Contrasting Conventional and Machine Learning Approaches to Optical Critical Dimension Measurements

Bryan M. Barnes*[1] and Mark-Alexander Henn[2]

[1]Nanoscale Device Characterization Division, Physical Measurement Laboratory
[2]Applied and Computational Mathematics Division, Information Technology Laboratory
National Institute of Standards and Technology 100 Bureau Drive MS 8423
Gaithersburg, MD USA 20899-8423

Abstract

Accurate, optics-based measurement of feature sizes at deep sub-wavelength dimensions has been conventionally challenged by improved manufacturing, including smaller linewidths, denser layouts, and greater materials complexity at near-atomic scales. Electromagnetic modeling is relied upon heavily for forward maps used to solve the inverse problem of optical measurements for parametric estimation. Machine learning (ML) approaches are continually under consideration, either as a means to bypass direct comparison to simulation or as a method to augment nonlinear regression. In this work, ML approaches are investigated using a well-characterized experimental data set and its simulation library that assumes a 2-D geometry. The benefits and limitations of ML for optical critical dimension (OCD) metrology are illustrated by comparing a straightforward library look-up method and two ML approaches, a data-driven surrogate model for nonlinear regression using radial basis functions (RBF) and multiple-output Gaussian process regression (GPR) that indirectly applies the simulated intensity data. Both RBF and GPR generally improve accuracy over the conventional method with as few as 32 training points. However, as measurement noise is decreased the uncertainties from RBF and GPR differ greatly as the GPR posterior estimate of the variance appears to overestimate parametric uncertainties. Both accuracy and uncertainty must be addressed in OCD while balancing simulation versus ML computational requirements.

1. INTRODUCTION

Inexpensive, non-destructive, and relatively fast optical approaches permeate overlay metrology,[1,2] defect inspection,[3,4] and critical dimension (CD) metrology[5] in advanced semiconductor manufacturing. Unlike defect and overlay metrologies, optical CD (OCD) metrology has been strongly dependent upon electromagnetic simulations to yield forward maps for the inverse problem of determining CDs and optical properties from intensity measurements. The downward scaling of semiconductor features has prompted not only improvements in electromagnetic simulations[6] but also reductions in measurement wavelength[7] and the rigorous incorporation of prior knowledge (e.g., hybrid metrology[8–12]) to extend the continued utility of OCD even as dimensions are deep sub-wavelength and approaching near-atomic scales.

The clear goal of OCD of sub-wavelength features is determining the set of parameters, $\mathbf{p} = \{p_1, \ldots p_T\}$, where T is the number of free parameters in the simulation, that best represent the measurand. Generally, optical data are collected for combinations of several measurement conditions (e.g., angle, wavelength, polarization, *etc.*) that can be labeled as $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$, where t is the total number of conditions. The optical data collected at each $\boldsymbol{\omega}_i$ can be classified as $\mathbf{I} = (I_1, \ldots, I_t)^\mathsf{T}$ and the simulated data are $\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) = [f(\boldsymbol{\omega}_1, \mathbf{p}), \ldots, f(\boldsymbol{\omega}_t, \mathbf{p})]^\mathsf{T}$. Most often, nonlinear regression is utilized to determine the estimate $\hat{\mathbf{p}}$ from the relationship

$$\mathbf{I} = \mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}$ is an error term.

To improve OCD further, researchers and technologists are contemplating the use of artificial intelligence (AI) and machine learning (ML) towards determining $\hat{\mathbf{p}}$ with two divergent approaches emerging: First, one

*bmbarnes@nist.gov

could abandon the long-held practice that experimental data need to be connected directly to simulation data. In simplest terms, the measurements $\mathbf{I}$ could be fed directly into ML to yield $\hat{\mathbf{p}}$. Second, ML could be centered on improving existing regression approaches for solving Eq. 1. In selecting between these two approaches for metrology, it is crucial that ML needs to yield not only $\hat{\mathbf{p}}$ but also the variances $\sigma^2_{\hat{\mathbf{p}}} = \{\sigma^2_{\hat{p}_1}, \ldots, \sigma^2_{\hat{p}_T}\}$. Of the first approach, Bischoff *et al.* reported fitting of 0.25 μm patterns using neural networks, but no uncertainties were reported.[13] Robert *et al.* reported an estimation of values using neural networks and also variance estimation using a second neutral network, but they clarify "variables are assumed to be not correlated". Rana *et al.* indicated industrial viability with reports of uncertainty using neural networks but without specific derivation.[14]

The second approach, the use of ML approaches to augment regression, has recently gained attention. Hammerschmidt *et al.* employed a Bayesian approach with a Newton-like method to quantify the parameters.[15] Heidenreich *et al.* reported a polynomial chaos based surrogate model for EUV scatterometry measurements with errors determined from Markov-Chain Monte Carlo sampling of the posterior distribution.[16] Later, Hammerschmidt *et al.* reported the use of Gaussian processes in Bayesian optimization, also referred to as Gaussian process regression in that work, using a Taylor expansion to quantify uncertainties in scatterometry.[17] In each of the works, results for the values were comparable with prior approaches at fitting.

In this work, previously published data are reprocessed to consider the challenges and opportunities of both the first and second approaches. The implementation of Gaussian process regression (GPR) presented here (unlike in Refs.[17,18]) circumvents direct application of Eq. 1 to proceed straight from $\mathbf{I}$ to $\hat{\mathbf{p}}$ through ML, although the ML training and validation are completely informed by rigorous electromagnetic simulations. GPR is a well-studied ML approach notable in that GPR can simultaneously calculate both a mean and a variance using multiple outputs (here, geometrical parameters) even with correlation.[19] Alternatively, radial basis functions (RBF) are trained and validated to augment the solution of Eq. 1 to yield improved values for the parametric means and variances.[20] These two approaches will be compared against a straightforward application of a conventional library look-up method. For the ML approaches, the complexity and computational expense of modern OCD modeling will be considered through the lens of data scarcity, a reduction in the number of training points, or n_{TP}, that are available in the existing library. Comparisons will show that even with smaller values of n_{TP}, both GPR and RBF yield generally good values for $\hat{\mathbf{p}}$, that RBF yields relatively small uncertainties based on its variance, but that this GPR method as presented appears to overestimate parametric uncertainties.

2. EXPERIMENTAL TECHNIQUE AND SAMPLE DETAILS

The experimental data used in this work were collected from scatterfield microscopy measurements in an angular scan mode. Scatterfield microscopy combines sophisticated illumination engineering in a high-magnification imaging platform with optimized information collection about targets of interest from the full 3-D electromagnetic scattered field. Overviews of the method are available as Ref.[22,23] Similar approaches have been deemed "micro-scatterometry".[24,25] Specifically, using Köhler illumination as illustrated in Fig. 1, an aperture in the conjugate to the back focal plane of the objective lens leads to an angularly resolved illumination beam (illumination NA

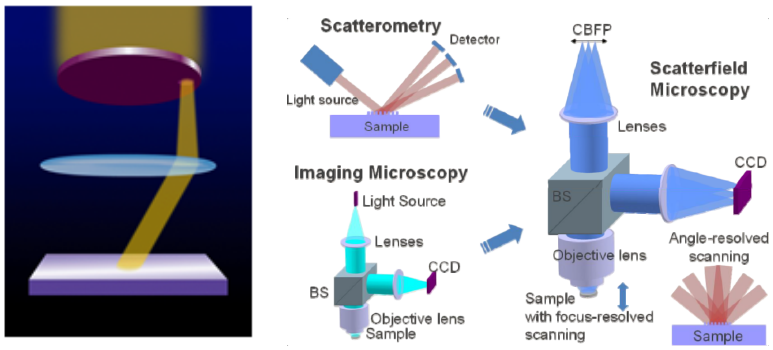

Figure 1. Angularly resolved scatterfield microscopy. (left) Schematic showing angularly resolved illumination. (right) Angle-scanning capabilities due to an aperture in the conjugate to the back focal plane (CBFP). Reprinted from Ref.[21]

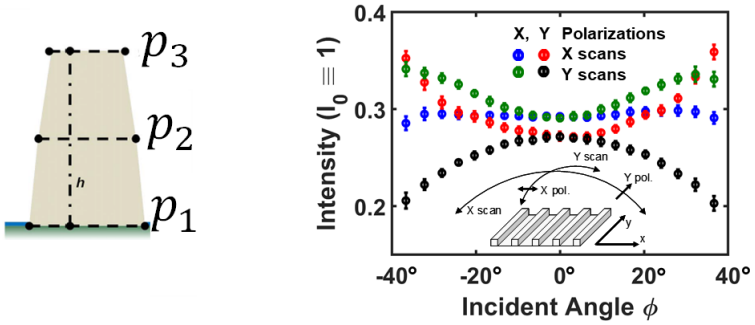

Figure 2. L100P300 Parameterized geometry and example data set. Measured intensities as functions of incident polarization and scan direction, with X, Y defined relative to line direction. Intensity here is unitless after normalization by the incident intensity $I_0 \equiv 1$. Error bars are 1σ uncertainties in the measured values. After Ref.[27]

≈ 0.13) at the sample plane. This light is either reflected or scattered depending on the characteristics of the sample. The large collection NA ≈ 0.95 allows the capture of scattered light between $\phi = -72°$ to $72°$. For these experiments, the field-of-view of collection path was focused solely on the target of interest, a periodic array of patterned lines. The experimental wavelength is $\lambda = 450$ nm

The patterned structures under test are three scatterometry targets patterned using a focus/exposure matrix to yield variations in linewidth. The specific target is the "L100P300" target produced by SEMATECH,[26] with the notation implying a nominal line width of 100 nm and a period of 300 nm. These data have been of great utility, having been reported for initial demonstrations of Scatterfield Microscopy,[22] in pioneering work on hybrid metrology,[10,11] as well as in a recent investigation of the treatment of potential experimental bias.[27] In these works, the geometry has been parameterized using a dual trapezoid model yielding three floating parameters, with a fixed height. Dimensional measurements from this sample have been reported from scanning electron microscopy (SEM)[22] and from atomic force microscopy (AFM).[8] Although its dimensions far exceed those of current technology nodes, comparisons can be made among the ML results as well as these prior data. Furthermore, with a large library of over 2000 simulated $\mathbf{f_I}(\boldsymbol{\omega}_i, \mathbf{p})$ in-hand, multiple realizations of sub-sampling of that library have been performed to yield additional rigor to the observed trends in accuracy and uncertainty.

3. METHOD DERIVATIONS

In this section, summaries describing the estimation of the parametric values $\hat{\mathbf{p}}$ and their uncertainties $\sigma_{\hat{\mathbf{p}}}$ are presented. Other sources cover the derivation of the RBF and GPR approaches more thoroughly (see the documentation for Ref.[20] and Ref.,[19] respectively).

3.1 Commonalities between Library look-up and RBF

Both library look-up and RBF utilize a weighted least squares fit. The weights are based on the known measurement errors, $\sigma_{\mathsf{meas},i}$, at each of the $\boldsymbol{\omega}_i, i = 1, 2, \cdots, t$ for the measurements I_i. In this work, $T = 84$ by concatenating the four angle scans in Fig. 2 into a single vector. This approach leads to the function

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{t} \frac{1}{\sigma^2_{\mathsf{meas},i}} |f(\boldsymbol{\omega}_i, \mathbf{p}) - I_i|^2, \tag{2}$$

the weighted χ^2 (chi-square) function, which can also be expressed as:

$$\chi^2(\mathbf{p}) = \left[\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) - \mathbf{I}\right]^\intercal \mathbf{V}^{-1} \left[\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) - \mathbf{I}\right], \tag{3}$$

with

$$\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) = \left[f(\boldsymbol{\omega}_1, \mathbf{p}), f(\boldsymbol{\omega}_2, \mathbf{p}), \ldots, f(\boldsymbol{\omega}_t, \mathbf{p})\right]^\intercal, \ \mathbf{I} = (I_1, I_2, \cdots, I_t)^\intercal \tag{4}$$

and for this work,

$$\mathbf{V}^{-1} = \left(\delta_{ij} \frac{1}{\sigma^2_{\mathsf{meas},i}} \right)_{i,j=1,\ldots,t} \in \mathbb{R}^{t \times t}, \tag{5}$$

and as uncorrelated errors are assumed here $\mathbf{V}^{-1}$ is a diagonal matrix.

In addition, it can be shown that the uncertainties in $\hat{\mathbf{p}}$ can be calculated using the covariance matrix

$$\mathbf{\Sigma} = \left(\mathbf{J}_{\mathbf{f_I}}^{\intercal} \mathbf{V}^{-1} \mathbf{J}_{\mathbf{f_I}} \right)^{-1}, \tag{6}$$

where $\mathbf{J}_{\mathbf{f_I}}$ is the Jacobian of the model function. It is through Eqs. 5 and 6 that Library Look-up and RBF are both sensitive directly to measurement noise.

3.2 Library look-up

In the library look-up scheme, χ^2 is evaluated at a subset of $\mathbf{f_I}(\boldsymbol{\omega}_i, \mathbf{p})$ for which $\mathbf{p}$ are aligned on a grid such that for a given p_j there exists nearest neighbors (NN_-) and (NN_+) such that $p_{j(NN_-)} < p_j < p_{j(NN_+)}$ with $p_{i \neq j}$ being equal. The value of $\mathbf{p}$ which yields the smallest χ^2 is in this scheme the best fit $\hat{\mathbf{p}}$. The Jacobian is approximated using

$$\mathbf{J}_{\mathbf{f_I}} = \left(\frac{\partial f_i}{\partial p_j} \right)_{i=1,\ldots,t, j=1,\ldots,T} \in \mathbb{R}^{t \times T}, \frac{\partial f_i}{\partial p_j} \approx \frac{\Delta f_i}{\Delta p_j} \approx \frac{f_{i(NN_+)} - f_{i(NN_-)}}{p_{j(NN_+)} - p_{j(NN_-)}}. \tag{7}$$

Using additional nearest neighbors (if available) may improve this ratio through averaging of f_i, but here these simple approximations are used to quickly assess $\hat{\mathbf{p}}$ and $\sigma_{\hat{\mathbf{p}}}$.

3.3 Radial Basis Function Interpolation

Library look-up does not interpolate the function among points in the simulation domain, and for modern OCD such an extensive library seems impractical. One solution is to avoid evaluating the function directly by interpolating on a grid of already calculated values. Again, the RBFs are calculated using a subset of the entire library made up of n_{TP} entries where TP represents "training points"; this nomenclature comes from the ML community and shall be used for both RBF and GPR. The parameters for which $\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p})$ is evaluated are $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iT}), i = 1, \ldots, n_{\mathsf{TP}}$ with

$$\mathbf{\Pi} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_{n_{\mathsf{TP}}} \end{pmatrix} \in \mathbb{R}^{n_{\mathsf{TP}} \times T}, \mathbf{f_I}(\mathbf{\Pi}) = \mathbf{\Phi} = \begin{pmatrix} \mathbf{f_I}(\mathbf{p}_1)^{\intercal} \\ \vdots \\ \mathbf{f_I}(\mathbf{p}_{n_{\mathsf{TP}}})^{\intercal} \end{pmatrix} \in \mathbb{R}^{n_{\mathsf{TP}} \times t}, \tag{8}$$

where T remains the number of parameters and t the number of measurement conditions, i.e., the length of an individual measurement vector.

The interpolation problem then amounts to find an approximation $\widetilde{\mathbf{f}}_{\mathbf{I}}$ to the function $\mathbf{f_I}$ such that

$$\widetilde{\mathbf{f}}_{\mathbf{I}}(\mathbf{p}, \mathbf{\Pi}, \mathbf{\Phi}) \approx \mathbf{f_I}(\mathbf{p}). \tag{9}$$

For the sake of clarity we will write $\widetilde{\mathbf{f}}_{\mathbf{I}}(\mathbf{p})$ for the function in Eq. 9 if it is clear what library $\{\mathbf{\Pi}, \mathbf{\Phi}\}$ is being used.

Assume we have a set of functions $\rho_i : \mathbb{R}^T \to \mathbb{R},\ i = 1, \ldots, N$, such that for all $j \in \{1, \cdots, t\}$, we can approximate the j-th component of $\mathbf{f_I}$, i.e. the scalar function $f_j(\mathbf{p})$ by

$$f_j(\mathbf{p}) \approx \sum_{i=1}^{N} a_{ji} \rho_i(\mathbf{p}). \tag{10}$$

In the RBF approach specific functions are utilized that depend only on the distance of the parameter vector $\mathbf{p}$ from the different grid points $\mathbf{p}_i$, and an additional hyperparameter r as

$$\rho_i(\mathbf{p}, r) = \rho(\|\mathbf{p} - \mathbf{p}_i\|, r), \ i = 1, \ldots, n_{\mathsf{TP}}, \tag{11}$$

hence $N = n_{\mathsf{TP}}$ in Eq. 10. Specifically, the multiquadratic radial basis functions,

$$\rho_i(\mathbf{p}, r) = \sqrt{\|\mathbf{p} - \mathbf{p}_i\|^2 + r^2}, \ i = 1, \ldots, n_{\mathsf{TP}} \tag{12}$$

have been utilized with the hyperparameter r determined from leave-one-out cross-validation (LOOCV)[28] and the optimal hyperparameter $\hat{r}$ found using a particle-swarm optimization (PSO) algorithm, see Ref.[29] The matrix $\mathbf{A}$ (derived elsewhere[20]), consisting of the t coefficient vectors $\mathbf{a}_i = (a_{i1}, \ldots, a_{in_{\mathsf{TP}}})^\intercal \in \mathbb{R}^{n_{\mathsf{TP}}}$ determined from Eq. 10, given as

$$\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_t)^\intercal \in \mathbb{R}^{t \times n_{\mathsf{TP}}}, \tag{13}$$

is also optimized using PSO. Using $\mathbf{A}$ we can calculate the approximation of all t entries to the model function for an arbitrary parameter vector $\mathbf{p}$ by

$$\widetilde{\mathbf{f}}_{\mathbf{I}}(\mathbf{p}) = \mathbf{A} \cdot \mathbf{P}(\mathbf{p}), \text{ with } \mathbf{P}(\mathbf{p}) = [\rho_1(\mathbf{p}), \ldots, \rho_{n_{\mathsf{TP}}}(\mathbf{p})]^\intercal \in \mathbb{R}^{n_{\mathsf{TP}}}. \tag{14}$$

To quickly summarize, the training step for RBF solves for $\hat{r}$ and $\mathbf{A}$ that are used in the determination of $\widetilde{\mathbf{f}}_{\mathbf{I}}(\mathbf{p})$ which can be shown to yield

$$\hat{\mathbf{p}} = \operatorname{argmin} \left\{ \left[\widetilde{\mathbf{f}}_{\mathbf{I}}(\mathbf{p}) - \mathbf{y} \right]^\intercal \mathbf{V}^{-1} \left[\widetilde{\mathbf{f}}_{\mathbf{I}}(\mathbf{p}) - \mathbf{y} \right] \right\}, \tag{15}$$

where $\mathbf{y}$ is a general form, and is $\mathbf{I}$ specifically here as in Eq. 3. The derivation of the Jacobian of $\widetilde{\mathbf{f}}_{\mathbf{I}}(\mathbf{p})$, $\mathbf{J}_{\widetilde{\mathbf{f}}_{\mathbf{I}}}$, can be found elsewhere,[20] but as with the library look-up, the parametric mean and uncertainty are both directly tied to the measurement uncertainty through $\mathbf{V}^{-1}$.

3.4 Gaussian Process Regression

In this implementation of Gaussian process regression, which has followed the derivation set forth by Liu *et al.*,[19] we rethink the relationships between the geometry and its scattering and attempt a vastly different ML approach. Instead of trying to establish a functional relationship between the geometry parameters and the optical response, i.e., trying to find the function

$$\mathbf{f}_{\mathbf{I}} : \mathbf{p} \mapsto \mathbf{f}_{\mathbf{I}}(\mathbf{p}), \tag{16}$$

and minimize the difference between a particular measurement $\mathbf{y}$ and the function value to determine the optimal value of geometry parameters, we want to establish a direct relationship between measured intensities and geometry parameters, i.e., find a function such that

$$\mathbf{g} : \mathbf{I} \mapsto \mathbf{g}_{\mathbf{p}}(\mathbf{I}), \tag{17}$$

informed by observed measurements and/or simulated values $\mathbf{I}$ at various measurement conditions $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$.

In order to keep the notation simple we will in the following assume that we only want to determine a single parameter value p instead of a vector of parameter values $\mathbf{p}$, before giving a brief explanation on how to address the latter with a multi-output Gaussian process (MOGP). Similar to the previous approach we assume that the observations, that are now the values of the parameter of interest, are noisy realizations of the underlying model, such that

$$p = g_p(\mathbf{I}) + \epsilon_s, \ \epsilon_s \sim \mathcal{N}(0, \sigma_s^2). \tag{18}$$

Note, that there is no relationship between ϵ_s in Eq. 18 and ϵ in Eq. 1, nor is there a relationship between σ_s and the measurement noise σ_{meas} in Eq. 2.

In Gaussian process regression the target function in Eq. 18 is approximated by interpreting it as a probability distribution in a function space, more precisely as a collection of random variables, such that any finite number

of which have a joint Gaussian distribution. As such it is completely defined by its mean function $m(\mathbf{I})$ and the covariance function $k(\mathbf{I}, \mathbf{I}')$. We can therefore write a Gaussian process as

$$g_p(\mathbf{I}) \sim \mathcal{GP}\left[m(\mathbf{I}), k(\mathbf{I}, \mathbf{I}')\right], \tag{19}$$

and furthermore, w.l.o.g., assume the mean function to be zero. In order to reduce the complexity of the problem the choice of the covariance function is limited to a certain class of functions that can be characterized by a few parameters.

In this study, a simple exponential (SE) kernel has been applied, defined here as

$$k_{\mathsf{SE}}\left(\mathbf{I}, \mathbf{I}'\right) = \sigma^2 \mathsf{exp}\left(-\frac{||\mathbf{I} - \mathbf{I}'||}{2l}\right), \tag{20}$$

where the signal variance σ^2 represents an output scale amplitude and l represents a characteristic length scale. Here, σ and l are the hyperparameters ω for this kernel that must be fitted in order to realize the optimal GPR. Indeed, not only is this similar to RBF, but this kernel is itself a radial basis function also.

Because of the Euclidian norm in Eq. 20, a function of the vector difference between two intensity vectors with t elements is reduced to a scalar value. This is both a benefit (by speeding up the calculation) and a challenge (by removing information from the regression). Contrast this with the RBF interpolation, which in Eqn. 11 reduces a function of two vectors of length n_{TP} into a scalar value also. Alternative kernels including non-radial kernels should be explored further for use in GPR-based OCD metrology.

Thus, in the context of Eq. 17, there exists a set of training points $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_{n_{\mathsf{TP}}}\}$ from the input domain with associated output observations $\mathcal{P} = \{p(\mathbf{I}_1), \ldots, p(\mathbf{I}_{n_{\mathsf{TP}}})\}$, that we can arrange in a vector $\mathbf{p} = [p(\mathbf{I}_1), \ldots, p(\mathbf{I}_{n_{\mathsf{TP}}})]^{\mathsf{T}}$. Assume now that $\mathbf{I}$ need not (as defined above) be from simulation but rather comes instead from experiment as $\mathbf{y}$. Again following Ref.[19] and since a Gaussian process is a stochastic process wherein a finite subset of random variables follows a joint Gaussian distribution, the joint prior distribution of the observations $\mathbf{p}$ together with $g_p(\mathbf{y})$ is

$$\begin{bmatrix} \mathbf{p} \\ g_p(\mathbf{y}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathcal{I}, \mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}} & \mathbf{k}(\mathcal{I}, \mathbf{y}) \\ \mathbf{k}(\mathbf{y}, \mathcal{I}) & k(\mathbf{y}, \mathbf{y}) \end{bmatrix}\right), \tag{21}$$

where $\mathbf{K}(\mathcal{I}, \mathcal{I}) \in \mathbb{R}^{n_{\mathsf{TP}} \times n_{\mathsf{TP}}}$ is the symmetric and positive semi-definite covariance matrix with the element $\mathbf{K}_{ij} = k(\mathbf{I}_i, \mathbf{I}_j)$, $\mathbf{k}(\mathcal{I}, \mathbf{y}) \in \mathbb{R}^{n_{\mathsf{TP}}}$ denotes the vector of covariances between the n_{TP} training points and the test point $\mathbf{y}$, and $\mathbf{E}_{n_{\mathsf{TP}}}$ is the identity matrix in $\mathbb{R}^{n_{\mathsf{TP}} \times n_{\mathsf{TP}}}$. Be aware that we understand the term point in a topological sense, since both the training points and the test point are vector quantities.

It can be shown that for a single-output Gaussian process regression, the prediction mean $\hat{g}_p(\mathbf{y})$ and prediction variance $\sigma_p^2(\mathbf{y})$ are

$$\hat{g}_p(\mathbf{y}) = \mathbf{K}(\mathbf{y}, \mathcal{I})^{\mathsf{T}} \left[\mathbf{K}(\mathcal{I}, \mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}\right]^{-1} \mathbf{p}, \tag{22}$$

and

$$\sigma_p^2(\mathbf{y}) = k(\mathbf{y}, \mathbf{y}) - \mathbf{K}(\mathbf{y}, \mathcal{I})^{\mathsf{T}} \left[\mathbf{K}(\mathcal{I}, \mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}\right]^{-1} \mathbf{K}(\mathbf{y}, \mathcal{I}). \tag{23}$$

To use Eqs. 22 and 23 for prediction, we need to infer the hyperparameters $\boldsymbol{\theta}$ in the covariance function k by minimizing the negative log marginal likelihood (NLML) as

$$\boldsymbol{\theta}_{\mathsf{opt}} = \arg_{\boldsymbol{\theta}} \mathsf{min}\ \mathsf{NLML}, \tag{24}$$

where

$$\mathsf{NLML} = -\log \pi(\mathbf{p}|\mathcal{I}, \boldsymbol{\theta}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}[\mathbf{K}(\mathcal{I}, \mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}]^{-1}\mathbf{p} + \frac{1}{2}\log|\mathbf{K}(\mathcal{I}, \mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}| + \frac{n_{\mathrm{TP}}}{2}\log 2\pi. \tag{25}$$

After inferring $\boldsymbol{\theta}$, GPR can be applied to test data sets.

For brevity, we also state without additional derivation the formulation of a multi-output GPR in which we want to determine a T-dimensional parameter vector. We will assume that we have observations of the multiple parameters that share the same underlying set of training points, i.e., we have one set $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_{n_{\mathsf{TP}}}\}$ from the input domain and n sets of corresponding output observations $\mathcal{P}_i = \{p_i(\mathbf{I}_1), \ldots, p_i(\mathbf{I}_{n_{\mathsf{TP}}})\}$, each of which organized in a vector $\mathbf{p}_i = [p_i(\mathbf{I}_1), \ldots, p_i(\mathbf{I}_{n_{\mathsf{TP}}})]^\intercal$, and define the concatenated vector $\mathbf{p} = [\mathbf{p}_1^\intercal, \ldots, \mathbf{p}_T^\intercal]^\intercal$. For the MOGP we will furthermore assume that we have a single kernel function k that accounts for the different parameters by different coefficients a_{ii} for the different output parameters. Furthermore using a single underlying kernel function allows us to take correlations between the different output parameters into consideration by defining:

$$\mathbf{K}_M(\mathcal{I},\mathcal{I}) = \mathbf{A} \otimes \mathbf{K}(\mathcal{I},\mathcal{I}) = \begin{bmatrix} a_{11}\cdot\mathbf{K}(\mathcal{I},\mathcal{I}) & \cdots & a_{1T}\cdot\mathbf{K}(\mathcal{I},\mathcal{I}) \\ \vdots & \ddots & \vdots \\ a_{T1}\cdot\mathbf{K}(\mathcal{I},\mathcal{I}) & \cdots & a_{TT}\cdot\mathbf{K}(\mathcal{I},\mathcal{I}) \end{bmatrix} \in \mathbb{R}^{n_{\mathsf{TP}}\cdot T \times n_{\mathsf{TP}}\cdot T}, \tag{26}$$

with $\mathbf{K}(\mathcal{I},\mathcal{I})$ as in the single-output GPR. In this case the prediction mean and variance are given as

$$\hat{g}_{\mathbf{p}}(\mathbf{y}) = \mathbf{K}_M(\mathbf{y},\mathcal{I})^\intercal \left[\mathbf{K}_M(\mathcal{I},\mathcal{I}) + \mathbf{\Sigma}_s\right]^{-1}\mathbf{p}, \tag{27}$$

and

$$\mathbf{\Sigma}(\mathbf{y}) = \mathbf{K}_M(\mathbf{y},\mathbf{y}) - \mathbf{K}_M(\mathbf{y},\mathcal{I})^\intercal \left[\mathbf{K}_M(\mathcal{I},\mathcal{I}) + \mathbf{\Sigma}_s\right]^{-1}\mathbf{K}_M(\mathbf{y},\mathcal{I}), \tag{28}$$

respectively. Here

$$\mathbf{\Sigma}_s = \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}} \otimes \mathbf{E}_T \in \mathbb{R}^{n_{\mathsf{TP}}\cdot T \times n_{\mathsf{TP}}\cdot T},$$

and $\mathbf{K}_M(\mathbf{y},\mathbf{y})$ and $\mathbf{K}_M(\mathbf{y},\mathcal{I})$ are defined analogous to the single-output case taking into account Eq. 26. Note that in the learning phase of the MOGP we therefore need to determine the entries $\{a_{ij}\}_{i,j=1,\ldots,T}$ of the matrix $\mathbf{A}$ in addition to the kernel's hyperparameters $\boldsymbol{\theta}$. However, requiring the matrix to be symmetric and positive semi-definite helps to reduce the total number of floated variables.

4. TRAINING AND VALIDATION DATA

4.1 Methodology

For the two ML approaches, RBF and GPR, the simulation data will be treated as if these data are computationally expensive, while utilizing over 1000 library points on a grid in library look-up as a comparison. For the ML approaches, this data scarcity will appear similar to Fig. 3, except that at each n_{TP} the actual points in the parameter space will be varied at each realization. Specifically, for $n_{\mathsf{TP}} = 16, 32, \ldots, 128$ there are 112 realizations for each of five values of the simulated "measurement" noise $\sigma_{\mathsf{noise}} = 10^i, i = \{-6, \ldots, -2\}$ where the incident intensity $I_0 \equiv 1$. For $n_{\mathsf{TP}} = 256$, the number of realizations is smaller, between 64 and 80. In training and validation, $\sigma_{\mathsf{meas}} \equiv \sigma_{\mathsf{noise}}$. For all methods, this noise is applied to the validation data while for RBF and library-look up, it is also used in the formation of the matrix $\mathbf{V}^{-1}$.

As can be inferred from the small markers in Fig. 3, the simulation data library is not equally distributed throughout the simulation domain. In each realization, the validation set is indexed randomly while the n_{TP}

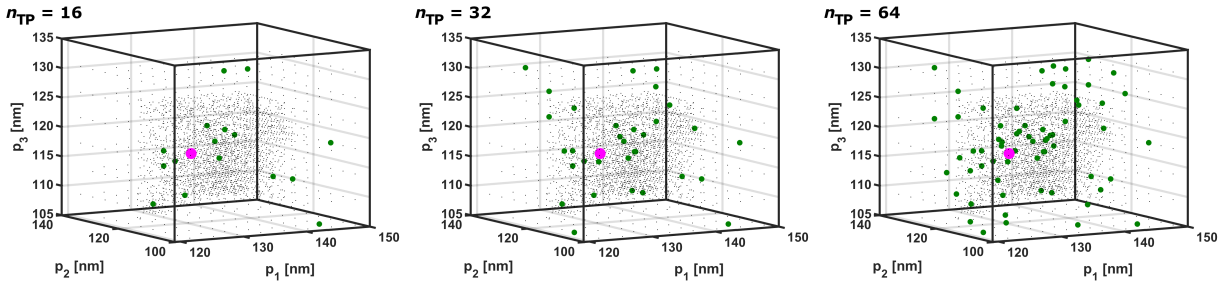

Figure 3. Examples of data scarcity in the simulation space. Validation, training points vary among n_{TP} except in Sec. 4.4

.

training data are selected through draws of random values for the parameters, normally distributed about the center of the simulation domain. With each triplet of values drawn, the closest library entry is indexed. If the candidate index matches the validation point or is repeated, a new candidate is drawn until n_{TP} points are determined.

Two key metrics are required to evaluate these approaches. It is clear that the parametric uncertainties of each parameter must be compared against each other, as well as the parametric values. A single value for comparing the accuracy among the methods can be taken from the GPR literature,[19] the relative average absolute error (RAAE). The RAAE is defined for a general function $f(\mathbf{x})$ and its expectation value $\hat{f}$ as

$$\mathrm{RAAE} = \frac{\Sigma_{i=1}^{n_{\mathsf{TP}}} |f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|}{n_{\mathsf{TP}} \times \sigma(f_{\mathsf{train}}(\mathbf{x}))}, \tag{29}$$

where $\sigma(f_{\mathsf{train}}(\mathbf{x}))$ is the standard deviation of the training data.

In the analysis that follows, the RAAE and parametric uncertainties are compared after hundreds of realizations. Fig. 4 illustrates one such metric, the uncertainty in p_2 computed using all entries in the library look-up at $\sigma_{\mathsf{noise}} = 10^{-3}$. To compare among the different approaches, bar and whisker plots will illustrate the distribution of the RAAE and uncertainty from the training and validation steps and the parametric values from RBF and GPR in the testing step.

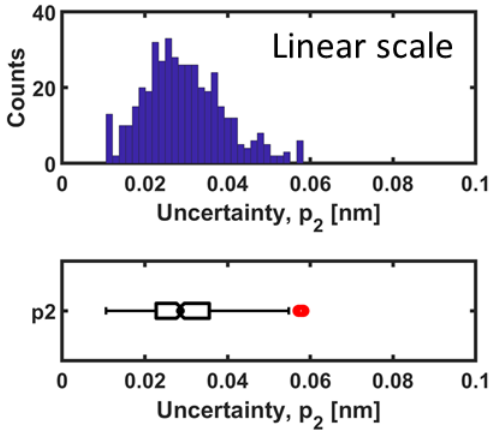

Figure 4. (top) Histogram of uncertainty values from library look-up for $\sigma_{\mathsf{noise}} = 10^{-3}$. (bottom) Example box-and-whisker plot for quickly conveying the distribution from these multiple realizations. In this work, lower and upper box edges show the 25th and 75th percentiles respectively (q_{25}, q_{75}); narrowest notch point is the median (50th percentile); minimum whisker length extends to the outermost data point beyond the box (see left side) while maximum whisker length is $1.5 \times (q_{75} - q_{25})$ (see right side); each outlier beyond the whisker range is shown individually using a symbol.

4.2 Accuracy

The accuracy of the library look-up mode is illustrated at the bottom of in Fig. 4. These data are to be compared against the resulting RAAE values from RBF and GPR as functions of n_{TP} with $\sigma_{\mathsf{noise}} = 10^{-3}$.

Figure 5 shows that for this three-parameter model and sample, only $n_{\mathsf{TP}} \geq 32$ is required to attain a lower (thus better) RAAE than from library look-up. Both the augmenting of the regression through RBF and the indirect use of simulation data through GPR allow far fewer simulations for improved validation set accuracy over the more straightforward conventional approach. Accuracy as expected improves with increases in n_{TP}. Values for the RAAE between GPR and RBF are similar.

4.3 Uncertainty

The major difference observed in this investigation is among the uncertainties in the validation set. As seen in Section 3, both the estimated parametric mean and parametric uncertainties can be determined from RBF and GPR without disregarding correlations among parameters.

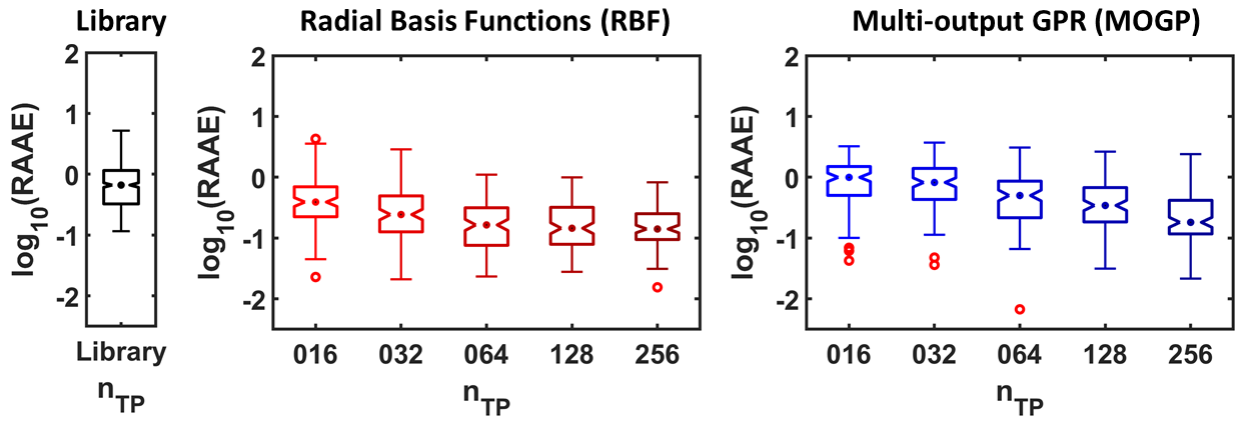

Figure 5. Accuracy of RBF and GPR versus library look-up as a function of n_{TP} for these training and validation data. In general, increasing n_{TP} decreases RAAE, indicating improved accuracy. If notches between any two of these box plot do not overlap, one may conclude that their two medians differ with 95 % confidence.

A key test is the scaling of the uncertainty with the measurement noise. As also established in Section 3, the measurement uncertainty is incorporated into the covariance matrix for both library look-up and RBF such that the parametric uncertainties should scale linearly with σ_{noise}. However, σ_{noise} only indirectly enters into the estimation of the covariance matrix through the application of this noise to the "measurement" data. It was uncertain if this would translate into changes in uncertainty.

In Fig. 6, the expected behavior is observed from library look-up and RBF. Furthermore, the RBF interpolation consistently appears to reduce the uncertainty relative to library look-up by about two orders of magnitude. However, it is unclear if the measurement noise affects the GPR similarly. Here, there is a slight uptick in uncertainty between $\sigma_{\mathsf{noise}} = 10^{-3}$ and $\sigma_{\mathsf{noise}} = 10^{-2}$, but the uncertainty remains unchanged for $\sigma_{\mathsf{noise}} \leq 10^{-3}$. Notably, GPR yields a higher estimate of the uncertainty than either library look-up or RBF interpolation, suggesting a near constant factor in the uncertainty that is independent of measurement noise except for large values of σ_{noise}. Note also, training of the RBF involves solving for both $\mathbf{A}$ in Eq. 13 and hyperparameter r from Eqn. 11, a total of $1 + (t \times n_{\mathsf{TP}})$ parameters, while here GPR training solves for the two hyperparameters σ and l in Eqn. 20. Incorporating additional hyperparameters in the GPR might positively influence the uncertainty from GPR.

4.4 Monte Carlo Assessment of GPR Uncertainty

To assess this nature of the uncertainty estimated from GPR, the numerical experiment was repeated for the multi-output GPR, but instead of picking a new subset of points from the library of size n_{TP} on each realization,

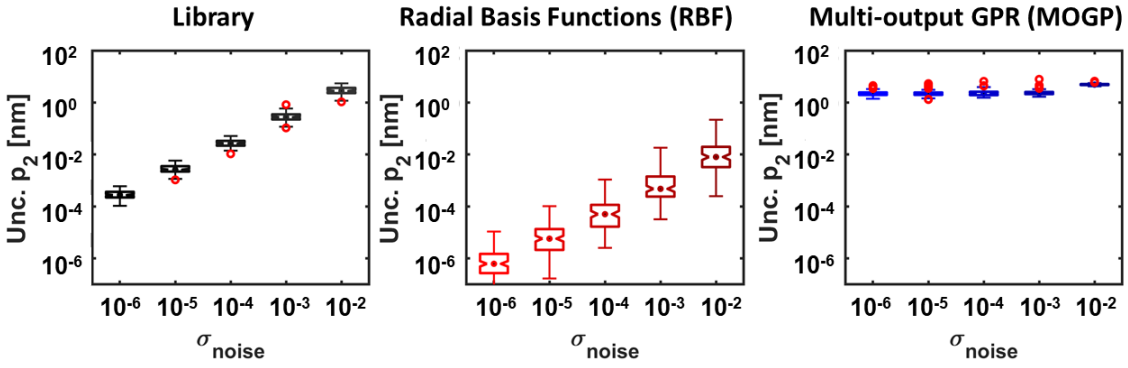

Figure 6. Uncertainty of parameter 2 (middle width) of RBF and GPR versus library look-up as a function of σ_{noise} for these training and validation data. As intensities have been normalized to the incident intensity, σ_{noise} is also unitless.

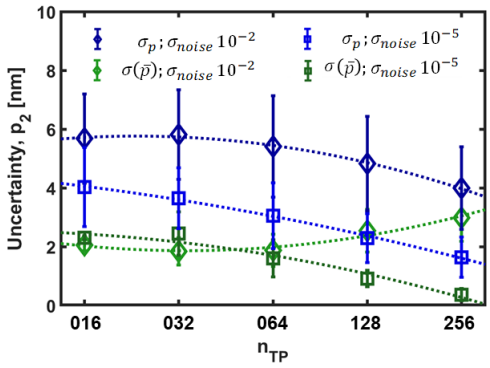

Figure 7. Comparisons of the variance of the middle width estimated by the GPR, σ_p^2, to the variance of the middle width $\sigma^2(\bar{\mathbf{p}})$ from Monte Carlo as a function of n_{TP}, the number of training points. Error bars are one standard deviation from 112 realizations.

multiple realizations were performed using the exact same points at each realization for n_{TP} points. Furthermore, as n_{TP} increased, the previous points were retained. That is, there were not new random draws for $n_{\mathsf{TP}} = 64$ compared to $n_{\mathsf{TP}} = 32$ but rather the subset for $n_{\mathsf{TP}} = 64$ contained 32 new points as well as the points considered at $n_{\mathsf{TP}} = 32$ (which also contained the points for $n_{\mathsf{TP}} = 16$.) For all these simulations, the validation point also remained the same with only its measurement noise varied on each realization, applied at each of the experimental measurement conditions $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$ independently assuming a Gaussian distribution. GPR was performed for 112 realizations for each combination of n_{TP} and σ_{noise}. This approach not only allows the estimation of the posterior mean with its variance, $\sigma(\mathbf{p})$, but also allows computation of the variance of the mean parametric value $\bar{\mathbf{p}}$, or $\sigma^2_{\bar{\mathbf{p}}}$. The uncertainties are shown in Fig. 7 for $\sigma_{\mathsf{noise}} = 10^{-2}$, a relatively large amount of measurement noise and also a much lower value, $\sigma_{\mathsf{noise}} = 10^{-5}$.

Figure 7 illustrates a nearly factor of two difference between the uncertainties $\sigma(\bar{\mathbf{p}})$ and σ_p for low noise. From this additional investigation, GPR as presented here may overestimate its variance except for high noise and a relatively large number of training points. Note that the mean of the parameter estimate only approaches zero for low noise and $n_{\mathsf{TP}} = 256$. Additional work is required to further close the two orders-of-magnitude or more gap in uncertainties between RBF and GPR observed in Fig. 6.

5. RBF AND GPR TESTING

With the discrepancies among RBF and GPR uncertainties, only the parametric values will be considered from these testing results. The experimental measurement noise varied for each measurement condition $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$ but on average, with the incident intensity $I_0 \equiv 1$, the magnitude fell between $\sigma_{\mathsf{noise}} = 10^{-2}$ and $\sigma_{\mathsf{noise}} = 10^{-3}$, for reference.

In Ref.[8] three specific target dies were measured and reported. For consistency, these same three die are measured here using the library look-up and ML approaches. In Fig. 8, the library look-up value is plotted as the dotted line in each individual panel. The leftmost column shows the measurement from each of the die and its label as identified in Ref.:[30] Dies (-1,-1), (0,0), and (1,1). The top row for each die shows the GPR fits to the three parameters identified in Fig. 2, while the bottom row for each die shows the RBF results. For Dies (1,1) and (0,0), the ML-based fits generally agree with the conventional library look-up approach.

Notable problems appear for the RBF fits for the third die, Die (-1,-1). Noted with ellipses on the bottom-most row, there is strong evidence of the distribution of parametric value splitting into two values across the range of n_{TP}. This is highly indicative of an additional local minimum to the weighted χ^2 fitting. Fortunately, the literature has shown a path towards a remedy in two steps. First, in the documentation to Ref.,[20] it is clear that the optimal parameter from RBF can be evaluated using and

$$\hat{\mathbf{p}} = \operatorname{argmin} \left[\left(\widetilde{\mathbf{f}}_{\mathbf{I}} (\mathbf{p}) - \mathbf{y} \right)^{\mathsf{T}} \mathbf{V}^{-1} \left(\widetilde{\mathbf{f}}_{\mathbf{I}} (\mathbf{p}) - \mathbf{y} \right) \right]. \tag{30}$$

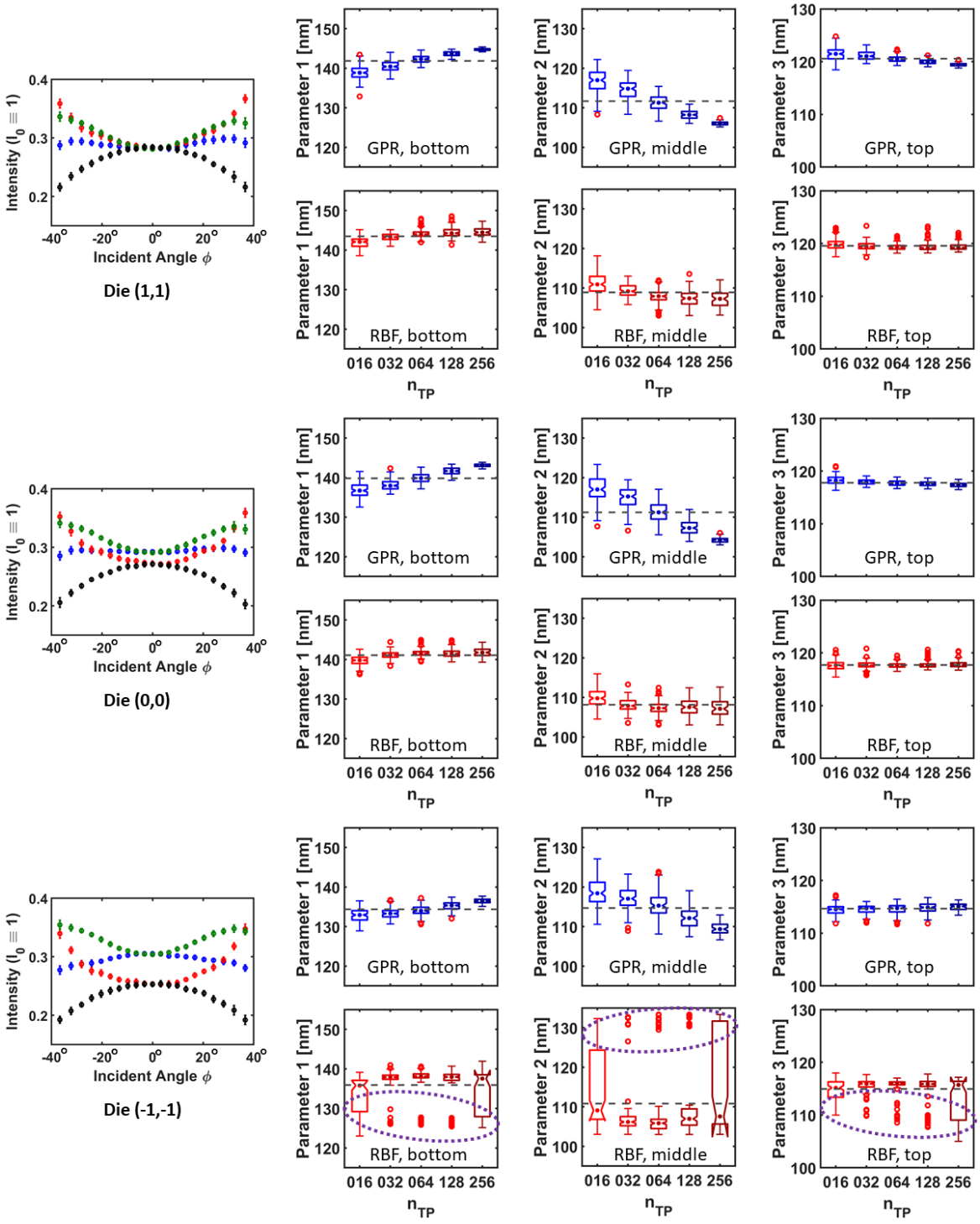

Figure 8. Realizations of the RBF and GPR compared to current library look-up best-fit values. In the left column, measurement error bars are 1σ uncertainties. Box and whisker plots follow definitions in Fig. 4. Ellipses in the bottom-most row indicate the presence of an additional local minimum in the global optimization.

Second, it has been reported that for Bayesian optimization, an objective function can be added to the maximum posterior estimate by means of Bayes' theorem as prior knowledge about non-physical self interactions (e.g., to prevent nonphysical geometries).[17] While the parametric values at this local minimum are physical, they contradict our prior knowledge of these samples from SEM and AFM measurements, and a similar penalty term can be applied. No such function was applied here to illustrate the challenges even for RBF of the inverse problem.

6. CONCLUSION

The need for robust, quantitative machine learning remains especially as sample complexity, improved electromagnetic modeling, and added parameters increase electromagnetic simulation time. While ML can add computational time, this may be smaller in comparison to simulation requirements. Two general approaches have been analyzed, with RBFs augmenting the use of simulated data, while GPR as presented was used to avoid direct application of simulation data. Conventional and ML techniques proved to generally agree on parametric values, but it is clear from the validation steps that uncertainties from GPR here are more problematic than those from RBFs. Testing on experimental data yielded general agreement but a local minimum complicated the fitting in RBF. There are potential solutions to be found for decreasing GPR uncertainty and for constraining the range of allowed fits in RBF due to prior information. These both suggest that additional study is warranted.

Acknowledgements

The authors thank Rick Silver of NIST for supervising the collection of these data and Hui Zhou, formerly of NIST, for initial data processing of the L100P300 data.

REFERENCES

1. den Boef, A. J., "Optical wafer metrology sensors for process-robust cd and overlay control in semiconductor device manufacturing," *Surface Topography: Metrology and Properties* **4**(2), 023001 (2016).
2. Barnes, B. M., Howard, L. P., Jun, J., Lipscomb, P., and Silver, R. M., "Zero-order imaging of device-sized overlay targets using scatterfield microscopy," *Proc SPIE* **6518**, 65180F (2007).
3. Crimmins, T. F., "Defect metrology challenges at the 11nm node and beyond," *Proc SPIE* **7638**, 76380H (2010).
4. Barnes, B. M., Sohn, Y.-J., Goasmat, F., Zhou, H., Silver, R. M., and Arceo, A., "Scatterfield microscopy of 22-nm node patterned defects using visible and duv light," *Proc SPIE* **8324**, 83240F (2012).
5. Silver, R. M., Barnes, B. M., Attota, R., Jun, J., Filliben, J., Soto, J., Stocker, M., Lipscomb, P., Marx, E., Patrick, H. J., et al., "The limits of image-based optical metrology," *Proc SPIE* **6152**, 61520Z (2006).
6. Pomplun, J., Burger, S., Zschiedrich, L., and Schmidt, F., "Adaptive finite element method for simulation of optical nano structures," *physica status solidi (b)* **244**(10), 3419–3434 (2007).
7. Sohn, Y. J., Quintanilha, R., Barnes, B. M., and Silver, R. M., "193 nm angle-resolved scatterfield microscope for semiconductor metrology," *Proc SPIE* **7405**, 74050R (2009).
8. Silver, R., Zhang, N., Barnes, B., Zhou, H., Heckert, A., Dixson, R., Germer, T., and Bunday, B., "Improving optical measurement accuracy using multi-technique nested uncertainties," *Proc SPIE* **7272**, 727202 (2009).
9. Rana, N. and Archie, C., "Hybrid reference metrology exploiting patterning simulation," *Proc SPIE* **7638**, 76380W (2010).
10. Silver, R. M., Zhang, N. F., Barnes, B. M., Zhou, H., Qin, J., and Dixson, R., "Nested uncertainties and hybrid metrology to improve measurement accuracy," *Proc SPIE* **7971**, 797116 (2011).
11. Zhang, N. F., Silver, R. M., Zhou, H., and Barnes, B. M., "Improving optical measurement uncertainty with combined multitool metrology using a Bayesian approach," *Applied Optics* **51**(25), 6196–6206 (2012).
12. Henn, M.-A., Silver, R. M., Villarrubia, J. S., Zhang, N. F., Zhou, H., Barnes, B. M., Ming, B., and Vladár, A. E., "Optimizing hybrid metrology: rigorous implementation of Bayesian and combined regression," *Journal of Micro/Nanolithography, MEMS, and MOEMS* **14**(4), 1 – 8 (2015).

13. Bischoff, J., Bauer, J. J., Haak, U., Hutschenreuther, L., and Truckenbrodt, H., "Optical scatterometry of quarter-micron patterns using neural regression," *Proc SPIE* **3332**, 526–537 (1998).
14. Rana, N., Zhang, Y., Kagalwala, T., and Bailey, T., "Leveraging advanced data analytics, machine learning, and metrology models to enable critical dimension metrology solutions for advanced integrated circuit nodes," *Journal of Micro/Nanolithography, MEMS, and MOEMS* **13**(4), 041415 (2014).
15. Hammerschmidt, M., Weiser, M., Santiago, X. G., Zschiedrich, L., Bodermann, B., and Burger, S., "Quantifying parameter uncertainties in optical scatterometry using Bayesian inversion," *Proc SPIE* **10330**, 1033004 (2017).
16. Heidenreich, S., Gross, H., and Bär, M., "Bayesian approach to determine critical dimensions from scatterometric measurements," *Metrologia* **55**(6), S201 (2018).
17. Hammerschmidt, M., Schneider, P.-I., Santiago, X. G., Zschiedrich, L., Weiser, M., and Burger, S., "Solving inverse problems appearing in design and metrology of diffractive optical elements by using Bayesian optimization," *Proc SPIE* **10694**, 1069407 (2018).
18. Schneider, P.-I., Hammerschmidt, M., Zschiedrich, L., and Burger, S., "Using Gaussian process regression for efficient parameter reconstruction," *Proc SPIE* **10959**, 1095911 (2019).
19. Liu, H., Cai, J., and Ong, Y.-S., "Remarks on multi-output Gaussian process regression," *Knowledge-Based Systems* **144**, 102 – 121 (2018).
20. Henn, M.-A. and Zhang, N.-F., "Model-Based Optical Metrology in R: M.o.R.." `http:/doi.org/10.18434/T4/1502429`. Accessed: 2020-02-28.
21. Sohn, M. Y., Barnes, B. M., and Silver, R. M., "Design of angle-resolved illumination optics using nonimaging bi-telecentricity for 193 nm scatterfield microscopy," *Optik (Stuttgart)* **156**, 635–645 (2018).
22. Silver, R. M., Barnes, B. M., Attota, R., Jun, J., Stocker, M., Marx, E., and Patrick, H. J., "Scatterfield microscopy for extending the limits of image-based optical metrology," *Applied Optics* **46**(20), 4248–4257 (2007).
23. Barnes, B. M., Henn, M.-A., Sohn, M. Y., Zhou, H., and Silver, R. M., "Appraising the extensibility of optics-based metrology for emerging materials," *ECS Transactions* **92**(1), 73 (2019).
24. Yoshioka, S., Matsuhana, B., Tanaka, S., Inouye, Y., Oshima, N., and Kinoshita, S., "Mechanism of variable structural colour in the neon tetra: quantitative evaluation of the venetian blind model," *Journal of the Royal Society Interface* **8**(54), 56–66 (2011).
25. Ehret, G., Pilarski, F., Bergmann, D., Bodermann, B., and Buhr, E., "A new high-aperture 193 nm microscope for the traceable dimensional characterization of micro-and nanostructures," *Measurement Science and Technology* **20**(8), 084010 (2009).
26. "Certain commercial materials are identified in this paper in order to specify the experimental procedure adequately. such identification is not intended to imply recommendation or endorsement by the national institute of standards and technology, nor is it intended to imply that the materials are necessarily the best available for the purpose."
27. Zhang, N. F., Barnes, B. M., Zhou, H., Henn, M.-A., and Silver, R. M., "Combining model-based measurement results of critical dimensions from multiple tools," *Measurement Science and Technology* **28**(6), 065002 (2017).
28. Mongillo, M., "Choosing basis functions and shape parameters for radial basis function methods," *SIAM Undergraduate Research Online* **4**(190-209), 2–6 (2011).
29. Eberhart, R. and Kennedy, J., "A new optimizer using particle swarm theory," in [*Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995. MHS'95.*], 39–43, IEEE (1995).
30. Silver, R. M., Barnes, B. M., Heckert, A., Attota, R., Dixson, R., and Jun, J., "Angle resolved optical metrology," *Proc SPIE* **6922**, 69221M (2008).