Convergence Behaviour of Numerical Measurement Uncertainty Evaluation using a Virtual Metrological Computed Tomography System

Florian Wohlgemuth, Tino Hausotte

1Institute of Manufacturing Metrology, Friedrich-Alexander-University Erlangen-Nürnberg, Nägelsbachstraße 25, Erlangen, Germany, e-mail: florian.wohlgemuth@fmt.fau.de

Abstract

Digital twins of measurement devices offer fascinating applications, e.g. systematic error correction [12] or numerical uncertainty evaluation according to GUM Supplement 1 [5]. The latter is state of the art for tactile coordinate measurements [3]. For Computed Tomography (CT), this is currently not the case [18]. Prior work at the Institute of Manufacturing Metrology showed the potential to numerically evaluate task-specific measurement uncertainties in good agreement with values determined experimentally in accordance with the VDI/VDE standard 2630 Part 2.1 [18] for an unproblematic measurement task (see [8]). Details on the digital twin and the Monte Carlo method used can be found in [19]. In this prior work, the number of simulations for the Monte Carlo trials was chosen based on computational feasibility and expert knowledge. A systematic study of the convergence behaviour is the object of this contribution. Other contributions towards the numerical evaluation of task-specific measurement uncertainties can be found in [9, 15].

Keywords: numerical measurement uncertainty evaluation, convergence, digital twin, dimensional metrology

1 Introduction

According to the Supplement 1 of the GUM [5], numerical evaluation of measurement uncertainties with a Monte Carlo approach is especially feasible if:

- ‘it is difficult or inconvenient to provide the partial derivatives of the model, as needed by the law of propagation of uncertainty’ (partial derivatives would require a full analytical model of the CT measurement process which is currently not available and whose feasibility can currently be gauged infeasible).

- ‘the models are arbitrarily complicated’ (the whole measuring chain of dimensional CT, including not only the attenuation of X-rays within the measurement object and the subsequent X-ray detection, but also all image processing, volume reconstruction and surface segmentation, can be judged a complex process).

Therefore, dimensional metrology using computed tomography systems is a prime candidate for the numerical evaluation of task-specific measurement uncertainties using digital twins (i.e., realistic simulation models of real CT devices). Radiographic simulation platforms such as aR Tist 2 by the German Federal Institute of Materials Research and Testing (BAM) [1], SimCT by the University of Applied Sciences Upper Austria in Wels [14] or Scorpion XLab by the Fraunhofer Development Center X-Ray Technology EZRT [9] do in principle offer the potential to construct a digital twin for a given CT measurement setup. An example for the use of aR Tist 2 for numerical measurement uncertainty evaluation can be found in [8]. The methodological description of this digital twin can be found in [19]. In this contribution, we want to address the necessary number of Monte Carlo trials for the convergence of the numerically evaluated measurement uncertainties. In section 2, we discuss in detail the application of the GUM Supplement 1 to CT measurements, the numerical experiments performed and the statistical analysis of the resulting data. In section 3, we present the results of this analysis and in section 4 we summarise and discuss the findings.

2 Methods

Section 2.1 discusses the approach to numerical uncertainty evaluation for dimensional measurements, subsequently section 2.2 describes the simulations performed and their evaluation and section 2.3 describes the statistical analysis of the results.

2.1 Numerical Measurement Uncertainty Evaluation

Numerical measurement uncertainty evaluation based on a Monte Carlo approach is described in Supplement 1 of the GUM [5]. The approach consists of assigning probability density functions to the input quantities of a numerical model of the measurement process and sampling this numerical model with input values chosen randomly from the respective probability distributions. The mean and standard deviation of the measurement values obtained by the Monte Carlo trials is taken as estimate for the measurement value and its associated standard uncertainty. The endpoints of the coverage interval for a selected probability (e.g. 95 %) need to be provided as part of the measurement uncertainty statement and are derived from the empirical distribution function of the numerical measurement values. [5] The GUM assumes correction of all systematic effects [4, Section 3.2 and 6.3.1] which is currently not state of the art for dimensional CT measurements, therefore, we discuss both the case of uncorrected measured

https://doi.org/10.58286/25082
quantity values (section 2.1.1) and corrected measured quantity values (section 2.1.2). The supplement suggests a number of $10^6$ Monte Carlo trials [5, 7.2.1: Note]. For computationally complex models, a lower number of Monte Carlo trials is allowed as approximation if large computing times occur. Large computational demand is a problem for many simulations of the CT measurement process, therefore a reduction of the number of trials seems desirable. The supplement explicitly states that values thus obtained are less reliable. For the numerical uncertainty evaluation of tactile measurements, a number of 200 trials was used until instabilities of the results were detected. Therefore, the approach has been amended by a dynamic stability criterion [3]. This shows the need of analysing convergence behaviour for numerical uncertainty evaluation - especially in the case of a reduced number of trials.

### 2.1.1 Systematic Uncertainty Contributions

The GUM [4, Section E.3] argues for treating uncorrected systematic effects the same way as random contributions to measurement uncertainty. This approach was also chosen for the VCMM for tactile measurements [17].

We can motivate this approach with the combined standard task-specific uncertainties using a calibrated work piece according to VDI/VDE standard 2630 Part 2.1 [18]:

$$u_{uncorrected} = \sqrt{u_{cal}^2 + u_{drift}^2 + u_p^2 + u_w^2 + b^2}$$  \hspace{1cm} (1)

Of the different contributions, $u_{cal}$ – the uncertainty of calibration of the calibrated work piece – as well as $u_{drift}$ – the uncertainty due to change in the work piece since calibration – and $u_p$ – the uncertainty due to variations in material and production – equal zero for numerical measurement uncertainty evaluation as a digital model has neither calibration error nor drift and a sufficiently similar representation of material and form needs to be assumed / assured. Therefore, only the factors $u_p$ – uncertainty due to the measurement process – and $b$ – the systematic error – remain to be determined.

Therefore, the combined standard task-specific uncertainty can be calculated as:

$$u_b = \sqrt{\sigma_{sim}^2 + b^2}$$  \hspace{1cm} (2)

1

Here, $\sigma_{sim}$ is the standard deviation of the simulated measurement values (a numerical estimate of $u_p$) and $b$ is estimated by $h_{sim}$, the deviation of the mean of the simulated measurement values to the nominal value. The nominal value is the value of the measurand measured on the model (e.g. triangulated surface) used for the simulation. If the simulation model should have a significant uncertainty itself, this uncertainty in the estimation of $b$ would need to be accounted for by setting $b^2 = b_{sim}^2 + u_{b}^2$, where $u_{b}^{sim}$ represents the uncertainty of $h_{sim}$. In general, $u_{b}^{sim} \ll h_{sim}$ should hold and therefore, this correction should not be necessary. Equation 2 is equal to the VCMM approach and can be derived for arbitrary scalar measurement models [7, 11, 17]. Further, it allows for a direct comparison of numerically and experimentally evaluated uncertainties if $u_{cal}$, $u_{drift}$ and $u_w$ are sufficiently small in comparison to $u_p$ and $b$.

According to GUM Supplement 1, the numerically determined standard uncertainty needs to be reported along with a coverage interval determined from the simulated cumulative density (compare section 2.1, [5, 5.1.1]). If this coverage interval is chosen such that it is symmetric, it can be expressed by an expanded uncertainty [5, 5.3.3: Note]. This is different from the framework of the GUM in which - due to the non-availability of a cumulative density for many experimental uncertainty evaluations - this expanded uncertainty is often determined through multiplying the standard uncertainty with a generic coverage factor [4, 2.3.6, Section 6, Annex G].

Obtaining an expanded uncertainty - or a derived coverage interval - for a measured quantity value with uncorrected systematic effects is problematic as the GUM offers no guidance on the topic (as correction is assumed within the GUM framework) and it is unclear what exactly a meaningful coverage interval would be. The GUM defines the expanded uncertainty as the ‘quantity defining an interval about the result of a measurement that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand’ [4, 2.3.5]. In accordance with the VIM [6, 2.36], we call this interval the coverage interval. The expanded uncertainty is obtained by multiplying the combined standard uncertainty with a numerical factor $k$, the coverage factor [4, 2.3.5]. In the case of relevant uncorrected systematic effects significantly larger than the width of the distribution of the measured quantity values (compare figure 1a for an example), any choice of $k \geq 1$ will lead to a 100% coverage interval (in the sense that all measured quantity values are included) and any choice of $k < 1$ will lead to a coverage interval that does not include the true value and is therefore not acceptable. This degenerate case is however not the only problematic case. Also in the case of a relatively small systematic effect, the choice is problematic (compare figure 1b). Reporting a 95% coverage interval based on the distribution of the measured quantity values does not indicate a lower reliability/quality for the measurement in comparison to a measurement with the same probability distribution but without systematic deviations - which is not constructive (compare also [16, p. 93]). Reporting ±2: $u_b$ again yields a 100% coverage interval.

For the case of tactile measurements, Rost has suggested a procedure for determining a coverage interval [16, Chapter 4]: the histogram of absolute deviations $|y - y_{input}|$ of the simulated values $y$ from the input value $y_{input}$ is used to calculate a symmetrical

\[\text{In principle, a similar additional uncertainty might be added to } \sigma_{sim}. \text{ It seems unlikely that a numerical uncertainty evaluation is constructive if that is necessary, though.} \]
Figure 1: Illustration of two cases of a measurement with significant systematic effects (not based on real data), one with the true value (marked by $x_{true}$) distributed within the range of the measured quantity values and one with the true value outside. The continuous line represents this distribution of the measured quantity values. $x_{mean}$ marks the mean of the measured quantity values. The double arrows represent two arbitrary choices for a coverage factor $k$. In case 1a, these lead to a similar percentage of the measured within the resulting coverage intervals. In case 1b this is not the case - however, reporting the measurement with the interval $\pm u_m$ for 95% coverage (assuming it was exactly 95% for the sake of the argument) would not indicate a worse measurement in comparison with a measurement with the same measured quantity value distribution but $x_{true} = x_{mean}$.

$p$% coverage interval centred around $y_{input}$. This is meaningful as the input value for the simulation is the actual measurement value and therefore, the coverage interval obtained is centred around the actual measurement value - this is not the case for CT simulations. For CT simulations, the histogram of the measured quantity values $m$ cannot be used to estimate a coverage interval as this would neglect known systematic effects. On the other hand, calculating a coverage interval centred around the nominal measurement value $m_{nominal}$ obtained from the model is nonsensical, there being no reason to identify the nominal value with the actual measurement value (which will include systematic effects and therefore have a different value even in the absence of deviations between virtual part model and real measured part).

Therefore, we suggest to only compare standard uncertainties if systematic effects are not corrected. Note that above discussion is independent of the distribution of $m$.

### 2.1.2 Correction for Systematic Effects

If the systematic effect $b$ is predicted accurately by the simulation, a correction of the real measurement based on the simulation is feasible (see also [12]). According to the GUM [4, Section 3.2], the uncertainty of this correction $u_b$ needs to be taken into account. This uncertainty has at least two contributions: the - generally unknown - contribution due to a digital twin not representing the real measurement device perfectly ($u_{bet}$ in section 2.1.1) and the contribution from the statistically limited knowledge of $b$. As $b$ needs to be calculated as:

$$b = \bar{y} - y_{nominal}$$

(3)

the statistically limited knowledge of $b$ equals the statistically limited knowledge of the mean of the simulated values $\bar{y}$ ($y_{nominal}$ being exactly known from the geometrical input model, e.g. the STL file). The standard deviation of the mean $\bar{y}$ is $\sigma_{sim} / \sqrt{n}$ [13, p. 99] which is a minimum estimate for $u_b$.

Based on the VDI/VDE standard 2630 Part 2.1 [18], the standard uncertainty for a corrected measurement can be calculated as in equation 4 (the meaning of the variables is identical to equation 1). Omitting the terms that are zero for simulations, the numerically evaluated combined uncertainty is given by equation 5. For the remains of this paper, we will assume $u_{cor, min} = u_{cor}$.

$$u_{corrected} = \sqrt{u_{cal}^2 + u_{drift}^2 + u_p^2 + u_{sim}^2 + u_{n}^2}$$

(4)

$$\Rightarrow u_{cor} = \sqrt{\sigma_{sim}^2 + u_{sim}^2} \geq \sigma_{sim} \sqrt{1 + \frac{1}{n}} = u_{cor, min}$$

(5)

Alternatively, if the corrected measurand value:

$$y_{cor} = y - b$$

(6)

is regarded as the measurand for which the associated measurement uncertainty should be evaluated, then the approach of GUM Supplement 1 [5] can be applied directly. If $p_{cor}$, $p_x$ and $p_{-b}$ are the respective probability densities, then [10, Theorem 14.19]:

$$p_{cor} = p_x * p_{-b}$$

(7)

\footnote{This might again include systematic effects and might in the future be evaluated based on experience in the sense of a Type B evaluation according to the GUM [4].}

\footnote{This implicitly assumes that $b$ has no systematic effects - which an imperfect digital twin of a measurement device can in general be expected to have.}
For $b$, the central limit theorem yields that $p_{-b}$ has a normal distribution in the limit of infinite Monte Carlo trials [10, Theorem 5.16]. In this limit, statistical independence of $b$ from $y$ can be assumed. But for $y$, only the empirical cumulative density is known and an estimate of $p_y$ - especially a non-parametric one - has a significant numeric uncertainty. Therefore, the convolution in equation 7 is in general not feasible. However, the procedure of GUM Supplement 1 for the propagation of distributions [5, Section 5.4] can be applied directly and yields an empirical cumulative density for $y_{cor}$. From this, a coverage interval can be calculated. For a sufficiently high number of simulations $n$, the contribution of $p_y$ should be small due to the $1/\sqrt{n}$-dependence of its standard deviation.

If $y$ should be normal distributed, $p_y = N(\bar{y}, \sigma_{sim})$ and $p_{-b} = N(-b, \sigma_{sim}/\sqrt{n})$ such that $p_{\text{cor}} = N(\bar{y} - b, \sigma_{sim} \sqrt{1 + \frac{1}{n}})$.

Here, $N(\mu, \sigma)$ denotes a normal distribution with mean $\mu$ and standard deviation $\sigma$. In this case, using $U = 2\sigma_{sim} \sqrt{1 + \frac{1}{n}}$ for a 95% coverage interval would be justified. The GUM Supplement 1 mentions assuming a normal distribution for complicated models and insufficient Monte Carlo trials, but emphasises the reduced reliability of such an approach [5, 7.2.3: Note]. The comparison of both approaches will be the subject of a future paper.

2.2 Numerical Experiments

The digital twin and Monte Carlo framework used in this study is adapted to the Werth Messtechnik GmbH (Gießen, Germany) TomoCheck 200 3D as described in [19]. The digital twin was developed using the software aRTist 2.8.3 (BAM Federal Institute of Materials Research and Science (Berlin, Germany)) [11]. It simulates measurements at 130 kV, 275 µA with a 0.5 mm Fe filter at a voxel size of 44 µm (magnification of 1.13) and 800 projections. 21 geometrical parameters of the setup are varied initially and in between each projection using random numbers (for details, see [19]). Therefore, the sampling space for the Monte Carlo trials is isomorphic to $[0, 1]^{21 \times 21000}$.

The measurement object used was the FMT stack cuboid (see Figure 2) in aluminum (developed in [21], dimensions $20 \text{mm} \times 20 \text{mm} \times 30 \text{mm}$). 2500 simulations were performed (run time $\approx 1$ hour/simulation on Intel Core i7-7900X; different computers were used for the computations). The CAD model was converted to a stl mesh file using VGStudioMax 3.1.1 using the mesh accuracy preset ‘extra high’. Every projection stack simulated by aRTist was also reconstructed by the Feldkamp reconstruction included in aRTist. The surface determination was calculated within VGStudioMax 3.2.1. The advanced, iterative surface determination with the automatically determined iso50-value as starting contour, a search distance of 4 voxel and the option ‘remove particles and all voids’ was chosen. The determined surface was subsequently registered to the CAD using a best fit registration with quality level 50, improved refinement and the option ‘consider surface orientation’. 11 cylinders and 29 circles were evaluated using the Gauss (least squares) fit method, yielding 11 resp. 29 radii and 24 cylinder-cylinder / 48 circle-circle distances as measurands. Additionally, the circularity of the 29 circles was evaluated.

2.3 Assessment Metric

We assess the convergence using relative deviations. The relative deviation $\eta$ of the numerically evaluated measurement uncertainty of a measurand $m$ in a simulation series $\{m_1, \ldots, m_N\}$ with $N$ simulations is:

$$\eta(m_1, \ldots, m_N) = \frac{u(m_1, \ldots, m_N) - u_{\text{tot}}(m)}{u_{\text{tot}}(m)}$$

Here, $\{m_1, \ldots, m_N\}$ is shorthand for the measured quantity value of the measurand $m$ in the different simulations and $u_{\text{tot}}(m)$ is the correct measurement uncertainty of the measurand. As this correct value is not known in this study, the uncertainty evaluated from all 2500 simulations is used as an estimate. We assume that the error between this value and the prediction in the limit of an infinite number of simulations is small. We are aware that this is a weakness of the study presented here, but given the simulation run times, more accurate values are not feasible. The value $u(m_1, \ldots, m_N)$ is the uncertainty derived from the simulation series values $\{m_1, \ldots, m_N\}$. The relative deviation metric has a limit value of 0 for $N = 2500$. A value of e.g. $\pm 0.1$ signifies a deviation of $\pm 10\%$ of the uncertainty value from the limit uncertainty $u_{\text{tot}}(m)$.
Figure 3: The plot shows the succession of $\eta$ for an arbitrary permutation along with $n_{\text{acc}}$ for a maximum deviation of 10%. $n_{\text{acc}}$ is marked by a vertical straight line, the dotted lines mark $\pm 0.1$. For simulation numbers greater than $n_{\text{acc}}$, the value of $\eta$ is within those dotted lines, meaning that the relative error stays below the maximum deviation chosen for all measurands.

Figure 4: Progression of $\eta$ with number of simulations for a distance between two circle-centres and 3000 permutations of length 1000. Figure 4a shows the case without correction for systematic effects, figure 4b the case including this correction.

We use $\eta_c$ for corrected systematic effects and $\eta_{\text{uc}}$ otherwise.

For a simulation series with indices $k = 1, \ldots, l$ and associated measurand values $m_k$, the relative errors after $j \leq l$ simulations can be calculated as $\eta(j) = \eta(m_1, \ldots, m_j)$. This yields a progression of the relative error $\eta(j)$ with the simulation number $j$ for $j = 1, \ldots, l$. As the simulations are all based on randomly generated simulation parameters, they are statistically independent and could, in principle, have occurred in any arbitrary order. To assess the convergence behaviour of the simulation, ideally, repeated simulation trials would be performed and evaluated to provide a statistic. Unfortunately, the CT simulation run times prohibit this. As a solution, we suggest that due to statistical independence, any permutation $\tau$ of the original index ordering of the simulations (i.e., any ‘virtual change’ of simulation sequence) is an approximation to a new simulation run and already allows some insight into the convergence behaviour. This should especially be a valid approach in the case of using a large data set (here, 2500 simulations) to choose smaller ‘virtual simulation trials’ of smaller size as arbitrary paths through the 2500 simulations. For a path length of $t$, by combinatorics, a number of $2500^t$ different such paths through the data set are available [20, p. 578], allowing for sufficient statistics. We are aware that this is not the perfect approach, but we are also convinced that a better solution is technically not easy to realise and a study into this aspect is a necessary step towards numerical measurement uncertainty determination for CT systems.

Let $\tau_1, \ldots, \tau_t$ be permutations of equal length $t$. Each permutation $\tau_k$ yields, for each measurand $m$, an associated relative error progression $\eta(\tau_k(1)), \eta(\tau_k(1), \tau_k(2)), \ldots, \eta(\tau_k(1), \ldots, \tau_k(t))$. From this error progression, we can obtain the number of simulations $n_{\text{acc}}$ necessary such that all relative errors - or the relative error for a subset of the measurands - are within a certain acceptance margin, e.g. $|\eta(l)| < 0.05$ for all $l > n_{\text{acc}}$ (compare figure 3 for an illustration). Each $\tau_i$ can yield a different value for $n_{\text{acc}}$ and we can either inspect the mean value, maximum value or different percentiles of $n_{\text{acc}}$. Similarly, for one permutation $\tau$ we can inspect the mean value, maximum value or different percentiles of $n_{\text{acc}}$ for a set of measurands $m$.

\[4\text{Strictly speaking, for } t < N, \text{ these are not permutations but the first } t \text{ elements of a permutation of the full set of all } N \text{ indices - or: ordered subsets with length } t \text{ of the original index set. For the purpose of this paper and notational simplicity, we will call them permutations.}\]

---

**Successive $\eta_c$**

![Successive $\eta_c$](image)

Figure 3: The plot shows the succession of $\eta$ for an arbitrary permutation along with $n_{\text{acc}}$ for a maximum deviation of 10%. $n_{\text{acc}}$ is marked by a vertical straight line, the dotted lines mark $\pm 0.1$. For simulation numbers greater than $n_{\text{acc}}$, the value of $\eta$ is within those dotted lines, meaning that the relative error stays below the maximum deviation chosen for all measurands.

![Progression of $\eta$](image)

Figure 4: Progression of $\eta$ with number of simulations for a distance between two circle-centres and 3000 permutations of length 1000. Figure 4a shows the case without correction for systematic effects, figure 4b the case including this correction.

We use $\eta_c$ for corrected systematic effects and $\eta_{\text{uc}}$ otherwise.

For a simulation series with indices $k = 1, \ldots, l$ and associated measurand values $m_k$, the relative errors after $j \leq l$ simulations can be calculated as $\eta(j) = \eta(m_1, \ldots, m_j)$. This yields a progression of the relative error $\eta(j)$ with the simulation number $j$ for $j = 1, \ldots, l$. As the simulations are all based on randomly generated simulation parameters, they are statistically independent and could, in principle, have occurred in any arbitrary order. To assess the convergence behaviour of the simulation, ideally, repeated simulation trials would be performed and evaluated to provide a statistic. Unfortunately, the CT simulation run times prohibit this. As a solution, we suggest that due to statistical independence, any permutation $\tau$ of the original index ordering of the simulations (i.e., any ‘virtual change’ of simulation sequence) is an approximation to a new simulation run and already allows some insight into the convergence behaviour. This should especially be a valid approach in the case of using a large data set (here, 2500 simulations) to choose smaller ‘virtual simulation trials’ of smaller size as arbitrary paths through the 2500 simulations. For a path length of $t$, by combinatorics, a number of $2500^t$ different such paths through the data set are available [20, p. 578], allowing for sufficient statistics. We are aware that this is not the perfect approach, but we are also convinced that a better solution is technically not easy to realise and a study into this aspect is a necessary step towards numerical measurement uncertainty determination for CT systems.

Let $\tau_1, \ldots, \tau_t$ be permutations of equal length $t$. Each permutation $\tau_k$ yields, for each measurand $m$, an associated relative error progression $\eta(\tau_k(1)), \eta(\tau_k(1), \tau_k(2)), \ldots, \eta(\tau_k(1), \ldots, \tau_k(t))$. From this error progression, we can obtain the number of simulations $n_{\text{acc}}$ necessary such that all relative errors - or the relative error for a subset of the measurands - are within a certain acceptance margin, e.g. $|\eta(l)| < 0.05$ for all $l > n_{\text{acc}}$ (compare figure 3 for an illustration). Each $\tau_i$ can yield a different value for $n_{\text{acc}}$ and we can either inspect the mean value, maximum value or different percentiles of $n_{\text{acc}}$. Similarly, for one permutation $\tau$ we can inspect the mean value, maximum value or different percentiles of $n_{\text{acc}}$ for a set of measurands $m$.

\[4\text{Strictly speaking, for } t < N, \text{ these are not permutations but the first } t \text{ elements of a permutation of the full set of all } N \text{ indices - or: ordered subsets with length } t \text{ of the original index set. For the purpose of this paper and notational simplicity, we will call them permutations.}\]
3 Results

Section 3.1 discusses the propagation of $\eta$ for different cases and section 3.2 concerns the necessary number of simulations for a given accuracy goal. All standard uncertainties given are calculated according to equations 2 and 5.

3.1 Plots of $\eta$

Figure 4 shows, for a distance between two circle-centres, the progression of $\eta$ up to a number of 1000 simulations for 3000 different permutations, both for the case of corrected and for uncorrected measured quantity values. It is noteworthy that there is no bias towards over- or underestimation. Adding a safety margin for a low number of Monte Carlo trials (in the sense of a factor to be multiplied with to account for too low a number of trials) could therefore severely overestimate the measurement uncertainty. Based on the data in this study, such a safety margin cannot be advised. A difference between corrected and uncorrected case can hardly be recognised from the plots (compare figure 4). A histogram of all $\eta$-values for all measurands (figure 5) shows a sharper peak for the case of uncorrected systematic effects - meaning that, on average, the prediction of the uncorrected measurement error is already closer to its ‘correct’ value.

Plots for the cylinder-based measurands (not included in the paper) show a similar picture with less pronounced differences between the case of corrected and uncorrected systematic effects.

3.2 Discussion of $n_{acc}$

For a condensed view of the plots and histograms from section 3.1, we suggest the quantity $n_{acc}$ as defined in 2.3. Generally, the value of $n_{acc}$ depends strongly on the required accuracy and on the metric used. The metric ‘abs’ (for ‘absolute’) gives the simulation number after which $|\eta|$ is below the required maximum deviation for all measurands for one permutation $\tau$. The metric ‘prc90’ is the 90% percentile of the $n_{acc}$ value of all measurands for one permutation $\tau$ (meaning that $|\eta|$ is below the required maximum deviation for 90% of the measurands) and the metric ‘mean’ measures the mean value of $n_{acc}$ over all measurands for one permutation $\tau$.

The plots (figures 6 - 10) show that a change from 5% to 10% accuracy goal costs more than a factor 2 in simulation numbers, meaning that additional statistical accuracy is computationally expensive. An increase to 20% allowed deviation already reduces all median values of $n_{acc}$ in the absolute metric below 100.

In general, the measurement uncertainty for uncorrected measured quantity values can be evaluated with a lower number of necessary Monte Carlo trials than the corrected one. The difference between the number of necessary Monte Carlo trials is high for the roundnesses evaluated, notable for the radii and relatively small for the distances. This difference between the different classes of measurands correlates with the ratio of systematic deviation and standard deviation, $|b_{sim}/\sigma_{sim}|$. This ratio has a mean of 6.10 for the roundnesses, 2.27/5.54 for radii (circles/cylinders) and 0.49/0.73 for distances (circles/cylinders). For the data in this work, a relatively high systematic deviation $b_{sim}$ therefore means that this systematic deviation dominates the measurement uncertainty for uncorrected measured quantity values and as it converges more quickly, the measurement uncertainty for uncorrected measured quantity values also converges notably quicker.

The median values of $n_{acc}$ are generally lower for cylinder radii than for the circle radii. This is especially strong for the uncorrected case and less pronounced for the corrected case, correlating with the difference in $|b_{sim}/\sigma_{sim}|$ between cylinders and circles. For distances, the cylinder measurands also have lower median values for $n_{acc}$ but the difference is not as strong as with the radii.

The difference between the three metrics used indicates an $\eta$-distribution with a pronounced tail for each $\tau$. Boxplots of the measurand-wise values for $n_{acc}$ (see as an example figure 11) do however show a quite uniform behaviour across the different measurands in each class, indicating that the cause for these tails are not single measurands but instead different measurands in different permutations.
Figure 6: Boxplot of $n_{acc}$-values for circle roundnesses. The labels read ‘uc’ for uncorrected, ‘c’ for corrected and ‘abs’ for absolute, ‘prc90’ for 90% percentile and ‘mean’ for mean value metric (compare section 3.2). The values are based on 1000 permutations of length 2500.

Figure 7: Boxplot of $n_{acc}$-values for circle radii. The labels read ‘uc’ for uncorrected, ‘c’ for corrected and ‘abs’ for absolute, ‘prc90’ for 90% percentile and ‘mean’ for mean value metric (compare section 3.2). The values are based on 1000 permutations of length 2500.

Figure 8: Boxplot of $n_{acc}$-values for circle distances. The labels read ‘uc’ for uncorrected, ‘c’ for corrected and ‘abs’ for absolute, ‘prc90’ for 90% percentile and ‘mean’ for mean value metric (compare section 3.2). The values are based on 1000 permutations of length 2500.
Figure 9: Boxplot of $n_{acc}$-values for cylinder radii. The labels read 'uc' for uncorrected, 'c' for corrected and 'abs' for absolute, 'prc90' for 90% percentile and 'mean' for mean value metric (compare section 3.2). The values are based on 1000 permutations of length 2500.

Figure 10: Boxplot of $n_{acc}$-values for cylinder distances. The labels read 'uc' for uncorrected, 'c' for corrected and 'abs' for absolute, 'prc90' for 90% percentile and 'mean' for mean value metric (compare section 3.2). The values are based on 1000 permutations of length 2500.

Figure 11: Boxplot of $n_{acc}$ for all distances between circle centres (uncorrected values, 5% accepted deviation, based on 1000 permutations of length 2500).
4 Discussion

The GUM states that: 'Even the experimental standard deviation of the mean of as many as 30 repeated observations of a quantity described by a normal distribution has itself an uncertainty of about 13 percent [...]’ [4, Annex G]. If the task-specific measurement uncertainty of a measurement is dominated by standard deviations, this means that an experimentally determined measurement uncertainty from 30 repeated measurements has an uncertainty of about ±7%. Therefore, the deviations of ±5% and ±10% used as accuracy goals in this paper are acceptable deviations in comparison to experimentally determined values. It should however be noted that deviations between digital twin and real CT device are potentially scaled by the factor set as allowed deviation $\eta$ due to statistical effects (if there is e.g. a systematic deviation leading to a deviation $\Delta u$ in the limit of infinite Monte Carlo trials, this can potentially rise to $1.1 \cdot \Delta u$ due to a finite number of trials if an allowed deviation $\eta$ of 10% is chosen). It is therefore important to either minimise these kind of deviations or to add a safety margin by reducing the accuracy goal accordingly.

Even though a correction of systematic effects is automatically possible if a digital twin is used (as the systematic deviation $b$ can be obtained directly from the simulated values), the results also show that this correction comes at the expense of an increased number of necessary Monte Carlo trials. The results from this work show that even with a digital twin well adapted to a real CT device, the high-dimensional sampling space still poses a challenge concerning necessary simulation number and is currently a limiting factor to applicability. Research both towards faster simulation algorithms and more intelligent sampling schemes is necessary for the numerical measurement uncertainty determination with digital twins to become a viable option in industrial contexts.

Author contributions according to CRediT taxonomy [2]

F. W. contributed Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization as well as Writing - original draft. Both T. H. and F. W. contributed to Funding acquisition and Project administration. T. H. contributed Resources, Supervision and Writing - review & editing.

Acknowledgements

This work was funded through the project CTSimU (WIPANO project 03TNH026A). WIPANO projects are financed by the German Federal Ministry for Economic Affairs and Energy and managed by Project Management Jülich.

References