International Journal for Uncertainty Quantification, 5 (6): 569–583 (2015)

ROBUSTNESS OF WILKS' CONSERVATIVE ESTIMATE OF CONFIDENCE INTERVALS

***Jan Peter Hessling*[1,*] *& Jeffrey Uhlmann*[2]**

[1] *SP Technical Research Institute of Sweden, Measurement Technology, Box 857, SE-50115 Borås, Sweden*

[2] *University of Missouri—Columbia, Department of Computer Science, 201 EBW, Columbia, Missouri 65211, USA*

[*] Address all correspondence to Jan Peter Hessling E-mail: jan.peter.hessling@gmail.com

Original Manuscript Submitted: 12/05/2014; Final Draft Received: 11/02/2015

The striking generality and simplicity of Wilks' method has made it popular for quantifying modeling uncertainty. A conservative estimate of the confidence interval is obtained from a very limited set of randomly drawn model sample values, with probability set by the assigned so-called stability. In contrast, the reproducibility of the estimated limits, or robustness, is beyond our control as it is strongly dependent on the probability distribution of model results. The inherent combination of random sampling and faithful estimation in Wilks' approach is here shown to often result in poor robustness. The estimated confidence interval is consequently not a well-defined measure of modeling uncertainty. To remedy this deficiency, adjustments of Wilks' approach as well as alternative novel, effective but less known approaches based on deterministic sampling are suggested. For illustration, the robustness of Wilks' estimate for uniform and normal model distributions are compared.

KEY WORDS: *propagation, evaluation, uncertainty, modeling, sampling*

1. INTRODUCTION

In critical applications where safety is a major concern, e.g., nuclear power plant operation, experiments may have fatal consequences. By manipulating experimental key variables and monitoring the response of the system, conditions under which the system becomes unstable, or fails entirely, may be revealed. However, such procedures are clearly impractical and also intolerable for safety-critical systems as, e.g., a nuclear power plant [1]. Experiments are also often very expensive and can only cover a part of the operational space. Consequently, modeling is an appealing alternative. A truthful model of the system may also reveal high-risk situations triggered by rare but fatal events. Their low probability of occurrence makes them unlikely to occur in the relatively few physical experiments that can be performed, but they may be systematically studied with a model. Calibrated models [2–4] are traceable [5] to physical observations, and/or vice versa, thereby contributing to a generalized traceability chain. Proper physical operation can then be verified against accumulated knowledge (like nuclear physics) condensed in the model. Furthermore, modeling can yield results on the uncertainties associated with the parameters of interest. Normally, in safety applications, the uncertainty must be conservatively evaluated. More true than ever after Fisher's pioneering work on statistical inference [6], the reliability of our assessments of uncertainty is central. The existence of robust and versatile methods to assess uncertainty are thus essential to justify the use of modeling. For wide utilization and acceptance such methods also have to be simple (understandable by non-specialists) and effective (to allow for large complex models).

In the early 1940's, Samuel Wilks derived conservative estimates of univariate tolerance limits of specimen characteristics in mass production [7]. Later, Abraham Wald extended Wilks' work to multivariate observations [8]. The

2152–5080/15/$35.00

proposed method has an extraordinary generality and simplicity. Translated into confidence intervals of modeling results [9–11], it has become a *de facto* standard in some fields of engineering. In its simplest form, the upper confidence limit is, with a given probability, less than or equal to the largest value of a sample[1] of a fairly small model ensemble of results (typically 59, see below). It is obtained by evaluating the model for different parameter sets, randomly drawn from a plausible probability distribution. The appealing simplicity of this approach has likely contributed to its wide adoption. As defined, the stability of the method was thoroughly explored by Wilks, while the robustness of the estimates was not addressed at all. Lack of robustness does *not* undermine Wilks' approach, but the potentially high volatility of its estimates of confidence limits may lead to practical challenges. For example, a regulatory authority may need to know how much credibility can be associated with an estimated confidence interval for purposes of establishing safety guidelines.

Faithful estimation of confidence intervals (CI) $[x^-_{[\alpha]}, x^+_{[\alpha]}]$ which directly addresses the enclosed probability, is difficult and to some extent ill-posed. The problem is often circumvented by *not* evaluating CIs faithfully, exchanging an unreliable analysis with assumptions: Typically, a coverage factor k_α is used to expand robust estimates of mean and variance to a confidence interval [5]. It is *assigned* on the basis of a selected plausible probability density function (pdf) $f(x)$, rather than inferred. Then there is no problem of robustness, but also no lower limit on the validity as the result is never better than the hypothesis behind the assignment. That is particularly important for a model with high complexity which might render virtually any multi-variate probability distribution of calibrated parameters.

The problem of low robustness of faithful estimation of CIs is readily understood by studying sampling statistics of the enclosed probability,

$$P(S) = \int_S f(x)dx, \tag{1}$$

where S contains a subdomain of the support $\Psi(f)$ of the pdf $f(x)$ for the random observation x. Now, estimate the CI $S = [\hat{x}^-_{[\alpha]}, \hat{x}^+_{[\alpha]}]$ and refer to the confidence level error $\delta\hat{P} \equiv P([\hat{x}^-_{[\alpha]}, \hat{x}^+_{[\alpha]}]) - \alpha$ due to estimation errors $\delta\hat{x}^\pm_{[\alpha]} \equiv \hat{x}^\pm_{[\alpha]} - x^\pm_{[\alpha]}$ as lack of *stability*. Let the *robustness* describe the uncertainty of the estimates $\hat{x}^\pm_{[\alpha]}$, i.e., the variability of the errors $\delta\hat{x}^\pm_{[\alpha]}$. The stability studied by Wilks has a regular dependence on $\delta\hat{x}^\pm_{[\alpha]}$, since according to the mean value theorem of integration,

$$\delta\hat{P} \equiv \sum_\pm \delta\hat{P}^\pm = \sum_\pm \pm\delta\hat{x}^\pm_{[\alpha]} f(\hat{x}^\pm_{[\alpha]} + \theta^\pm\delta\hat{x}^\pm_{[\alpha]}),\ \theta^\pm \in [0,1]. \tag{2}$$

Thus, for any $\epsilon > 0$ there is a $\eta = \epsilon / \sum_\pm |f(\hat{x}^\pm_{[\alpha]} + \theta^\pm\delta\hat{x}^\pm_{[\alpha]})| > 0$ such that $|\delta\hat{P}| < \epsilon$ if $|\delta\hat{x}^\pm_{[\alpha]}| < \eta$. The reverse does not hold however, since for $\max_\pm |f(\hat{x}^\pm_{[\alpha]} + \theta^\pm\delta\hat{x}^\pm_{[\alpha]})| < |\delta\hat{P}|/2\epsilon$ we find $\max_\pm |\delta\hat{x}^\pm_{[\alpha]}| > \epsilon$. Robustness is thus a stronger criterion than stability. If $f(x_{[\alpha]^\pm})$ has no finite (non-zero) lower bound, *no faithful* stochastic CI estimator can be robust, meaning that there is no upper bound on $|\delta\hat{x}^\pm_{[\alpha]}|$. In that case there exists no bound on how much a modeling error $\delta f(x)$ of the pdf $f(x)$ may be amplified to estimated CIs. Clearly, it may become critical when the *aspect ratio* $\lambda(\alpha) \equiv \min_\pm f(x^\pm_{[\alpha]}) / \max_x f(x)$ is low. Robustness thus calls for attention for the ubiquitous high-confidence estimates of shallow tail distributions, typically $\alpha = 95\%$ limits of normal distributions ($\lambda = 0.15$).

Wilks estimated the tolerance range $[x^-_{[\alpha]}, x^+_{[\alpha]}]$ conservatively, from a sampled set $\{x_k\}^n_1$ of independent observations of x. In contrast to a Bayesian approach it is applied in one step utilizing one finite set of sampled data but no prior information. Requiring the enclosed probability to be at least as large as the confidence level α with probability, or level of *stability* β,

$$\mathcal{P}\left(P([\hat{x}^-_{[\alpha]}(\{x_k\}^n_1), \hat{x}^+_{[\alpha]}(\{x_k\}^n_1)]) \geq \alpha\right) = \beta, \tag{3}$$

where $\mathcal{P}$ labels sampling probability. This is equivalent to evaluating the lower confidence limit of the random quantity $P(\hat{x}^\pm_{[\alpha]}(\{x_k\}^n_1))$, at level β. Given estimators $\hat{x}^\pm_{[\alpha]}$, Eq. (3) will fix the least required sample size n, which was the primary goal of Wilks. Random sampling is here combined with faithful estimation of CIs. A potentially low robustness is then further depleted by the slow convergence of random sampling [12]. The robustness, i.e., the reproducibility of estimates $\hat{x}^\pm_{[\alpha]}$ may thus be exceedingly poor if β is only moderate, λ is low, and the estimators $\hat{x}^\pm_{[\alpha]}$ are crude.

[1] To conform [7], a sample will here denote a complete set of values, not a single observation.

Wilks obtained an explicit form of the sampling distribution of $P(\hat{x}^{\pm}_{[\alpha]})$, which is general and entirely independent of $f(x)$. That makes statements of stability immune to faulty assumptions of $f(x)$, a very powerful aspect in perspective of the current practice of statistical modeling. In stark contrast, the robustness to be studied here is strongly dependent on $f(x)$.

The derivation and main results of Wilks' method will be briefly recapitulated (Section 2), before robustness is studied (Section 3). Suggestions of modifications (Section 4) and alternative approaches (Section 5) will then follow, before the conclusion (Section 6) summarizes our findings.

2. WILKS' METHOD

For risk assessment we are primarily interested in one-sided CIs, rather than the double-sided CIs addressed in the original work [7]. The superscripts $^{\pm}$ of $\hat{x}_{[\alpha]}$ will therefore often be omitted in the following discussion. The common practice described in Section 2.1 is a simplified version of the more general case discussed in Section 2.2.

2.1 Full Sampling Range

A conservatively estimated one-sided CI is readily found with a simple box counting experiment. First divide the sample space S of a stochastic scalar observation x into disjoint subspaces S_1 and S_2,

$$S = S_1 \cup S_2,\ S_1 \cap S_2 = 0,\ \mathcal{P}(x \in S_1) = \alpha,\ \mathcal{P}(x \in S_2) = 1 - \alpha. \tag{4}$$

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

$$\beta \equiv \mathcal{P}(n_2 \geq 1) = 1 - \mathcal{P}(n_2 = 0) = 1 - \mathcal{P}(n_1 = n) = 1 - \prod_{j=1}^{n} \mathcal{P}(x_j \in S_1) = 1 - \alpha^n. \tag{5}$$

Provided $x_k \in S_k$ implies $x_1 \leq (\geq) x_2$, $\mathcal{P}(\hat{x}_{[\alpha]} \equiv \max(\min) x_k \in S_2) = \beta$. The most extreme sample $\hat{x}_{[\alpha]}$ thus provides a conservative estimate of the true CI limit $x_{[\alpha]}$. Since no sample value is excluded, the original, or full sampling range is utilized. Equation (5) yields an explicit lower bound on the sample size, $n \geq \log(1-\beta)/\log(\alpha)$. For $\alpha = 0.95 \leq \beta$, $n \geq 59$.

2.2 Truncated Sample Range

In the original work of Wilks, the CI of interest was determined from a truncated sampling range. That is, $\hat{x}_{[\alpha]} \equiv \tilde{x}_r$, $r \geq 1$, where $\{\tilde{x}_k\}$ is the ordered set of values $\{x_j\}$, ascending or descending depending on whether the lower or upper bound is estimated. The sampling range spanned by all sample values is here reduced, or truncated for $r > 1$ since $r-1$ of the most extreme values are excluded. Truncation will generally result in better estimates, at the expense of larger samples.

It is not necessary to derive the sampling density function $g(P)$, even if Wilks did so. The sampling distribution obtained by integrating g can be found directly by generalizing the box counting exercise of Section 2.1. As before [Eq. (4)], divide the sample space S of a stochastic scalar observation x into disjoint subspaces S_1 and S_2. This time, require at least r values of the sample to belong to S_2. The conjugated event is that $0, 1, 2, \ldots r-1$ values fall into category S_2. 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 $\binom{n}{k} \equiv n!/k!(n-k)!$. The sampling probability is thus given by

$$\beta \equiv \mathcal{P}(n_2 \geq r) = 1 - \sum_{k=0}^{r-1} \mathcal{P}(n_2 = k) = 1 - \sum_{k=0}^{r-1} \binom{n}{k} \alpha^{n-k}(1-\alpha)^k. \tag{6}$$

There is evidently a cost of additional sampling as the degree of truncation r is increased, since $\mathcal{P}(n_2 \geq r_1) < \mathcal{P}(n_2 \geq r_2)$, if $r_1 > r_2$. The robustness is however improved because more of the most extreme values are removed. Equation (6) generalizes Eq. (5) in Section 2.1. For $\alpha = 0.95 \leq \beta$ and $r = 1, 5, 10$, $n \geq 59, 181, 311$, respectively.

2.3 Sampling Density Function

The derivation by Wilks is here reproduced for a truncated sample range, but for one-sided instead of double-sided CIs. The sampling pdf $g(P)$ of the enclosed probability P can be found by first dividing the sample space S of a stochastic scalar observation x with pdf $f(x)$ into three disjoint subspaces S_1, dS, and S_2, where the infinitesimal interval $dS = dx_{n-r+1}$ contains $\tilde{x}_{n-r+1} = \hat{x}_\alpha$. Intervals S_1 and S_2 contain the remaining $n - r$ and $r - 1$ samples, respectively. The sampling probability of this configuration is infinitesimal,

$$d\mathcal{P} = P^{n-r}(-\infty, \tilde{x}_{n-r+1}) f(\tilde{x}_{n-r+1}) d\tilde{x}_{n-r+1} P^{r-1}(\tilde{x}_{n-r+1}, \infty). \tag{7}$$

The number of equivalent sets of sampled values is given by the same type of box counting experiment practiced in Sections 2.1 and 2.2. For the current three subspaces, the binomial generalizes to the multinomial coefficient,

$$N = \frac{n!}{(n-r)!1!(r-1)!}. \tag{8}$$

The random variable of interest is the enclosed probability P associated with this sample,

$$P \equiv P(-\infty, \tilde{x}_{n-r+1}), \rightarrow dP = f(\tilde{x}_{n-r+1}) d\tilde{x}_{n-r+1}. \tag{9}$$

Collecting, Eqs. (7)–(9) results in

$$g(P) = \frac{n!}{(n-r)!(r-1)!} P^{n-r}(1-P)^{r-1}. \tag{10}$$

This one-sided sampling pdf corresponds to the double-sided one derived by Wilks [7, Eq. (1)]. The similarity and difference of Eq. (10) to the well-known binomial distribution [6] is worth mentioning: The latter is discrete in a number related to r, while the former is continuous in P; $r - 1$ is also substituted with r in the binomial pdf. The sampling distribution in Eq. (6) can be verified by integration,

$$\beta = \mathcal{P}(P \geq \alpha) = \int_\alpha^1 g(P) dP = G(1) - G(\alpha), \tag{11}$$

where the primitive function $G(P)$ is found by repeatedly integrating Eq. 10 by parts,

$$G(P) = \sum_{k=0}^{r-1} \left(\begin{array}{c} n \\ k \end{array} \right) P^{n-k}(1-P)^k. \tag{12}$$

Inserting Eq. (12) into Eq. (11), Eq. (6) is obtained.

The sampling mean μ_P and sampling standard deviation σ_P of P can be evaluated using Eq. (10),

$$\begin{aligned} \mu_P &\equiv \langle P \rangle = 1 - \frac{r}{n+1}, \\ \sigma_P &\equiv \sqrt{\langle \delta^2 P \rangle} = \sqrt{\frac{r(n+1-r)}{(n+1)^2(n+2)}}. \end{aligned} \tag{13}$$

Their dependencies on r, β displayed in Fig. 1 are complicated since n has a complex variation with r, β, implicitly given by Eq. (6). Nevertheless, the graphs $\sigma_P(r)$ and $\mu_P(r)$ are amazingly similar for $\alpha = 0.95, 0.99$ (top, bottom),

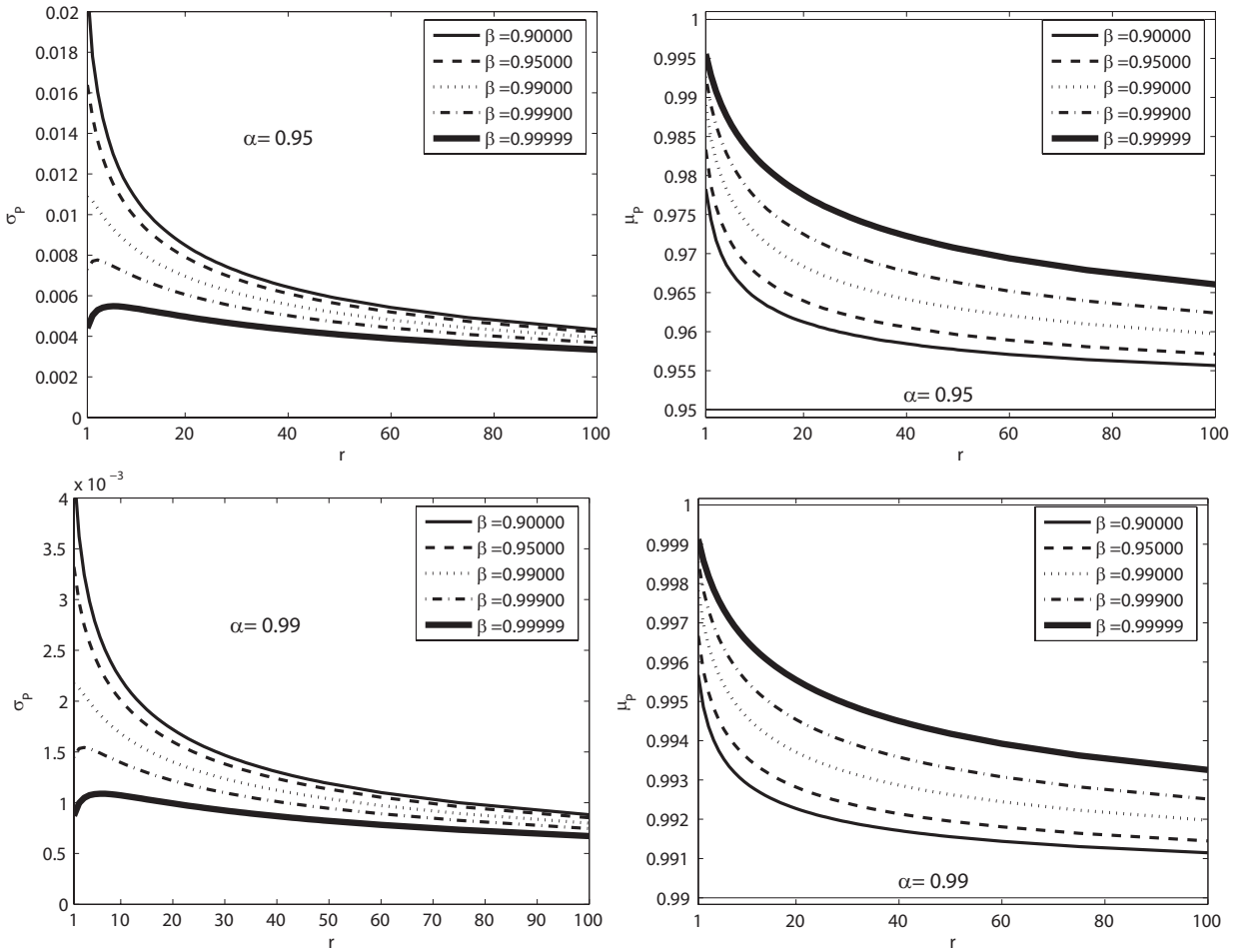

FIG. 1: Left: The sampling standard deviation σ_P (left) and sampling mean μ_P (right) of enclosed probability P [see Eq. (13)], for $\alpha = 0.95$ (top) and $\alpha = 0.99$ (bottom). The confidence level α is indicated for comparison (right, thin).

when scaled with the relevant range $1 - \alpha$. That is easily verified for the full sampling range ($r = 1$) for which there is an explicit expression for n (Section 2.1),

$$r = 1: \quad \mu_P = 1 - \sigma_P + R(1-\alpha)^2, \;\; \sigma_P = \frac{1-\alpha}{-\log(1-\beta)} + R(1-\alpha)^2, \tag{14}$$

where $R(x)$ labels rest terms of order x. Apparently there is a sum rule limited to $r = 1$, $\mu_P + \sigma_P = 1 + R(1-\alpha)^2$, for all values of β, which results in a significant skew for $g(P)$ since $P \leq 1$. Most importantly, the standard deviation σ_P appears to saturate at a finite non-zero level in the limit of very high β and r. Hence, the pdf $g(P)$ cannot be made arbitrarily narrow by selecting sufficiently large values of r, β, as first might be expected. This will have direct consequences for the robustness addressed in Section 3.

3. ROBUSTNESS

A general expression for the robustness, as defined by the confidence interval width of Wilks' estimate, is derived in Section 3.1. Examples of its magnitude are illustrated in Section 3.2. In addition, further insight into the issue of robustness is provided by relating to classical hypothesis testing in Section 3.3.

3.1 General

While the stability refers to the variability of the enclosed probability $P(-\infty, \hat{x}_{[\alpha]} = \tilde{x}_{n-r+1})$, the robustness describes that of $\hat{x}_{[\alpha]}$. Since the former is a function of the latter, the sampling or robustness pdf (h) is obtained from the stability pdf (g) by a change of variable,

$$G(P) = \int g(P)dP = \int g(P(\hat{x}_{[\alpha]}))\frac{dP}{d\hat{x}_{[\alpha]}}d\hat{x}_{[\alpha]} = \int g(P(\hat{x}_{[\alpha]}))f(\hat{x}_{[\alpha]})d\hat{x}_{[\alpha]} \equiv \int h(\hat{x}_{[\alpha]})d\hat{x}_{[\alpha]}. \tag{15}$$

Expressed in the pdf $f(x)$ of model results x,

$$h(\hat{x}^{(\pm)}_{[\alpha]}) = g\left(\pm\int_{\mp\infty}^{\hat{x}^{(\pm)}_{[\alpha]}} f(x)dx\right)f(\hat{x}_{[\alpha]}) = \frac{dG(P(\hat{x}^{(\pm)}_{[\alpha]}))}{d\hat{x}_{[\alpha]}}. \tag{16}$$

The estimated confidence limit $\hat{x}^{(s_\alpha)}_{[\alpha]}$ is a random quantity. Preferably, it is transformed into a *sampled scaled coverage factor* $q^{(s_\alpha)}_{[\alpha]}$, normalized to the true coverage factor $k^{(s_\alpha)}_{[\alpha]}$,

$$q^{(s_\alpha)}_{[\alpha]} \equiv s_\alpha\frac{\hat{x}^{(s_\alpha)}_{[\alpha]} - \mu_x}{k^{(s_\alpha)}_{[\alpha]}\sigma_x}, \tag{17}$$

where μ_x and σ_x are the true mean and standard deviation, respectively, of the model results x. Due to the normalization, perfect estimation corresponds to $q^{(s_\alpha)}_{[\alpha]} = 1$. Consequently, the estimate is conservative if $q^{(s_\alpha)}_{[\alpha]} > 1$, while it is invalid if $q^{(s_\alpha)}_{[\alpha]} < 1$. Its *normalized sampling confidence interval* $[q^{-(s_\alpha)}_{[\gamma,\alpha]}, q^{+(s_\alpha)}_{[\gamma,\alpha]}]$ for level γ is the equivalent of sampling variance of estimation [13] for confidence intervals. The limits $q^{s_\gamma(s_\alpha)}_{[\gamma,\alpha]}$ also reflect the recursive symmetry of evaluating a coverage factor of a coverage factor, with comparable meanings of confidence levels α and γ, as well as the signs $s_\alpha, s_\gamma = \pm 1$ indicate the respective upper and lower bounds. This suggests a universal measure $\Delta^{(s_\alpha)}_{[\gamma,\alpha]}$ of relative sampling variability,

$$\Delta^{(s_\alpha)}_{[\gamma,\alpha]} \equiv q^{+(s_\alpha)}_{[\gamma,\alpha]} - q^{-(s_\alpha)}_{[\gamma,\alpha]} : \quad \mathcal{P}\left(q^{(s_\alpha)}_{[\alpha]} \in \left[q^{-(s_\alpha)}_{[\gamma,\alpha]}, q^{+(s_\alpha)}_{[\gamma,\alpha]}\right]\right) = \gamma. \tag{18}$$

This unit-less width expresses the *robustness*, or precision[2] as defined by γ, of $q^{(s_\alpha)}_{[\alpha]}$ at level α. It is implicitly dependent on the stability β relating to the degree of *conservatism* of estimation accuracy, and the truncation r describing the *utilization* of the sampled model values. As defined, high robustness corresponds to $\Delta^{(s_\alpha)}_{[\gamma,\alpha]} \ll 1$.

The probabilistically symmetric double-sided CI $[q^{-}_{[\gamma,\alpha]}, q^{+}_{[\gamma,\alpha]}]$ must be determined implicitly,

$$G\left(P = \int_{-\infty}^{\mu_x \pm q^{\pm}_{[\gamma,\alpha]}k_{[\alpha]}\sigma_x} f(x)dx, r, \alpha, \beta\right) \quad = \quad \frac{1 \pm \gamma}{2}, \tag{19}$$

where $G(P, r, \alpha, \beta)$ given by Eq. (12) is indirectly dependent on r, α, β via $n(r, \alpha, \beta)$ and r according to Eq. (6).

For the bound of the corresponding one-sided interval, γ becomes equivalent to β. Up to a round-off error due to the limitation of integer values of n, r, the relation $G(P(\hat{x}^{(s_\alpha)}_{[\alpha]}), r, \alpha, \beta) = \beta$ is already fulfilled by the choice of sample size n [Eq. (6)], since this is the constraint of conservative estimation. That does not imply that the lack of robustness expressed by the width $\Delta_{[\gamma,\alpha]}$ of the double-sided CI is unimportant. A finite interval but not a single bound can illustrate the volatility, or reproducibility of the estimate.

[2]The distinction between precision and accuracy is frequently emphasized, see, e.g., [14, pp. 9–14].

3.2 Illustration

The robustness $\Delta^{(s_\alpha)}_{[\gamma,\alpha]}$ defined by Eq. (18) varies strongly with the pdf $f(x)$ of model results x. The key aspect is whether or not it is possible to achieve acceptable robustness, by setting the truncation r appropriately and require enough stability β.

As seen in Fig. 2 the robustness may be very low since the width of the pdf $h(q_{[0.95]})$ is large for the common normal distribution (top). It is however significantly higher for the uniform distribution (bottom). If the stability β is increased, $h(q_{[0.95]})$ is shifted upward to reduce the risk of underestimation. The truncation r acts in the opposite way, as it filters out the most extreme samples. To improve robustness, further truncation (r) is more effective than increasing the stability (β). The computational cost of additional sampling for extreme stability and truncation appears much higher than the gain (right). That is a direct consequence of the saturation of the width of $g(P)$ on a finite level, as also illustrated with the standard deviation σ_P in Fig. 1 (left). The fundamental problem of low robustness is apparently difficult to resolve by more extensive sampling.

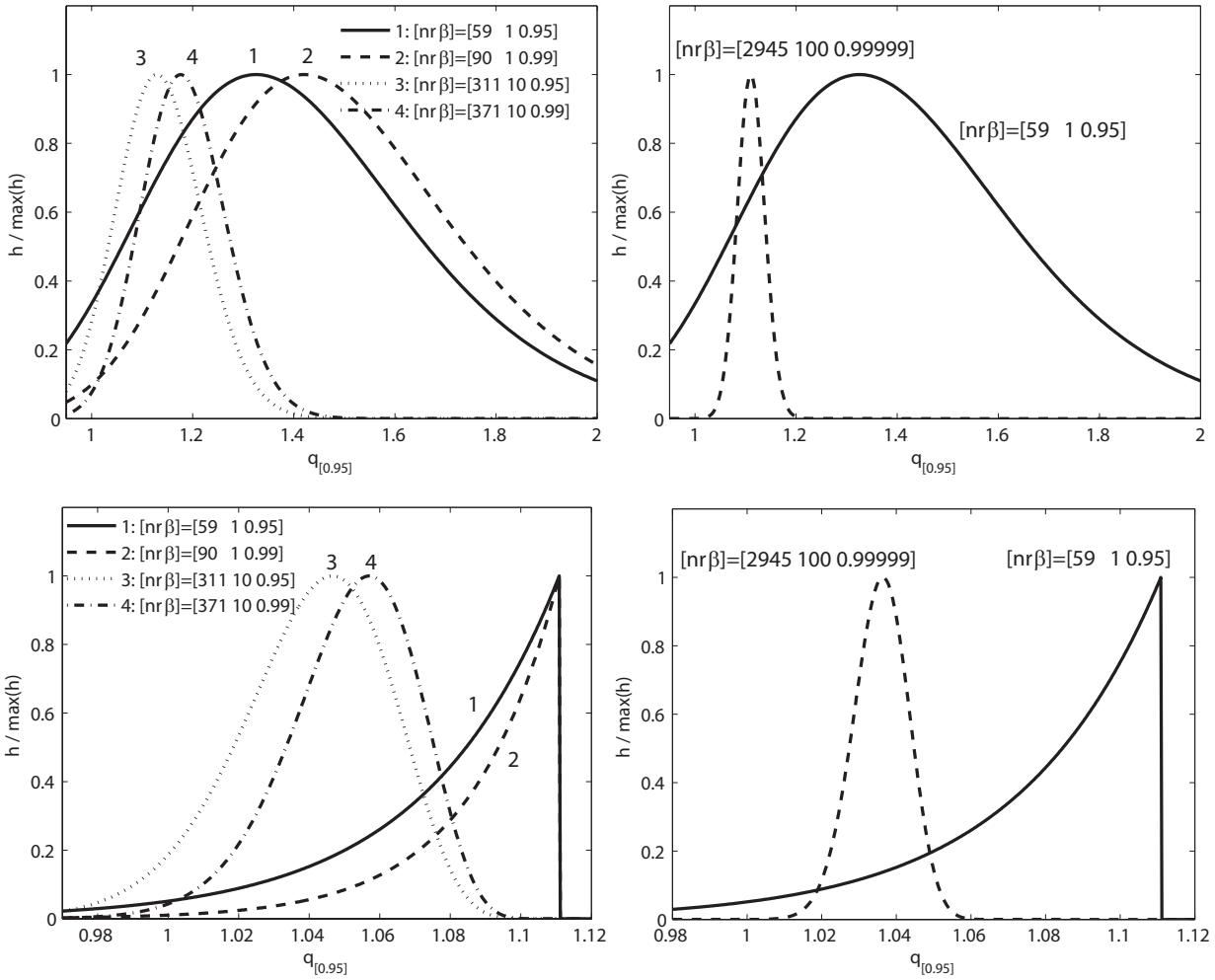

FIG. 2: Left: The robustness illustrated by the pdf $h(q_{[0.95]})$ of the scaled estimated one-sided upper coverage factor $q_{[0.95]}$ [see Eq. (17)] of model results x, for normal (top) and uniform (bottom) pdfs $f(x)$. It is evaluated for various values of stability β and truncation r, resulting in different minimum sample size n. Right: The limit of very large truncation and extreme stability.

Increasing the confidence level α, the least required sample size n will increase according to Eq. (6). The pdfs $h(q_{[0.99]})$ corresponding to Fig. 2 are shown in Fig. 3. Despite the indirect complex dependence on α via n, $h(q_{[\alpha]})$ is very well maintained after shifting and scaling according to Eqs. (13). This agrees with the observed scaling of μ_P and σ_P in Fig. 1 (top vs. bottom) and Eq. (14). Numerical values of the CI $([q^-_{0.95,\alpha}, q^+_{0.95,\alpha}]$ of robustness and minimal sample sizes n are summarized in Table A.1. For the common choice of $\alpha = 0.95 \leq \beta$ (Section 2.1), a variation of the estimated bound $q_{[0.95]}$ with no less than a factor of 2 is possible: For the normal distribution $([q^-_{[0.95,0.95]}, q^+_{[0.95,0.95]}] = [0.94, 2.03]$, while $([q^-_{[0.95,0.95]}, q^+_{[0.95,0.95]}] = [0.98, 1.11]$ for the uniform distribution. That is consistent with the general statement that more confined distributions with larger aspect ratios λ yield more robust estimates $q_{[\alpha]}$. The dependence is clearly very strong. The generic problem of low robustness caused by the combination of faithful estimation of CIs and random sampling is apparently difficult to resolve for distributions with low aspect ratios λ.

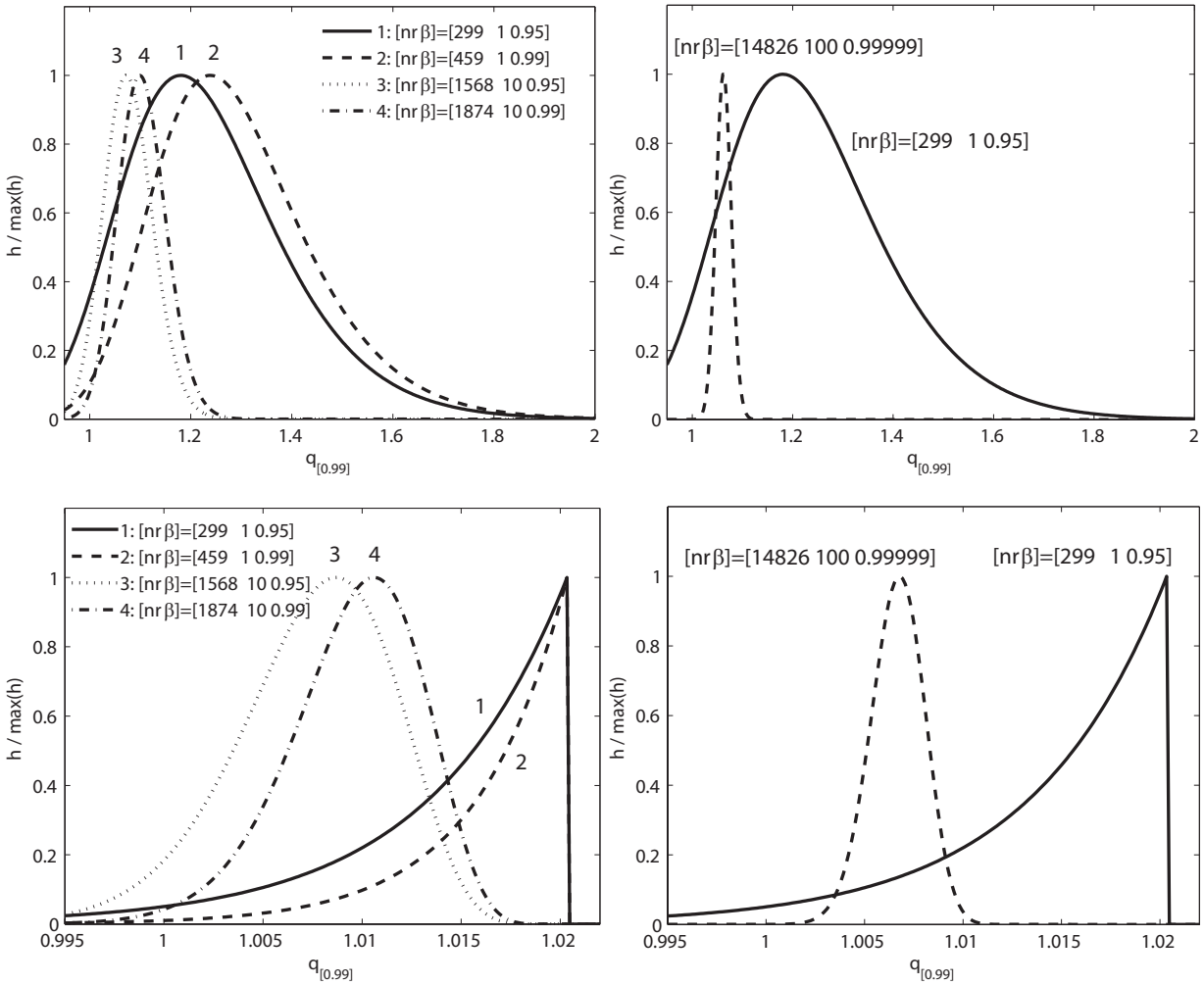

FIG. 3: Left: The robustness illustrated by the pdf $h(q_{[0.99]})$ of the scaled estimated one-sided upper coverage factor $q_{[0.99]}$ [see Eq. (17)] of model results x, for normal (top) and uniform (bottom) pdfs $f(x)$. It is evaluated for various values of stability β and truncation r, resulting in different minimum sample size n. Right: The limit of very large truncation and extreme stability.

3.3 Relation to Statistical Hypothesis Testing

On first sight, the problem of estimation and robustness of $x_{[\alpha]}$ appears to be closely related to classical hypothesis testing [6]. A sample of n values is drawn from the population of model results x. A characteristic of the infinite population is then inferred from statistics of the finite sample. Specifically, the probability $1-\beta$ of obtaining a non-conservative, i.e., failing estimate of the CI appears related to the so-called p value. It expresses the probability of obtaining an estimate as extreme as the actual, given a correct null hypothesis, and is the principal tool of rejection.

On second thought, there are distinct differences. The population mean μ is often the main attribute and the sample mean $\hat{\mu}$ the obvious algebraic statistic to be used for testing. Here it is instead the confidence interval limit $x_{[\alpha]}$ that is to be estimated, and Wilks' method provides an *order statistic*. Since the latter is a consequence of ordering, rather than algebraic summation as for the mean, the relation between the distributions of x and $\hat{x}_{[\alpha]}$ is much less obvious than between x and $\hat{\mu}$. Therefore, the standard deviation $\mathrm{std}(\hat{x}_{[\alpha]})$ over different samples cannot be determined from $\mathrm{std}(x)$ for a single sample. The standard procedure to directly relate intersample statistics to intrasample statistics, such as $\mathrm{std}(\hat{\mu}) = \mathrm{std}(x)/\sqrt{n}$, thus cannot be applied.

However, for $\hat{x}_{[\alpha]}$ the entire distribution $h(\hat{x}_{[\alpha]})$ in Eq. (16) has been derived from first principles. It is thus possible to calculate (not estimate) $\mathrm{std}(\hat{x}_{[\alpha]})$ explicitly, for the cost of deriving h. The standard deviation is thus known, provided $f(x)$ is correctly assigned. Similar assumptions of distributions are indeed required in conventional testing.

Generally, $h(\hat{x}_{[\alpha]})$ is not normal [see Figs. 2 and 3 of $h(q)$ above, and note that the affine mapping $\hat{x}_{[\alpha]} \to q$ in Eq. (17) only rescales and translates h]. This violates conventional procedures of testing, since the sample statistic usually is assumed normal distributed when the standard deviation is known.

The central limit theorem states that the distribution of the sample mean tends to normal, as the sample size increases. That is indeed a highly useful fact in conventional testing based on sample mean, as it makes assumptions of normal distribution legitimate. It is not applicable for Wilks' estimate based on ordering of sample values. Little can be said about $h(\hat{x}_{[\alpha]})$, except than it can be calculated and strongly depends on $f(x)$.

Formulated in terms of classical statistical testing, the null hypothesis should relate to the quantity of interest, i.e., the true CI of the model, $x_{[\alpha]} \leq \tilde{x}_{n-r+1}$ for an upper and $x_{[\alpha]} \geq \tilde{x}_{n-r+1}$ for a lower bound, respectively. Note that the statistical model represents the truth, with given certain CI limits. In classical testing, an appropriate statistic would be evaluated to determine the p value from the distribution associated with that statistic. The inverse problem here is to set $p = 1-\beta$ and not determine any corresponding one-sided statistic but the values of n, r which results in that p, not the reverse. This procedure is closely related to conventional determination of CIs using hypothesis testing, except that there is no statistic involved. Instead there is an explicit non-trivial relation to the CI limits. The least value of n, for given r, would then correspond to the highest allowed p value (in some literature denoted α) for rejecting the null hypothesis.

4. CREDIBLE APPLICATION OF WILKS' METHOD

In this section modifications of Wilks' approach are presented. The principal aspect is that the uncertainty of the estimate reduces its reliability well below the confidence level α. The acceptable margin α has in fact already been 'consumed' by the true bound of the model, which leaves nothing for its estimation: If the probability of obtaining a conservative estimate is β and the quantity itself describes a probability α of capturing the most extreme scenario, the total probability of making a conservative prediction will be at least $\tilde{\alpha} = \beta\alpha$. To distinguish it from the confidence level α, $\tilde{\alpha}$ will be labeled *credible level*.

4.1 Credibility

The credible level aggregates two sources of failing assessments, the uncertainty of model results and our limited ability of evaluating this uncertainty from a random sample. As in Bayesian estimation [15], the idea is to incorporate all known sources of uncertainty and describe the state of knowledge of the observer, rather than a physical variability. There is indeed also a mathematical similarity with Bayes' theorem: The overall goal is to guarantee, with a certain

probability (credibility $\tilde{\alpha}$), that the true but unknown physical result x is less (larger) than, or equal to an estimate $\hat{x}_{[\alpha]}$ of the true upper (lower) bound $x_{[\alpha]}$ of the model with confidence α,

$$\begin{aligned}\mathcal{P}\left(x \leq (\geq)\hat{x}_{[\alpha]}\right) &\geq \mathcal{P}\left(x \leq (\geq)\hat{x}_{[\alpha]} | \hat{x}_{[\alpha]} \geq (\leq) x_{[\alpha]}\right) \cdot \mathcal{P}(\hat{x}_{[\alpha]} \geq (\leq) x_{[\alpha]}) \\ &\geq \mathcal{P}(x \leq (\geq) x_{[\alpha]}) \cdot \mathcal{P}(\hat{x}_{[\alpha]} \geq (\leq) x_{[\alpha]}) = \alpha\beta \equiv \tilde{\alpha}. \end{aligned} \tag{20}$$

The first inequality stems from the fact that the possibility $\hat{x}_{[\alpha]} < (>) x_{[\alpha]}$ is excluded, while the second results from a reduction of the interval. The probability of making a correct assessment is thus *at least* $\tilde{\alpha}$, not necessarily or even likely equality. Several layers of conservatism are thus embedded in $\tilde{\alpha}$. That is acceptable in critical applications where a failing prediction often spells disaster. The second expression also resembles Bayes' theorem for the posterior probability, being a product of likelihood and prior information. Estimating $\hat{x}_{[\alpha]}$ is indeed prior to the targeted comparison between modeled and physical results. However, the similarity merely reflects that both analyze sequential events with conditional probabilities.

The renormalized probability of false prediction $1 - \tilde{\alpha}$ is close to twice as large as $1 - \alpha$ for $\alpha = \beta$. For instance, when estimating an $\alpha = 0.95$ confidence limit we obtain a $\tilde{\alpha} = 0.95^2 \approx 0.90$ credible bound. The assigned confidence levels of modeling (α) and estimation (β) must both *always* be set larger than the desired credible level ($\tilde{\alpha}$), $\alpha, \beta > \tilde{\alpha}$, e.g. $\beta = \alpha = \sqrt{0.95} \approx 0.975$ results in a credible level of 0.95. Since the larger value of α, β will increase the least number of samples (n), accounting for the limited robustness will require additional sampling.

4.2 Modifications

Following Wilks' approach as it is currently practiced (Section 2) there are at least three possible modifications of interpretation and/or application, which translate an acceptable confidence level into a credible level describing the probability of making false predictions. These adjustments will not alter the method as such, but change the interpretation (1), add safety margins (2), or modify the acceptance criterion (3):

1. Accept the elevated probability $1 - \hat{\alpha} = 1 - \alpha\beta$ of failing bounds.

2. Apply an additional safety margin $w_{[\alpha]}$ to Wilks' estimate $\hat{x}_{[\alpha]}$ to compensate for its limited robustness [see Eq. (21) below]. Since the least number of model evaluations n then is kept, the obvious advantage is that existing results easily can be adjusted without additional model evaluations. The disadvantages are that $w_{[\alpha]}$ depends strongly on the pdf $f(x)$ of model results and that $\tilde{\alpha} < \alpha$ makes it impossible to strictly maintain the officially accepted probability of failure. The ambition to restore that risk must be held back, since $\tilde{\alpha} \to \alpha$ generally implies $w_{[\alpha]} \to \infty$, if $f(x)$ does not have compact support.

3. Adjustment of α and β. That will increase the least number of samples n substantially. The advantages are that the result will apply for any pdf $f(x)$ of model results, and no extra safety factor needs to be included (as in 2 above). The obvious disadvantages are that the efficiency is drastically reduced and existing results cannot be recycled and needs to be complemented with further sampling of the model. If not all of the $r - 1$ excluded most extreme model results are known, full re-evaluation is required.

In alternative 2, the margin $w_{[\gamma]}$ is calculated for a different stability level $\gamma > \beta$, such that $\mathcal{P}(\hat{x}_{[\alpha]} \geq (\leq) x_{[\alpha]}) = \gamma$, from $h(\hat{x}_{[\alpha]})$ instead of increasing n in Eq. (6). To achieve $\tilde{\alpha} \approx \alpha$, $1 - \gamma \ll 1 - \alpha$. A reasonable compromise might be to set $\gamma = 1 - (1 - \alpha)/5$, or $\gamma = 1 - (1 - \alpha)/10$ (practiced in Table B.1 and B.2). The estimated bound is adjusted as

$$\hat{x}_{[\alpha]} \to \hat{\mu} + w_{[\gamma,\alpha,\beta,r]} \cdot \left(\hat{x}_{[\alpha]} - \hat{\mu}\right), \; w_{[\gamma,\alpha,\beta,r]} = \frac{1}{q_{[\gamma,\alpha]}}, \tag{21}$$

where additional subscripts of w indicate dependencies and $\hat{\mu}$ is the sample mean, or equivalent best estimate of model results. Examples of correction factors $w_{[\gamma,\alpha]}$ (fixed n) are given in Table B.1 (alt. 2), while the enlarged sample sizes are compared to the original in Table B.2 (alt. 3). Note that for a given credible level $\tilde{\alpha}$ some combinations of β, α are more efficient (lower n) than others. It should come as no surprise that it may be more profitable to distribute our margin of failure between α and β in one specific way, than any other.

5. ALTERNATIVE DETERMINISTIC METHODS

There is one principal distinction between applications such as the one addressed by Wilks, and the evaluation of modeling uncertainty considered here. Direct physical sampling as practiced by Wilks makes no reference to the sources of uncertainty whatsoever. It is drastically different for modeling uncertainty, which results from a known uncertain model. Such models provide detailed information of the sources of uncertainty. While Wilks addressed the *statistical* problem of analyzing a finite set of randomly drawn samples, the evaluation of modeling uncertainty constitutes the *deterministic* problem of *uncertainty propagation*. In the latter but not the former case, an exact result can in principle be found. Consequently, Wilks' method has here been applied to solve a problem it was *not* primarily designed for.

The existence and utilization of knowledge are key ingredients for the quality of any calculation. Specifically, uncertainty propagation but not physical sampling *requires* knowledge of a model structure and input uncertainties. Applying Wilks' method only a minor part of this information is explored by chance. A much better method would be to *systematically* explore all relevant pieces of information, as efficiently as possible. Sampling may still be used, but with all sample points calculated deterministically, using best available sampling *rules*. From the completely general perspective that the more and the better utilization of information the higher quality of the result, we expect that fully *deterministic sampling* (DS) is superior to random sampling (RS). The hallmark of DS is that repeated sampling yields identical samples without any variation between corresponding sample values, and represents the whole statistical population from which any RS sample is drawn. Thus in contrast to RS, DS does not suffer from sampling variance [6] of any kind.

Indeed, partial DS [16, Fig. 2, p. 61] is a common way to improve brute force RS. Any kind of stratification is a deterministic operation. In Latin hypercube sampling techniques (LHS) [17], stratification is combined with exclusion criteria (any stratum of any parameter should be sampled once and only once). Hence, LHS contains two deterministic operations, stratification and exclusion. Orthogonal sampling [18] makes use of an additional second layer of stratification, with subspaces containing several strata. As the degree of determinism increases, the distribution of samples improves, which reduces the sampling variance and consequently increases the efficiency. It is plausible to assume that the quality of the result continues to improve as the degree of determinism increase further, providing the rules are good. As long as no apparent limit is known beyond which further determinism is not profitable, it is quite plausible that the best sampling strategy is entirely deterministic. A well-known example of such a method is the unscented Kalman filter [19].

The issue of robustness is avoided entirely with DS. Estimated bounds will be perfectly repeatable and reproducible. Excessive noise systematically fed into the analysis by the process of randomized sampling is avoided. Effectively, a finite sampling variance is substituted with a finite sampling error. Provided that error can be controlled, such an approach is superior for regulatory authorities since any specific modeling task can be repeated by anyone with identical results.

An extension to Wilks' approach beyond the scope of our current study is sequential processing of information by means of Bayesian analysis. It can be evaluated with RS [3] as well as DS [4] methods. Especially for small samples, Bayesian methods are often superior provided assumptions of probability distributions hold.

By setting acceptance limits according to expected errors as in Section 4.2, the risk of failing risk assessment should be manageable. An elevated safety margin is thereby traded for fairness. For comparisons of different calculations or models, a minor common systematic error of evaluation is much easier to accept than a large sampling variance, which arbitrarily and unpredictably penalizes some calculations more than others.

In the original application of Wilks' method for evaluating tolerance limits, the independence of the pdf $f(x)$ is a major advantage since it provides immunity against false assignments of input statistics. For evaluating modeling uncertainty though, the samples are generated from an assigned distribution. The result is therefore nevertheless strongly dependent on $f(x)$, which lifts the immunity completely. Then there is no generic preference of generality for Wilks' method, compared to other methods of uncertainty propagation requiring explicit assignment of input statistics.

As many methods based on RS, Wilks' does not suffer from the curse of dimensionality, which is the main obstacle for sampling on large grids or integration by conventional quadrature methods. The proposed efficient methods of

deterministic sampling [4, 12, 20] have comparatively mild dependencies on dimensionality, due to very sparse sampling and low or moderate requirements on the fidelity of surrogate models. Often, the number of samples increases linearly with dimensionality. There is also a natural limit of how many uncertain parameters that may contribute substantially to the resulting modeling uncertainty. The advantage of Wilks' method of being essentially independent on dimensionality is thus relatively insignificant in practice.

The uncertainty can be deterministically propagated through any monotonic non-linear univariate model $y(\theta)$ without any error by evaluating the model at the confidence limits, $[y_-, y_+] = [\min_\pm y(\theta_\pm), \max_\pm y(\theta_\pm)]$. For multi-variate models with j parameters, the confidence limits are generalized into confidence boundaries (CB) in parameter space of dimension $j-1$. There is one CB for one-sided and two CBs for double-sided intervals. For many parameters, the topology is non-trivial and it is generally impossible to evaluate $[y_-, y_+]$ exactly, but often with remarkable accuracy. The only error sources are the assignment of the pdf $f(x)$, which in the present context is unavoidable, and the error of the method to determine the CB(-s). Recently, we proposed such a novel approach based on DS on CBs [20]. In that case, the CBs were determined with surrogate models found by linear regression. The main disadvantage of the methodology is that the CBs are unique for each point of the result. For instance, for a signal evaluated in 100 time instants no less than 200 CBs must be determined. However, for integral quantities like total energy, there will only be one (one-sided) or two (double-sided) CBs. The number of model samples required to find a linear surrogate model and sample on the CB(-s) is not larger than $j+3$. For a typical number of relevant parameters $j=10$, only 12 model evaluations are needed for one-sided CIs. Besides absolute robustness the efficiency is in this case about five times better than the default application of Wilks' method today (described in the end of Section 2.1).

For evaluating CIs of high-dimensional field quantities calculated with typical computational fluid dynamics, electro-magnetic, or structural mechanics models, it is more efficient and convenient to propagate parameter statistics than sample on CBs. Any CI will then have to be calculated non-faithfully by expansion with coverage factors, as mentioned in Section 1. That inherent limitation of accuracy is of minor importance for two reasons. Firstly, the distribution of model parameters is generically multi-variate. Properly identified models will, almost without exception, exhibit relatively strong dependencies. Such complex pdfs will hardly ever be accurately known beyond the lowest statistical moments, typically the first (mean) and second (covariance). Anything close to faithful evaluation of CIs is then genuinely impossible because of a fundamental lack of information. Secondly, non-faithful evaluation of CIs is an accepted and established *de facto* standard practiced in many fields of science and technology. The knowledge of coverage factors is most often exceedingly vague. The only motivation to report CIs and not low-order statistics appears to be that CIs often are considered more "understandable," than for instance a conservative bound scaled with standard deviation (σ), like 2σ or 3σ. Non-faithful CIs should be regarded as what they are, hypothetically expanded statistics with no measure of reliability comparable to stability β in Wilks' method.

There are numerous deterministic methods for non-linear propagation of model statistics [12, 19], which could provide viable alternatives to Wilks' method. Complex analysis such as identification of models and evaluation of reliability comparable to β is of major importance and supported by DS [4]. The methodology is novel and currently in development and has not yet been widely employed. A complete survey of selected deterministic alternatives to Wilks' method is beyond the scope of the present study and will be presented elsewhere.

6. CONCLUSIONS

Wilks' approach for conservative estimates may be simple and practical but low robustness makes them volatile measures of modeling uncertainty. Our study illustrates that the difference between a one-sided confidence bound estimated with Wilks' method and the expected result may be no less than twice as large as the true one, equivalent to about 100% relative uncertainty. That may cause problems in the evaluation of different competing models.

The potentially low robustness of Wilks' methodology is caused by the combination of faithful evaluation of CIs and slowly converging random sampling. To remedy this deficiency, a novel perspective on Wilks' approach was proposed. It focuses on the composite set of comparisons, the estimated with the true bound of the model, and the

model bound with the physical value. Combining the probability of success of both, the true physical result is covered by the estimate with a probability at least as large as its so-called *credibility level*.

Expected levels of robustness for conventional application of Wilks' method were given for some common cases. Robustness can only be improved to a certain extent by excessive sampling. When the robustness of Wilks' method is unsatisfactory, our methods of choice belongs to the novel class of highly efficient deterministic uncertainty propagation techniques. These are completely robust, but may suffer from systematic errors. A brief survey was included.

ACKNOWLEDGMENT

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

REFERENCES

1. Marples, D. R., *The Social Impact of the Chernobyl Disaster*, St Martin's, New York, 1988.
2. Ljung, L., *System Identification: Theory for the User*, 2 ed., Prentice Hall, Englewood Cliffs, NJ, 1999.
3. Kennedy, M. C. and O'Hagan, A., Bayesian calibration of computer models, *J. R. Stat. Soc. Series B, Stat. Methodol.*, 63(3):425–464, 2001.
4. Hessling, J. P., Identification of complex models, *SIAM/ASA J. Uncertainty Quantification*, 2(1):717–744, 2014.
5. ISO GUM, Guide to the Expression of Uncertainty in Measurement, Tech. Rep., International Organisation for Standardisation, Geneva, Switzerland, 1995.
6. Bennett, J. and Fisher, R. A., *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford University Press, Oxford, 1995.
7. Wilks, S. S., Determination of sample sizes for setting tolerance limits, *Ann. Math. Stat.*, 12(1):91–96, 1941.
8. Wald, A., An extension of Wilks' method for setting tolerance limits, *Ann. Math. Stat.*, 14(1):45–55, 1943.
9. Radhakrishna, S., Murthy, B., Nair, N., Jayabal, P., and Jayasri, R., Confidence intervals in medical research, *Indian J. Med. Res.*, 96:199–205, 1992.
10. Braga, P. L., Oliveira, A. L., and Meira, S. R., Software effort estimation using machine learning techniques with robust confidence intervals, in Tools with Artificial Intelligence, *19th IEEE Int. Conf.*, Patras, Greece, 29–31 Oct., Vol. 1, pp. 181–185, 2007.
11. Le Boudec, J.-Y., *Performance Evaluation of Computer and Communication Systems*, EPFL Press, Lausanne, Switzerland, 2010. (accessed 2015-12-03; http://perfeval.epfl.ch)
12. Hessling, J. P., Deterministic sampling for propagating model covariance, *SIAM/ASA J. Uncertainty Quantification*, 1(1):297–318, 2013.
13. Kay, S., *Fundamentals of Statistical Signal Processing, Estimation Theory*, Vol. 1, Prentice Hall, Englewood Cliffs, NJ, 1993.
14. VIM, ISO, International vocabulary of basic and general terms in metrology (VIM), Tech. Rep., International Organisation for Standardisation, Geneva, Switzerland, 2004.
15. Sivia, D. and Skilling, J., *Data Analysis—A Bayesian Tutorial*, Oxford University Press, Oxford, UK, 2006.
16. Hessling, J. P., *Digital Filters and Signal Processing*, Chapter *Deterministic Sampling for Quantification of Modeling Uncertainty of Signals*, pp. 53–79, INTECH, Rijeka, Croatia, 2013.
17. Helton, J. and Davis, F., Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, *Reliab. Eng. Syst. Safety*, 81:23–69, 2003.
18. Tang, B., Orthogonal array-based latin hypercubes, *J. Am. Stat. Assoc.*, 88(424):1392–1397, 1993.
19. Julier, S. and Uhlmann, J., Unscented filtering and nonlinear estimation, in *Proc. IEEE*, 92:401–422, 2004.
20. Hessling, J. P. and Svensson, T., Propagation of uncertainty by sampling on confidence boundaries, *Int. J. Uncertainty Quantification*, 3(5):421–444, 2013.

APPENDIX A. TYPICAL VALUES OF ROBUSTNESS AND SAMPLE SIZE

Expected levels of robustness ($\Delta_{[\gamma,\alpha]}$) and efficiency (n) of Wilks' conservative estimate of one-side CIs are indicated in Table A.1, for a selection of cases. Coverage factors are normalized to their true values to clearly display the relative errors.

TABLE A.1: Least required sample size n, the sampling interval $[q^-_{[0.95,\alpha]}, q^+_{[0.95,\alpha]}]$ and the robustness $\Delta_{[0.95,\alpha]}$, defined in Eq. (18). The evaluation is made for different pdfs $f(x)$ (NRM: Normal, UNI: Uniform) with aspect ratio λ defined in Section 1, levels of truncation r, stability β, and confidence levels α of modeling (x)

$f(x)$	λ	r	β	α	n	$[q^-_{[0.95,\alpha]}, q^+_{[0.95,\alpha]}]$	$\Delta_{[0.95,\alpha]}$
NRM	0.15	1	0.95	0.95	59	$[0.94, 2.03]$	1.08
NRM	0.15	1	0.99	0.95	90	$[1.06, 2.10]$	1.03
NRM	0.15	1	0.95	0.99	299	$[0.97, 1.62]$	0.65
NRM	0.15	1	0.99	0.99	459	$[1.04, 1.66]$	0.63
NRM	0.15	5	0.95	0.95	181	$[0.97, 1.44]$	0.47
NRM	0.15	5	0.99	0.95	229	$[1.04, 1.49]$	0.45
NRM	0.15	5	0.95	0.99	913	$[0.98, 1.25]$	0.27
NRM	0.15	5	0.99	0.99	1157	$[1.02, 1.28]$	0.26
NRM	0.15	10	0.95	0.95	311	$[0.98, 1.31]$	0.34
NRM	0.15	10	0.99	0.95	371	$[1.03, 1.35]$	0.33
NRM	0.15	10	0.95	0.99	1568	$[0.99, 1.18]$	0.19
NRM	0.15	10	0.99	0.99	1874	$[1.02, 1.20]$	0.19
UNI	1.00	1	0.95	0.95	59	$[0.98, 1.11]$	0.13
UNI	1.00	1	0.99	0.95	90	$[1.02, 1.11]$	0.09
UNI	1.00	1	0.95	0.99	299	$[1.00, 1.02]$	0.02
UNI	1.00	1	0.99	0.99	459	$[1.00, 1.02]$	0.02
UNI	1.00	5	0.95	0.95	181	$[0.99, 1.09]$	0.10
UNI	1.00	5	0.99	0.95	229	$[1.01, 1.10]$	0.08
UNI	1.00	5	0.95	0.99	913	$[1.00, 1.02]$	0.02
UNI	1.00	5	0.99	0.99	1157	$[1.00, 1.02]$	0.02
UNI	1.00	10	0.95	0.95	311	$[0.99, 1.08]$	0.09
UNI	1.00	10	0.99	0.95	371	$[1.01, 1.08]$	0.07
UNI	1.00	10	0.95	0.99	1568	$[1.00, 1.01]$	0.02
UNI	1.00	10	0.99	0.99	1874	$[1.00, 1.02]$	0.01

APPENDIX B. CREDIBLE BOUNDS

The adjustment factors $w_{[\alpha]}$ for sub-sampling of the model are indicated in Table B.1, while the enlarged sample sizes n for full sampling are listed in Table B.2, for a selection of cases. For comparison, numbers of the corresponding current application of Wilks' method are included, where confidence levels play the role of credible levels.

TABLE B.1: Adjustment factor $w_{[\alpha]}$ for credible application of Wilks' method [Eq. (21)], for letting the credible level $\tilde{\alpha} = \alpha \cdot \gamma$ approach the accepted confidence level with the proposed choice $\gamma^{(k)} \equiv 1 - (1 - \alpha)/k$. The evaluation is made for the different pdfs $f(x)$ in Table A.1 (NRM, UNI), levels of truncation r, stability β, and confidence levels $\alpha : \alpha = \beta$. The least required sample size n is then preserved

r	γ	α	$\tilde{\alpha}$	$w_{[\alpha]}$, NRM	$w_{[\alpha]}$, UNI
1	0.950	0.950	0.902	1.00	1.00
1	$0.990 = \gamma^{(5)}$	0.950	0.940	1.14	1.06
1	$0.995 = \gamma^{(10)}$	0.950	0.945	1.20	1.09
1	0.990	0.990	0.980	1.00	1.00
1	$0.998 = \gamma^{(5)}$	0.990	0.988	1.05	1.01
1	$0.999 = \gamma^{(10)}$	0.990	0.989	1.07	1.01
5	0.950	0.950	0.902	1.00	1.00
5	$0.990 = \gamma^{(5)}$	0.950	0.940	1.07	1.03
5	$0.995 = \gamma^{(10)}$	0.950	0.945	1.10	1.04
5	0.990	0.990	0.980	1.00	1.00
5	$0.998 = \gamma^{(5)}$	0.990	0.988	1.03	1.00
5	$0.999 = \gamma^{(10)}$	0.990	0.989	1.04	1.01
10	0.950	0.950	0.902	1.00	1.00
10	$0.990 = \gamma^{(5)}$	0.950	0.940	1.05	1.02
10	$0.995 = \gamma^{(10)}$	0.950	0.945	1.08	1.03
10	0.990	0.990	0.980	1.00	1.00
10	$0.998 = \gamma^{(5)}$	0.990	0.988	1.02	1.00
10	$0.999 = \gamma^{(10)}$	0.990	0.989	1.03	1.00

TABLE B.2: The credible level $\tilde{\alpha} = \alpha \cdot \beta$ and the least required sample size n, for Wilks' conservative estimate $\hat{x}_{[\alpha]}$ of the one-sided confidence limit $x_{[\alpha]}$, for different levels of truncation r, stability β, and confidence levels α. Since $\gamma = \beta$, $w_{[\alpha]} = 1$ (see Table B.1). A given $\alpha^\diamond$ is almost transformed into $\tilde{\alpha}$ with the proposed choice $\gamma^{(k)} \equiv 1 - (1 - \alpha^\diamond)/k$. The supplementary case $\alpha = \beta = \gamma = \sqrt{\alpha^\diamond \gamma^{(5)}}$ illustrates that different n may yield the same $\tilde{\alpha}$

$\beta = \gamma$	α	$\tilde{\alpha}$	$n(r=1)$	$n(r=5)$	$n(r=10)$
0.950	$0.950 = \alpha^\diamond$	0.902	59	181	311
$0.990 = \gamma^{(5)}$	0.950	0.940	90	229	371
$0.970 = \sqrt{\alpha^\diamond \gamma^{(5)}}$	$0.970 = \sqrt{\alpha^\diamond \gamma^{(5)}}$	0.940	115	327	550
$0.995 = \gamma^{(10)}$	0.950	0.945	104	248	395
0.990	$0.990 = \alpha^\diamond$	0.980	459	1157	1874
$0.998 = \gamma^{(5)}$	0.990	0.988	619	1382	2148
$0.994 = \sqrt{\alpha^\diamond \gamma^{(5)}}$	$0.994 = \sqrt{\alpha^\diamond \gamma^{(5)}}$	0.988	849	2049	3271
$0.999 = \gamma^{(10)}$	0.990	0.989	688	1475	2259