International Journal for Uncertainty Quantification, 5 (6): 511–526 (2015)

BAYESIAN APPROACH TO THE STATISTICAL INVERSE PROBLEM OF SCATTEROMETRY: COMPARISON OF THREE SURROGATE MODELS

Sebastian Heidenreich, [*] ***Hermann Gross, & Markus Bär***

Physikalisch-Technische Bundesanstalt, Abbestr 2-12, 10587 Berlin, Germany

Original Manuscript Submitted: 12/21/2014; Final Draft Received: 08/03/2015

Scatterometry provides a fast indirect optical method for the determination of grating geometry parameters of photomasks and is used in mask metrology. To obtain a desired parameter, inverse methods like least squares or the maximum likelihood method are frequently used. A different method, the Bayesian approach, has many advantages against the others, but it is often not used for scatterometry due to the large computational costs. In this paper, we introduce different surrogate models to approximate computationally expensive calculations by fast function evaluations, which enable the Bayesian approach to scatterometry. We introduce the nearest neighbor interpolation, the response surface methodology and a method based on a polynomial chaos expansion. For every surrogate model, we discuss the approximation error and the convergence. Moreover, we apply Markov Chain Monte Carlo sampling to determine desired geometry parameters, and its uncertainties form simulated measurement values based on Bayesian inference. We show that the surrogate model involving polynomial chaos is the most effective.

KEY WORDS: *scatterometry, Bayesian inference, stochastic partial differential equations, collocation, inverse problem*

1. INTRODUCTION

Scatterometry is a frequently used metrology technique for the characterization of critical dimensions on wafers and masks used in photolithography [1, 2]. It provides a non-destructive alternative to direct optical methods like electron microscopy and atomic force microscopy. In a simple case, a line grating is illuminated by monochromatic, coherent light and diffraction patterns are detected. By solving an inverse problem, desired geometry parameters are determined from the diffraction signals [see Fig. 1(a) for an illustration]. Recently, geometry parameters of rough absorption lines have been reconstructed from extreme ultraviolet light (EUV) measurements by employing a maximum likelihood method for parameter estimations and a Fisher information matrix method for the determination of uncertainties [3–6]. This method provides accurate geometry parameters, but associated uncertainties are calculated only locally. A more reliable technique to derive uncertainties is based on sampling methods. In the Supplement 1 of the guide to the evaluation of uncertainties in measurement (GUM-S1) the use of Monte Carlo methods in a Bayesian context is recommended [7]. Monte Carlo methods provide a full probability distribution function (PDF) for desired quantities. Uncertainties can be determined by certain credible intervals of the PDFs. A further advantage of the Bayesian approach is the possibility of the combination of different measurements and the possibility to include prior knowledge of the parameters of interest. The prior knowledge is formalized by specific PDFs and the gain of information from the measurement is expressed by the likelihood function. Both together result in the posterior distribution, which includes the joint information about the prior knowledge and the measurement information. In the Bayesian approach, parameters and uncertainties are estimated by Markov chain Monte Carlo (MCMC) sampling methods from the posterior distribution. However, the convergence of MCMC sampling methods is slow and many sampling steps are often needed for acceptable results.

[*]Correspond to Sebastian Heidenreich, E-mail: sebastian.heidenreich@ptb.de

2152–5080/15/$35.00

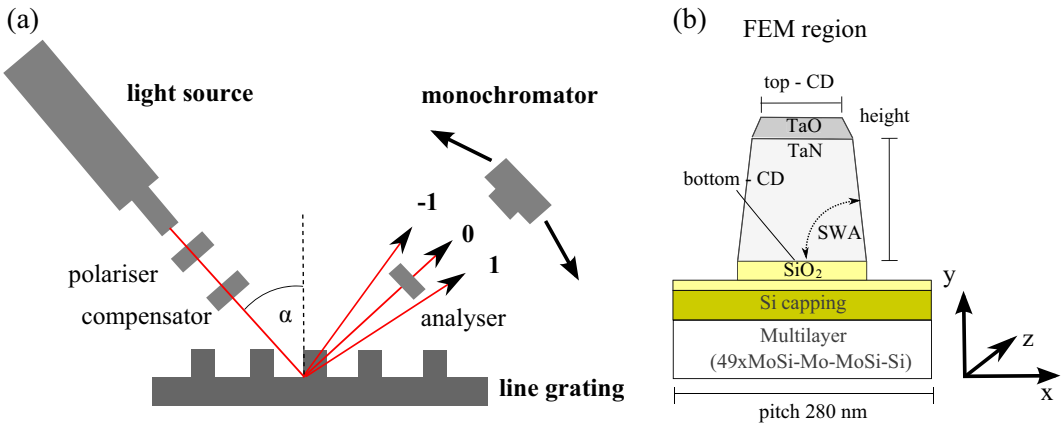

FIG. 1: (a) Principal scheme of the measurement setup is shown. Diffraction orders are labeled by $-1...1$. (b) Cross section (in the xy plane) of the EUV photomask is depicted.

In this paper, we replace time-demanding calculations, which are required for the solutions of the inverse problem by many less-expensive numerical solutions by the use of surrogate models. In particular, for a certain parameter range, we replace a differential operator by a superposition of nonlinear functions.

Related to this work, preliminary results have been published elsewhere [8]. While the previous work demonstrated the feasibility of the polynomial chaos expansion as a surrogate model, this paper gives a detailed comparison of three models and a comprehensive analysis of the approximation errors.

In this paper we proceed as follows. In the next paragraphs, we introduce the mathematical model of scatterometry, the Bayesian approach, and three different surrogate models, i.e., nearest neighbor interpolation, the response surface methodology, and a polynomial chaos-based method. Then we discuss the approximation error of each surrogate model by a comparison with rigorous finite-element solutions (FEM) of the partial differential equation (PDE) behind. We discuss applicability, precision, accuracy and convergence. Moreover, we employ each surrogate model to the Bayesian approach, and we reconstruct desired geometry parameters and associated uncertainties with a MCMC method. Finally, we discuss the effect of the approximation error on the uncertainties.

2. MASK GEOMETRY

The grating used consists of periodic absorber lines, two capping layers, and a multilayer stack. The stack functions like a mirror for waves with a wavelength of $\lambda = 13.4$ nm. The pitch is 280 nm and geometry parameters of interest are the bottom-CD (width at bottom of the absorber line), the height of the second absorber line, and the SWA of the TaN layer (SWA). A cross section of the presented photomask is shown in Fig. 1(b).

In our study, we have chosen geometry parameters with the following minimal and maximal values: bottom-CD 140 nm $\pm$ 11.2 nm, height = 80 nm $\pm$ 5.6 nm, and the sidewall angle $= 87.5° \pm 5°$.

3. MATHEMATICAL MODEL

The mathematical model is characterized by the propagation of electromagnetic waves through the material and is described by Maxwell's equations. The particular line structure introduced in the previous paragraph implies invariance of optical properties in one spatial direction (z direction). In this specific case, the time-harmonic Maxwell equations reduce to the two-dimensional Helmholtz equation in the $x - y$ plane,

$$\Delta u\,(x,y) + k^2\,(x,y)\,u\,(x,y) = 0, \tag{1}$$

where u is the transversal field component that oscillates in the z direction, and $k\,(x,y)$ is the wave number

$$k(x,y) = \omega\left(\mu_0 \epsilon(x,y)\right)^{1/2}, \tag{2}$$

that is assumed to be constant for areas filled with the same material. The symbol Δ is the Laplace operator in two dimensions, i.e.,

$$\Delta = \partial_x^2 + \partial_y^2. \tag{3}$$

The boundary conditions imposed on the PDE are periodic on the lateral boundaries due to the repeated structure and usual outgoing wave conditions in the infinite regions [9]. For rigorous calculations, we used the finite element (FEM) solver DIPOG.[1] We fixed the wavelength of the incident light $\lambda = 13.4$ nm, choose an incident angle of $\alpha = 6^\circ$ and optical parameters from Henn et al. [3]. For every set of geometry parameters joined in the vector $p_1 ... p_N$, there is a set of efficiencies $f_1(\boldsymbol{p}) ... f_M(\boldsymbol{p})$, labeled by the diffraction orders $1...M$ [see Fig. 1(a)], which defines a nonlinear map (defined as forward model)

$$\begin{aligned} \mathrm{I}_1 \times \mathrm{I}_2 \times \mathrm{I}_3 &\to \mathbb{R}_+^{22} \\ \boldsymbol{p} &\mapsto f_j(\boldsymbol{p}). \end{aligned} \tag{4}$$

Efficiencies obtained in experiments include additional noise.

3.1 Error Model

In our approach, measurement values y_j are modeled by the forward model $f_j(\boldsymbol{p})$ and a measurement error ϵ_j^{m}, i.e.,

$$y_j = f_j(\boldsymbol{p}) + \epsilon_j^m. \tag{5}$$

If there are no correlations between measurements and no further systematic errors, it can be assumed that ϵ_j is distinct for different measurements. For simplicity, we assume a Gaussian distributed error with zero mean. In principle, there are many error sources, but from experiments, two main contributions are known [3]. The detection of diffraction pattern in scatterometry takes a while and power fluctuations during the record process are a significant contribution to the measurement error. This error depends on the diffraction order. The error of the detector noise is for every order the same. To include the two main error contributions the following approach has been approved [3]:

$$\left(\sigma_j^m\right)^2 = \left(a_{\mathrm{m}} f_j(\boldsymbol{p})\right)^2 + b_{\mathrm{m}}^2, \tag{6}$$

where $\left(\sigma_j^m\right)^2$ is the variance of the jth diffraction order. The symbols a_{m} and b_{m} are the error parameters. In our approach, we aim to approximate rigorous FEM calculations by a surrogate model. This approximation introduces an additional error, i.e.,

$$f_j(\boldsymbol{p})^{\mathrm{FEM}} = f_j(\boldsymbol{p})^{\mathrm{surrogate}} + \epsilon_j^{\mathrm{approx}}. \tag{7}$$

We assume that the approximation error has the same properties as the measurement error. For the measurement values we obtain

$$y_j = f_j(\boldsymbol{p})^{\mathrm{surrogate}} + \epsilon_j^{\mathrm{approx}} + \epsilon_j^{\mathrm{m}}, \tag{8}$$

with

$$\left(\sigma_j^{\mathrm{approx}}\right)^2 = \left(a^{\mathrm{approx}} f_j^{\mathrm{surrogate}}(\boldsymbol{p})\right)^2 + \left(b^{\mathrm{approx}}\right)^2. \tag{9}$$

We define the joined error $\epsilon_j := \epsilon_j^{\mathrm{m}} + \epsilon_j^{\mathrm{approx}}$ and rewrite Eq. (8),

$$y_j = f_j^{\mathrm{surrogate}} + \epsilon_j, \tag{10}$$

where ϵ_j is Gaussian distributed with zero mean and variance $\left(\sigma_j\right)^2 = \left(a f_j\right)^2 + b^2$. The noise parameters are then given by $a^2 = a_{\mathrm{m}}^2 + a_{\mathrm{approx}}^2$ and $b^2 = b_{\mathrm{m}}^2 + b_{\mathrm{approx}}^2$.

[1] Developed at WIAS; http://www.wias-berlin.de/software /DIPOG

4. THE BAYESIAN APPROACH TO THE STATISTICAL INVERSE PROBLEM IN SCATTEROMETRY

The reconstruction of general geometries from the diffraction pattern is an ill-posed problem and therefore, prior information or regularization techniques are necessary [10]. We have chosen prior information from experimental knowledge and the mask design to restrict the solution to a certain geometry class and to fix the optical parameters of the materials used. From simple least square methods to maximum likelihood, there are many methods to find geometry parameters and related diffraction pattern with optimal fit to measured values. A powerful method to include prior knowledge and to obtain reliable uncertainties is based on the Bayesian theorem. The Bayes theorem links the prior distribution π_0 to the posterior distribution π via the likelihood function $\mathcal{L}(\mathbf{p}, a, b; \mathbf{y})$, i.e.,

$$\pi(\boldsymbol{p}, a, b|\boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{p}, a, b; \boldsymbol{y})\pi_0(\boldsymbol{p}, a, b)}{\int \mathcal{L}(\boldsymbol{p}, a, b; \boldsymbol{y})\pi_0(\boldsymbol{p}, a, b)\ \mathrm{d}\mathbf{p}\ \mathrm{d}a\ \mathrm{d}b}. \tag{11}$$

The prior distribution expresses the prior knowledge about parameters; the likelihood functions the information gain from experiments, and the posterior distributions are the related joined distributions. As introduced in the previous paragraph, the measurement errors are assumed to be Gaussian distributed. In this case, the likelihood function is given by

$$\mathcal{L}(\boldsymbol{p}, a, b; \boldsymbol{y}) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma_j^2(\boldsymbol{p}, a, b)}} \exp\left[-\frac{(y_j - f_j(\boldsymbol{p}))^2}{2\sigma_j^2(\boldsymbol{p}, a, b)}\right], \tag{12}$$

$$\text{with}\ \ \sigma_j^2 = (a \cdot f_j(\boldsymbol{p}))^2 + b^2. \tag{13}$$

Estimates of parameters are calculated by the help of the posterior distribution. If we define $\theta := (\boldsymbol{p}, a, b)^{\mathrm{T}}$, the expectation values of parameters are given by $\bar{\theta}_i = E[\theta_i]_\pi$ and the associated variances by $\mathrm{var}(\theta_i) = E[\theta_i^2]_\pi - E[\theta_i]_\pi^2$ with $E[\theta_i]_\pi = \int \theta_i d\pi$, where $d\pi$ denotes the posterior measure. In this paper, we have chosen uniform prior distributions for the geometry parameters: $\mathrm{U}[128.8, 151.2]/nm$ for the bottom CD, $\mathrm{U}[74.4, 85.6]/nm$ for the height and $\mathrm{U}[82.5, 92.5]/\circ$ for the sidewall angle.

5. DIFFERENT SURROGATE MODELS FOR THE FORWARD MODEL

Surrogate models are applied in many fields like aerospace science [11, 12], geostatistics [13–15], engineering design [16, 17], risk analysis [18–20], optimization [21] and machine learning techniques [22]. The application areas listed are all computationally expensive and in many cases the software used is proprietary and hence can be treated as a black box for the construction of the surrogate models. The huge application range of surrogate models implies many approaches. Here, the focus is on the uncertainty analysis in metrology where the propagation of uncertainties is of special interest. For convenience, we selected the most simple surrogate model, the nearest neighbor interpolation. The second model we have chosen is the widely used response surface methodology [23–25] and the last model selected is based on the polynomial chaos expansion. The latter was successfully employed in many applications to investigate the propagation of uncertainties [26–31].

In our investigations, we assume that the parameter ranges of the surrogate models chosen are given by the support of the prior distributions [32]. We have selected the intervals: $I_1 = [128.8, 151.2]$nm for the bottom-CD, $I_2 = [74.4, 85.6]$nm for the height, $I_3 = [82.5, 92.5]^\circ$ for the SWA, $I_4 = [0.01, 0.1]$ for a and $I_5 = [0.001, 0.1]\%$ for b. Moreover, the independences of random input variables are assumed.

5.1 Nearest-Neighbor Interpolation

Nearest-neighbor (NN) interpolation is one of the simplest interpolation techniques. For example, a nonlinear function $f(p)$ is approximated by the rigorous calculation of function values at specific training points p_i. An arbitrary function value $\tilde{f}_k := f(p_k)$ is given by the values at the training point close to p_i. In particular, $\tilde{f}_k = f(p_k)$ if and only if $p_k \in U(\mathbf{p}_i)$, where U denotes the nearest neighbor of $\mathbf{p}_i$. This approximation leads to a piecewise constant discontinuous approximation of the forward map. The NN interpolation is simple to implement, but in many cases,

the convergence speed is slow. It is particularly inaccurate for functions that change rapidly. In our study, we divided the three-dimensional parameter space into 729 cubes and calculated a value of the forward map for each cube with the FEM solver rigorously.

5.2 Response Surface Methodology

The next surrogate model is the response surface methodology (RSM). Here, model outputs of the forward model are represented as a polynomial function g_k and a truncation error ϵ_k,

$$f_k(\boldsymbol{p}) = g_k(\boldsymbol{p}, \boldsymbol{\beta}) + \epsilon_k. \tag{14}$$

The polynomial function can have any order, but typically only terms up to the second order are used (quadratic response surface). In this case, the model is expressed as

$$\tilde{f}_k(\boldsymbol{p}) = \beta_0 + \sum_{i=1}^{N} \beta_i p_i + \sum_{i=1}^{N} \sum_{j \geq i}^{N} \beta_{ij} p_i p_j, \tag{15}$$

where the coefficients β represent unknown hyperparameters of the surrogate model which are determined by rigorous FEM calculations at some training points. Once hyperparameters are determined, the training set is no longer used and only the hyperparameters of the model characterize the response at new points. For our study, we used 609 training points and calculated all hyperparameters by projecting $f_k(\boldsymbol{p})$ onto the polynomial basis (of quadratic order).

5.3 Polynomial Chaos Expansion

The most sophisticated surrogate model presented here is based on polynomial chaos (PC). The idea of PC goes back to Wiener and his 1938 paper [33]. In his theoretical investigation about ergodic systems, he used polynomial approximations for a Gaussian stochastic process. In 1947, Cameron and Martin [34] proved that it is possible to approximate every stochastic process with a finite second moment by an infinite convergent sum of orthogonal polynomials. The approach became popular for uncertainty quantification after Xiu and Karniadakis [35] introduced the Wiener-Askey scheme. The Wiener-Askey scheme relates different stochastic processes to orthogonal polynomials such that the approximation error is small (cf. Table 1).

In general, the PC expansion of the forward map for uncertain input parameters p_i is given by

$$f_j(p_{1\ldots n}) = f_j(\mu_{1\ldots n}) + \sum_{i=1}^{n} \sum_{k=1}^{R} f_{jk}^{i}(\mu_{1\ldots n}) \Psi_k(\omega_i) \tag{16}$$

$$+ \sum_{\alpha \neq \beta} \sum_{k=1}^{S} \sum_{l=1}^{k} f_{jkl}^{\alpha\beta}(\mu_{1\ldots n}) \Psi_k(\omega_{\sigma(\alpha)}) \Psi_l(\omega_{\sigma(\beta)}) + \ldots \tag{17}$$

$$\int_{-1}^{1} \Psi_k(\omega_\alpha) \Psi_l(\omega_\alpha) d\omega_\alpha = \int_{-1}^{1} \Psi_k^2(\omega_\alpha) d\omega_\alpha \; \delta_{kl} \tag{18}$$

TABLE 1: The Wiener-Askey scheme scheme relates polynomial basis functions to distributions of random variables

Distribution of random variable	Polynomial type	Support
Gaussian	Hermite	$(-\infty, \infty)$
Gamma	Laguerre	$[0, \infty)$
Beta	Jacobi	$[a, b]$
Uniform	Legendre	$[a, b]$

where μ_i is the mean value of the prior distribution and ω_i are random numbers such that $p_i = \mu_i + \omega_i$ is a random number drawn from the prior distribution. Here, n is the number of geometry parameters and R (respectively, S) the maximal order of the polynomial approximation. f_{jkl} are the expansion coefficients, and the symbol $\sigma(\alpha)$ stands for all permutations of the index set $\alpha = 1...n$. According to the Wiener-Askey scheme we have chosen Legendre polynomials as orthogonal basis functions Ψ_k. The expansion coefficients are determined by projections onto functions Ψ_k, i.e.,

$$f^i_{jk}(\mu_{1...n}) = \frac{\int f_j(p_{1...n})\Psi_k(\omega_i)\rho(\omega_i)d\omega_i}{\int \Psi^2_k(\omega_i)\rho(\omega_i)d\omega_i} \tag{19}$$

$$\approx \frac{1}{\int \Psi^2_k(\omega_i)\rho(\omega_i)d\omega_i}\sum_m \lambda^i_m f_j(\omega^m_i)\Psi_k(\omega^m_i)\rho(\omega^m_i) \tag{20}$$

and

$$f^{\alpha\beta}_{jkl}(\mu_{1...n}) = \frac{\int f_j(p_{1...n})\Psi_k(\omega_\alpha)\Psi_l(\omega_\beta)\rho(\omega_\alpha)\rho(\omega_\beta)d\omega_\alpha d\omega_\beta}{\int \Psi^2_k(\omega_\alpha)\Psi^2_l(\omega_\beta)\rho(\omega_\alpha)\rho(\omega_\beta)d\omega_\alpha d\omega_\beta} \tag{21}$$

$$\approx \frac{\sum_m \lambda^i_{mn} f_j(\omega^m_i,\omega^n_i)\Psi_k(\omega^m_i)\Psi_l(\omega^n_i)\rho(\omega^m_i)\rho(\omega^n_i)}{\int \Psi^2_k(\omega_\alpha)\Psi^2_l(\omega_\beta)\rho(\omega_\alpha)\rho(\omega_\beta)d\omega_\alpha d\omega_\beta}, \tag{22}$$

where ρ is the specific probability distribution. In many cases, the denominator can be integrated analytically. The integration of the numerator can be done by numerical integrations using quadrature or cubature rules. In the one-dimensional case, training points ω^m_i are given by zeros of the expansion polynomials. Higher-dimensional functions are treated by the tensor product of lower-dimensional functions, such that the number of training points increase with the power of the parameter space dimension (curse of dimensionality). For high-dimensional problems, alternative approaches such as the Smolyak quadrature [36] are useful to reduce the number of training points and therefore, the computational time for the construction of the surrogate model. In our study, we solved the denominator analytically and the numerator by a sparse Legendre grid with the index m and weights λ^i_m (respectively, λ^i_{mn}).

6. RESULTS

6.1 Comparison of Surrogate Models

In the presented study, all three surrogate models are constructed from rigorous calculations with 600 to 700 training points and validated by FEM calculations. In Fig. 2, a comparison of efficiencies is shown. The approximation by the NN interpolation method has large errors, in particular, at higher diffraction orders. The approximation for RSM and PC is much better [see Figs. 2(b) and 2(c)].

To quantify the approximation error in detail, we have drawn 100 sample geometries from the prior distributions and calculated rigorously 22 efficiencies. We compared the results with approximations obtained from each surrogate model. For every geometry, we computed the L^2 norm of the difference between rigorous calculations and the approximation, i.e., $\sqrt{\sum_j (f_j^{\text{FEM}}(\boldsymbol{p}) - f_j^{\text{surrogate}}(\boldsymbol{p}))^2}$. Figure 3 depicts boxplots of the errors for all surrogate models. Errors of NN and RSM are of the same magnitude, but differ by a factor of four. The approximation error of the surrogate model based on the PC expansion is more than one magnitude smaller ($\mathcal{O}(10^{-3})$).

The approximation error is determined by the number of training points and, in the limit of infinitely many points, the error should vanish. To study the convergence, we constructed our surrogate models from coarse to finer meshes, calculated the difference between approximated and rigorous calculated efficiencies within the L^2 norm, and we took its maximum value for 100 randomly chosen geometries. Figure 4 depicts the convergence behavior of the different surrogate models. As we expected, the convergence of NN interpolations is very slow, whereas the convergence of the PC-based surrogate model is quite fast. The convergence of RSM is in between the two others, but saturates after about 40 training points. This shows that hyperparameters are well approximated with less training points, but by further increasing the number of training points yields not substantially better approximations.

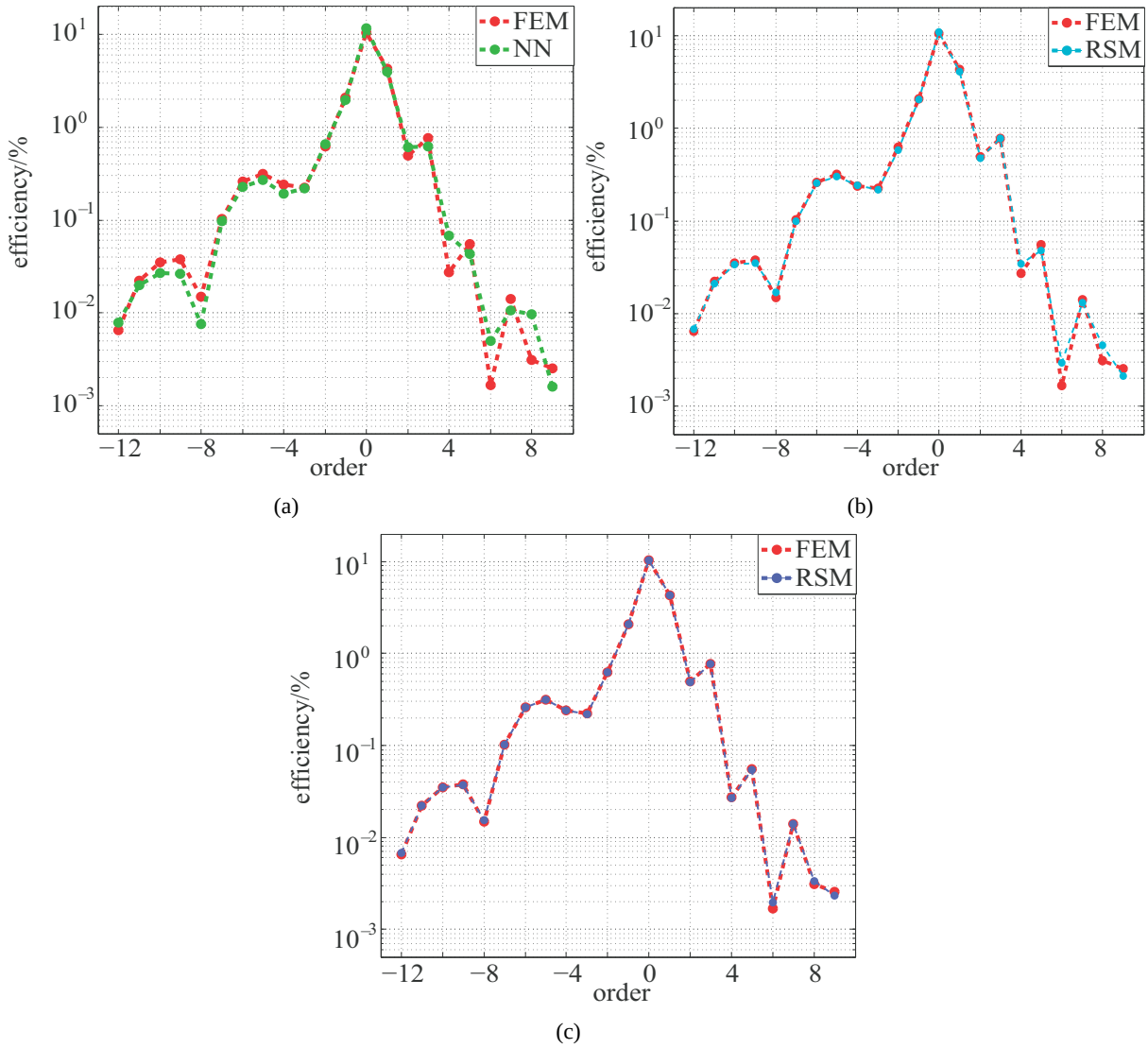

FIG. 2: Comparison of efficiencies calculated rigorously by FEM and by a surrogate approximation: (a) Nearest-neighbor interpolations, (b) response surface methodology, and (c) polynomial chaos expansion.

To estimate the speedup of each model, we measured the time of function evaluations. Since the construction of the surrogate model requires the rigorous calculation of function values of the 700 training points, there is a pre-calculation time of about 20 h for each model. For applications like Bayesian inference or sensitivity analysis, large sampling sizes often have to be used, and the computational cost of the forward model becomes significant. In Table 2 we compare the time duration of all models with rigorous FEM calculations. For the estimation of one parameter set within the Bayesian approach using FEM calculations, a time of more than 6 months (one core) is needed. In contrast, all surrogate models need less than an hour. Note, there is a strong speedup of PC for the sensitivity analysis due to the special structure of the expansion [20].

The advantage of a surrogate model over rigorous FEM calculations is the speedup, but the disadvantage is the approximation error. In Fig. 5 we compared the approximation error of each surrogate model with a typical measurement error. The green line shows the standard deviation ($1.96\sigma_j$) of the measurement error with parameter values $a = 0.02$

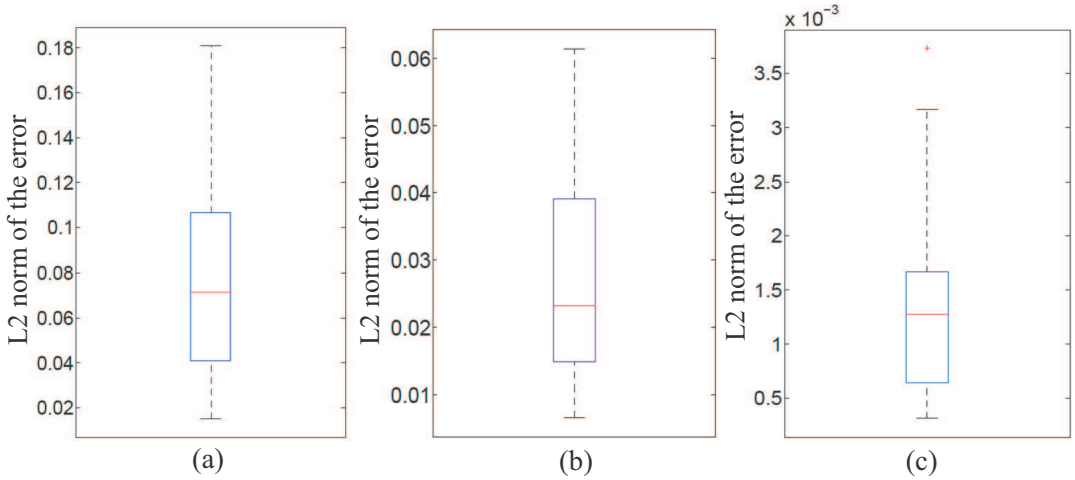

FIG. 3: Boxplots of the L^2 norm of the difference between rigorous FEM calculations and surrogate approximations for 100 geometries chosen randomly from the prior distributions: (a) nearest-neighbor interpolation, (b) response surface methodology, and (c) polynomial chaos expansion.

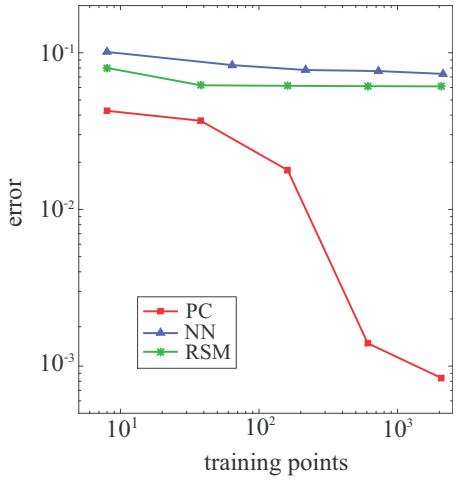

FIG. 4: Convergence rate of each surrogate model. The error was determined as the maximum of 100 sample geometries of the L^2 norm between rigorous FEM calculations and surrogate approximations.

TABLE 2: Speedup for different methods. All surrogate methods need a pre-calculation time of about 20 h to determine all hyperparameters

Method	Function evaluations	FEM	PC	RSM	NN
Sensitivity analysis	3×10^4	41 d	20 h + 0.1 s	20 h +9.0 min	20 h + 5 min
Bayesian MCMC	$\sim 10^5$	139 d	20 h + 47 min	20 h + 30 min	20 h + 17 min
Maximum likelihood	30	1 h	20 h + 0.84 s	20 h + 0.54 s	20 h + 0.3 s
Least squares	15	0.6 h	20 h + 0.42 s	20 h + 0.27 s	20 h + 0.15 s

and $b = 0.002\%$ of the model given by Eq. (12). The red curve shows the absolute value of the approximation error. With the number of training points chosen, the approximation error of the RSM and the NN method is comparable or sometimes even larger than the typical measurement error in EUV scatterometry. We obtained similar results for

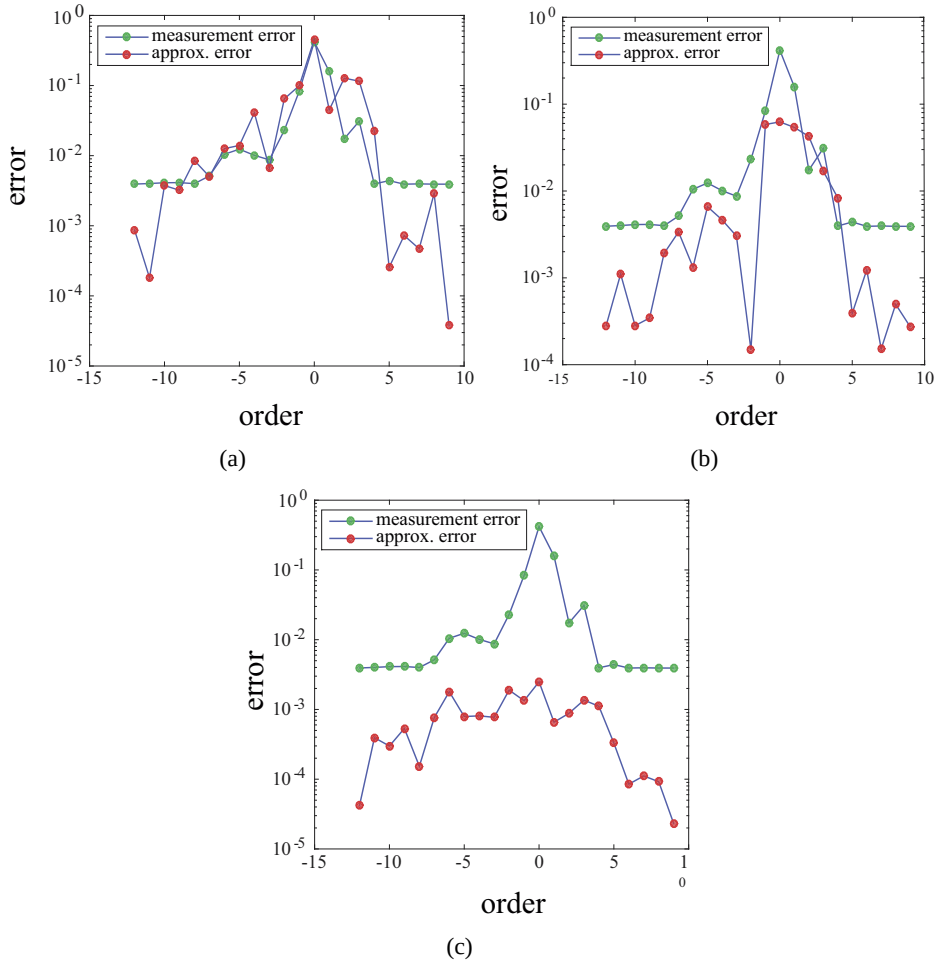

FIG. 5: Comparison of the approximation error with the measurement error. The approximation error is given by error $= |y_j^{\text{FEM}} - y_j^{\text{surrogate}}|$ and the measurement error is given by error $= 1.96\sigma_j$ [Eq. (12) with $a = 0.02$ and $b = 0.002\%$]. We used the geometry as described in the next section. (a) Nearest-neighbor interpolations, (b) response surface methodology, and (c) polynomial chaos expansion.

different geometries (results are not shown here). The comparison shows that only the surrogate model based on PC provides approximation errors smaller than typical measurement errors.

6.2 Parameter Estimation and the Determination of Uncertainties

In the previous section, we have studied the approximation error of our surrogate models. In this paragraph, we apply the surrogate models to Bayesian inference. For fast computations, we approximate the forward model by a surrogate model. For testing, we simulated a set of measurement data by choosing randomly a geometry from the prior distribution, calculated the corresponding efficiencies rigorously, and superimposed noise accordingly to the error model Eq. (12). We have chosen $a_{\text{m}} = 0.02$ and $b_{\text{m}} = 0.002\%$ accordingly to experiments in EUV-scatterometry [3]. We used MCMC to estimate geometry parameters and determine its uncertainties from the approximated posterior distribution. MCMC methods are algorithms for sampling from a probability distribution by the construction of a Markov chain which has the desired distribution as its equilibrium. There are different methods to construct a Markov chain. We have chosen the Metropolis-Hastings algorithm [37, 38]. For diagnostics, we have inspected the rejection

rates and autocorrelation functions. We observed a good mixing and no artifacts for all chains. Furthermore, we applied the Gelman-Rubin criterion [39] to our MCMC simulations. Here, different Markov chains are constructed and the variances in the chains are compared with the variance between chains. When $\bar{y}_j$ is the mean of a chain j and $\bar{y}$ the mean of all chains, we can determine the variance within the chain

$$B_n = \frac{1}{M-1}\sum_{j=1}^{M}(\bar{y}_j - \bar{y})^2 \tag{23}$$

and between chains

$$W = \frac{1}{M(N-1)}\sum_{ij}\sum_{ij}(y_j^i - \bar{y}_j)^2, \tag{24}$$

where N and M are the number of samplings and the number of chains, respectively. The symbol y_j^i is the value of the ith sampling step and of the jth chain. The Gelman-Rubin coefficient R is then determined by

$$R = \frac{[(N-1)/N]W + [(M+1)/M]B_n}{W}. \tag{25}$$

R is always larger than 1 by construction. For convergence $R < 1.03$ is needed. Note that the Gelman-Rubin criterion is not a sufficient criterion to decide if the equilibrium distribution is reached. In Table 3 we summarized the Gelman-Rubin coefficients for Markov chains constructed for the approximated posterior distributions for each surrogate model. We used four chains and 25×10^3 sampling steps for each chain. All Gelman-Rubin coefficients are clearly below 1.03.

To estimate the geometry parameters and uncertainties from the posterior distribution we construct Markov chains with 2×10^5 steps. For the burn-in phase a sampling size of 25,000 was used according to our diagnostics result.

We estimated all parameters by calculating the means of the posterior distributions, and we determined the uncertainties by the 95% credible intervals of the distributions for 20 reconstructions. Figure 6 depicts results of the estimates and the associated uncertainties for the bottom-CD, height and the SWA. The error bars indicate credible intervals and can be asymmetric if the underlying distribution is asymmetric. For all data sets, the geometry was the same (true geometry). Differences appear due to randomly distributed error according to Eq. (12). The true geometry is marked by the red dashed lines. The green dashed lines give the mean value for the set of estimates.

In the case of the NN method, Markov chains are drifting to the boundary of the support of the prior. The true bottom-CD was not estimated, and uncertainties are not consistent with the true values. Estimates using the RSM method are consistent, but the results are not very precise and exhibit large standard uncertainties. In addition, there is a large bias which is independent of the simulated measurement noise. In the case of the PC expansion results are almost all consistent, standard uncertainties are small and there is no bias.

So far, we discussed estimates of the geometry parameters and credible intervals, but the Markov chain encodes also information about the entire posterior distribution. In Fig. 7, projections of the posterior distribution onto different parameter axes are depicted. The supports of the distributions are rescaled to $[-1, 1]$ for convenience. The green dashed lines show the true values of parameters. All distributions are non-Gaussian, which makes the application of the

TABLE 3: Application of the Gelman-Rubin convergence criterion. For the calculation, we used four chains with 25000 sampling steps. For convergence a value of $R < 1.03$ should be reached

Surrogate	R
Nearest-neighbor interpolation	1.0014
Response surface methodology	1.0052
Polynomial chaos approximation	1.0157

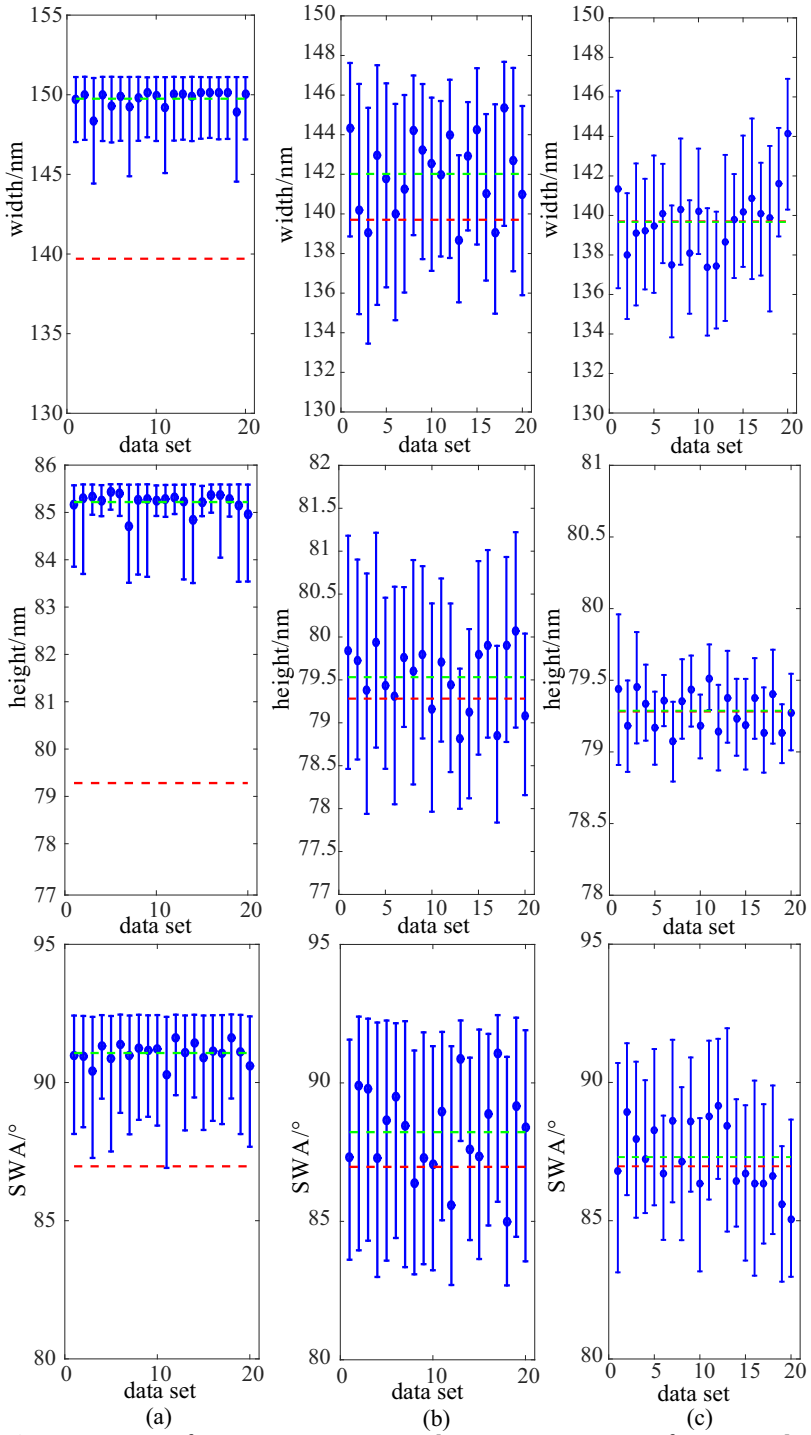

FIG. 6: (color online) Estimations of geometry parameters by using Bayesian inference and surrogate models: (a) nearest neighbor method, (b) response surface methodology, and (c) polynomial chaos expansion. For each surrogate model, 2×10^5 MCMC samplings are used. The burn-in was set to 25,000 samplings. Estimates are obtained from a data set of 20 simulated measurements. Red dashed line is the true geometry, and the green dashed line is the average of 20 estimates.

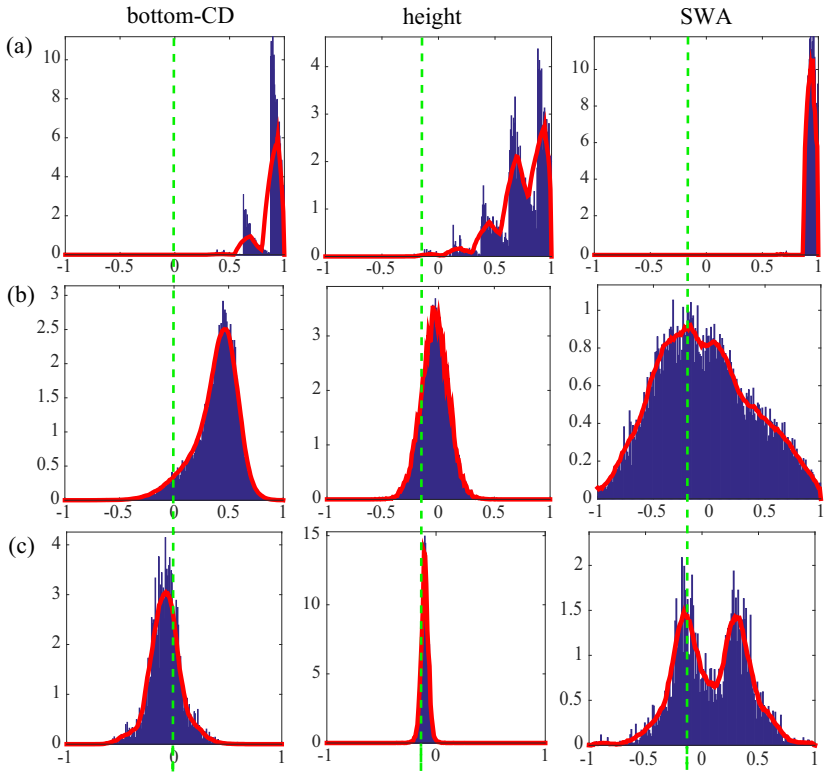

FIG. 7: (color online) Projections of the posterior distribution displayed by histograms for (a) nearest-neighbor method, (b) response surface methodology, and (c) polynomial chaos expansion. The parameter intervals are scaled to $[-1, 1]$ for convenience by 11.2 nm for the bottom-CD, 5.6 nm for the height, and 5° for the SWA. The true parameter values are shown by dashed green lines.

standard deviation as a measure for uncertainty disputable. The distributions obtained by NN interpolations are close to the boundary of the imposed approximation interval and show some maxima. This indicates that approximation errors are too large, and a sufficient solution of the inverse problem is not possible. The resolution of reconstructed parameters is determined by the cubes' length. We have chosen 729 cubes for the three-dimensional parameter space which defines intervals of about 2.5 nm for the bottom-CD. This means that the error of the estimated geometries is up to 2.5 nm (0.22 when scaled to $[-1, 1]$).

Distributions obtained from the RSM display mean values which are close to the true values of geometrical parameters, but show still large variations. For the bottom-CD the distribution is asymmetric and shifted to larger values. The asymmetry comes not from a distortion of the Gaussian distribution of the measurement error, but from the approximation error of the forward model. The approximation error of the RSM was larger than that of the PC expansion (see Fig. 3) which yields larger uncertainties incorporated in wider distributions. The distributions for the PC expansion have only a very small bias and smaller standard deviations then the RSM method. Surprisingly, the distribution of the sidewall angle is bimodal, which comes from the approximation error (see next section). Even if this phenomenon yields larger credible intervals and may lead to a bias of 2.5°, the related uncertainties cover a range that includes the true value.

The projections of the posterior distribution on a surface spanned by two parameters are shown in Fig. 8. The cross depicts the value of the true geometry. All the aspects discussed above are recovered. In addition, the anisotropy of the distribution reflects different sensitivity to the parameters. It is known from previous studies that the sensitivity of the light-diffracted intensities against variations of the sidewall angle is small. Therefore, the uncertainties of the SWA are always large.

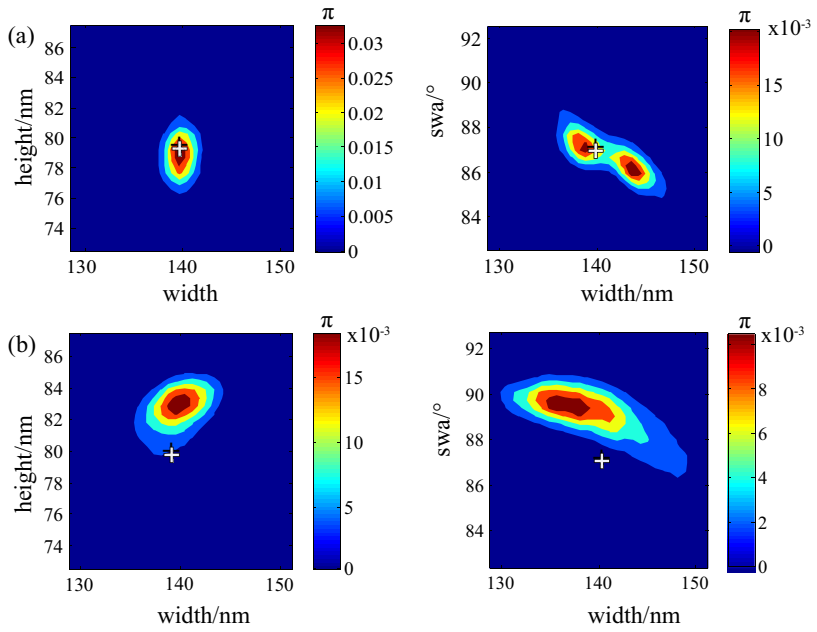

FIG. 8: Examples of projections onto the plane spanned by two parameters. The posterior was approximated by (a) response surface methodology and (b) polynomial chaos expansion. The cross labels the true geometry.

7. APPROXIMATION ERROR

So far, we obtained grating geometry parameters from the posterior distribution and identified related uncertainties by its 95% credible intervals. These uncertainties include the measurement noise as well as the approximation error. Now, the effect of the approximation error on the uncertainties is briefly discussed. For this purpose, we set $a_m = b_m = 0$ and use MCMC sampling to obtain the posterior distribution in the same way as explained in the previous section. Figure 9 shows estimates of the height together with its 95% credible intervals. It is clearly seen that the approximation errors of the RSM method yield large uncertainties of the estimated parameters. In addition the bias

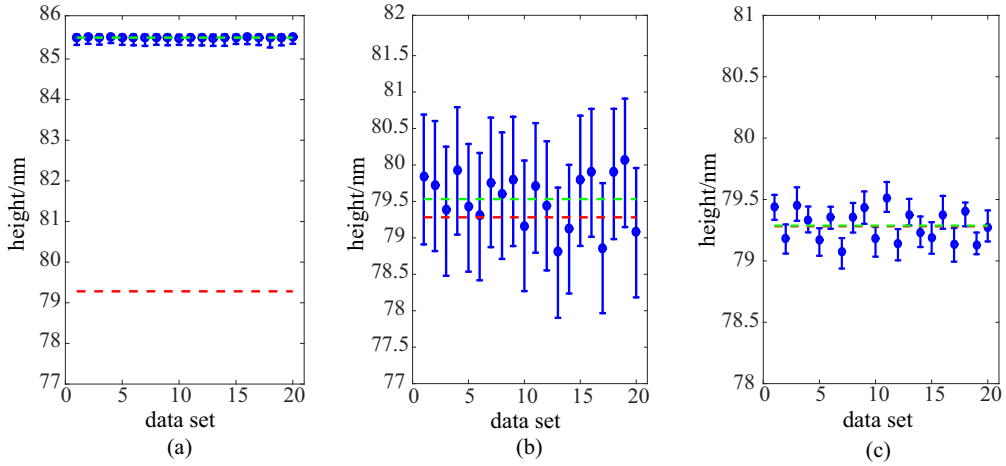

FIG. 9: Reconstructions of the height with uncertainty (95% credible interval) from simulation data without noise. The uncertainties shown are a result of the approximation error. (a) Nearest-neighbor method, (b) response surface methodology, and (c) polynomial chaos expansion.

appears already when only approximation errors are present. This shows that approximation errors yield a contribution of uncertainties and an additional bias. The approximation error of NN yield already to a strong bias which yields to failed reconstructions. Uncertainties are inconsistent with the true values.

In Table 4 the results of the error parameters are shown. We calculated the error parameters from the posterior distribution with measurement error ($a_{\text{m}} = 0.02$ and $b_{\text{m}} = 0.002\%$) and without measurement error ($a_{\text{m}} = 0.0$ and $b_{\text{m}} = 0.0\%$), respectively. Using Eq. (10), we can estimate the error of the simulated measurements and can compare it to the chosen values. The values of a_{approx} and b_{approx} for the RSM and NN are large due to the approximation error. However, estimates for a_m and b_m are in a good agreement with the chosen values for all methods.

8. CONCLUSIONS

In this article, we introduced and constructed three different surrogate models (NN, RSM, PC) to approximate the forward model in scatterometry. We validated the surrogate models by FEM calculations. We calculated the approximation error for each surrogate model and studied its convergence. For the same number of training points the approximation error was large for NN interpolations, moderate for the RSM, and small for the PC expansion. In the second part of the article, we used the constructed surrogate models to accelerate Bayesian inference and to estimate geometry and noise parameters from 20 simulated measurements by using a MCMC method (Metropolis-Hastings). Uncertainties are determined by the calculation of 95% credible intervals. We found a reasonable agreement and consistency with the geometry used for simulated measurements for the RSM and the PC method. The NN method fails completely. This shows that the amount of training points was not enough for a proper approximation. The RSM method shows large uncertainties for all parameters and a bias for some of them. The uncertainties of the estimated parameters obtained by the surrogate model based on the PC expansion are quite small and reconstructed parameters are accurate and without any additional bias.

Furthermore, we investigated the posterior distribution by the consideration of its projections. The distributions display the same features as discussed above. The distributions obtained by the NN method have several peaks and MCMC chains are drifting to the boundaries of the prior. Hence, the posterior is dense close to the interval boundaries, and the algorithm fails to reconstruct the geometry parameters properly. The PC surrogate model has shown the best performance. The posterior distribution was close by the true parameter values. However, surprisingly, the distribution of the side wall angle was bimodal with narrow peaks. The bimodality introduces an additional uncertainty of about 2.5°. We observed also bimodal distributions for the RSM method not shown here. The discussion shows that the approximation errors affect the shape of the posterior distribution in all cases.

In this article, we have shown that the Bayesian approach to computationally expensive systems like scatterometry is feasible if surrogate models are used. Among the three models analyzed, the polynomial chaos expansion was the most efficient approximation. More future work is needed to extend parameter intervals and to apply the algorithms to experimental measurements.

ACKNOWLEDGMENTS

This work was funded through the European Metrology Research Program (EMRP) Project NEW04 Uncertainty. The EMRP is jointly funded by EMRP partnering countires within EURAMET and the European Union. S.H. tanks Katy Klauenberg for helpful discussions.

TABLE 4: Error parameters a, b and $a_{\text{approx}}, b_{\text{approx}}$ are calculated from the posterior distribution. a_{m} and b_{m} are calculated with Eq. (10). For simulations of the measurements we used an error of $a_{\text{m}} = 0.02$ and $b_{\text{m}} = 0.002\%$

Surrogate	a	a_{approx}	a_{m}	b	b_{approx}	b_{m}
RSM	0.03666	0.03096	0.01932	0.002670%	0.00159%	0.002146%
PC	0.02467	0.01160	0.02177	0.002334%	0.00113%	0.0020434%
NN	0.02634	0.01210	0.02340	0.002613%	0.00114%	0.0023513%

REFERENCES

1. Huang, H. and Terry, Jr., F., Spectroscopic ellipsometry and reflectometry from gratings (Scatterometry) for critical dimension measurement and in situ, real-time process monitoring, *Thin Solid Films*, 455-456:828–836, 2004.
2. Raymond, C., Murane, M., Prins, S., Sohail, S., Navi, H., Mc Neil, J., and Horsch, J., Multiparameter grating metrology using optical scatterometry, *J. Vac. Sci. Technol.*, B15:361–368, 1997.
3. Henn, M.-A., Gross, H., Scholze, F., Wurm, M., Elster, C., and Bär, M., A maximum likelihood approach to the inverse problem of scatterometry, *Opt. Express*, 20:12771–12786, 2012.
4. Henn, M.-A., Heidenreich, S., Gross, H., Rathsfeld, A., Scholze, F., and Bär, M., Improved grating reconstruction by determination of line roughness in extreme ultraviolet scatterometry, *Opt. Lett.*, 37(24):5229–5231, 2012.
5. Henn, M.-A., Gross, H., Heidenreich, S., Scholze, F., Elster, C., and Bär, M., Improved reconstruction of critical dimensions in extreme ultraviolet scatterometry by modeling systematic errors, *Meas. Sci. Technol*, 25(4):044003, 2014.
6. Gross, H., Henn, M.-A., Heidenreich, S., Rathsfeld, A., and Bär, M., Modeling of line roughness and its impact on the diffraction intensities and the reconstructed critical dimensions in scatterometry, *Appl. Optics*, 51:7384–7394, 2012.
7. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, Joint Committee for Guides in Metrology, JCGM, Evaluation of measurement data—Supplement 1 to the Guide to the expression of uncertainty in measurement—Propagation of distributions using a Monte Carlo method, 100:2008, 2008.
8. Heidenreich, S., Gross, H., Henn, M.-A., Elster, C., and Bär, M., A surrogate model enables a Bayesian approach to the inverse problem of scatterometry, *J. Phys. Conf. Series*, 490:012007, 2014.
9. Gross, H., Model, R., Bär, M., Wurm, M., Bodermann, B., and Rathsfeld, A., Mathematical modelling of indirect measurements in scatterometry, *Measurement*, 39:782–794, 2006.
10. Tarantola, A., *Inverse Problem Theory*, SIAM, Philadelphia, 2005.
11. Ahmed, M. Y. M. and Quin, N., Comparison of Response Surface and Kriging Surrogates in Aerodynamic Design Optimization of Hypersonic Spiked Blunt Bodies, in *13th Intl. Conf. on Aerospace Sciences and Aviation Technol.*, ASAT-13-AE-15, May 26–28, 2009.
12. Simpson, T. W., Comparison of Response Surface and Kriging Models in the Multidisciplinary Design of an Aerospike Nozzle, NASA/CR-1998-206935, ICASE98-16, Technical Report 19980046640, 1998.
13. Matheron, G., Principles geostatistics, *Econ. Geol.*, 58:1246–1266, 1963.
14. Keane, A. J. and Prasanth, B. N., *Computational Approaches for Aerospace Design: The Pursuit of Excellence*, John Wiley, West Sussex, 2005.
15. Cressie, N., The origins of Kriging, *Math. Geol.*, 22(3):239–252, 1990.
16. Wang, G. G., Dong, Z., and Aitchison, P., Adaptive response surface method—A global optimization scheme for computation-intensive design problems, *J. Eng. Optim.*, 33:707–734, 2001.
17. Wang, G. G., Adaptive response surface method using inherited Latin hypercube design points, *J. Mech. Des.*, 125:210–220, 2003.
18. Isukapalli, S. S., Roy, A., and Georgopoulos, P. G., Stochastic response surface methods (SRSMs) for uncertainty propagation: Application to environmental and biological systems, *Risk Anal.*, 18(3):351–363, 1998.
19. Iooss, B., Boussouf L., Feuillard, V., and Marrel, A., *Safety, Reliability and Risk Analysis: Theory Methods and Applications*, Vol. 1-4, pp. 2135–2141, 2009.
20. Sudret, B., Global sensitivity analysis using polynomial chaos expansions, *Reliab. Eng. Syst. Safety*, 93(7):964–979, 2008.
21. Jeong, S., Murayama, M., and Yamamoto, K., Efficient optimization design method using Kriging model, *J. Aircraft*, 42(2):413–420, 2005.
22. Rasmussen, C. E. and Williams, C. K. I., *Gaussian Process for Machine Learning*, MIT Press, Cambridge, MA, 2006.
23. Box, G. E. P. and Draper, N., *Response Surfaces, Mixtures, and Ridge Analysis*, Wiley, Hoboken, NJ, 2007.
24. Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M., *Response Surface Methodology*, Wiley, Hoboken, NJ, 2009.
25. Myers, R. H., Khuri, A. I., and Carter, W. H., Response surface methodology: 1966–1988, *Technometrics*, 31(2):137–157, 1989.

26. Debusschere, B. J., Najm, H. N., Pebay, P. P., Knio, O. M., Ghanem, R. G., and Le Maitre, O. P., Numerical challenges in the use of polynomial chaos representations for stochastic processes, *SIAM J. Sci. Comput.*, 26(2):698–719, 2004.
27. Le Maitre, O. P., Knio, O. M., Najm, H. N., and Ghanem, R. G., A stochastic projection method for fluid flow: I. Basic formulation, *J. Comput. Phys.*, 173(2):481–511, 2001.
28. Wan, X. and Karniadakis, G. E., An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, *J. Comput. Phys.*, 209:617–642, 2005.
29. Lovett, T. E., Ponci, F., and Monti, A., A polynomial chaos approach to measurement uncertainty, *IEEE Trans. Instrumentation Meas.*, 55(3):729–736, 2006.
30. Marzouk, Y. M., Najm, H. N., and Rahn, L. A., Stochastic spectral methods for efficient Bayesian solution of inverse problems, *J. Comput. Phys.*, 224(2):560–586, 2007.
31. Marzouk, Y. M. and Najm, H. N., Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems, *J. Comput. Phys.*, 228(6):1852–1902, 2009.
32. Marzouk, Y. M. and Xiu, D., A stochastic collocation approach to Bayesian inference in inverse problems, *Commun. Comput. Phys.*, 6(4):826–847, 2009.
33. Wiener, N., The homogeneous chaos, *Am. J. Math.*, 60(4):897–935, 1938.
34. Cameron, R. H. and Martin, W. T., The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals, *Ann. Math.*, 48(2):385–392, 1947.
35. Xiu, D. and Karniadakis, G. E., Modeling uncertainty in flow simulations via generalized polynomial chaos, *J. Comput. Phys.*, 187:137–167, 2002.
36. Nobile, F., Tempone, R., and Webster, C. G., A sparse grid stochastic collocation method for partial differential equations with random input data, *SIAM J. Numer. Anal.*, 46(5):2309–2345, 2008.
37. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., Equations of state calculations by fast computing machines, *J. Chem Phys.*, 21:1087–1091, 1953.
38. Hastings, W. K., Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57:97–109, 1970.
39. Gelman, A. and Rubin, D. B., Inference from iterative simulation using multiple sequences, *Stat. Sci.*, 7:457–472, 1992.