedberg
Swedish Radiation Safety Authority,
SE-171 16 Stockholm, Sweden
peter.hedberg@ssm.se

ABSTRACT

For evaluation of the uncertainty of nuclear power calculations, Wilks' approach has the appearance of an ideal tool. A conservatively estimated bound is obtained as the $r-th$ most extreme model result, of a random sample of size determined by r. The methodology is non-invasive, simple and seems efficient and adequate. However, as is shown it comes with a high price of large bias and substantial sampling variance. This jeopardizes its utilization as well as lowers its credibility and perceived efficiency.
The unfortunate combination of random sampling and faithful estimation may result in a relative sampling uncertainty of the estimated bound(-s) no less than 100%. What is defined credibility, i.e. the probability that the *estimated* bound is conservative relative to the true result, is well below the confidence relating the *targeted* bound(-s) to the true result. For the default application of Wilks' method, that translates into an expected failure rate of up to 10% (instead of 5%) of estimated bounds. To compensate for this deficit in credibility compared to the chosen level of confidence, adjustments of current practice are proposed.
The application to modeling uncertainty is to be clearly distinguished from the original experimental sampling problem addressed by Wilks. Here, more is known but not utilized. A viable novel alternative based on so-called deterministic sampling with higher accuracy, precision and efficiency will therefore be briefly discussed and illustrated.

KEYWORDS
Wilks, confidence, uncertainty, sampling, model

1. INTRODUCTION

In critical nuclear power applications it is crucial to add sufficient margins in calculation results to account for their lack of accuracy and precision. While accuracy refers to a systematic deviation, the precision expresses the variability around the expected result. Historically, margins have been added to account for any source of misjudgment by inclusion of safety factors. The analyses were completely deterministic, meaning that any calculation is completely reproducible with a well-defined reproducible error. No distinction was made between systematic deviation and variability. The current prevailing statistical perspective on calculations is different and boosted by frequent use of random sampling (RS) of computer models, often denoted Monte Carlo simulations. Any such method enforces a statistical perspective on possible errors. Systematic deviations and statistical variations should then be distinguished and are usually characterized with *bias* (ε) and *sampling variance* (σ^2), defined as the expected error of the calculated result and the second statistical moment around its mean.

A method of quantifying any well-defined number from a finite sample is often labeled *estimator*. Here, a sample consists of a set of results generated by random sampling of uncertain parameters of a complex model. The model could for instance describe the core temperature of a nuclear power reactor when cooling is lost. It should be emphatically emphasized that no single estimation or specific model is of interest, it is rather the *expected* outcome and variability of *repeated* estimation for a wide range of models that is in focus. These are classical topics in the fields of mathematical statistics and statistical signal processing [1], where some of the adopted terminology is frequently used.

For instance, assume it is of interest to estimate the expected model result $y = \langle x \rangle$. For this purpose define $\hat{y} = 1/n \sum_{k=1}^{n} x\left(\theta^{(k)}\right)$, where $\theta^{(k)}$ is the $k-th$ random instance of n model parameter sets. Clearly, $\hat{y}$ is a function of random parameters and consequently itself a random quantity. Hence, it has an expected value $\langle \hat{y} \rangle$ resulting in a bias $\varepsilon = \langle \hat{y} \rangle - y = \langle \hat{y} \rangle - \langle x \rangle$ as well as sampling variance $\sigma^2 = \langle \delta^2 \hat{y} \rangle, \delta\hat{y} \equiv \hat{y} - \langle \hat{y} \rangle = \hat{y} - (\langle x \rangle + \varepsilon)$, which in turn can be estimated by evaluating corresponding expectations over the sample $\{\Theta\}_{m=1}^{N} = \{\theta^{(1,k)}\}_{k=1}^{n}, \{\theta^{(2,k)}\}_{k=1}^{n}, \ldots \{\theta^{(N,k)}\}_{k=1}^{n}$ of N sets of n parameter sets, in total requiring nN model evaluations $x\left(\theta^{(jk)}\right)$. As any random quantity $\hat{y}$ has a *sampling probability density function* $g(\hat{y})$ as well as a cumulative *probability distribution* $G(\hat{y}) = \int_{-\infty}^{\hat{y}} g(y)dy$. The formal statistical analysis of $\hat{y}$ is not different from that of x, only the interpretations differ. Now if N is finite, $\{\Theta\}_{m=1}^{N}$ will itself be a random quantity with bias, finite sampling variance and its own sampling distribution. This chain of random quantities will continue forever, unless it is truncated at a certain level by evaluating expectations over *all* possible model results, e.g. let $N \to \infty$ above. If an analytical evaluation is not possible numerical estimation may be satisfactory, provided N is sufficiently large. For many examples in statistics analytical evaluation is possible [2], but hardly ever for realistic computer models for which the computational burden of evaluating a very large number nN of models is well beyond reach. This might explain why sampling statistics of Wilks' approach [3] has not been discussed in the literature in the context of advanced computations. Our goal is to improve the understanding of the magnitude of bias and sampling variance of Wilks' method, so it may be utilized with greater confidence.

The statistician Ronald Fisher [2] was one of the first to address bias and sampling variance. These concepts are vital to *statistical hypothesis testing*, today widely practiced in virtually all fields of science and engineering since almost a century. Prior to his work no distinction was generally made between estimator statistics of a finite *sample* and the entire *population* $(N \to \infty)$. That is very important since the uncertainty of the sample exceeds that of the population, and the infinite population always is inaccessible. Samuel Wilks [3] was one successor of Fisher's who fully acknowledged his distinction between sample and population statistics. Active in the field of manufacturing, rather than agriculture as Fisher, he targeted 'safe', or *conservative* bounds instead of expected outcomes. This focus made his approach well adapted for risk assessment. However, the asymmetry of conservatism naturally enforces bias. Its magnitude appeared to be of minor interest to Wilks, compared to the degree of conservatism. The least bias required is a consequence of the sample variance. The level of conservatism, bias and sampling variance are thus intimately related. On one hand, a large bias is *economically* costly since results suggest excessive physical margins. On the other hand, enforcing a small bias of estimation usually requires a large *computationally* costly model sample.

While Fisher [2] focused on the quality of estimated mean, variance and correlation, Wilks' [3] addressed *faithful* estimation (FE) of confidence intervals (CI), following the definition based on ordering of outcomes. The difference is larger than it may appear, since the latter sorts sample values (as in section 3) and the former involves statistical expectation over the population (as practiced in section 4). As currently practiced in measurement science [4] the estimated mean $\langle x\rangle$ and variance σ_x^2 are often related to CIs $\left[x_\alpha^-, x_\alpha^+\right]$ of confidence α by more or less hypothetical expansions with *coverage factors* $k_\alpha^\pm$ satisfying $x_\alpha^\pm \equiv \langle x\rangle \pm k_\alpha^\pm \sigma_x$. This is *not* FE of CI's since possible outcomes are *not* ordered, as prescribed by the definition. Most importantly, FE is nearly ill-posed for limited samples of commonly assigned shallow tail (normal) distributions [5] and thus might be of poor quality. For evaluation of modeling uncertainty (MU) the value of FE is more questionable than for the experimental sampling practiced by Wilks, as it is merely a question whether the pdf $f_\theta(\theta)$ of model parameters or the coverage factor k_α of the model result $f_x(x)$ is assigned. By no means, it is obvious that the best estimator of MU utilizes FE.

Wilks' approach estimates CI *bounds* of a *statistical model* with a *statistical method,* for confidence α and probability β expressing the degree of *conservatism*. It is central that two statistical perspectives are imposed on a *deterministic* physical problem, as illustrated in Fig. 1 below. The uncertainty reflects our limited knowledge of the true result $\bar{x}$. Note that bias and sampling variance are *not* deficiencies of the sampling generator. On the contrary, their definitions rely upon perfect random sampling.

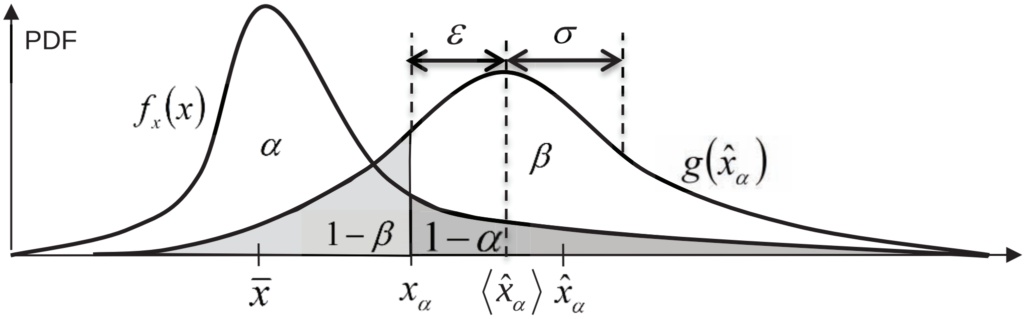

Figure 1. Wilks' estimate $\hat{x}_\alpha$ of the upper bound x_α for confidence α follows the sampling pdf $g(\hat{x}_\alpha)$, has bias ε and sampling variance σ, with given probability bound, or conservatism $P(\hat{x}_\alpha \geq x_\alpha) \geq \beta$. The model pdf $f_x(x)$ expresses our limited knowledge of the true value $\bar{x}$.

The addressed problem is first illustrated with a trivial example (section 2), before the derivation of Wilks' method is given (section 3). The main part of statistical evaluation (section 4) is then followed by possible modifications and presentation of some viable alternatives (section 5). Lastly, conclusions (section 6) summarize our findings.

2. PROBLEM ILLUSTRATION

Suppose we are interested in the core temperature $(\bar{x})$ of a nuclear power reactor. Before high risk experiments are performed it is desirable to simulate possible random outcomes (x) with a dedicated software model (Ω). For the purpose of illustration, let's assume this model is given by,

$$\bar{x} \leftrightarrow x = \Omega(\theta) = \Omega_0 + 53\theta_1 - 38\theta_2, \quad \Omega_0 = 1000, \quad \begin{cases}\theta_1 \sim UNI(0,1)\\ \theta_2 \sim NRM(0,1)\end{cases}, \tag{1}$$

where $UNI(0,1)$ and $NRM(0,1)$ refer to zero mean unit variance uniform and normal probability distributions, respectively. Of course, any realistic model would be much more complex and would, e.g. include solutions to partial differential equations of heat diffusion and describe non-linear effects. These complications must be properly accounted for the result to have any real meaning, but will be irrelevant here. The quantity $\bar{x}$ describes the true temperature but our understanding represented by the random variable x is not perfect, neither with respect to the structure (Ω), nor the parameters (θ) of the model generating it. We are usually confident enough about the model structure not to associate any uncertainty to the *form* (Ω) of these equations. It is different for the parameters θ, which may consist of universal physical constants, initial and/or boundary constraints, and sometimes inferred from experiments [6], or just plausible assumptions. Using a *statistical* representation of our limited knowledge of θ, as is most common today, is our active choice.

The obvious safety aspect is how large the true temperature $\bar{x}$ possibly can be. The uncertain model should be used to find a safety margin, or provide a bound x_α which is at least extreme as the true physical temperature with a given probability, or confidence level α, $P(x_\alpha \geq \bar{x}) \geq \alpha$. Wilks' statistical method estimates a bound $\hat{x}_\alpha$ of this bound x_α with a given probability, or degree of conservatism β, $P(\hat{x}_\alpha \geq x_\alpha) \geq \beta$. The probability $P(\hat{x}_\alpha \geq \bar{x})$ of a correctly estimated bound with respect to the true value is however lower than both α and β. This probability describing the credibility of estimation should be our main concern in risk analysis, a fact that will receive special attention in section 5.1.

A common application of Wilks' method (section 3.1) is to let $\hat{x}_\alpha$ be given by the most extreme result of one sample comprising $n = 59$ random values of $\Omega(\theta)$, $\hat{x}_\alpha = \max\limits_{k=1,2,\ldots 59}\left(\Omega\left(\theta^{(k)}\right)\right)$, for $\alpha = \beta = 0.95$. The generic problem to be addressed is for the model $\Omega(\theta)$ illustrated in Fig. 2: ***Every time Wilks' method is repeated, a different estimate*** $\hat{x}_\alpha$ ***is obtained***. To reduce the variance, the $r-th$ most extreme value can be chosen, provided the number of samples n is adjusted according to r, α, β (section 3.2).

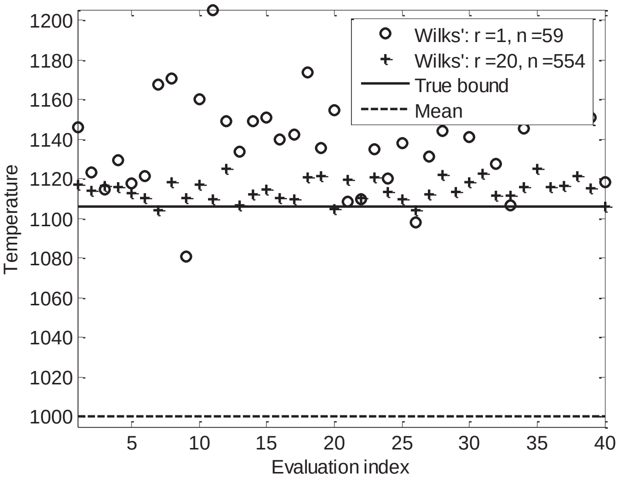

Figure 2. Results of repeated application of Wilks' method on the model $\Omega(\theta)$ (Eq. 1), for $\alpha = \beta = 0.95$ and $r = 1{,}20$, with the true bound (solid) and mean result (dashed) indicated.

For $\beta = 0.95$, the *expected* number of failing estimates of 40 repetitions is 2. This will generally not hold for any finite number of repetitions, due to sampling variance of the failure ratio (see section 1). To verify such high levels β accurately, the number of repetitions must be exceedingly large. The targeted bias and sampling variance of $\hat{x}_\alpha$ describe its *expected* deviation from the true bound and variation, respectively, evaluated for an infinite number of repetitions.

3. WILKS' METHOD

3.1. Full Sampling Range

A conservative estimate of a one-sided CI is readily found with a simple box counting experiment. First divide the sample space S of model results x into disjoint subspaces S_1 and S_2,

$$x \in S = S_1 \cup S_2, \quad S_1 \cap S_2 = 0, \quad P(x \in S_1) = \alpha \, . \tag{2}$$

Then, draw a sample of n independent model results $\{x^{(j)}\}_{j=1}^{n}$. Let n_k denote the number of values in subspace S_k. The probability β of finding at least one value $x \in S_2$ is given by,

$$\beta \equiv P(n_2 \geq 1) = 1 - P(n_1 = n) = 1 - \alpha^n \, . \tag{3}$$

Now define the estimator $\hat{x}_\alpha$ of the confidence bound x_α for a one-sided CI to be given by the most extreme sample value: Provided $x_k \in S_k$ implies $x_1 \leq x_2$, $P(\hat{x}_\alpha \equiv \max(x_k) \in S_2) = \beta$ for the upper and $P(\hat{x}_\alpha \equiv \min(x_k) \in S_1) = \beta$ for the lower bound. The probability of obtaining a true bound is hence β, representing the degree of *conservatism*. This is a safety margin measured in probability of success. Equation 3 yields an explicit expression for the least possible sample size, $n \geq \log(1-\beta)/\log(\alpha)$. For the common choice $\alpha = \beta = 0.95$, $n \geq 59$.

3.2. Truncated Sampling Range

The selection of the most extreme sample value may be generalized to the $r \geq 1$ most extreme value. This will exclude extreme sample values for $r > 1$ and thus *truncate* the sampling range. That is, $\hat{x}_\alpha \equiv \tilde{x}^{(r)}, r \geq 1, \{\tilde{x}^{(j)}\} = O\{x^{(j)}\}$, where the operator $O\{x^{(j)}\}$ sorts values $\{x^{(j)}\}$, in ascending or descending order depending on whether the lower or upper bound is estimated.

As in section 3.1, divide the sampling range S of model results into disjoint subspaces S_1 and S_2. This time however, require at least r values to belong to S_2 or S_1, for estimating the upper or lower bound, respectively. The conjugated event is that $0,1,2,\ldots \text{or}\, r-1$ values fall into category S_2 or S_1. The probability of each such configuration with $n_2 = k$ is given by $\alpha^{n-k}(1-\alpha)^k$. Since the order the successive values are obtained is irrelevant their number is given by the binomial coefficient $\begin{pmatrix} n \\ k \end{pmatrix} \equiv \dfrac{n!}{k!(n-k)!}$. The sampling probability of successful estimation, or degree of conservatism is thus,

$$\beta \equiv P(n_2 \geq r) = 1 - \sum_{k=0}^{r-1} P(n_2 = k) = 1 - \sum_{k=0}^{r-1} \begin{pmatrix} n \\ k \end{pmatrix} \alpha^{n-k} (1-\alpha)^k \, . \tag{4}$$

As required, for the full sampling range $r=1$ Eq. 4 is identical to Eq. 3. A difference is that it is here necessary to determine sample size n from α, β, r numerically for $r>1$. According to Fig. 3 below, there is an almost linearly increasing cost of additional sampling for truncation. Fortunately, as shown in section 4 the bias ε and variance σ^2 of $\hat{x}_\alpha$ decreases with r.

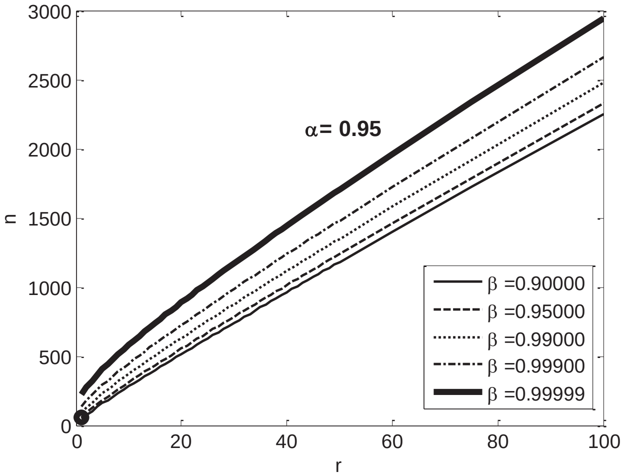

Figure 3. The number n of samples as function of truncation level r, for different conservatism β and confidence $\alpha=0.95$. The default $r=1, \beta=0.95, n=59$ is indicated ('o' close to origin).

4. STATISTICAL EVALUATION

4.1. Bias and variance of sampling

The bias and sampling variance of any estimator $\hat{y}$ of a well-defined quantity y is defined as,

$$\begin{aligned} \varepsilon &\equiv \langle \hat{y} \rangle - y, \\ \sigma^2 &\equiv \left\langle \left(\hat{y} - \langle \hat{y} \rangle \right)^2 \right\rangle = \langle \hat{y}^2 \rangle - \langle \hat{y} \rangle^2, \end{aligned} \tag{5}$$

where $\langle \cdot \rangle$ denotes statistical expectation over the population of all possible results. While ε reflects systematic errors, σ^2 is perhaps the most frequently used measure of variation. Consequently, 'minimum variance unbiased' (MVU) estimators are obvious targets [1]. For instance, the mean (y_1) and variance (y_2) of model results are usually estimated with $\hat{y}_1$ and $\hat{y}_2$, respectively,

$$\begin{aligned} y_1 &\equiv \langle x \rangle \approx \hat{y}_1(n) \equiv \frac{1}{n} \sum_{k=1}^{n} x^{(k)}, \\ y_2 &\equiv \left\langle \left(x - \langle x \rangle \right)^2 \right\rangle \approx \hat{y}_2(n) \equiv \frac{1}{n-1} \sum_{k=1}^{n} \left(x^{(k)} - \frac{1}{n} \sum_{k=1}^{n} x^{(k)} \right)^2, \end{aligned} \tag{6}$$

from a finite sample $\left\{ x^{(k)} \right\}_{k=1}^{n}$ of n model results. The normalization $1/(n-1)$ [instead of $1/n$] of $\hat{y}_2$ is here specifically chosen to eliminate its bias ε, a very well-known result [2].

Evaluation of the bias ε and the sampling variance σ^2 requires knowledge of the sampling pdf $g(\hat{y})$ of $\hat{y}$, or corresponding information. Integrating by parts, equivalent expressions with the probability distribution function $G(\hat{y})$ can be found,

$$\left\langle \hat{y}^k \right\rangle = \int_a^b \hat{y}^k g(\hat{y}) d\hat{y} = \left[\hat{y}^k G(\hat{y})\right]_a^b - k\int_a^b \hat{y}^{k-1} G(\hat{y}) d\hat{y}, \quad G(\hat{y}) = \int g(\hat{y}) d\hat{y}, \quad \hat{y} \in [a,b]. \tag{7}$$

Estimation of ε and σ with $\hat{\varepsilon}$ and $\hat{\sigma}$ requires a set $\left\{\hat{y}^{(j)}(n)\right\}_{j=1}^{N}$ of N results, each from a unique finite sample $\left\{x^{(k)}\right\}_{k=1}^{n}$ of n model results. The size N should be large enough to suppress the residual sampling variance of $\hat{\varepsilon}$ and $\hat{\sigma}$ below recognition. According to Eq. 5 and analogously to Eq. 6, but this time for $\hat{y}$ instead of x,

$$\begin{aligned} \varepsilon &= \left\langle \hat{y}(n) \right\rangle - y \approx \hat{\varepsilon} \equiv \frac{1}{N}\sum_{j=1}^{N} \hat{y}^{(j)}(n) - y, \\ \sigma^2 &= \left\langle \left(\hat{y}(n) - \left\langle \hat{y}(n) \right\rangle\right)^2 \right\rangle \approx \hat{\sigma}^2 \equiv \frac{1}{N-1}\sum_{j=1}^{N} \left(\hat{y}^{(j)}(n) - \frac{1}{N}\sum_{k=1}^{N} \hat{y}^{(k)}(n) \right)^2 . \end{aligned} \tag{8}$$

Wilks' approach estimates CI's. Two closely related quantities (corresponding to y) are involved: The confidence level α is estimated by the enclosed probability (section 4.2) and x_α with selected sample value(-s) (section 4.3). Both quantities have statistically well-defined bias ε and sampling variance σ^2.

4.2. Enclosed probability

The probability $\hat{\alpha}$ enclosed by Wilks' estimator of the CI S is given by

$$\alpha \approx \hat{\alpha} \equiv \int_S f_x(x) dx, \quad S = \left[\hat{x}_\alpha^-, \hat{x}_\alpha^+\right], \tag{9}$$

where $f_x(x)$ is the pdf of model results x. As indicated, perfect estimation corresponds to an enclosed probability $\hat{\alpha}$ equal to the assigned confidence level α. Since the CI is determined from a sample of random model results $\left\{x^{(j)}\right\}_{j=1}^{n}$ as described in section 2, $\hat{\alpha}$ is itself a random variable. According to Eqs. 5 and 7, the bias ε_α and sampling variance σ_α^2 of $\hat{\alpha}$ is calculable from the statistical moments,

$$\left\langle \hat{\alpha}^k \right\rangle = 1 - k\int_0^1 \hat{\alpha}^{k-1} G(\hat{\alpha}) d\hat{\alpha}, \quad k = 1,2, \tag{10}$$

here expressed in $G(\hat{\alpha})$ instead of $g(\hat{\alpha})$ for reasons to become apparent: The condition of conservative estimation states that the *expected* enclosed probability $\hat{\alpha}$ should be at least as large as α,

$$\beta = \int_\alpha^1 g(\hat{\alpha}) d\hat{\alpha} = G(1) - G(\alpha) = 1 - G(\alpha). \tag{11}$$

Now, instead of finding the pdf $g(\hat{\alpha})$ by evaluating the derivative $g(\hat{\alpha}) = -\partial\beta/\partial\alpha$ of Eq. 4 at $\alpha = \hat{\alpha}$, the sampling distribution $G(\hat{\alpha})$ may be directly read off from the same equation by comparison to Eq. 11,

$$G(\hat{\alpha}) = \sum_{k=0}^{r-1} \binom{n}{k} \hat{\alpha}^{n-k} (1-\hat{\alpha})^k . \tag{12}$$

Using this expression the desired expectations of Eq. 10 can be readily evaluated by repeatedly integrating by parts. Inserting the result into Eq. 5, the targeted bias and sampling variance are found,

$$\begin{aligned} \varepsilon_\alpha &= \langle\hat{\alpha}\rangle - \alpha = 1 - \alpha - \frac{r}{n+1}, \\ \sigma_\alpha^2 &\equiv \left\langle (\hat{\alpha} - \langle\hat{\alpha}\rangle)^2 \right\rangle = \sqrt{\frac{r(n+1-r)}{(n+1)^2(n+2)}} . \end{aligned} \tag{13}$$

where n is determined from r, α, β according to Eq. 4. Incidentally, the results for $\varepsilon_\alpha/1-\alpha$ as well as $\sigma_\alpha/1-\alpha$ shown in Fig. 4 are almost indistinguishable for $\alpha = 0.95$ and $\alpha = 0.99$. This exceedingly weak dependence on α remain to be understood.

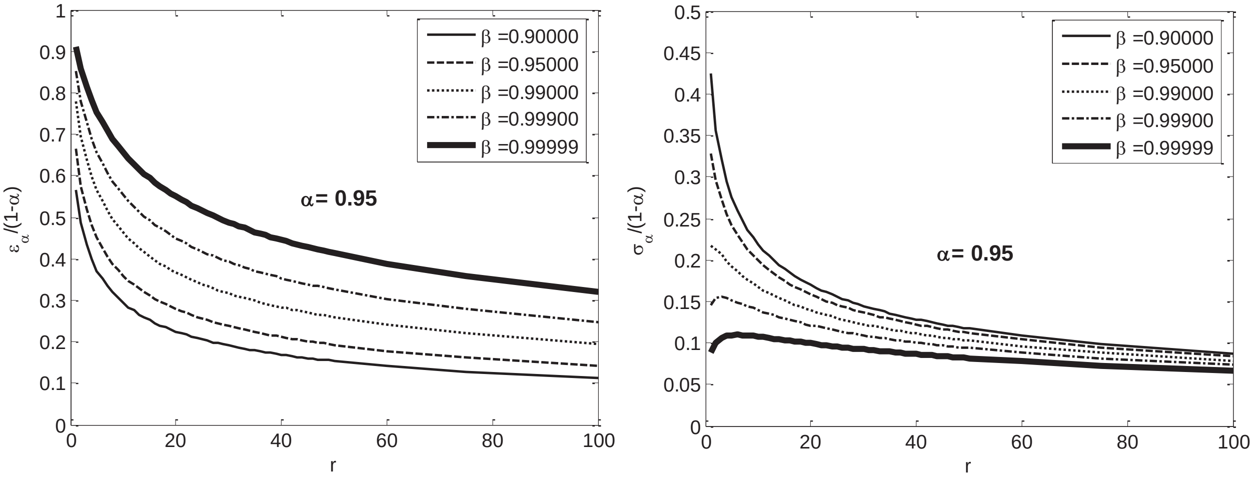

Figure 4. The scaled bias $\varepsilon_\alpha/(1-\alpha)$ (left) and standard deviation $\sigma_\alpha/(1-\alpha)$ (right) of the enclosed probability of Wilks' estimator, as function of truncation level r, for various degrees β of conservatism and confidence level $\alpha = 0.95$.

The bias ε_α is large due to the asymmetric requirement of conservative estimation. In principle, no matter how large the sampling variance is, any given level of conservatism can be enforced by adding a large enough bias with appropriate choices of parameters such as the truncation level r and sample size n.

Up to this point all results apply for *any* distribution $f_x(x)$ of model result. This is expected since only ordering and sorting of model results has been involved. Whether one result is larger or smaller than any other is entirely independent of their distribution, or the pdf of model results. Besides its simplicity, this generality of Wilks' approach likely contributes to its popularity.

4.3. Confidence interval

In stark contrast to the enclosed probability $\hat{\alpha}$ analyzed in section 4.2, the true and estimated confidence interval bounds $\hat{x}_\alpha^\pm$ are strongly dependent on how model results are distributed. The bias and sampling variance of Wilks' estimator of the CI will consequently by unique for every pdf $f_x(x)$ of model results. The sampling distribution $G(\hat{x}_\alpha^\pm)$ can be obtained from that of $\hat{\alpha}$, $G(\hat{\alpha})$. First observe that $\hat{\alpha}$ is a monotonic function of $\hat{x}_\alpha^\pm$,

$$\hat{\alpha} = \int_{\hat{x}_\alpha^-}^{\hat{x}_\alpha^+} f(x)dx = F(\hat{x}_\alpha^+) - F(\hat{x}_\alpha^-), \qquad (14)$$

since the probability distribution $F(x)$ is monotonically increasing. Changing focus from $\hat{\alpha}$ to $\hat{x}_\alpha = \hat{x}_\alpha^\pm$ is hence equivalent to the application of an invertible transformation $\hat{\alpha}(\hat{x}_\alpha)$. The probability distribution $G(\hat{x}_\alpha)$ is therefore obtained by simple substitution in Eq. 12, to evaluate the expectations of Eq. 7,

$$\langle \hat{x}_\alpha^k \rangle = b - k\int_a^b \hat{x}_\alpha^{k-1} G(\hat{\alpha}(\hat{x}_\alpha))d\hat{x}_\alpha, \quad k = 1{,}2, \quad \hat{x}_\alpha \in [a,b]. \qquad (15)$$

Evaluating these statistical moments numerically will allow for determination of the bias ε_x and sampling variance σ_x^2 of $\hat{x}_\alpha$ defined by Eq. 5. Clearly, the resulting bias shown in Fig. 5 is much less for the uniform (right, UNI), than for the normal (left, NRM) distribution of model results. That is mainly due to the infinite support (range) of the NRM distribution. As expected, the sampling variance shown in Fig. 6 is correspondingly much larger for NRM (left), than for UNI (right). That agrees with the statement in section 1 that a large bias is required for a large sampling variance.

By scaling up the expected variation σ to a measure of uncertainty (around 2σ) representing a plausible width of a confidence interval of $\hat{x}_\alpha$, it is seen that the relative error $|\hat{x}_\alpha - x_\alpha| / |x_\alpha - \langle x \rangle|$ (with bias) can be as large as 100% for the default application $r = 1, \alpha = \beta = 0.95$ of Wilks' method. The bias and uncertainty 2σ are almost equal, as conservatism implies for symmetric sampling pdfs.

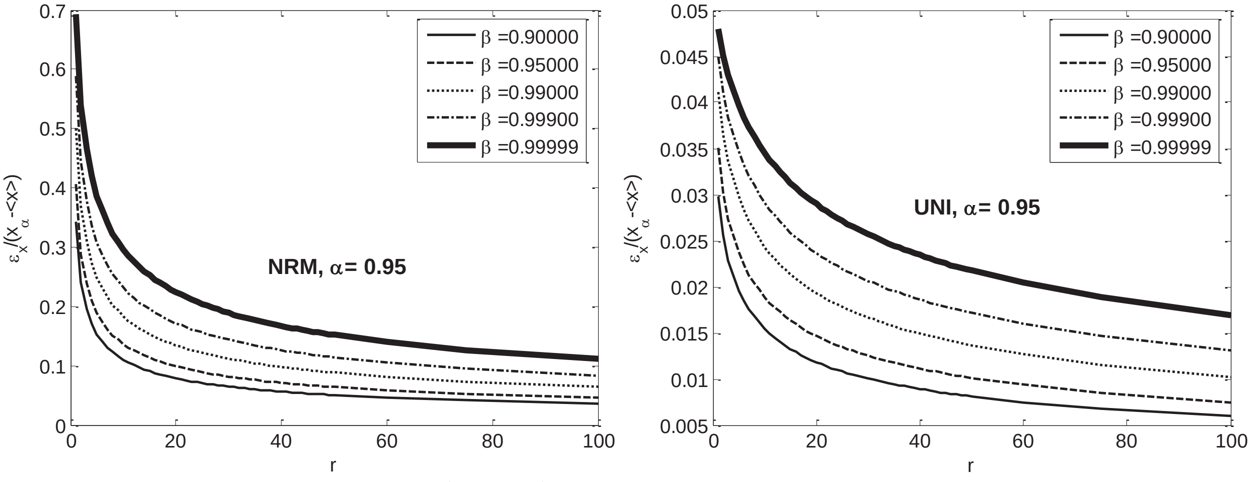

Figure 5. The normalized bias $\varepsilon_x/(x_\alpha^+ - \langle x \rangle)$ of the estimated upper bound $\hat{x}_\alpha^+$, for normal (left) and uniform (right) model pdfs $f_x(x)$, for conservatism β and confidence level $\alpha = 0.95$.

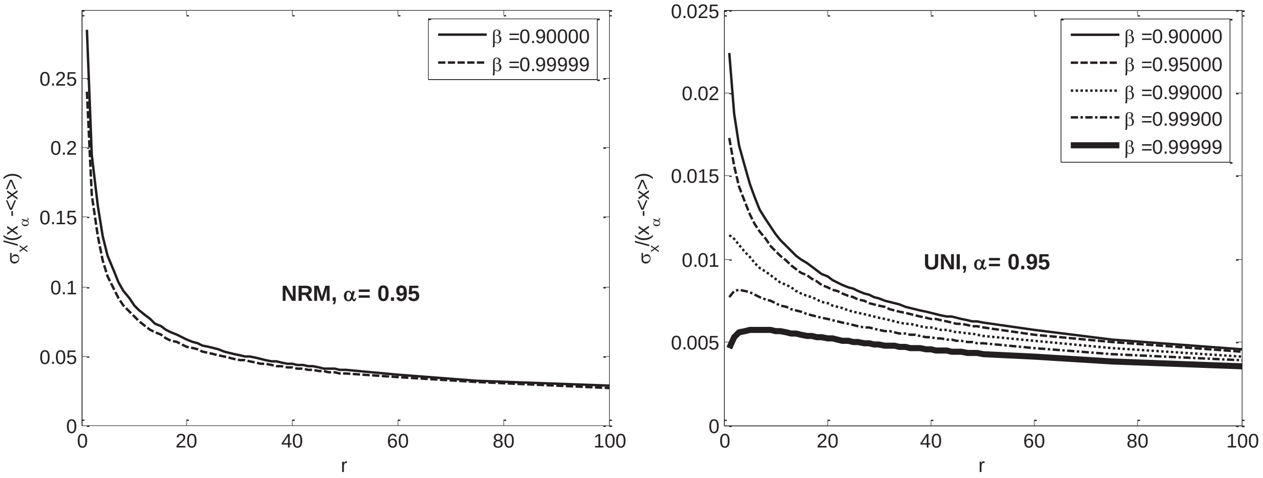

Figure 6. The normalized standard deviation $\sigma_x / \left(x_\alpha^+ - \langle x \rangle\right)$ of the estimated upper bound $\hat{x}_\alpha^+$, for normal (left) and uniform (right) model pdfs $f_x(x)$, for conservatism β and confidence level $\alpha = 0.95$.

5. MODIFICATIONS AND ALTERNATIVES

5.1. Credible application

A problem with current de-facto standard for applying Wilk's approach is that the acceptable uncertainty effectively has already been consumed by the true confidence bound $x_\alpha^\pm$ of the model, which leaves no uncertainty for its estimate $\hat{x}_\alpha^\pm$. To distinguish it from the confidence level, the probability of successful estimation will be denoted *credible level*. This will aggregate two sources of failing assessments, the uncertainty of model results and our limited ability of evaluating this uncertainty from a finite random sample. Using conditional probabilities $P(a|b)$ describing the probability statement a will occur, given b is satisfied,

$$\begin{aligned} P\left(\hat{x}_\alpha^+ \ge \bar{x}\right) &\ge P\left(\hat{x}_\alpha^+ \ge \bar{x} \middle| \hat{x}_\alpha^+ \ge x_\alpha^+\right) \cdot P\left(\hat{x}_\alpha^+ \ge x_\alpha^+\right) + P\left(\hat{x}_\alpha^+ \ge \bar{x} \middle| \hat{x}_\alpha^+ < x_\alpha^+\right) \cdot P\left(\hat{x}_\alpha^+ < x_\alpha^+\right) \\ &\ge P\left(x_\alpha^+ \ge \bar{x}\right) \cdot P\left(\hat{x}_\alpha^+ \ge x_\alpha^+\right) = \alpha\beta \end{aligned} , \qquad (16)$$

for the upper bound, and similarly for the lower one. A lower bound of the probability of success $\left(\hat{x}_\alpha^+ \ge \bar{x}\right)$, or credibility of estimation, is hence given by $\alpha\beta$, not α. Setting a level $\alpha = 0.95$ as a vague limit of perceived 'always' stems from early development of hypothesis testing [2] but might be unsatisfactory for critical risk assessment in nuclear power applications. Applying this level twice for both α and β in Wilks' approach, the notion of 'always' translates into credible bound no better than $\alpha\beta \approx 0.90$, which is an even more questionable probability of failing bounds.

To remedy this situation, the obvious change of perspective is to start by defining a level of credibility $\alpha\beta$ which is reasonable for the targeted application. How that is distributed between α and β is irrelevant for the bound on credibility, but has influence on the required sample size n. Effectively, the result of this modification is that n increases substantially. That will reduce the efficiency below what is perceived assuming α describes the risk of failure. In comparison, other alternatives like deterministic sampling (section 5.2) might then become more attractive. Examples of how given credibility levels translates into sample size is presented in Table 1, where the entry $\alpha\beta = 0.95^2 \approx 0.90$ is included for comparison to current practice. Assigning an accepted probability of failure to the credibility bound $\alpha\beta$ instead of the confidence level α the sample size increases substantially, compare cases 1 and 2, 5 and 6 or 9 and 10. How the credibility is shared between confidence α and conservatism β does indeed matter, as seen from cases 2 and 3, 6 and 7, or 10 and 11. Sharing them equally (see *), $\alpha = \beta \Leftarrow \sqrt{\alpha\beta}$, appears to be a fair strategy to limit n. In that case, n is roughly doubled.

Table I. Sample sizes n for given lower bound $\alpha\beta$ on credibility and confidence level α.

Case	Truncation r	Credibility bound $\alpha\beta$	Confidence level α	Sample size n
1	1	0.902	0.950	59
2	1	0.950	0.975*	145
3	1	0.950	0.990	320
4	1	0.980	0.990	458
5	5	0.902	0.950	181
6	5	0.950	0.975*	405
7	5	0.950	0.990	947
8	5	0.980	0.990	1156
9	10	0.902	0.950	311
10	10	0.950	0.975*	678
11	10	0.950	0.990	1611
12	10	0.980	0.990	1872

5.2. Deterministic sampling

There is one clear distinction between the statistical estimation addressed by Wilks and the evaluation of MU addressed here. Physical sampling as practiced by Wilks makes no reference to the sources of uncertainty whatsoever. It is drastically different for how MU is determined. Completely specified uncertain models provide detailed information of all sources of uncertainty. While Wilks' addressed the *statistical* problem of analyzing a finite set of randomly drawn sample values, the evaluation of MU constitutes an entirely *deterministic* problem of *uncertainty propagation*. In the latter but not the former case, the result is well defined and can in principle be found with arbitrary high accuracy. Consequently, Wilks' method is *not* primarily derived for the addressed task of evaluating MU.

A high level of utilization of available information is generally required for optimizing the quality of any calculation. Uncertainty propagation but not physical sampling requires knowledge of model structure (Ω in section 2). Wilks' approach explores a minor portion of this information *by chance* (random sampling of uncertain model). A presumably superior approach would be to explore all available relevant pieces of information *systematically*. Sampling as a general idea of representing information of finite dimensionality may still be used. Though, systematic exploration of information requires such samples to calculated deterministically using specific *sampling rules*, preferably optimized for the task considered. This defines an infinitely large class of methods that for obvious reasons can be labeled *deterministic sampling* (DS), to contrast them from the ubiquitous set of random sampling (RS) techniques to which Wilks' approach belong. The perhaps best known example of DS is the unscented Kalman filter [7], in which the covariance is propagated with specific sampling rules with entirely reproducible results.

Virtually all DS methods targets highest possible efficiency by using a minimal calculated model sample. For small samples of any kind, adding a single sample value will modify the statistics considerably. Consequently, to any sample rule there is a definite sample size that must not be modified without changing the rule completely. The sampling rule and sample size are intimately related for DS, in contrast to most RS methods. The computational efficiency of DS is precisely what is needed to evaluate MU of many complex models.

Let's briefly illustrate some principles of DS by revisiting the trivial example of section 2. The relevant question is then how the statistical information of Eq. 1, i.e. the statistics of θ, can be *represented,* or described in an efficient way with as few sample values as possible. These should not primarily be typical or probable (as in RS) but rather *encode*, or carry the available *reliable* pieces of information. There are many ways to satisfy the first $\mu_\theta = \langle \theta \rangle$ and second statistical moments around the mean $\sigma_\theta^2 = \left\langle \left(\theta - \langle \theta \rangle \right)^2 \right\rangle$ of θ. For instance, the two DS samples

$$\left\{ \theta^{(k)} \right\} = \begin{cases} \left(\mu_\theta - \sigma_\theta \quad \mu_\theta + \sigma_\theta \right), \quad \text{or} \\ \left(\mu_\theta - \sqrt{2}\cos(\varphi)\sigma_\theta \quad \mu_\theta + \sqrt{2}\cos(\varphi)\sigma_\theta \quad \mu_\theta - \sqrt{2}\sin(\varphi)\sigma_\theta \quad \sqrt{2}\sin(\varphi)\sigma_\theta \right) \end{cases}, \tag{17}$$

use two different sampling rules but both represent the first and second moments $\left(\mu_\theta, \sigma_\theta \right)$ correctly, no matter what φ is. Their fourth moments are however generally different,

$$\gamma_\theta^4 = \left\langle \left(\theta - \langle \theta \rangle \right)^4 \right\rangle = \begin{cases} \sigma_\theta^4, \quad \text{or} \\ 2\left(\cos^4 \varphi + \sin^4 \varphi \right) \end{cases}. \tag{18}$$

By modifying φ it is hence possible to vary γ to satisfy a given fourth moment. The computational cost to achieve this control is a doubled sample size. In practice, reliable statistical information above the second moment is rarely known, often making it superfluous to represent higher moments. Any finite skew $\left\langle \left(\theta - \langle \theta \rangle \right)^3 \right\rangle$ often has a large influence and may need to be controlled. Without further ado, one DS sample which encodes the first and all second moments of the two independent parameters $\theta = \left(\theta_1 \quad \theta_2 \right)^T$ of Ω with zero skew is given by [8],

$$\left\{ \theta^{(k)} \right\} = \begin{pmatrix} +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \end{pmatrix}. \tag{19}$$

Here, column k contains the $k-th$ set of sample values of parameters θ_1 (top) and θ_2 (bottom). The numerical values are very simple in this case since both parameters have zero mean and unit variance. As a matter of fact, for this particular case of an *affine* (linear combinations of parameters) model $\Omega(\theta)$, any moment of θ higher than the second is completely irrelevant for the propagated variance $\text{var}(\Omega(\theta))$. This will not hold for models non-linear in parameters, but often be remarkably accurate for modest non-linearity. How much information that *needs* to be encoded depends on the available information (of θ), the model ($\Omega(\theta)$) the resulting statistics ($\text{var}(\Omega(\theta))$) of interest, and desirable accuracy. Even though DS methods for FE of CI can be found [9], it is often more efficient and practical to rely upon the current practice of non-FE [4], using a coverage factor k_α to expand propagated mean and covariance to confidence bounds,

$$\begin{aligned}\hat{x}_\alpha^\pm &= \langle x\rangle \pm k_\alpha \sigma_x \\ \langle x\rangle &\approx \left\langle \Omega\left(\theta^{(k)}\right)\right\rangle_k \equiv \frac{1}{4}\sum_{k=1}^{4}\Omega\left(\theta^{(k)}\right),\quad \sigma_x^2 \approx \frac{1}{4}\sum_{k=1}^{4}\delta^2\Omega\left(\theta^{(k)}\right),\quad \delta\Omega\left(\theta^{(k)}\right) \equiv \Omega\left(\theta^{(k)}\right) - \left\langle \Omega\left(\theta^{(k)}\right)\right\rangle_k \end{aligned} \quad . \qquad (20)$$

Note that since the expectations are evaluated with a representation of the whole population obtained with DS rather than estimated for a random sample, all expectations $\langle\cdot\rangle_k$ utilize normalization $1/n = 1/4$ and not $1/(n-1)$ as some in Eqs. 6 and 8. Instead of bias and sampling variance, the *evaluated* (deterministic DS method) upper confidence bound $\hat{x}_\alpha = 1107.3$ has relative *error* $e \equiv (\hat{x}_\alpha - x_\alpha)/|x_\alpha - \langle x\rangle| = 1.4\%$ for a coverage factor $k_\alpha = 1.64$, assuming $\Omega(\theta)$ to be normally distributed. Thus for this example and particular DS sampling rule the efficiency as well as the accuracy is superior to that of Wilks' approach: Only four model samples were utilized and the error $e = 1.4\%$ is much less than the bias and sampling variance indicated by Figs. 5 and 6. The result of DS is compared to that of Wilks' method in Fig. 7.

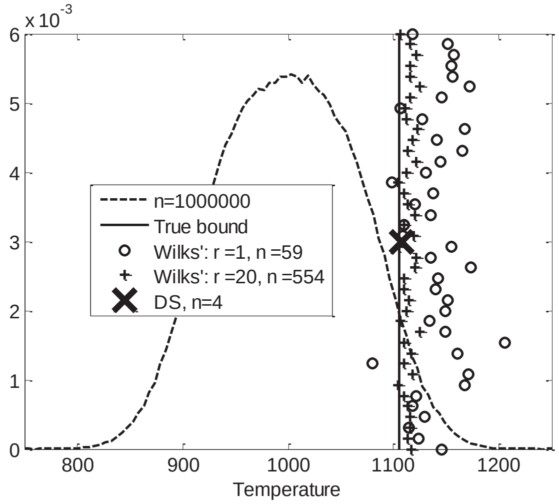

Figure 7. The upper confidence bound, evaluated with *one* DS method (x) and Wilks' method for $r = 1, 20$ (o,+), and true bound (solid) with approximate pdf (dashed) estimated from 10^6 samples.

The evaluation of covariance with DS rules typically requires around twice as many samples as the number of parameters [8]. Since the number of influential parameters of parametric models often is limited, DS generally utilizes significantly smaller samples than Wilks' method. Since all DS methods are completely reproducible, it is pointless to repeat the calculation as in RS to assess the magnitude of the error. However, there are usually many possible sampling rules, or DS ensembles which all represent the reliable pieces of statistical information correctly but produce slightly different results. The high efficiency allows for the use of several DS methods, to indicate a spread in the result. This spread is a consequence of encoding *in*-complete statistical information and that different ensembles represent the omitted information differently. By no means, it is related to any statistical effect like sampling variance.

Finally, it is worth mentioning that the mean and the covariance of n variables can be represented *exactly* by just $n+1$ samples (simplex ensemble [8]). This is *not* possible for RS, sampling variance will always render a finite error, no matter how large the sample is. To find the variance of an affine model, which is an obvious starting point of most approximations, this is *all* that matters. All DS methods but no RS technique will then provide the exact correct answer. DS methods are thus on par with the method of linearization. For moderate non-linearities, DS is often superior to model approximating techniques since it approximates model statistics, which almost without exception is considerably simpler. Further illustrations of DS with larger and more realistic examples can be found elsewhere [6,8-10].

6. CONCLUSIONS

In Wilks' approach for assessing modeling uncertainty (MU), the large sampling variance and the requirement of conservative estimation results in strong bias. It is not unlikely to obtain an error of estimated confidence bounds close to the true MU, i.e. 100% relative error. For risk management in nuclear power applications this might render excessive economical costs.

Focusing on the credibility of Wilks' estimate, i.e. the probability of failing bounds rather than confidence levels of the model, the number of samples should roughly be doubled at the same acceptance level. Current de-facto standard of 95% probability of estimating a proper bound of the 95% confidence limit(-s) with the largest sample value implies failure in up to 10% of all cases.

Wilks' approach of estimating MU is frequently outperformed by competing techniques such as the proposed novel class of deterministic sampling (DS) methods. In an illustrating example, one DS method was shown to give superior results with just four calculated sample values relying upon common practice utilizing coverage factors, compared to Wilks' method with 554 samples.

The reasons why Wilks' approach does not perform better is an unfortunate combination of faithful estimation and random sampling giving a large sampling variance, and no explicit utilization of the available information of the sources of resulting uncertainty, i.e. model structure and parameter distributions. That kind of information is *never* known in the original application of physical sampling proposed by Wilks, making it a comparatively much better choice in that case.

ACKNOWLEDGMENTS

Financial support from the Swedish Radiation Safety Authority, project 2037012-71, is gratefully acknowledged.

REFERENCES

1. S. Kay, *Fundamentals of Statistical Signal Processing, Estimation theory*, Vol. 1, Prentice Hall, New Jersey (1993).
2. J. Bennett, R.A. Fisher, *Statistical methods, experimental design, and scientific inference*, Oxford University Press (1995).
3. S.S. Wilks, "Determination of sample sizes for setting tolerance limits," *The Annals of Mathematical Statistics*, **12(1)**, pp. 91-96 (1941).
4. ISO GUM, Guide to the Expression of Uncertainty in Measurement, International Organisation for Standardisation, Geneva, Switzerland, 1995.
5. J.P. Hessling and J. Uhlmann, "Robustness of Wilks' Conservative Estimate of Confidence Intervals," Submitted to (dec 2014): *International Journal for Uncertainty Quantification*.
6. J.P. Hessling, "Identification of Complex Models," *SIAM/ASA Journal for Uncertainty Quantification*, **2(1)**, pp. 717-744 (2014).
7. S. Julier, J. Uhlmann, "Unscented Filtering and nonlinear estimation," *Proceedings IEEE,* **92**, pp. 401-422 (2004).
8. J.P. Hessling, "Deterministic Sampling for Propagating Model Covariance," *SIAM/ASA Journal for Uncertainty Quantification*, **1(1)**, pp. 297-318 (2013).
9. J.P. Hessling, T. Svensson, "Propagation of Uncertainty by Sampling on Confidence Boundaries," *International Journal for Uncertainty Quantification*, **3(5)**, pp. 421-444 (2013).
10. J.P. Hessling, *Digital Filters and Signal Processing*, chapter "Deterministic sampling for Quantification of Modeling Uncertainty of Signals", pp. 53-79, INTECH, Rijeka, Croatia (2013).