Evaluating Uncertainty of Microwave Calibration Models With Regression Residuals

Dylan Williams, *Fellow, IEEE*, Benjamin Jamroz, Jacob D. Rezac and Robert Jones

***Abstract*—We present a sensitivity-analysis and a Monte-Carlo algorithm for evaluating the uncertainty of multivariate microwave calibration models with regression residuals. We then use synthetic data to verify the performance of the algorithms and explore their limitations in the presence of correlated errors. The uncertainties we evaluate can be used to estimate the total uncertainty of a calibrated measurement when combined with the prediction intervals for that measurement.**

***Index Terms*— Coupling corrections, microwave calibration, on-wafer measurement, uncertainty.**

I. Introduction

WHILE the two current supplements to The Guide to the Expression of Uncertainty in Measurement [1] treat Monte-Carlo techniques [2] and multivariate data sets [3], neither treats the use of regression residuals to evaluate uncertainty. However, regression residuals are extremely important in the microwave measurement and other fields for estimating errors in incomplete models, which often are attributed to either incompletely characterized calibration artifacts or incompletely understood sources of error in an experiment or a data set. This paper helps fill this gap by 1) developing algorithms for evaluating uncertainty with regression residuals arising from incomplete multivariate regression models, 2) accounting for the uncertainty of the algorithms' input quantities *and* 3) exploring the advantages and disadvantages of these algorithms for evaluating the uncertainty of microwave calibrations.

Often microwave calibration models are well-enough understood to accurately predict their performance from first principles. In these situations, a complete understanding of the behavior of the calibration artifacts and their uncertainties and the calibration model can be used to predict the error in calibrated measurements of a device under test (DUT) with the procedures recommended in [1].

However, many microwave calibration models are not completely understood, and it becomes very difficult to accurately predict their performance from first principles. Examples include on-wafer calibrations in the presence of probe-to-probe coupling [4] and calibrations for over-the-air tests [5].

Even when the calibrations are well understood, the sources of errors in the calibrations may be difficult to characterize, particularly when there are many error mechanisms to be considered. For example, while classic vector-network-analyzer (VNA) calibrations are quite well understood, characterizing even a simple 50 Ω load requires measuring all of the connector dimensions, dimensions of the access line, and the dimensions and material parameters of the resistive elements and the substrates on which they are fabricated, and performing complex simulations to characterize from first principles [6]. Establishing calibrations in an on-wafer setting can be just as challenging [7]. In these situations, evaluating uncertainty from regression residuals is often useful.

In [8], we discussed a general approach for using regression residuals to calculate the uncertainty in corrected microwave results. In that work we discussed three types of error that must be combined to arrive at the total uncertainty of the measurement of a DUT: 1) errors present in the input data used in the calibrations that do not lead to regression residuals, 2) uncertainty in the calibration model due to uncertainty in the coefficients describing the calibration model based on the calibration's regression residuals, and 3) prediction intervals characterizing the remaining uncertainty in the measurement of the DUT based on a set of independent experiments.

In this paper, we focus on algorithms that evaluate the uncertainty of microwave calibration models (*i.e.* error type 2 in the last paragraph), by which we mean the impact on the modeled values of the uncertainty of the coefficients describing the model. These uncertainties are commonly characterized with "confidence intervals" or "confidence bands" in the regression literature (*e.g.* [8-17]) to differentiate the error in the model associated with prediction intervals, which characterize the uncertainty of the value of a new data point around the regressed curve (*i.e.* the estimated variation of the DUT results around the calibrated result) and are typically determined with Type B methods in The Guide to the Expression of Uncertainty in Measurement [1]. However, to avoid confusion with the conventional statistical use of the term confidence interval, we will simply refer to the uncertainty in the calibration model due to uncertainty in the coefficients describing the calibration model as the "uncertainty of the calibration model" in this paper. Likewise, we will reserve the use of the term confidence interval to refer to the probability that a population parameter falls inside the interval. This is the conventional use of the term

Submitted March 3, 2018.

The authors are with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: dylan.williams@nist.gov, ben.jamroz@nist.gov, jacob.rezac@nist.gov, robert.jones@nist.gov).

in the statistics community [1].

When applied to regression problems, Monte-Carlo jackknife and bootstrap algorithms are often used to evaluate the uncertainty of models (*i.e.* error type 2 above) because they can leverage regression residuals *and simultaneously* account for nonlinearity in the propagation of uncertainty. For example, bootstrap methods are very useful for evaluating the statistics of economic models [18] and models in biology [19], where models are also difficult to understand and quantify.

Jackknife and bootstrap algorithms are based on repeatedly sampling from regression residuals to evaluate the variability of the underlying regression models. However, they differ with respect to their resampling plans. The jackknife was one of the earliest approaches to evaluating the variance of a model from regression residuals. The jackknife is based on systematically removing each observation used in a regression problem and recalculating the result to evaluate the statistics of the model parameters. Bootstrapping was developed later and uses random sampling with replacement. Parametric bootstrapping is based on resampling under the assumption that the original data set is a realization of a random sample from a distribution of a specific parametric type. This is the approach that we follow in this paper. We remark that bootstrap and jackknife methods are used commonly-enough in statistics that even Wikipedia contains introductions to the methods [20, 21]. References [22-24] offer more in-depth explanations of the jackknife and bootstrap approaches to utilizing regression residuals and a variety of other statistical estimation problems.

In [25], Wu presents an excellent summary of jackknife and bootstrapping algorithms for univariate regression problems, and discusses contributions made by Efron [26], Freedman [27], and others. One of Wu's most important contributions in [25] was the development of "local" weighting functions used to adjust the variance of the residuals used in jackknife and bootstrapping algorithms to better reproduce the actual variance of the quantities calculated by the algorithm. While Wu primarily focused on algorithms for linear problems, he also briefly discussed extensions to nonlinear problems. Freedman [27, 28] also extended some of his univariate results to multivariate regression problems. Building on the contributions of Wu [25] and Freedman [27, 28], Eck [29] formalized Freedman's extensions of bootstrap procedures to multivariate linear regression problems and offered detailed proofs of their validity.

Finding the sample mean and evaluating the uncertainty in this value as an estimate of the population mean can be considered a special case of a linear regression problem with a single variable in the model (*i.e.*, the mean of the inputs). References [30] and [31] treat the problem of evaluating the uncertainty in the mean of multivariate input quantities with associated Monte-Carlo samples that approximate their uncertainty, which is not treated in The Guide to the Expression of Uncertainty in Measurement [1] or its supplements [2, 3]. Because the algorithm allows for inputs with associated Monte-Carlo samples that approximate their uncertainty, it is suitable for use in the NIST Microwave Uncertainty Framework [32] and other similar software packages [33-35] designed for microwave calibration problems. These software packages are designed for ease of use and allow uncertainties to be propagated through complicated nonlinear problems [36, 37] when constructing complex traceability paths. This contrasts with Eck, Freedman and Wu, who do not incorporate inputs with associated uncertainty.

TABLE I
COMPARISON TO PRIOR BOOTSTRAP ALGORITHMS

	Wu [25]	Eck [29]	Frey [30]	This Work
Dimension	Univariate	Multivariate	Multivariate	Multivariate
Problem type	Regression	Regression	Mean	Regression
Regression	Nonlinear	Linear	Linear	Nonlinear
Input uncert.	No	No	Yes	Yes
Weighting	Local	None	Global	Local

While the sensitivity analysis we summarize in Section II-A is quite straight forward, the Monte-Carlo bootstrap algorithm we develop in Section II-B is more complex and departs from prior work. Table I compares some of the prior work on bootstrap algorithms to the bootstrap algorithm we develop in this paper, which is the only one of the four that accepts inputs with uncertainty, is applicable to linear and nonlinear multivariate regression problems and supports local weighting to account for limited degrees of freedom in the input data (see Section III-C).

As we mentioned earlier, in this paper we will focus primarily on uncertainty of microwave calibration models (*i.e.* error type 2 above) and test the sensitivity-analysis and Monte-Carlo bootstrap algorithms with synthetic data on an important problem in on-wafer microwave calibration: the evaluation of uncertainty in microwave on-wafer coupling corrections as formulated in [4]. We chose this problem because there are no well-established models for determining the errors in these calibrations from first principles.

We will also verify some of the considerations discussed in [8] for summing the uncertainty of the microwave calibration models (*i.e.* error type 2 above) and the remaining uncertainty (*i.e.* error type 3 above) associated with the prediction intervals. While we will construct our algorithms to accurately propagate input errors through the algorithms (*i.e.* error type 1 above), we will examine and verify this aspect of the performance of these algorithms in a separate publication.

II. Uncertainty-Evaluation Algorithms

Most microwave calibration models, and many calibration models in other fields, can be written in the form $y = f(\beta; x) + \varepsilon$, where f is the calibration model and the β are usually referred to as calibration coefficients. In microwave problems, the calibration model f is almost always a vector function describing the way in which uncorrected vectors of measurements x are transformed into corrected vectors measurements y. The error term ε captures the inability of the calibration model f to perfectly calibrate the corrected measurements y. Regression algorithms are often used to derive an estimate $\hat{\beta}$ for the calibration coefficients β that define the

calibration model.

In calibration problems, the goal is to evaluate the uncertainty in an estimate $\hat{y}_{\mathrm{DUT}}$ of a calibrated measurement from a specific measurement x_{DUT} using $y_{\mathrm{DUT}} \approx f(\hat{\beta}; x_{\mathrm{DUT}})$. Evaluating the uncertainty in the calibration model *f*, which arises from the uncertainty of $\hat{\beta}$, is the principal goal of this paper. However, the uncertainty of $\hat{\beta}$ is of direct interest when regression is used to solve for physical parameters, such as occurs in transistor-model parameter extraction and calculating dielectric constants from resonator and transmission-line measurements. The methods that we discuss here are equally applicable to this set of problems.

While sensitivity analyses are unable to evaluate probability distribution functions and statistical bias in nonlinear statistical problems, they allow us to evaluate the impact of each uncertainty mechanism separately and can be made very efficient. Monte-Carlo analyses, on the other hand, are ideally suited to non-linear statistical problems[1] and evaluate the probability distributions of variables they determine, even if they cannot be performed as efficiently as sensitivity analyses. These considerations motivated the development of both a sensitivity-analysis and a Monte-Carlo algorithm to support the broad range of calibration algorithms supported by the NIST Microwave Uncertainty Framework [32] and other similar packages [33-35].

A. Sensitivity-Analysis Algorithm

For most microwave calibration problems, we perform the calibration by collecting frequency-dependent measurements x_i of *I* calibration artifacts, where $i = 1, 2, \ldots, I$. We then optimize the calibration coefficients β at each frequency to best map the *I* measurements x_i of the calibration artifacts into the *I* responses y_i that we expect based on our physical understanding of the electrical behavior of the calibration artifacts. Thus, the y_i, which are column vectors of size *R*, are usually well understood. In the case of vector network analyzer calibrations, models for the calibration artifacts are often derived from electrical models of the behavior of the artifact based on its physical characteristics, such as its mechanical dimensions, metal conductivity and dielectric constants.

Most often, we find our nominal estimate $\hat{\beta}$ for the calibration coefficients β by minimizing the sum of squares of the vector-residuals ε_i in

$$y_i = f(\hat{\beta};\, x_i) + \varepsilon_i \qquad (1)$$

for the $i = 1, 2, \ldots, I$ calibration artifacts. Here *f* is a known vector function of size *R* describing the way in which the calibration transforms uncorrected measurements x_{DUT} of a device under test into corrected measurements y_{DUT} of the device under test. Furthermore, we are usually interested in evaluating the uncertainty of the corrected measurements *y* given the accuracy with which we know the actual electrical properties of the calibration artifacts. Thus, we see that the definitions y_i of the calibration artifacts share the same space as the corrected measurements y_{DUT} of a device under test and we are interested in the uncertainty of both the y_i and the y_{DUT}, as well as any uncertainty of the uncorrected measurements x_i and the $x_{\mathrm{DUT.}}$ that may give rise to additional uncertainty in $\hat{\beta}$ and y_{DUT}.

The estimated values $\hat{y}_i$ of the y_i based on $\hat{\beta}$ are given by

$$\hat{y}_i = f(\hat{\beta};\, x_i)\,. \qquad (2)$$

Thus, we can express the regression residuals ε_i as

$$\varepsilon_i = y_i - \hat{y}_i\,. \qquad (3)$$

We now use the regression residuals ε_i to approximate the variance of y_i. If *P* is the total number of complex values we estimate in β at each frequency, *I* is the number of measurement pairs (x_i, y_i) and *R* is the dimension of the y_i, we have *IR* complex observations and *IR* - *P* degrees of freedom that we can use to estimate β. Thus, $(IR/(IR-P))\,\mathrm{var}(\varepsilon_i)$ is an unbiased estimator for $\mathrm{var}(y_i)$ [38, 39]. Finally, if we use the regression residual ε_i to approximate the square root of the variance of $y_i - \hat{y}_i$, we can estimate the standard uncertainty of y_i with $\sqrt{IR/(IR-P)}\;\varepsilon_i$ and associate 2(*IR* - *P*) degrees of freedom with that estimate [38, 40, 41]. The factor of two in the associated degrees of freedom is needed to account for the real and imaginary parts of each complex number.

1) Sensitivity-Analysis-Algorithm Implementation

We implemented the sensitivity-analysis algorithm by, for each *i*, finding the $\hat{\beta}_i$ that minimizes the sum of squares of the *I* residuals Δ_{ij} in

$$y_j + \delta_{ij}\sqrt{IR/(IR-P)}\;\varepsilon_i = f(\hat{\beta}_i;\, x_j) + \Delta_{ij}\,, \qquad (4)$$

where $j = 1, 2, \ldots, I$, δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$), the residuals ε_i were minimized in (1) and (3) during the regression that solved for $\hat{\beta}$, and the new residuals Δ_{ij} are the residuals of the regression used to find the $\hat{\beta}_i$ in (4). Note that $\hat{\beta}_i - \hat{\beta}$ estimates the impact of the i^{th} residual on $\hat{\beta}$. As we are already estimating the standard uncertainty of y_i with $\sqrt{IR/(IR-P)}\;\varepsilon_i$ and assuming linearity, we can identify the $\hat{\beta}_i - \hat{\beta}$ as the linear approximation of the standard uncertainty of the component of uncertainty of $\hat{\beta}$ due to the i^{th} residual. Given that $\sqrt{IR/(IR-P)}\;\varepsilon_i$ had 2(*IR* - *P*) degrees of freedom associated with it, this component of uncertainty in $\hat{\beta}$ also has 2(*IR* - *P*) degrees of freedom associated with it.

Finally, we note that the newer statistical tools used in the microwave community to evaluate uncertainties and track correlations [32-35] simply add components of the uncertainty in the inputs from the sensitivity-analysis to the components of the uncertainty at the output, and we do not need to include this functionality explicitly in the sensitivity-analysis algorithm we outlined in this section. This is justified because the sensitivity analysis is based on assumptions of linearity. The Monte-Carlo

[1] The importance of nonlinear statistical analyses in microwave engineering should not be underestimated, as even linear electrical circuits display non-linear behavior in the statistical sense. This is because electrical linearity between the input and output of the circuit is not enough to ensure that the response of the circuit is linear *in each* of the error mechanisms included in the error analysis, which is the criterion for linearity in the statistical sense, not just linearity in the inputs to the electrical circuit.

bootstrap algorithm we discuss next is specifically designed to avoid assumptions of linear error propagation.

B. Monte-Carlo-Analysis Algorithm

We also implemented a frequentist (as contrasted to a Bayesian) Monte-Carlo bootstrap algorithm to evaluate the uncertainty in calibration models. The bootstrap algorithm must account not only for the residuals observed in the regression process, but also for input quantities with associated Monte-Carlo samples approximating their uncertainty that do not manifest themselves as regression residuals.

We start by indexing both the Q Monte-Carlo samples associated with the input quantities to this algorithm and the Q Monte-Carlo bootstrap realizations this algorithm produces with the same superscript $q = 1, 2, \ldots, Q$. Thus, x_i^q corresponds to the q^{th} Monte-Carlo sample approximating the uncertainty of the uncorrected incoming measurements, while y_i^q corresponds to the q^{th} Monte-Carlo sample approximating the uncertainty of the incoming model for the electrical behavior of the i^{th} calibration artifact.

Wu [25], Efron [26], Freedman [27, 28], Eck [29], and others use the nominal residuals ε_i to form the Monte-Carlo bootstrap samples. As we wish to account for possible statistical bias introduced by the Monte-Carlo samples associated with the algorithm's inputs, we depart from this conventional approach and follow Frey, *et al*. [30, 31] to account for the variability of the x_i^q and y_i^q as we form Monte-Carlo samples that will simulate the randomness of the multivariate residuals ε_i we observe in the calibration procedure. Thus, we use a two-step procedure. In the first step, we use the Monte-Carlo samples x_i^q and y_i^q associated with the x_i and the y_i to evaluate the residuals. In the second step, we use the residuals evaluated in the first step to perform the Monte-Carlo bootstrap.

1) Residual Evaluation

Neither Wu [25], Efron [26], Freedman [27, 28] or Eck [29] account for the Monte-Carlo samples x_i^q and y_i^q that are input into the regression problem. This forces us to change our resampling scheme to account for statistical bias and problems in which we have a limited number of measurements, as described below.

To evaluate the residuals while accounting for possible statistical bias in the input Monte-Carlo samples x_i^q and y_i^q, we follow [30, 31] and perform Q regressions to find the Q model parameters $\hat{\beta}^q$ by minimizing the sum of squares of the vector-residuals ε_i^q in

$$y_i^q = f\left(\hat{\beta}^q;\, x_i^q\right) + \varepsilon_i^q \tag{5}$$

for each q. We then form the q^{th} Monte-Carlo estimate $\hat{y}_i^q$ for y_i from the q^{th} set of calibration parameters $\hat{\beta}^q$ with

$$\hat{y}_i^q = f\left(\hat{\beta}^q;\, x_i^q\right). \tag{6}$$

Now, instead of relying on the nominal residuals $\varepsilon_i = y_i - \hat{y}_i$ from (3), as was done by Wu [25], Efron [26], Freedman [27, 28] and Eck [29], or the residuals $\varepsilon_i^q = y_i^q - \hat{y}_i^q$ calculated from (6) at each Monte-Carlo iteration, we calculate the mean $\dot{\varepsilon}_i$ of the Q Monte-Carlo realizations of the i^{th} residual

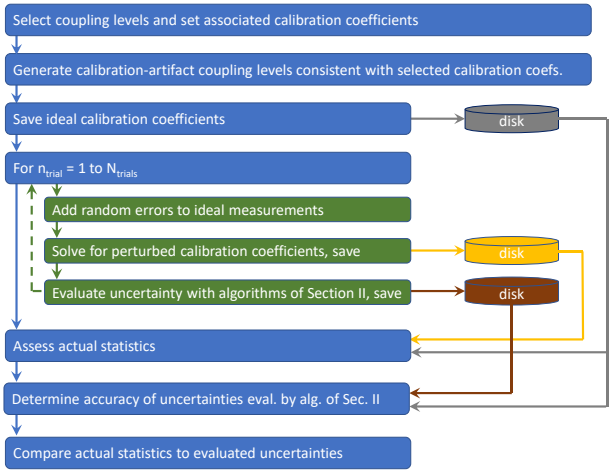

Fig. 1. Simulator flow chart.

$$\dot{\varepsilon}_i = \frac{1}{Q}\sum_q\left(y_i^q - \hat{y}_i^q\right) - \dot{\varepsilon}, \tag{7}$$

where

$$\dot{\varepsilon} = \frac{1}{IQ}\sum_{i,q}\left(y_i^q - \hat{y}_i^q\right) = \frac{1}{IQ}\sum_{i,q}\varepsilon_i^q. \tag{8}$$

The means $\dot{\varepsilon}_i$ of the residuals ε_i^q allow us to capture the statistical bias missing in the nominal residuals $\varepsilon_i = y_i - \hat{y}_i$ from (3) and have lower variance than the ε_i^q calculated from (6) [30]. The $\dot{\varepsilon}_i$ will be used in the next step of the algorithm to form Monte-Carlo bootstrap samples that simulate the randomness of the residuals in independent measurements.

2) Monte-Carlo Bootstrap Procedure

Now that we have determined the $\dot{\varepsilon}_i$, we perform an additional Monte-Carlo parametric bootstrap procedure across the q. Because we often don't have many measurements, we do not use sampling with replacement, as is typically done [26]. Instead, we form the q^{th} Monte-Carlo parametric bootstrap sample y_i^{q*} (see Section 5.2 of [24]) with

$$y_i^{q*} = y_i^q + \sqrt{\frac{IR}{IR-P}}\,\dot{\varepsilon}_i N_i^q(0,1), \tag{9}$$

where the y_i^{q*}, y_i^q and $\dot{\varepsilon}_i$ are all vectors of the same dimension R, N_i^q is the q^{th} draw from the i^{th} scalar Gaussian distribution (see Appendix I for a discussion of other sampling distributions) with mean 0 and variance 1, and we use the factor $\sqrt{IR/(IR-P)}$ in (9) to again compensate for the finite number of degrees of freedom available in the IR observations.

The first Monte-Carlo sample y_i^q in (9) captures uncertainty associated with the incoming y_i, while the second term in (9) is used to simulate the randomness of the residuals. The two terms in equation (9) capture the overall uncertainty in the model. The first term in (9) must be included to account for systematic error present in the calibration. The second term in (9) must be included to account for the error due to the lack of fit to the model, the principle goal of the algorithm. Neither can be neglected.

Finally, the y_i^{q*} are used to find the bootstrap solutions $\hat{\beta}^{q*}$ by minimizing the sum of the squares of the Δ_i^q in[2]

$$y_i^{q*} = f\left(\hat{\beta}^{q*};\, x_i^q\right) + \Delta_i^q \,. \tag{10}$$

In summary, the incoming Monte-Carlo samples y_i^q in (9) and x_i^q in (10) serve to propagate the uncertainty in the x_i and y_i through the bootstrap algorithm to the $\hat{\beta}^{q*}$. The second term $\sqrt{IR/(IR-P)}\,\varepsilon_i^{\cdot} N_i^q(0,1)$ in (9) simulates the randomness of the residuals in the model. And, finally, we use the means $\varepsilon_i^{\cdot}$ in (9) to include the systematic bias captured in the Monte-Carlo samples x_i^q and y_i^q approximating the uncertainty in the inputs x_i and y_i. This also takes advantage of the fact that the means $\varepsilon_i^{\cdot}$ have a higher number of degrees of freedom than the $\varepsilon_i^q = y_i^q - \hat{y}_i^q$, which may add excess variability to the residuals [30, 31].

Now, not only can we evaluate the uncertainty of the model with the Monte-Carlo bootstrap results $\hat{\beta}^{q*}$, but we can also evaluate the uncertainty of calibrated measurements of a DUT due to uncertainty in the model with the bootstrap samples $\hat{y}_{\mathrm{DUT}}^{q*}$ calculated from

$$\hat{y}_{\mathrm{DUT}}^{q*} = f\left(\hat{\beta}^{q*};\, x_{\mathrm{DUT}}^q\right) . \tag{11}$$

Furthermore, the uncertainties we evaluate in both $\hat{\beta}^{q*}$ and $\hat{y}_{\mathrm{DUT}}^{q*}$ include components of uncertainty in the inputs to the calibration algorithm and statistical bias the uncertainty in these inputs may add to the residuals.

III. Simulations

After integrating the two general algorithms for estimating the uncertainty of calibration models described in the last section into the Microwave Uncertainty Framework [32], we applied the algorithms to the specific problem of correcting for coupling in on-wafer calibrations treated in [4]. This coupling-correction algorithm estimates the four complex coupling coefficients of the 16-term VNA error-correction model of [42]. Thus, the total number of complex numbers P we estimate in β at each frequency is four.

We don't expect the two algorithms we developed in the last section to evaluate the uncertainties we are looking for perfectly. Thus, we wrote a simulator tool to investigate the ability of the algorithms to evaluate the uncertainty of the coupling-correction calibration model. The flowchart for the simulator is sketched in Fig. 1 and details of the "internal" coupling model we used may be found in [4]. The simulator generates a synthetic calibration and DUT data with random errors designed to simulate various physical errors, allowing us to verify our implementation of the sensitivity-analysis and Monte-Carlo algorithms described in the preceding section and explore their advantages and disadvantages. The simulator then tests the ability of the algorithms to accurately evaluate, on average, calibration coefficients and transmission through a DUT and the variance of these quantities when on-wafer coupling corrections are applied to noisy synthetic data.

[2] Note that the residuals Δ_i^q in (10) are not equal to the residuals $\varepsilon_i^q = y_i^q - \hat{y}_i^q$ calculated in (6). The residuals ε_i^q were minimized in (6) to find the calibration parameters $\hat{\beta}^q$, while the residuals Δ_i^q are minimized in (10) to find bootstrap samples $\hat{\beta}^{q*}$.

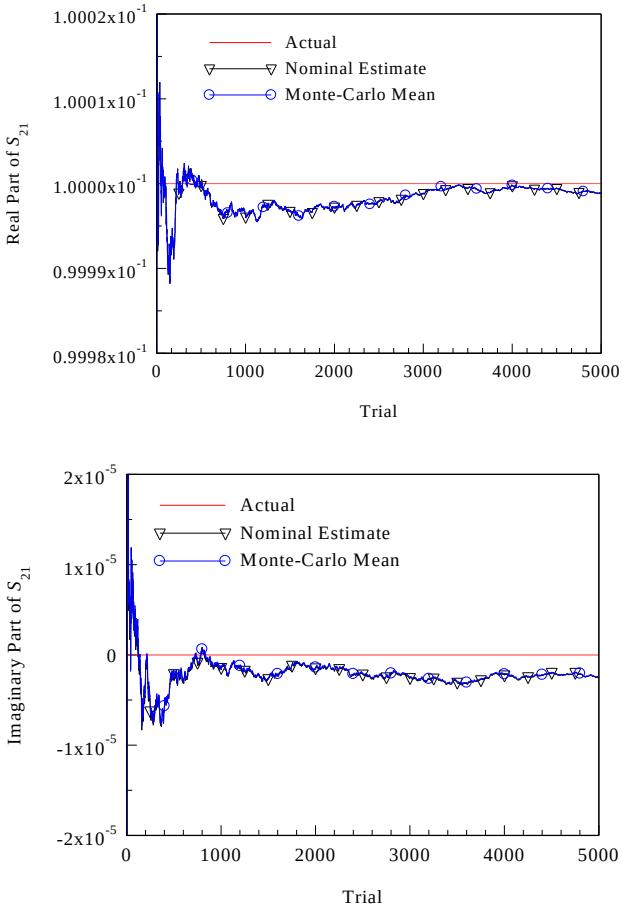

Fig. 2. Real (top) and imaginary (bottom) parts of a nominal single-frequency estimate and the Monte-Carlo mean of a calibrated DUT transmission coefficient compared to actual values in single-frequency simulations based on synthetic coupling-correction data. The first 9 coupling-correction artifacts in Table II were used in the simulations.

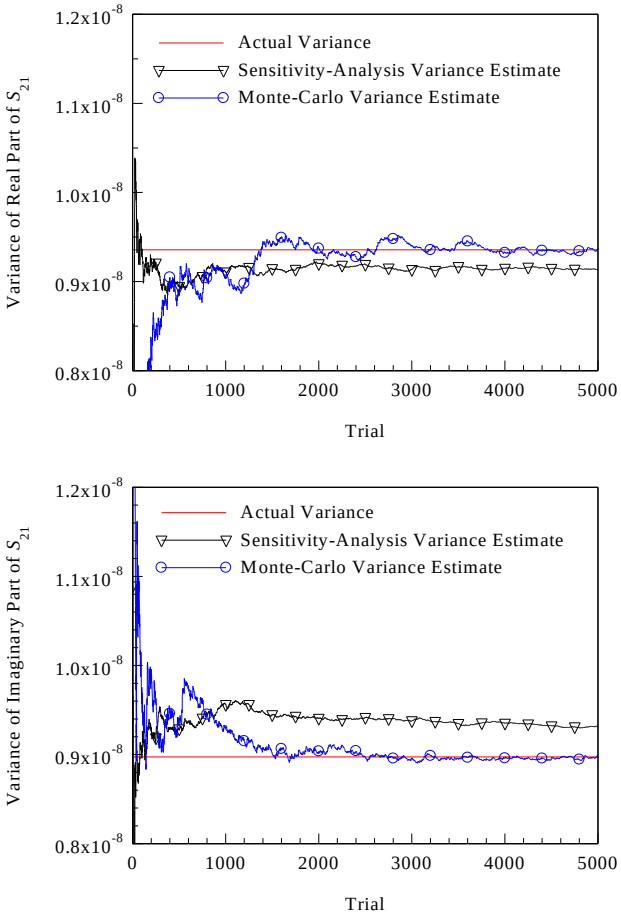

Fig. 3. Variance of the real (top) and imaginary (bottom) parts of the calibrated transmission term of the DUT evaluated by the algorithms presented in Section II compared to the actual variance in single-frequency simulations based on synthetic coupling-correction data. The first 9 coupling-correction artifacts in Table II were used in the simulations.

Mundform, *et al*. [43] performed a comparative study of prior simulations of this type, and suggested that between 5,000 and 10,000 Monte-Carlo replicates were appropriate in the simulation tools they reviewed. However, the coupling corrections we performed were too computationally intensive on the desktop computer platform using the current version of the Microwave Uncertainty Framework, which was programmed in Visual studio with un-optimized algorithms, to complete 5,000 simulation trials with each trial testing the algorithms we presented in Section II with 5,000 Monte-Carlo replicates, in less than 24 hours. Thus, we had to place some limitations on the number of Monte-Carlo replicates and trials used in our results. In Section III-B-1 we will examine the performance of the algorithms as a function of the number of Monte-Carlo replicates utilize by the algorithm.

For these simulations we selected a rectangular distribution with a width of 0.001 for the real and imaginary parts of the errors in the coupling-correction artifacts we used in the simulation (see Table II) and used a perfect 20 dB attenuator as the DUT in these numerical experiments. The simulator was configured to first generate ideal data for the algorithms of Section II to solve, and then perturbed data for each trial for the algorithms of Section II to re-solve. The simulator was configured to add errors into either the uncalibrated (x) calibration-artifact and/or DUT measurements *or* the calibration-artifact definitions and/or calibrated DUT measurements (y). However, in practice, we found that the algorithms of Section II performed similarly when errors were added into the uncalibrated (x) measurements and when errors were added into the definitions and calibrated DUT measurements (y), and thus only present results from the former case here.

TABLE II
COUPLING-CORRECTION ARTIFACTS

Artifacts	Type	S_{11}	S_{22}
1, 10, 18-23	Loads	0	0
2, 11	Opens	1	1
3, 12	Shorts	-1	-1
4, 13, 24, 25	Load/Open	1	1
5, 14, 26, 27	Load/Short	0	-1
6, 15, 28, 29	Open/Load	1	0
7, 16, 30, 31	Short/Load	-1	0
8, 17	Open/Short	1	-1
9, 32	Short/Open	-1	1

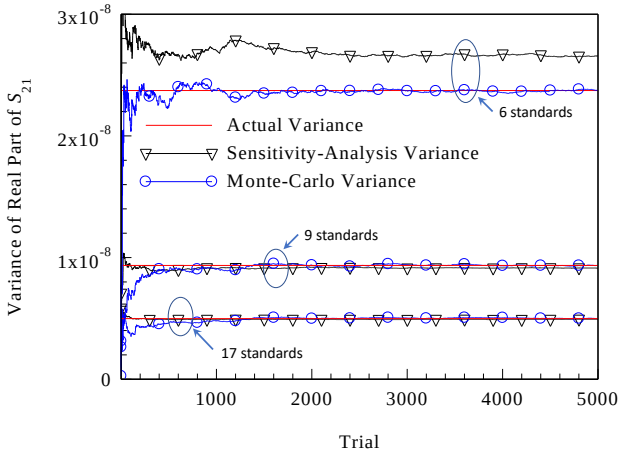

Fig. 4. Variance of the real part of the calibrated transmission term of the DUT evaluated by the algorithms presented in Section II compared to the actual variance in the simulation as a function of the number of coupling-correction artifacts used in the coupling-correction calibration. The coupling-correction artifacts in Table II were used in the simulations. Variances of the imaginary parts are not similar. Also see Table III for further results.

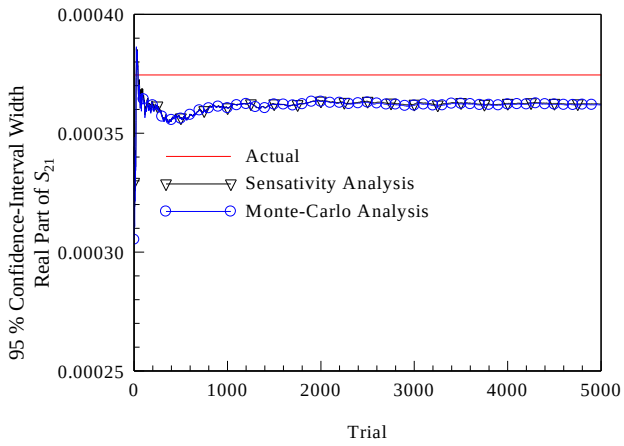

Fig. 5. Width of the 95 % confidence intervals[3] of the real part of the calibrated transmission term of the DUT evaluated by the algorithms presented in Section II using 1000 Monte-Carlo replicates compared to the actual 95 % confidence intervals in the simulation. The first 9 coupling-correction artifacts in Table II were used in the simulations. Confidence intervals for the imaginary part is similar.

A. Single-Frequency Simulations

The sensitivity-analysis and Monte-Carlo algorithms we presented in Section II were designed to evaluate the uncertainty of the coupling-correction calibrations we investigated. We first used our simulation tool to verify the ability of these algorithms to correctly estimate coupling corrections and evaluate the uncertainty of those corrections for single-frequency problems.

Fig. 2 compares the nominal estimate (*i.e.* the estimate based on the simulated errors introduced into the data for that trial) from the coupling-correction algorithm of the transmission through the 20 dB attenuator from the sensitivity analysis and the mean of 1000 realizations from the Monte-Carlo analysis to the actual (true) value of 0.1 used in the simulations. The mean value for each of these quantities is plotted in the figure as a function of the number of trials included in the average. The nominal estimate from the sensitivity analysis algorithm of Section II and the means of the realizations from the Monte-Carlo analysis algorithm of Section II converge to the actual value used in the simulations. This shows that the algorithms of Section II do not add significant bias into the coupling-corrected data.

Fig. 3 compares the variances of the corrected DUT transmission coefficients evaluated with the sensitivity-analysis and Monte-Carlo algorithms we presented in Section II to the actual variance of the transmission coefficients around their true values specified in the numerical simulation tool. As can be seen in the figures, the variance evaluated by the two algorithms converge reasonably well to the actual variances in the synthetic data sets. (The slight bias in sensitivity-analysis evaluations may be due to mild nonlinearity in the problem.) From this we conclude that the two algorithms, when applied to our on-wafer coupling-correction problem, provide reasonable estimates of the actual transmission-coefficients and their variance in our synthetic data sets. That is, the algorithms yield reasonable estimates of their own accuracy from the residuals in the calibrations.

1) Number of Calibration Artifacts

In practice, most engineers typically use only the minimum number of artifacts necessary to perform a calibration (*i.e.* four coupling-correction standards in this two-port case) because it is often difficult to devise large numbers of calibration artifacts. Thus, engineers are sometimes reluctant to use more than the

TABLE III
SUMMARY STATISTICS FOR PERFORMANCE OF ALGORITHMS OF SECTION II

Number of Calibration Artifacts	Degrees of Freedom	Actual Model Variance v_a (x10^{-9})	Normalized Stand. Dev. SD$(v_e)/v_a$			
			Sensitivity Analysis	Monte-Carlo (vs. Q)		
				100	1,000	10,000
6	4	23.6	0.909	0.896	0.863	0.878
9	10	9.8	0.353	0.367	0.346	0.344
17	30	5.6	0.207	0.227	0.220	0.218
32	56	2.5	0.132	0.169	0.131	0.130

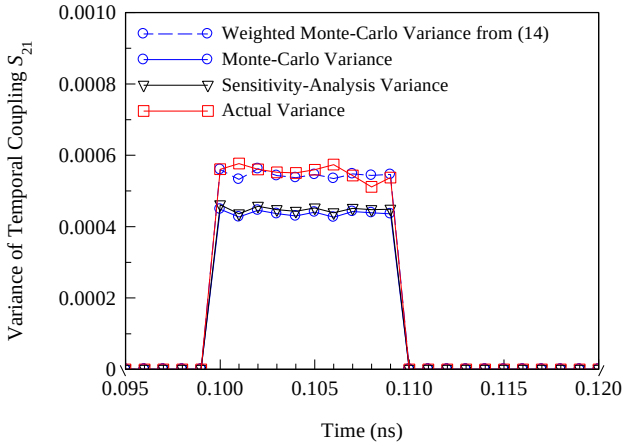

Fig. 6. Variance of the calibrated transmission term of the DUT evaluated by the algorithms presented in Section II compared to the actual variance in the simulation of coupling viewed in the time domain for frequency-independent calibration artifacts. The first 9 coupling-correction artifacts in Table II were used in the simulations.

minimum number of calibration artifacts required to obtain a nominal result and it is important to investigate how the algorithms perform as a function of the number of calibration artifacts.

Fig. 4 compares the variances of the transmission coefficients of the 20 dB attenuator due to the uncertainty of the model evaluated by the algorithms of Section II to the actual variance of that same quantity when 6, 9 and 17 coupling-correction artifacts were included in the simulations. The figure shows that both the actual variance v_a due to the errors added by the simulator and the variance v_e evaluated by the algorithms of Section II increase considerably when the number of coupling-correction artifacts approaches the minimum number of 4 required to find the 4 coupling-correction coefficients in the calibration, as we expect. That is to say, using more artifacts decreases the variance of the uncertainties we are able to evaluate.

Table III lists summary statistics quantifying the performance of the sensitivity-analysis algorithm of Section II and the Monte-Carlo bootstrap algorithm of Section II using 100, 1000 and 10,000 Monte-Carlo replicates in a 100-trial simulation as a function of the number of calibration artifacts used in the calibration listed in the first column. The second column lists the number of degrees of freedom in the residuals and third column lists the actual variance v_a in the nominal transmission coefficients of the 20 dB attenuator as a function of the number standards in the calibration in the simulations.

The fourth column of the table lists $\mathrm{SD}(v_\mathrm{e})/v_a$. This is the standard deviation $\mathrm{SD}(v_\mathrm{e}) = \sqrt{(1/I)\ \sum_i \left(v_{\mathrm{e},i} - v_a\right)^2}$ of the variances $v_{\mathrm{e},i}$ of the transmission coefficients of the 20 dB attenuator evaluated by the sensitivity-analysis algorithm of Section II about the actual variance v_a normalized to the actual variance v_a due to the errors injected into the data by the simulator. This is a measure of the relative accuracy of the variance evaluated by the sensitivity-analysis algorithm as a function of the number of calibration artifacts. The table shows that the normalized standard deviation of the variance estimated by the algorithms rises as the number of calibration artifacts becomes small and the number of degrees of freedom in the observations fall.

Table III also lists this last metric for the Monte-Carlo algorithm of Section II when 100, 1000 and 10,000 Monte-Carlo replicates Q are used by the algorithm in the evaluation of variance. Here we see that, for this fairly linear problem, the ability of the two algorithms to evaluate the uncertainty in the 20 dB attenuator is comparable. The table shows that, for the low number of degrees of freedom we investigated with these simulations, the Monte-Carlo algorithms ability to evaluate the uncertainty of the model is only weakly dependent on the number of Monte-Carlo replicates used in that evaluation.

Finally, the relatively large values of $\mathrm{SD}(v_\mathrm{e})/v_a$ listed in Table III add context to Figs. 3 and 4. In fact, the mean offsets in the variances evaluated by the algorithms of Section II and shown in the two figures are well below the relative standard deviations of these quantities listed in Table III.

2) 95 % Confidence Intervals

Fig. 5 compares the width of the 95 % confidence intervals[3] for the calibrated transmission term of the 20 dB attenuator evaluated by the algorithms presented in Section II to the actual 95 % confidence intervals used in the simulations when 9 coupling-correction artifacts are used in the simulations. Determining this range from just 100 Monte-Carlo realizations is quite difficult, so in this numerical experiment the Monte-Carlo algorithm presented in Section II was configured to generate and use 1000 Monte-Carlo replicates. While the evaluated widths of the 95 % confidence intervals are systematically lower than expected, the agreement is reasonable with these 1000 Monte-Carlo replicates.

B. Multi-Frequency Simulations with Frequency-Independent Calibration Artifacts

Correctly treating correlations in multivariate quantities is essential for evaluating the uncertainty of microwave calibrations. For example, the efficiency term in microwave power meters is typically very close to constant as a function of frequency. Thus, its errors are usually very highly correlated and errors in the efficiency term in the power-meter calibration lead to an overall increase or decrease of the measured power at all frequencies. As a result, this error does not impact the measurement of the communication metric error vector magnitude, which is independent of the amplitude of a signal. However, the same error may dominate the overall error in measurements of receiver sensitivity or antenna efficiency, illustrating how essential it is to track correlations in microwave measurements.

[3] As discussed in the introduction, here we use the term confidence intervals in its conventional statistical sense to refer to the probability that a population parameter will fall between the confidence intervals, as opposed to the way it is often used in the regression community to refer broadly to what we refer to here as the uncertainty of a model.

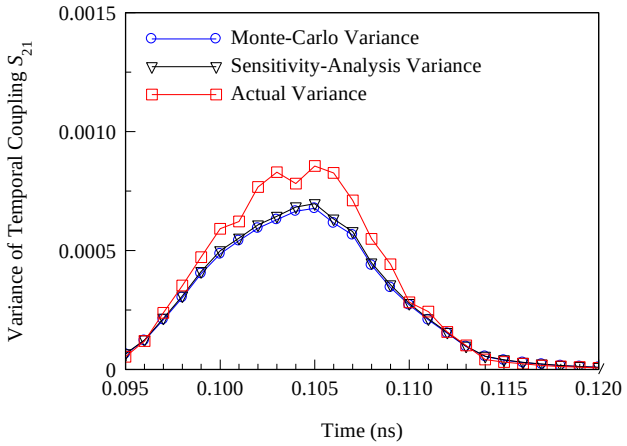

Fig. 7. Variance of the calibrated transmission term of the DUT evaluated by the algorithms presented in Section II compared to the actual variance of corrected DUT measurements as a function of time for frequency-dependent calibration artifacts. The first 9 coupling-correction artifacts in Table II were used in the simulations.

We propagate uncertainties and maintain correlations between them by treating the quantities *x*, *y*, β and ε in Section II as vectors of complex numbers when dealing with microwave vector quantities, such as the frequency-dependent coupling coefficients we treat here. Now we demonstrate the ability of our algorithms to preserve correlations by introducing errors consisting of sinusoidal ripples in the frequency-domain transmission terms of coupling-correction artifacts generated by our simulator.

In the numerical experiment, we simulated 1000 transmission coefficients on a 1 GHz grid to 1 THz. We then introduced ten sinusoidal ripples with random amplitudes and periods of 10/(1+0.01*k*) GHz, for k = 0, 1, …, 9, into the transmission terms of the coupling-correction artifacts and performed a Fourier Transform of the corrected transmission coefficient of the DUT to map these errors into the time domain.

Fig. 6 compares the variance evaluated by the algorithms we presented in Section II to the actual variance of the errors in the 1000-trial numerical simulation we performed. In Fig. 6, the calibration standards were assumed to be frequency independent, as shown in Table II. As we expected, the energy in the ten simulated frequency-domain ripples mapped into actual errors concentrated at 0.1 ns, 0.101 ns, …, and 0.109 ns in the time domain. This confirmed that our procedures maintain the correlations in the errors we used to test the algorithms. We also see from the figure that the variance evaluated by the two algorithms of Section II somewhat underestimate the actual variance of the errors in the temporal transmission coefficients of the DUT.

C. *Local Weighting*

The term $\sqrt{IR/(IR-P)}$ in (4) and (9) is a first-order correction for the finite number of degrees of freedom 2(*IR* - *P*) in the regression problem [38]. This term is intended to correct for the fact that the observed residuals $\varepsilon_i = y_i - \hat{y}_i$ in (3) and their means $\bar{\varepsilon}_i$ of the ε_i^q over the *Q* Monte-Carlo realizations in (7) only reflect the difference between the y_i and the $\hat{y}_i$, while it is the variance of the y_i that we would like to use in (4) and (9).

However, in practice, the sensitivity of the model estimated by regression algorithms are too complicated to be captured by the single term $\sqrt{IR/(IR-P)}$. Wu developed an improved estimate for the ratio of the variance of the $\hat{y}_i$ to the variance of the y_i that is "local" in i for univariate regression problems in Section 7 of [25]. We extended Wu's local approach to the two-port coupling corrections, which only contains a single complex scalar rather than a multivariate function.

First, we linearized the coupling-correction calibration around the nominal solution of the nonlinear calibration problem presented in (2) by forming the data matrix

$$\tilde{X}_{ij} = \frac{\partial f(\beta; x_i)}{\partial \beta_j}, \tag{12}$$

as suggested by Wu in [25], except that the $\tilde{X}_{ij}$ are now complex vectors, not real scalars. We then extended the scalar Wu weights of [25] to complex vector results by calculating the w_i from the data matrix $\tilde{X}$ with

$$w_i = \tilde{x}_i^{\dagger} \left(\tilde{X}^{\dagger}\tilde{X}\right)^{-1} \tilde{x}_i \ , \tag{13}$$

where the superscript † indicates the conjugate transpose and $\tilde{x}_i$ is the i^{th} column of $\tilde{X}^{\text{T}}$.

The w_i capture the relative level of variance of the y_i and the estimates $\hat{y}_i$, and can be used to correct the residuals for the underestimation or overestimation of the contribution of the variance of the $\hat{y}_i$ to the measured residuals by the term $\sqrt{IR/(IR-P)}$ in (4) and (9). By developing the complex form of the linearization captured in the data matrix $\tilde{X}$, the w_i in the coupling correction we investigate here reduce to a single complex number with no imaginary part at each frequency. Thus, we can simply follow Wu and replace the term $\sqrt{IR/(IR-P)}$ in (4) and (9) with $\sqrt{1/(1-w_i)}$. That is, (4) is replaced by

$$y_j + \delta_{ij}\sqrt{1/(1-w_i)}\ \varepsilon_i = f\left(\hat{\beta}_i;\ x_j\right) + \Delta_{ij} \tag{14}$$

and (9) is replace by

$$y_i^{q*} = y_i^q + \sqrt{\frac{1}{1-w_i}}\ \bar{\varepsilon}_i N_i^q(0,1)\ . \tag{15}$$

In cases where the number of degrees of freedom are small and the specific quantity one is interested in has greater variation than would be expected from the conventional factor $\sqrt{IR/(IR-P)}$, the Wu weights can be quite effective. An example is shown in a dashed line in Fig. 6, which shows that the variance evaluated with the Wu weights is much closer to the actual than the result using the conventional factors in (4) and (9). While the Wu weights were not always as effective as shown in Fig. 6, the results we saw were always either comparable or superior to those we obtain from the conventional factors in (4) and (9).

IV. Structural Transformations

Microwave calibration artifacts typically vary with frequency, wrapping around the Smith chart as frequency

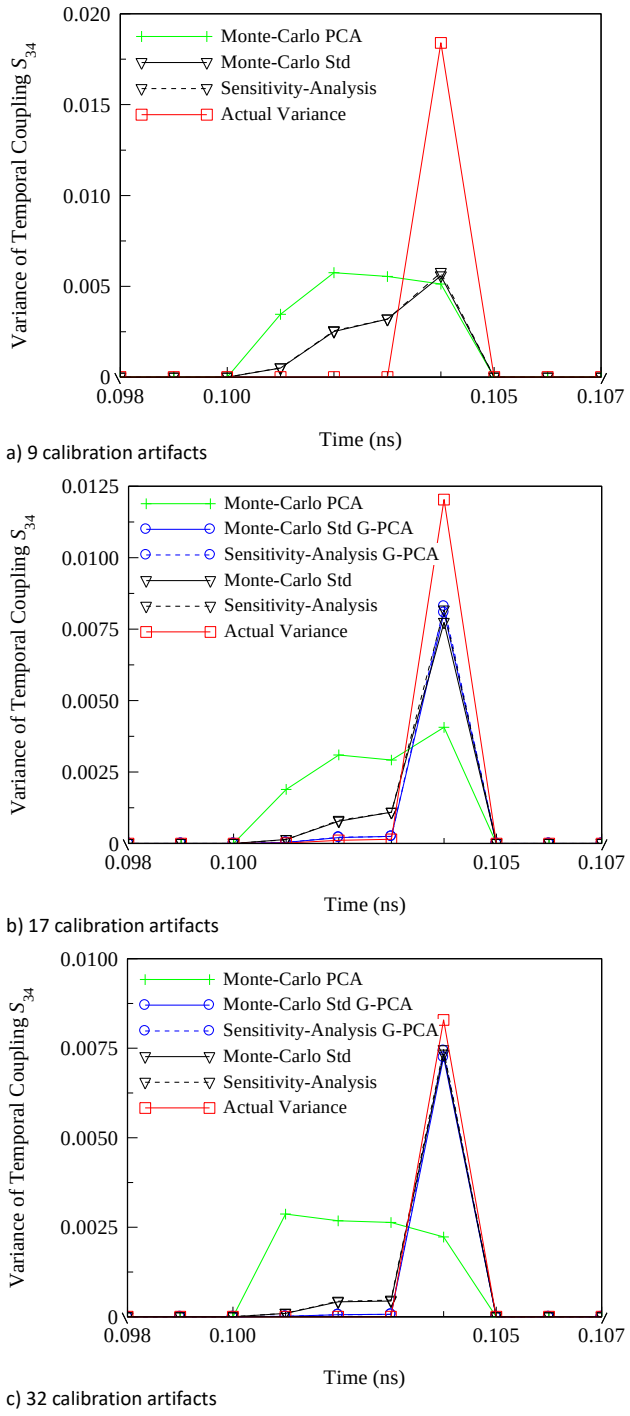

Fig. 8. Comparison of evaluated and actual variances of the S_{34} coupling term in the cross-talk-correction calibration model for 9, 17 and 32 calibration artifacts. The uncertainties were constructed in the simulation to be entirely concentrated at 0.104 ns.

increases. This behavior significantly complicates the propagation of uncertainties through microwave calibrations when correlations must be accounted for and transforms both the structure of the nominal results and their uncertainties.

To illustrate this, we repeated the experiment described in Section III-B with the same 9 calibration artifacts from Table II, except that in each trial we not only assigned a random sinusoidal error to each calibration artifact, but also added a linear phase rotation to each artifact and its definition (supplied in the form of a model) that varied from 0 to four complete phase rotations around the Smith Chart over the entire 1 THz frequency range.

Fig. 7 shows that the frequency-varying calibration artifacts fundamentally change the structure of the uncertainties, spreading the energy in the sinusoidal errors we introduced in the frequency domain in time. This is expected, as the calibration-artifacts simulated sinusoidal errors in frequency interact with the frequency-varying standard definitions as they wrap around the Smith chart, "mixing" the frequency variation of the simulated errors and the frequency dependence of the calibration artifacts in the output of the calibration.

Nevertheless, Fig. 7 shows that the uncertainties evaluated by the two algorithms of Section II track the changes in the structure of the error and broaden their temporal range to better match the actual errors shown in the figure.

V. Structural Leakage

We have just seen that correlations in errors, which reflect their underlying structure, can mix with the frequency variation of the calibration artifacts, changing the structure of the errors as they propagate through the calibration problem. Fig. 6 showed that the algorithms we presented in Section II do a good job of capturing this behavior.

However, the regression algorithm itself can transfer uncertainties with structure from one residual to other residuals for which that structure may not be appropriate. We will call this problematic phenomenon "structural leakage." Structural leakage occurs *between* residuals in regression problems and can result in unwanted structure in the residuals that are used to evaluate uncertainty in regression results and thereby introduce unwanted structure in the evaluated uncertainties.

Structural leakage is not an issue in univariant regression where the residuals are indistinguishable scalars and have no structure. However, as we shall soon see, structural leakage can be a serious concern in multivariate regression.

A. Origin of Structural Leakage

Ideally, the regression residuals themselves would be equal to the difference between the measured and true values of the y_i. But, as we have already discussed, we can only observe the residuals $\varepsilon_i = y_i - \hat{y}_i$, the difference between the measured y_i and the $\hat{y}_i$, which are the *estimated* values of the y_i, not their true values.

However, the $\hat{y}_i$ depend on all the calibration artifacts and have a complex structure that contains a mixture of all the structure found in all the y_i. Thus, the residuals $\varepsilon_i = y_i - \hat{y}_i$ we observe contain structure from residuals that may not be present in the difference between the measured and true values of a particular y_i. This, in turn, leads to what we call structural leakage, unwanted structure in the uncertainties evaluated using the algorithms we presented in Section II.

To illustrate structural leakage, we organized our calibration artifacts into four groups. Group 1 was comprised of the calibration artifacts that absorb energy incident on either port of the artifact, Group 2 was comprised of those artifacts that only

reflect energy incident on port 2 of the artifact, Group 3 was comprised of those artifacts that only reflect energy incident on port 1 of the artifact, and Group 4 was comprised of those artifacts that reflect energy incident on either port. Then we simulated sinusoidal errors in the artifacts in Group 1 with a period of (10/1.01) GHz, sinusoidal errors in the artifacts in Group 2 with a period of (10/1.02) GHz, sinusoidal errors in the artifacts in Group 3 with a period of (10/1.03) GHz and sinusoidal errors in the artifacts in Group 4 with a period of (10/1.04) GHz.[4]

Organizing the errors in our simulations in this way gives rise to temporal errors at 0.101 ns, 0.102 ns, 0.103 ns and 0.104 ns in the coupling-correction term S_{12} of the "internal-coupling-correction" error model of [4], to temporal errors at 0.02 ns and 0.104 ns in the coupling-correction term S_{14} of that error model, to temporal errors at 0.03 ns and 0.104 ns in the coupling-correction term S_{23} of that error model and to temporal errors at only 0.104 ns in the coupling-correction term S_{34} of that error model. This latter behavior is illustrated by the red curves marked with squares in Fig. 7, which indicate the actual errors introduced into the coupling-correction term S_{34} occur at only 0.104 ns.

Fig. 8 also shows that the uncertainties based on the residuals $\varepsilon_i = y_i - \hat{y}_i$ we use in the algorithms presented in Section II have unwanted structure not present in the actual variance of S_{34}. For example, in Fig. 8a we only expect to see uncertainty in the coupling-correction term S_{34} of the error model at 0.104 ns. However, the algorithms presented in Section II predict varying degrees of temporal error at 0.101 ns, 0.102 ns and 0.103 ns as well due to structural leakage.

B. Number of Calibration Artifacts

The structural leakage illustrated by Fig. 8 can be mitigated by increasing the number of calibration artifacts. This is because the $\hat{y}_i$ approach the true values of the y_i as the number of calibration artifacts is increased, improving the accuracy of the residuals used by the algorithms presented in Section II to evaluate the uncertainty of the calibration model. This is seen clearly in Fig. 8 by comparing the variance of 1) the coupling-correction term S_{34} evaluated by the standard sensitivity-analysis and Monte-Carlo algorithms of Section II plotted in the black curves and labeled with inverted triangles at 0.104 ns to 2) the actual variances plotted in the red curves and labeled with squares when using 9, 17 and 32 calibration artifacts at the same time. That is, adding calibration artifacts clearly raises the magnitude of the evaluated variances at 0.104 ns, nearly attaining the actual values of variance there when 32 calibration standards were used.

C. Structural Leakage and Principle-Component Analysis

Principal-Component Analysis (PCA) is used to help identify lower-dimensional subspaces that maintain predictive ability as well as possible [44, 45]. Principal-Component-Regression and Partial-Least-Squares-Regression algorithms are used to help identify "latent variables" with great predictive power in large multivariate data sets [46, 47]. However, indiscriminate use of PCA can unnecessarily introduce unwanted structural leakage into the residuals. This unwanted introduction of structural leakage is illustrated by the curves labeled "PCA" in Fig. 8. The curves show the uncertainty in the coupling-correction term S_{34} evaluated with the principal components of the regression residuals when they are distributed uniformly among the calibration artifacts. The increase in structural leakage is clear.[5]

In the Appendix we describe a PCA algorithm that controls structural leakage. However, the PCA algorithm does not reduce structural leakage below what is obtained with the algorithms of Section II and more study will be needed to prove its utility.

D. Reducing Structural Leakage with PCA

Finally, we found that when groups of calibration artifacts share the same structure, PCA can be used effectively within each group to identify the dominant error structures appropriate to that group and provide significant rejection of structural leakage from other groups. This is illustrated by the curves labeled "G-PCA" in Fig. 8, which shows the result of applying this approach to the four groups of calibration artifacts we used to illustrate structural leakage in this section. The figure shows that this approach can indeed be used to reject much of the structural leakage in regression residuals. Here again, the suppression of unwanted structural leakage continues to improve as the number of calibration artifacts are increased. Furthermore, using PCA in this way should allow the principle components to be grouped, increasing their degrees of freedom.

Unfortunately, rejecting structural leakage in this way requires additional study by the user to understand the distributions of the residuals in the problem and we were not able to find a way to automate this approach in a general way. This would make it difficult to incorporate this and related approaches into the Microwave Uncertainty Framework and other similar packages that emphasize ease of use.

VI. Prediction Uncertainty

As discussed in [8], the total variance of DUT measurements can be expressed as a sum of the variance associated with the calibration model and the variance associated with prediction errors, which are often expressed in terms of prediction intervals (see [8-17]). Predication uncertainty captures the remaining error due to factors that are not included in the errors corrected for by the calibration model (or, more generally, the regression model). For example, prediction uncertainty may capture errors in the data that are inconsistent with the

[4] The NIST Microwave Uncertainty Framework is designed to evaluate uncertainty in both linear and nonlinear settings. Thus, it makes use of an internal real-imaginary representation that is more general than a covariance matrix, which is limited to linear problems. This internal representation insensitive to discontinuities in phase as complex vectors wrap around the Smith chart.

[5] Freedman points to the advantages of "centering" the residuals in [19] before sampling from them in bootstrap algorithms. We accomplish this by subtracting $\bar{\varepsilon}$ from the ε_i in (7). However, the coupling term S_{21} in the coupling corrections we studied here centers the residuals in (7) even before the term $\bar{\varepsilon}$ is subtracted. Nevertheless, we note that for regression and calibration problems for which the residuals are not already centered, care should be taken when centering residuals as the mean of the residuals generally contains an equal mixture of structure from all the residuals and may introduce additional unwanted structural leakage into the uncertainty of the model.

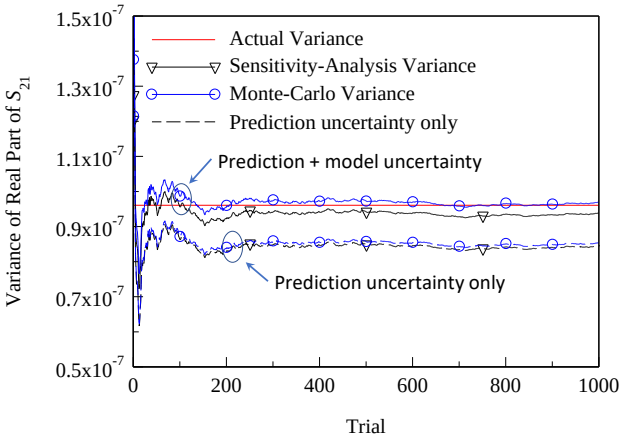

Fig. 9. Variance of the real part of the calibrated transmission term of the DUT evaluated with the sensitivity-analysis and Monte-Carlo algorithms presented in Section II compared to actual variances in numerical simulations performed with synthetic coupling-correction data. Calibration errors between experiments are fully correlated while errors in the DUTs between calibrations are uncorrelated. Variance of the imaginary parts are similar.

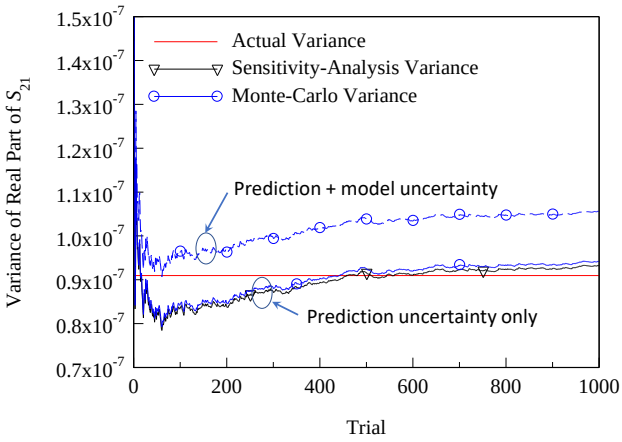

Fig. 10. Variance of the real part of the calibrated transmission term of the DUT evaluated with the sensitivity-analysis and Monte-Carlo algorithms presented in Section II compared to actual variances in experiments performed with synthetic coupling-correction data. Calibration errors and errors in the DUTs between calibrations are completely uncorrelated. Variance of the imaginary parts are similar.

calibration model or errors that are observed as a lack of reproducibility in repeated experiments that cannot be explained and corrected for with the calibration model.

Approaches for evaluating the variance associated with prediction uncertainty may or may not capture the variance associated with the calibration models, depending on how the errors in the calibration models are correlated across different experiments used to evaluate the prediction uncertainty. While the subject of prediction uncertainty and prediction intervals is well beyond the scope of this paper, here we use our simulator to briefly examine the question of when the variance associated with the prediction uncertainty will or will not include the variance associated with the calibration model, and how the two cases should be handled. This verifies several assertions made in [8].

A. *Correlated Model Errors*

Given that the calibrations share the same coupling-correction artifacts, it is quite likely that the errors introduced by the coupling-correction artifacts in repeated calibrations and measurements will be quite similar. In [8] we argued that if the errors captured by the residuals in the calibration were fully correlated (*i.e.* they were identical shared systematic errors and impacted all the DUT measurements in the same way), the calibration errors will not give rise to differences in repeated experiments. We concluded that, in this circumstance, the uncertainty of the model would have to be explicitly added to the uncertainty associated with the variance of repeated measurements from different experiments and calibrations, which we refer to as the prediction uncertainty.

We used our simulator to illustrate this. Fig. 9 compares the actual variance from the true values to the variance estimated by our algorithms when added to the calibration-model variance due to repeated calibrations with fully correlated errors. The agreement is close and confirms the observations in [8]. The figure also shows that not including the uncertainty in the calibration model in the total uncertainty underestimates the actual variance of the DUT measurements.

B. *Uncorrelated Model Errors*

Fig. 10 shows a similar comparison for the case in which the errors associated with the calibration model in each experiment are uncorrelated (*i.e.* different). This might occur if the errors introduced by the coupling-correction artifacts themselves in each calibration were insignificant compared to other sources of error in the repeated measurements, as might occur when thermal noise dominated differences between measurements. In this case, we argued in [8] that the variance of the different DUT measurements from each experiment would capture the total variance in the DUT measurements from their true values. These results are shown in solid lines in Fig. 10 and confirm the argument we presented in [8].

Fig. 10 also shows, as a dashed line, the result of explicitly adding the Monte-Carlo variance associated with the uncertainty in the model to the Monte-Carlo variance associated with the prediction uncertainty. Here, we see that explicitly adding the variance associated with the model "double-counts" that error and results in a variance that exceeds the actual variance, as was suggested would happen in [8].

VII. Conclusion

We presented two algorithms for evaluating the uncertainty of microwave calibrations from regression residuals, both of which extend methods discussed in the Guide to the Expression of Uncertainty in Measurements [1-3] to multivariant regression. The first was based on a straight-forward sensitivity analysis, which can be made efficient and can be used to

separately estimate the impact of each error mechanism. The second was based on a Monte-Carlo bootstrap procedure that deviates from prior work, can solve nonlinear statistical problems and evaluate probability distributions.

We verified that both algorithms estimate nominal single-frequency values and evaluate the uncertainty of on-wafer probe-to-probe coupling corrections with reasonable accuracy. This is just one example of a class of calibration problems in which calibration models are incomplete, a situation that often arises when it is either not possible to completely characterize calibration artifacts or when the sources of error in an experiment are incompletely understood. In the companion paper [48], we explore the ability of the algorithms to evaluate uncertainty in nonlinear VNA calibrations and provide experimental confirmation. When we introduced correlated sinusoidal errors into the frequency-domain calibrations, we found that our algorithms maintained those correlations as expected, even when those correlations are modified by propagating them through significant structural changes.

We noted that the most difficult aspect of using the algorithms we explored here is, as illustrated by Table III, Fig. 4 and Fig. 8, creating and measuring enough calibration artifacts to obtain a sufficient number of degrees of freedom in the residuals to suppress structural leakage and accurately evaluate the uncertainties in the calibration model. This points to a strategy of trying to develop models for calibration errors whenever possible and, when it is not possible to do that, creating and measuring as many calibration artifacts as possible in order to best evaluate the uncertainty in the calibration model with regression residuals. That is, obtaining as many independent regression residuals as possible should be the primary concern when applying the algorithms of Sections II to the evaluation of uncertainty from regression residuals.

Finally, we investigated structural leakage due to the way that regression algorithms mix errors associated with different residuals in their solutions. We showed that it is possible to employ PCA to help identify underlying latent variables and even to reduce structural leakage. However, more work is needed on the development of easy-to-use algorithms that automate the process of discovering underlying latent variables in multivariate regression problems and applying them appropriately.

APPENDIX I
CHOICE OF SAMPLING DISTRIBUTION

In the parametric Monte-Carlo Bootstrap algorithm of Section II we resampled from a Gaussian distribution (see Eqn. (9)). However, Wu uses a variety of resampling distributions with a mean of 0 and a variance of 1 in [25] and Eck proves convergence in the multivariant case when bootstrap residuals are resampled from the original distribution of the ε [29], indicating some flexibility in the choice of distribution from which one resamples.

Wu points out in [25] the resampling distribution must have a mean of 0 and variance of 1 to avoid introducing systematic bias in evaluated variances. However, many calibration algorithms fail to converge when resampled residuals are large. For this reason, the Microwave Uncertainty Framework allows for a variety of resampling distributions, including truncated Gaussian distributions. Table IV below tabulates the variance of truncated Gaussian distributions as a function of the truncation limit determined by numerical integration. We found this reduction in variance to be a good estimate of the overall reduction in evaluated variances when employing the Monte-Carlo bootstrap algorithm of Section II when using truncated Gaussians in (9). We found that we had to use truncation limits of ± 5 σ in the simulations shown in Figs. 2-5 to verify convergence. However, we used truncation limits of ± 3 σ elsewhere where this level of convergence was not required.

TABLE IV
VARIANCE OF TRUNCATED GAUSSIAN DISTRIBUTIONS

Truncation Limit	Variance of Truncated Gaussian
± 3 σ	0.9733369247
± 4 σ	0.9989292904
± 5 σ	0.9999851328
± 6 σ	0.9999999271

APPENDIX II
DIRECTED APPROACH TO PCA

We developed a "directed" approach to PCA that significantly reduces the excess structural leakage of standard PCA algorithms. To implement the algorithm, we first organized our residuals[6] in the matrix

$$Z = \frac{1}{\sqrt{I}}[\varepsilon_1^{\cdot} \varepsilon_2^{\cdot} \dots \varepsilon_I^{\cdot}] \ , \qquad (16)$$

where the $\varepsilon_i^{\cdot}$ are column vectors of dimension R, and applied Singular-Value Decomposition to decompose Z as

$$Z = U \, \Sigma \, V^{\mathrm{T}} , \qquad (17)$$

where $V = [v_1 v_2 \dots v_I]$ and $U = [u_1 u_2 \dots u_I]$ [45]. Now, in the Monte-Carlo analysis, we can draw the directed PCA residuals r_i from

$$r_i = \sum_{j=1}^{I} N_j(0,1) \left(u_j^{\mathrm{T}} \, \varepsilon_i^{\cdot}\right) u_j \ , \qquad (18)$$

where the $N_j(0,1)$ are independent random Gaussian variables with mean 0 and variance 1. We can avoid recalculating the $u_j^{\mathrm{T}} \, \varepsilon_i^{\cdot}$ in (17) by noting that $U^{\mathrm{T}}U$ is equal to the identity matrix, so that $U^{\mathrm{T}} \, Z = \Sigma \, V^{\mathrm{T}}$. Then we can find the $u_j^{\mathrm{T}} \, \varepsilon_i^{\cdot}$ from

$$u_j^{\mathrm{T}} \, \varepsilon_i^{\cdot} = \sqrt{I} \, (\Sigma \, V^{\mathrm{T}})_{ji} \ . \qquad (19)$$

Here we have weighted the component u_j introduced into the i^{th} residual with the dot product $u_j^{\mathrm{T}} \, \varepsilon_i^{\cdot}$ to mitigate unwanted structural leakage in the residuals. In essence, we are tempering the distribution of the components in each residual by the "amount" of that component already observed in that residual.

[6] Here we show the Monte-Carlo residuals $\varepsilon_i^{\cdot}$ organized as columns of Z. The same approach can be applied to the residuals ε_i used in the sensitivity analysis.

The directed approach to PCA we developed is easy to automate and we found that it greatly reduced the unwanted structural leakage of standard PCA algorithm we discussed in Section V-C. However, the directed approach to PCA did not reduce structural leakage below that of the standard algorithm, as we found was possible with the G-PCA algorithm of Section V-D.

ACKNOWLEDGMENT

We thank M. Frey, A. Pintar and F. Jimenez for explaining many of the concepts that underpin this work.

REFERENCES

[1] BIPM, "Evaluation of measurement data-Guide to the expression of uncertainty in measurement," *Int. Org. for Standardization*, vol. JCGM 100, 2008. [Online]. Available: http://www.bipm.org/en/publications/guides/gum.html.

[2] BIPM, "Evaluation of measurement data-Supplement 1 to the 'Guide to the expression of uncertinay in measurement'-Propagation of distributions using a Monte Carlo method," *Int. Org. for Standardization*, vol. JCGM 101, 2008. [Online]. Available: http://www.bipm.org/en/publications/guides/gum.html.

[3] BIPM, "Evaluation of measurement data – Supplement 2 to the "Guide to the expression of uncertainty in measurement" – Extension to any number of output quantities," *Int. Org. for Standardization*, vol. JCGM 102, 2011. [Online]. Available: http://www.bipm.org/en/publications/guides/gum.html.

[4] D. F. Williams, F. J. Schmuckle, R. Doerner, U. Arz, and W. Heinrich, "Crosstalk Corrections for Coplanar-Waveguide Scattering-Parameter Calibrations," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 8, pp. 1748-1761, 2014.

[5] S. Detomasi. (2018) Why 5G is going to over-the-air testing. *Test Meas. Tips*. Available: https://www.testandmeasurementtips.com/why-5g-is-going-to-over-the-air-testing-faq/

[6] F. Mubarak, E. Dierikx, and G. Rietveld, "Traceable DC − 18 GHz characterization of coaxial 50 Ω impedance standards," in *Conf. Precision Electromagn. Meas.*, 10-15 July 2016 2016, pp. 1-2, doi: 10.1109/CPEM.2016.7540479.

[7] U. Arz *et al.*, "Traceable Coplanar Waveguide Calibrations on Fused Silica Substrates up to 110 GHz," *IEEE Trans. Microw. Theory Techn.*, pp. 1-10, 2019, doi: 10.1109/TMTT.2019.2908857.

[8] D. Williams, B. Jamroz, and J. Rezac, "Confidence and Prediction Intervals for Microwave Calibrations and Measurements," in *ARFTG Microw. Meas. Conf.*, Boston, MA, June 7 2019, vol. 93.

[9] T. Heskes, "Practical confidence and prediction intervals," presented at the Advances in neural information processing systems, Pittsburgh, PA, December 2-4, 1997.

[10] W. Liu, S. Lin, and W. W. Piegorsch, "Construction of Exact Simultaneous Confidence Bands for a Simple Linear Regression Model," *Int. Statistical Review*, vol. 76, no. 1, pp. 39-57, 2008/04/01 2008, doi: 10.1111/j.1751-5823.2007.00027.x.

[11] K. D. Brabanter, J. D. Brabanter, J. A. K. Suykens, and B. D. Moor, "Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression," *IEEE Trans. Neural Networks*, vol. 22, no. 1, pp. 110-120, 2011, doi: 10.1109/TNN.2010.2087769.

[12] A. V. Vecchia and R. L. Cooley, "Simultaneous confidence and prediction intervals for nonlinear regression models with application to a groundwater flow model," *Water Resources Research*, vol. 23, no. 7, pp. 1237-1250, 1987, doi: 10.1029/WR023i007p01237.

[13] A. B. Owen, "Nonparametric Likelihood Confidence Bands for a Distribution Function," *J. American Statistical Association*, vol. 90, no. 430, pp. 516-521, 1995/06/01 1995, doi: 10.1080/01621459.1995.10476543.

[14] M. H. Neumann and E. Paparoditis, "Simultaneous confidence bands in spectral density estimation," *Biometrika*, vol. 95, no. 2, pp. 381-397, 2008, doi: 10.1093/biomet/asn005.

[15] W. Hardle, M. Muller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models*. Springer, 2004.

[16] R. J. Hyndman. "The difference between prediction intervals and confidence intervals." https://robjhyndman.com/hyndsight/intervals/ (accessed 2019).

[17] "Confidence and prediction bands." Wikipedia. https://en.wikipedia.org/wiki/Confidence_and_prediction_bands (accessed 2019).

[18] J. L. Horowitz, "Bootstrap Methods in Econometrics," *Annual Rev. Economics*, vol. 11, no. 1, pp. 193-224, 2019/08/02 2019, doi: 10.1146/annurev-economics-080218-025651.

[19] B. Efron, E. Halloran, and S. Holmes, "Bootstrap confidence levels for phylogenetic trees," *Proc. National Academy of Sciences*, vol. 93, no. 23, pp. 13429-13429, 1996, doi: 10.1073/pnas.93.23.13429.

[20] "Bootstrapping (statistics)." Wikipedia. https://en.wikipedia.org/wiki/Bootstrapping_(statistics) (accessed 2020).

[21] "Jackknife resampling." Wikipedia. https://en.wikipedia.org/wiki/Jackknife_resampling (accessed 2020).

[22] B. Efron and T. Hastie, *Computer age statistical inference*. Cambridge University Press, 2016.

[23] B. Efron and G. Gong, "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, vol. 37, no. 1, pp. 36-48, 1983, doi: 10.2307/2685844.

[24] B. Efron, *The Jacknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, 1982.

[25] C. F. J. Wu, "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *The Annals of Statistics*, vol. 14, no. 4, pp. 1261-1295, 1986. [Online]. Available: http://www.jstor.org/stable/2241454.

[26] B. Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol. 7, pp. 1-26, 1979.

[27] D. A. Freedman and S. C. Peters, "Bootstrapping a Regression Equation: Some Empirical Results," *J. American Statistical Association*, vol. 79, no. 385, pp. 97-106, 1984, doi: 10.1080/01621459.1984.10477069.

[28] D. A. Freedman, "Bootstrapping Regression Models," *The Annals of Statistics*, vol. 9, no. 6, pp. 1218-1228, 1981.

[29] D. J. Eck, "Bootstrapping for multivariate linear regression models," *Statistics & Probability Letters*, vol. 134, pp. 141-149, 2018, doi: 10.1016/j.spl.2017.11.001.

[30] M. J. Frey, B. F. Jamroz, A. A. Koepke, J. D. Rezac, and D. Williams, "Monte-Carlo Sampling Bias in the Microwave Uncertainty Framework," *Metrologia*, vol. 56, no. 5, p. 13, 2019, doi: 10.1088/1681-7575/ab2c18.

[31] B. F. Jamroz, D. F. Williams, J. D. Rezac, M. Frey, and A. A. Koepke, "Accurate Monte Carlo Uncertainty Analysis for Multiple Measurements of Microwave Systems," in *IEEE Int. Microw. Symp.*, Boston, MA, June 2-7 2019.

[32] *NIST Microwave Uncertainty Framework*. (2011). National Institute of Standards and Technology, http://www.nist.gov/ctl/rf-technology/related-software.cfm. [Online]. Available: http://www.nist.gov/ctl/rf-technology/related-software.cfm

[33] "VNA Tools II." Federal Institute of Metrology METAS. https://www.metas.ch/metas/en/home/fabe/hochfrequenz/vna-tools.html (accessed 2018).

[34] M. Garelli and A. Ferrero, "A Unified Theory for S-Parameter Uncertainty Evaluation," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 12, pp. 3844-3855, 2012, doi: 10.1109/TMTT.2012.2221733.

[35] G. Avolio *et al.*, "Software tools for uncertainty evaluation in VNA measurements: A comparative study," in *ARFTG Microw. Meas. Conf.*, 9 June 2017, pp. 1-7, doi: 10.1109/ARFTG.2017.8000820.

[36] B. F. Jamroz, D. F. Williams, K. A. Remley, and R. D. Horansky, "Importance of Preserving Correlations in Error-Vector-Magnitude Uncertainty," in *ARFTG Microw. Meas. Conf.*, Philadelphia, PA, June 15 2018.

[37] K. A. Remley, D. F. Williams, P. D. Hale, C. M. Wang, J. Jargon, and Y. Park, "Millimeter-Wave Modulated-Signal and Error-Vector-Magnitude Measurement With Uncertainty," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 5, pp. 1710-1720, 2015, doi: 10.1109/TMTT.2015.2416180.

[38] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*, Three ed. Wiley, 2012.

[39] F. A. Graybill, *Theory and application of the linear model*. Belmont, CA: Dunbury Press, 1976.

[40] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*, 3 ed. San Francisco: Dellen Publishing Company, 1988.
[41] D. C. R. Montgomery, G.C.;Hubele,N.F., *Engineering Statistics*, 5 ed. Hoboken, NJ: John Wiley and Sons, 2011.
[42] J. V. Butler, D. Rytting, M. F. Iskander, R. D. Pollard, and M. Vanden Bossche, "16-term error model and calibration procedure for on-wafer network analysis measurements," *IEEE Trans. Microw. Theory Techn.*, vol. 39, no. 12, pp. 2211-2217, December 1991.
[43] D. J. Mundform, J. Schaffer, M. J. Kim, D. Shaw, and A. Thongteeraparp, "Number of Replications Required in Monte Carlo Simulation Studies: A Synthesis of Four Studies," *J. Modern Applied Statistical Methods*, vol. 10, no. 1, pp. 19-28, 5 January 2011.
[44] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433-459, 15 July 2010, doi: 10.1002/wics.101.
[45] J. Shlens, "A Tutorial on Principal Component Analysis," *arXiv*, 3 April. [Online]. Available: https://arxiv.org/abs/1404.1100
[46] H. Abdi, "Partial Least Square Regression," in *Encyclopedia of Measurement and Statistics*, N. Salkind, Ed., ed, 2007.
[47] P. D. Wentzell and L. V. Montoto, "Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures," *Chemometrics and Intelligent Laboratory Systems*, vol. 65, pp. 257– 279, 2003.
[48] D. Williams, B. Jamroz, and J. Rezac, "Evaluating Uncertainty of Nonlinear Microwave Calibration Models with Regression Residuals," submitted for review by the NIST Editorial Review Board.

Dylan F. Williams (M'80-SM'90-F'02) received a Ph.D. in Electrical Engineering from the University of California, Berkeley in 1986. He joined the Electromagnetic Fields Division of the National Institute of Standards and Technology in 1989 where he develops electrical waveform and microwave metrology. He has published over 100 technical papers and is a Fellow of the IEEE. He is the recipient of the Department of Commerce Bronze and Silver Medals, the Astin Measurement Science Award, two Electrical Engineering Laboratory's Outstanding Paper Awards, three Automatic RF Techniques Group (ARFTG) Best Paper Awards, the ARFTG Automated Measurements Technology Award, the IEEE Morris E. Leeds Award, the European Microwave Prize and the 2013 IEEE Joseph F. Keithley Award. Dylan also served as Editor of the IEEE Transactions on Microwave Theory and Techniques from 2006 to 2010, as the Executive Editor of the IEEE Transactions on Terahertz Science and Technology, and as the 2017 President of the IEEE Microwave Theory and Techniques Society.

Benjamin F. Jamroz received the Ph. D. degree in applied mathematics from the University of Colorado, Boulder CO, USA, in 2009, where he developed new analytical and numerical models for plasma physics. As a computational scientist his work includes modeling physical phenomena including electromagnetics and fluid dynamics as well as applying machine learning techniques. In 2017, he joined the Communications Technology Laboratory at the National Institute of Standards and Technology where he models and analyzes the complex systems required for high-frequency communications.

Jacob D. Rezac received a B.S. (2011) and M.S. (2012) in Applied Mathematics from the Colorado School of Mines and a Ph.D in Applied Mathematics from the University of Delaware in 2017. Between 2017 and 2019 he was a Postdoctoral Researcher working at the Communications Technology Laboratory of the National Institute of Standards and Technology (NIST) and at the University of Colorado-Boulder. He has been a Mathematical Statistician at NIST since 2019. His research interests include theoretical and computational aspects of inverse problems as applied to problems in communications and in the physical sciences.

Robert D. Jones received dual B.S. degrees in electrical and mechanical engineering from the Colorado School of Mines in 2019, where he is currently pursuing his master's degree. Since 2017, he has been a student researcher at the National Institute of Standards and Technology (NIST), conducting experiments with loaded reverberation chambers. His current research interests are in computational electromagnetics, antenna design, and loaded reverberation chamber metrology.