Evaluating Uncertainty of Nonlinear Microwave Calibration Models with Regression Residuals

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

***Abstract*—We investigate the performance of a recently developed algorithm that evaluates the uncertainty of nonlinear multivariate microwave calibration models using regression residuals. We apply the algorithm to synthetic data consisting of both random and systematic errors and show that the algorithm can account for both types of errors even in the absence of accurate models for the random errors. We also verify the algorithm with measured data.**

***Index Terms*—Calibration, measurement, microwave, uncertainty, vector network analyzer.**

I. Introduction

We proposed a Monte-Carlo bootstrap algorithm in [1] for evaluating the uncertainty of nonlinear multivariate microwave calibration models from regression residuals. Regression residuals are extremely important in the microwave measurement and other fields for estimating errors in incomplete models, which often are associated with either incompletely characterized calibration artifacts or incompletely understood sources of error in an experiment or a data set.

The algorithm of [1] was based on an extension of the statistical approaches in [2] and [3] developed for evaluating the uncertainty of the mean of multivariate results. The algorithms of [1-3] were designed to fill a gap in the two current supplements to The Guide to the Expression of Uncertainty in Measurement [4], which treat Monte-Carlo techniques [5] and multivariate data sets [6], but do not discuss the use of regression residuals to evaluate uncertainty. The algorithms of [1-3] allow for inputs with associated Monte-Carlo samples that approximate their uncertainty, and thus are suitable for use in the NIST Microwave Uncertainty Framework [7] and other similar software packages [8-10].

Monte-Carlo bootstrap algorithms are often used to evaluate the uncertainty of models because they can account for the propagation of uncertainty through nonlinear models. In [11], Wu presents an excellent summary of jackknife and bootstrapping algorithms for univariate regression problems, and discusses contributions made by Efron [12], Freedman [13], and others. While Wu primarily focused on algorithms for linear problems, he also briefly discussed extensions to nonlinear problems. Freedman [13, 14] also extended some of his univariate results to multivariate regression problems. Building on the contributions of Wu [11] and Freedman [13, 14], Eck [15] 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). Yet even [2] and [3], of which the Monte-Carlo bootstrap algorithm of [1] is an extension, showed some unavoidable bias when mean results were later used in nonlinear problems.[1] For this reason, we investigate the ability of the Monte-Carlo bootstrap algorithm of [1] to evaluate uncertainty from regression residuals when applied to a nonlinear microwave vector-network-analyzer (VNA) calibration problem in rectangular waveguide (RWG).

In this paper, we first apply the Monte-Carlo bootstrap algorithm of [1] to synthetic data for VNA calibrations containing both random and systematic calibration errors. This allows us to investigate the ability of the algorithm of [1] to accurately evaluate the uncertainty of the VNA calibration models in an automated fashion in the presence of both types of errors and in the absence of models for the random errors. The simulations allow for both qualitative comparisons of single results and quantitative comparisons of averaged results from the Monte-Carlo bootstrap algorithm and the conventional algorithm.

We then apply the Monte-Carlo bootstrap algorithm of [1] to analogous measurements with deliberately introduced errors to verify the simulations and demonstrate the algorithm's suitability for use in actual measurement situations.

In this paper, we use random vertical E-plane offsets at the interfaces between test ports in rectangular-waveguide calibrations to demonstrate the ability of the Monte-Carlo bootstrap algorithm of [1] to estimate uncertainty in the VNA calibration. We chose to study vertical E-plane offsets as error mechanisms both because the errors are nonlinear in the offset

Submitted XXXX.

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).

[1] As discussed in [1], the importance of nonlinear statistical analyses in microwave engineering should not be underestimated, as the impact of measurement errors can be nonlinear even when the electrical circuit is linear. 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 electrical input to the circuit.

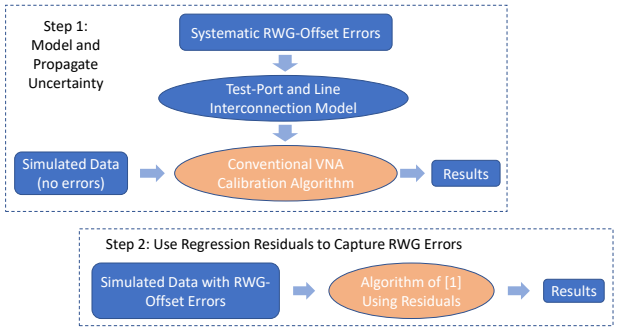

Fig. 1. Simulation approach for RWG calibration line offset errors. 1000 trials of the Monte-Carlo bootstrap algorithm of [1] shown in the lower-right are compared to the conventional Monte-Carlo algorithm illustrated in step one in the upper-left.

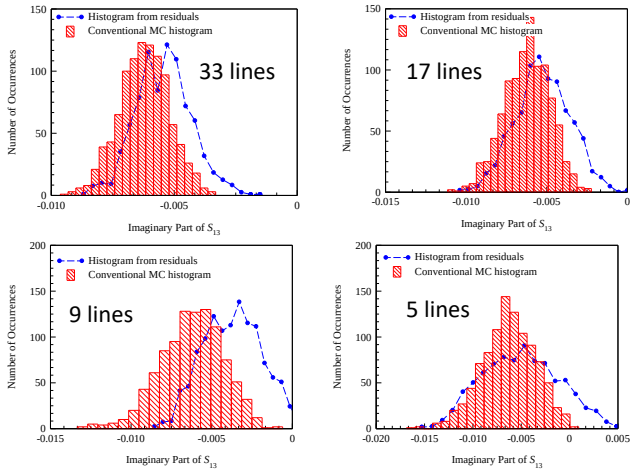

Fig. 2. Histograms of the imaginary part of S_{13} of the calibration model as a function of the number of RWG lines used in the calibration with random RWG line errors only. The histograms in red are evaluated by the conventional approach. The blue dashed curves with solid dots correspond to values of the histograms evaluated during the first of the 1000 trials of the algorithm of [1] based on the measurement residuals.

variable, but also because we have well-understood closed-form models in the NIST Microwave Uncertainty Framework [7] for those errors to simplify the analysis.

However, while we make use of closed-form models to evaluate the uncertainty introduced into VNA calibrations due to deliberately introduced vertical offsets at the interfaces between the VNA test ports and the VNA calibration artifacts, we want to emphasize that the algorithm of [1] we test does not make use of these error models. This allows us to illustrate the utility of the Monte-Carlo bootstrap algorithm of [1], which can estimate uncertainty even in the absence of models for the errors in the problem. This might arise when the user, for example, does not have access to the dimensional measurement equipment required to evaluate the size of the E-plane offsets giving rise to the uncertainty in the calibration.

However, the Monte-Carlo bootstrap algorithm of [1] can be useful in other situations as well. For example, the bootstrap algorithm of [1] could also be used to evaluate the uncertainty of a rectangular-waveguide calibration due to uneven surfaces at the interfaces between the VNA test ports and the VNA calibration artifacts, where models for the resultant discontinuities may be very difficult to come by. Other examples include evaluating for uncertainty due to difficult-to-characterize dimensional variations deep in the interior of rectangular-waveguide artifacts where it is difficult to reach them with coordinate-measurement machines and air gauges are unable to accurately determine the details of the geometric discontinuities.

II. Calibration Algorithm

Here we use the NIST Microwave Uncertainty Framework's Levenberg-Marquardt two-port VNA Calibration Algorithm to investigate the accuracy of the uncertainties evaluated by the Monte-Carlo bootstrap algorithm of [1] from regression residuals. We used the algorithm with 1000 Monte-Carlo replicates in all the simulations discussed in Sections II and IV, as well as the measurement verification discussed in Section V.

We based the calibration artifacts on perfect shorts and up to 33 sections of RWG line with simulated or actual vertical E-plane offset errors between the VNA test ports and the lines, depending on whether we were performing quantitative simulations or more qualitative measurement verification. The vertical E-plane offsets lead to statistical bias because, while there is no capacitance at a junction with a zero vertical E-plane offset, both positive *and* negative offsets at the junction between two RWG sections result in a positive and parabolically increasing excess capacitance at the junction. This is illustrated nicely in Fig. 4 of [16] and Figs. 1 and 2 of [17].

The variable excess capacitance at the junctions, which is always positive, introduces both statistical bias and random variation into the parameters of the VNA calibration model. Here we focus on the imaginary part of the transmission term S_{13} in the port-one calibration error box, which is impacted directly by changes in the excess capacitance at the junctions between the test ports and the RWG lines. Results for the other elements of the two calibration error boxes impacted by the junction capacitance displayed similar behavior.

III. Random RWG Line Offsets

We first simulated a WR-10 calibration with perfect test ports having no systematic offset between the test-ports' rectangular 2.54 mm horizontal H-plane by 1.27 mm vertical E-plane apertures and their alignment pins. We also assigned a uniform probability distribution function (PDF) with a range of ±0.1 mm to the offset in the vertical E-plane direction between the test-port pins and the rectangular aperture of each calibration artifact. We chose the range ±0.1 mm to exaggerate the vertical E-plane offsets and make them comparable to those sometimes encountered at submillimeter frequencies. This made the errors

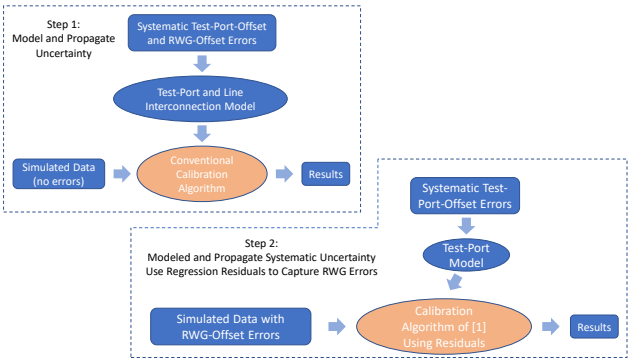

Fig. 3. Simulation approach for RWG test-port and calibration line offset errors. 1000 trials of the Monte-Carlo bootstrap algorithm of [1] shown in the lower-right are compared to the conventional Monte-Carlo algorithm illustrated in step one in the upper-left.

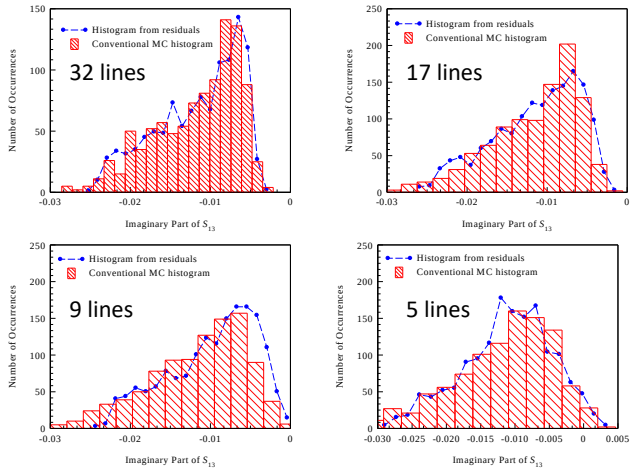

Fig. 4. Histograms of the imaginary part of S_{13} of the calibration model as a function of the number of RWG lines used in the calibration with both systematic test-port and random RWG line errors.

easier to analyze while not exceeding the region of validity of the analytic expressions from Hunter in [17], used in the NIST Microwave Uncertainty Framework to model the impact of these offsets. We will present measurement-verification results for this problem performed in a lower-frequency WR-90 RWG in Section V.

The analysis proceeded in two steps, as illustrated in Fig. 1. In the first step, we solved the RWG VNA calibration problem in the conventional way, as shown in the upper-left of Fig. 1. In the second step, we ran 1000 trials of the Monte-Carlo bootstrap algorithm of [1], and compared the results to the results from the conventional algorithm, which was based on straight-forward Monte-Carlo simulations using the analytic models for vertical RWG offsets implemented in the NIST Microwave Uncertainty Framework [7]. We expect the conventional results to be accurate, given the straight-forward algorithm employed and the 1000 Monte-Carlo simulations employed in the evaluation.

A. *Representative results*

Fig. 2 compares histograms of the imaginary part of S_{13} evaluated by the conventional approach (the red histogram in the figure) to the first histogram based on the measurement residuals, which was representative of other histograms evaluated during the 1000 trials of the algorithm of [1]. The conventional approach is based on straightforward Monte-Carlo uncertainty propagation through the models for the calibration artifacts and serves as our baseline result. While Fig. 2 only shows one result from the 1000 trials we ran, and thus is necessarily qualitative in nature, it does illustrate the ability of the algorithm of [1] to reasonably evaluate the mean and standard uncertainty obtained from the conventional algorithm.

Nevertheless, the results in Fig. 2, while necessarily qualitative in nature due to the single trial on which the figure is based, do indicate a systematic offset in the probability distribution function evaluated by the algorithm of [1] when compared to the conventional algorithm. Next, we quantify this and other differences between the algorithm of [1] and the conventional algorithm by summarizing the statistics we gathered over the full 1000 trials we ran.

TABLE I
SUMMARY STATISTICS FOR MONTE-CARLO RESIDUAL LINE ERRORS

Total Number RWG Lines	Normalized Std. Dev. MC Mean $SD(\mu_e)/\mu_a$	Actual Model Variance v_a (x10^{-6})	Statistical Bias of Eval. Variance (%)	Normalized Std. Dev. Variance $SD(v_e)/v_a$
5	0.433	7.6	83	1.603
9	0.311	3.8	43	0.976
17	0.227	2.0	18	0.412
33	0.163	1.0	8	0.259

B. *Complete Analysis*

Table I presents a quantitative summary of the key statistics related to the performance of the algorithm of [1] in this portion of the study over all 1000 trials as a function of the number of RWG lines included in the calibration. The second column of the table lists the normalized standard deviation of the Monte-Carlo means μ_e of the imaginary part of S_{13} over the 1000 trials as a function of the number of RWG lines used in the calibration listed in the first column of the table. The standard deviation $SD(\mu_e)$ was evaluated around the actual mean μ_a evaluated by the conventional Monte-Carlo method and then normalized by the actual mean μ_a to obtain the normalized standard deviation of the mean $SD(\mu_e)/\mu_a$. Here we defined the standard deviation of the mean μ_e evaluated by the algorithm of [1] around the conventional mean μ_a with the formula $SD(\mu_e) = \sqrt{(1/n)\sum_{i=1}^{n}(\mu_{e,i}-\mu_a)^2}$, where the mean $\mu_{e,i}$ of the imaginary part of the transmission coefficients S_{13} of the calibration model was evaluated in the i^{th} trial of the n=1000 trials.

The table shows that the normalized standard deviation of the mean $SD(\mu_e)/\mu_a$ decreases as the number of RWG calibration lines increases. This indicates that the algorithm of [1] does a better job of evaluating the Monte-Carlo mean μ_e as the

number of RWG calibration lines is increased, as we would expect to happen due to the increased averaging when the number of calibration artifacts are increased.

The third column of the table lists the actual variances of the of the imaginary part of the transmission coefficients S_{13}. Here we see that the actual variance in the calibration model drops as the number of RWG calibration lines is increased. Again, we see the overall benefits of using more calibration artifacts to reduce uncertainty levels.

The last two columns of Table I are more indicative of the ability of the algorithm of [1] to correctly evaluate uncertainty as a function of the number of calibration artifacts. The fourth column lists the statistical bias of the variances evaluated by the algorithm of [1], which decreases as the number of RWG calibration lines is increased. This shows that the accuracy of the uncertainties evaluated by the algorithm of [1] improves with the number calibration artifacts. Thus, not only are the uncertainties lowered by increasing the number of artifacts, but the relative accuracy with which that uncertainty can be evaluated from the residuals improves as well.

Finally, the table lists the normalized standard deviation of the evaluated variances of the imaginary part of S_{13} in its fifth column. These variances are evaluated around the actual variance v_a evaluated by the conventional method and normalized by the actual variance v_a to obtain $\mathrm{SD}(v_\mathrm{e})/v_\mathrm{a}$, where $\mathrm{SD}(v_\mathrm{e}) = \sqrt{(1/n)\sum_{i=1}^{n}\left(v_{\mathrm{e},i} - v_\mathrm{a}\right)^2}$ and the variances $v_{\mathrm{e},i}$ are the variances evaluated by the algorithm of [1] in the i^{th} trial. Again, while not perfect, Table I shows that the ability of the algorithm of [1] to evaluate uncertainty improves as the number of calibration artifacts is increased.

IV. Systematic Test-Port and Random RWG Line Offsets

Now we consider the same WR-10 calibration with random offsets *and* with imperfect test ports having a systematic offset with a range of roughly ±0.1 mm in the vertical E-plane direction between each of the test-ports' rectangular apertures and their alignment pins, which were not considered in Section III.

Here, the range of variation of the E-plane offsets in the test ports were chosen to be consistent with those in Section III. In each trial these two test-port offsets were fixed and added systematic error to the overall result. However, the Monte-Carlo bootstrap algorithm of [1] is provided with a set of Monte-Carlo replicates characterizing the PDFs of these test-port offsets, as shown in Fig. 2. We then assessed the ability of the algorithm of [1] to evaluate, in each trial, the uncertainty evaluated from the regression residuals when both the systematic test-port offsets and random RWG calibration line offsets vary (see Fig. 2). We performed the evaluation by comparing the uncertainties evaluated from the regression residuals to the uncertainties evaluated by the conventional approach.

Fig. 4 compares histograms of the imaginary part of S_{13} evaluated by the conventional approach (red histogram in the figure) to the first histogram when both systematic test-port and random RWG line offsets were included in the simulation. Here the uncertainty due to the systematic test-port offsets tends to dominate that due to the line offsets, which tend to average out when the number of RWG lines in the calibration is large.

The agreement between the histograms in Fig. 4 indicates that the algorithm of [1] does an excellent job overall of leveraging the Monte-Carlo replicates associated with the systematic test-port offsets and correctly evaluating the PDF of the imaginary part of S_{13}. Furthermore, as the number of lines is increased, the PDFs of the imaginary part of S_{13} more and more closely resemble the PDF of the underlying systematic component of the PDF due to the single test-port offset, which has a peak at 0 and decays in a fashion similar to a Rayleigh distribution. However, the systematic component of the distribution is also broadened by the Gaussian-like PDF of the random line offsets and translated to somewhat lower values of $\mathrm{Im}(S_{13})$ by the additional statistical bias in the random component of the PDF. This illustrates nicely the way in which the Monte-Carlo bootstrap algorithm of [1] preserves variation due to systematic error associated with its input quantities as it combines that systematic variation with variation and additional statistical bias originating from to the random error it captures with the regression residuals.

Table II summarizes the key statistics we collected over the 1000 trials. Here we see that, while not perfect, the accuracy of the algorithm of [1] is quite high and continues to improve in every aspect as the number of calibration lines is increased.

TABLE II
Summary Statistics For Monte-Carlo
Systematic Test-Port Errors and Residual Line Errors

Total Number RWG Lines	Normalized Std. Dev. MC Mean $\mathrm{SD}(\mu_\mathrm{e})/\mu_\mathrm{a}$	Actual Model Variance v_a (x10^{-5})	Statistical Bias of Eval. Variance (%)	Normalized Std. Dev. Variance $\mathrm{SD}(v_\mathrm{e})/v_\mathrm{a}$
5	0.241	5.7	-27	0.324
9	0.180	3.9	-14	0.161
17	0.144	3.1	-2.4	0.049
33	0.112	2.8	3.3	0.050

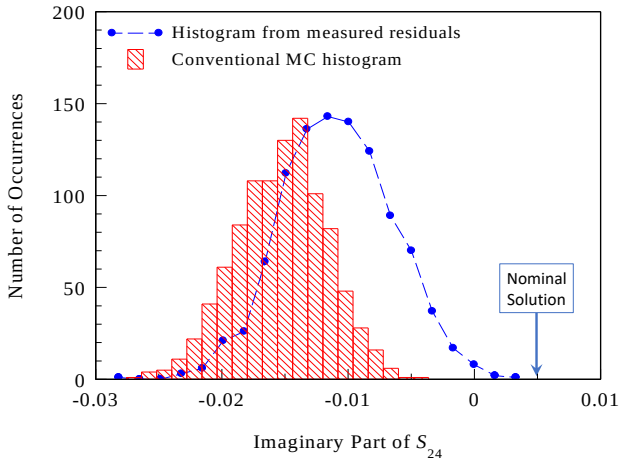

Fig. 5. Typical histograms of the imaginary part of S_{24} of the calibration model due to random RWG line errors evaluated by the conventional algorithm (red bars) and the Monte-Carlo bootstrap algorithm using regression residuals (blue dots).

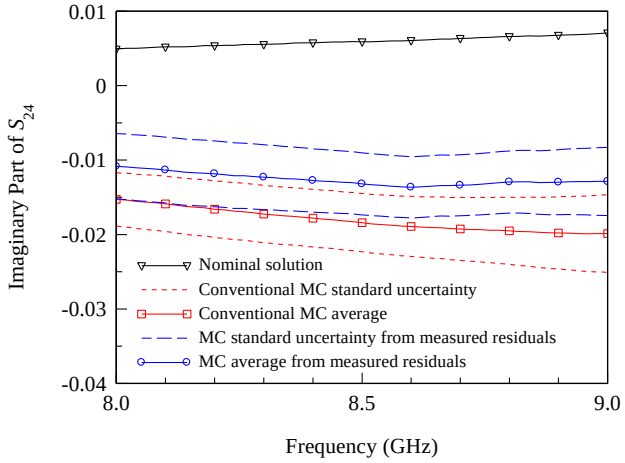

Fig. 6. The Monte-Carlo averages and standard deviations of the imaginary part of S_{24} of the calibration model due to random RWG line errors evaluated by the conventional algorithm (red curves) and the Monte-Carlo bootstrap algorithm using regression residuals (blue curves).

V. Measurement Verification

We used measurements of WR 90 RWG transmission lines to verify the simulations we performed in Section III using the Monte-Carlo bootstrap algorithm of [1]. We followed the procedure shown in Fig. 1, except that we used real measured data with deliberately introduced E-plane offsets. We chose WR-90 because it increased all the mechanical waveguide dimensions in the experiment by roughly a factor of 10, simplifying the introduction of the physical E-plane offsets before performing each measurement. We also chose the ratio of these offsets to the height of the WR 90 line to be comparable to the ratio of the simulated offsets to the height of the WR 10 lines used in Sections III and IV to amplify the impact of the E-plane offsets over other errors in the experiment and make them dominant while not exceeding the range of offsets supported by the Microwave Uncertainty Framework, as before. Finally, we chose to use a 3.18 cm long and a 5 cm long section of WR 90 in the calibration instead of a line and a thru. We did this because the E-plane offsets we introduced in the lines were easier to control in our experimental setup with these longer lines than the offsets between the test ports.

Due to experimental limitations, we were only able to perform the equivalent of one trial of the simulation study performed in Section III. Thus, this verification study, being based on a single experiment, is necessarily much more qualitative than the rigorous 1000-trial simulation studies in the previous sections and in [1]. In addition, also due to experimental limitations, we were only able to introduce vertical offset at the second port of each line.

We began the measurements by performing a first-tier thru-reflect-line (TRL) calibration performed with a pair of shorts, a straight thru connection, the 3.18 cm long WR 90 line and the 5 cm long WR 90 line. This calibration simplified the analysis by setting the calibration coefficients determined by subsequent calibrations to values close to those of an ideal "transparent" calibration.

Next, we measured each of the two WR 90 lines 14 times with deliberately introduced vertical E-plane offsets of up to 2.4 mm in port 2 of the line. We calculated the standard deviation of these E-plane offsets to be 1.29 mm around 0. Finally, we built models of the two lines from their lengths and dimensions and used these models to construct two second-tier Levenberg-Marquardt calibrations using the same models in the NIST Microwave Uncertainty Framework [7].

As in Section III, the first of the two second-tier calibrations we applied was a conventional algorithm based on the models of the two lines. Fig. 5 indicates the nominal solution of the imaginary part of the calibration coefficient S_{24} with a blue arrow, which had a value of 0.0049.[2] This non-zero nominal solution was primarily due to small inconsistencies in the calibrated measurements of the lines derived from the first-tier TRL calibration and calibration models we developed for the two lines based on their lengths.

The uncertainties evaluated by this conventional calibration were based on propagating the uncertainty of the offsets directly through the calibration models and algorithm using the Microwave Uncertainty Framework. Fig. 5. shows a 1000-

TABLE III
Comparison of Conventional Simulation to Measured Results

Total Number Measured RWG Lines	Normalized Difference of MC Means $(\mu_e - \mu_a)/\lvert\mu_a - \mu_n\rvert$	Normalized Difference of MC Variances $(v_e - v_a)/v_a$
28	0.220	0.491

[2] We chose to look at S_{24} because this is the transmission coefficient in the error box on port 2 where we introduced the vertical E-plane offsets. However, all the calibration coefficients show similar behavior.

sample histogram of the Monte-Carlo calculations from this conventional algorithm in red. The histogram is clearly shifted towards negative values by the E-plane offsets, which always introduce a positive parasitic capacitance at the junction between the lines and the second test port.

The second second-tier calibration was based on the same line models and measurements used in the conventional calibration, but used the Monte-Carlo bootstrap algorithm of [1], as implemented in the Microwave Uncertainty Framework, to evaluate the uncertainty in the imaginary part of S_{24} from the regression residuals. The resulting histogram evaluated from the regression residuals is marked with blue dots in Fig. 5. While based on a single calibration and thus necessarily qualitative, the two histograms in Fig. 5 are indeed comparable.

Fig. 6 compares the Monte-Carlo averages and standard uncertainties evaluated by the two calibrations as a function of frequency. Here again the results are in comparable agreement.

Finally, Table III lists the normalized difference of the Monte-Carlo mean μ_e evaluated by the algorithm of [1] from the regression residuals to the Monte-Carlo mean μ_a evaluated by the conventional algorithm normalized by the total statistical bias $\mu_a - \mu_n$, where μ_n is the nominal solution determined by the conventional algorithm in the second column. Table III also lists the normalized difference of the Monte-Carlo variance υ_e evaluated by the algorithm of [1] from the regression residuals to the Monte-Carlo variance υ_a evaluated by the conventional algorithm normalized by the variance υ_a evaluated by the conventional algorithm in the third column. The table shows that, while the statistical bias evaluated by the conventional algorithm and the algorithm of [1] based on the regression residuals are comparable, the differences of the variances are somewhat higher. However, variations like this are not unexpected given that we only were able to perform a single verification experiment and illustrate the utility of the more quantitative 1000-trial simulation studies we performed.

VI. Conclusion

We investigated the ability of the Monte-Carlo bootstrap algorithm of [1] to evaluate uncertainty of nonlinear microwave calibration models from regression residuals both quantitatively with simulated data and qualitatively with measured data. We quantified the ability of the algorithm to estimate the random uncertainty in the calibration and showed that the performance of the algorithm improves as the calibration becomes more overdetermined, even in the presence of significant nonlinearity in the calibrations. This illustrates that, even when models for calibration errors are not available, which often occurs when it is not possible to completely characterized calibration artifacts or understand all of the sources of error in an experiment or a data set, regression residuals to be used to evaluate the uncertainty in the calibration models. Furthermore, we showed that increasing the number of calibration artifacts not only reduces the overall level of uncertainty, but is also improves the relative accuracy of the estimates of that uncertainty provided by the algorithm of [1].

We also demonstrated the ability of the algorithm of [1] to combine random uncertainties evaluated from measurement residuals with systematic errors associated with the input quantities, even in the presence of nonlinear interactions between the random and systematic error mechanisms. Thus, we demonstrated that the algorithm of [1] can be used to automate the evaluation of uncertainty even when both random errors of unknown origin coexist with better-understood systematic errors in the problem. This greatly simplifies the use of regression residuals when they are needed.

Finally, the application of the algorithm to measured data showed that the method produces realistic estimates and that the evaluated uncertainties from the algorithm qualitatively matched those of the conventional approach.

References

[1] D. Williams, B. F. Jamroz, J. Rezac, and R. Jones, "Evaluating Uncertainty of Microwave Calibrations with Regression Residuals," *IEEE Transactions on Microwave Theory and Techniques*, accepted for publication.

[2] 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.

[3] 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 *International Microwave Symposium*, Boston, MA, June 2-7 2019.

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

[5] 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," *International Organization for Standardization*, vol. JCGM 101, 2008. [Online]. Available: http://www.bipm.org/en/publications/guides/gum.html.

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

[7] *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

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

[9] M. Garelli and A. Ferrero, "A Unified Theory for S-Parameter Uncertainty Evaluation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 12, pp. 3844-3855, 2012, doi: 10.1109/TMTT.2012.2221733.

[10] G. Avolio *et al.*, "Software tools for uncertainty evaluation in VNA measurements: A comparative study," in *2017 89th ARFTG Microwave Measurement Conference (ARFTG)*, 9 June 2017, pp. 1-7, doi: 10.1109/ARFTG.2017.8000820.

[11] 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.

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

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

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

[15] 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.

[16] D. F. Williams, "500 GHz – 750 GHz rectangular-waveguide vector-network-analyzer calibrations," *IEEE Trans.Terahertz Sci.Technol.*, vol. 1, no. 2, pp. 364-377, November 2011.

[17] J. D. Hunter, "The displaced rectangular waveguide junction and its use as an adjustable reference reflection," *IEEE Trans.Microw.Theory Techn.*, vol. 32, no. 4, pp. 387-394, 4/1984 1984.