Evaluation Framework for Assessing Validation Methods on Modeling and Simulation Models

Stephanie Y. Su, Samantha K. McCarty, Joseph D. Warfield Jr., Eric J. Uthoff, and Simone M. Youngblood

ABSTRACT

Modeling and simulation (M&S) is a critical step throughout the systems engineering process for developing and fielding a combat system. Verification and, more specifically, validation are essential to determining whether a simulation is credible and reliable. Although policy and guidance increasingly emphasizes the importance of rigorous validation founded in the application of strong statistical analysis, implementation continues to be challenging. As a result, test organizations and statisticians have been interested in developing a robust approach for measuring the performance of the validation methods used to assess model accuracy. The Johns Hopkins University Applied Physics Laboratory (APL) developed a flexible and extensible framework to evaluate the performance of the validation methods. The framework provides the modularity to evaluate multiple validation methods and is sufficiently generic to support assessment of multiple simulation models. This article details the framework design and the analysis of multiple statistical validation methods, including an exemplar assessment of the methods applied for a recently accredited missile system simulation.

INTRODUCTION

The Department of Defense (DOD) and the armed forces have recognized the growing significance of modeling and simulation (M&S) for many aspects of their operations. They have prepared directives and guidelines to provide general instructions on how, when, and under what circumstances formal verification, validation, and accreditation (VV&A) procedures should be employed. These three processes are defined as follows:

1. Verification—*"Did I build the thing right?"*[1]

   "The process of determining that a model implementation and its associated data accurately represent the developer's conceptual description and specifications."[2]

2. Validation—*"Did I build the right thing?"*[1]

   "The process of determining the degree to which a model and its associated data provide an accurate representation of the real world from the perspective of the intended uses of the model."[2]

3. Accreditation—*"Is it believable enough to be used?"*[1]

   "The official certification that a model or simulation and its associated data are acceptable for use for a specific purpose."[2]

APL supports the systems engineering life cycle for multiple combat systems. Of increasing importance is the application of high-fidelity, system-of-systems (SoS) distributed simulations. These SoS simulations consist of a mixture of both tactical code-based models and physics-based environmental models to represent combat system performance.

As part of the VV&A effort, models are tested, and data, such as telemetry data from tests events, are collected. These data are then compared with the SoS simulation results to assess the simulation's ability to accurately represent the actual behavior and performance of the combat system. It is challenging to select the most efficient and effective validation methods to apply in a given situation. To aid in overcoming this challenge, APL developed a modular, flexible, and transferable evaluation framework that expands validation methods for time-series data and enables better assessment of the validation methods to be applied.

VALIDATION METHOD PERFORMANCE

Validation methods seek to determine whether data derived from a physical test are consistent with data output by a simulation of that same test. This can be done by comparing the data from the test (test data set) with one or more data sets generated by the simulation (comparison data sets). The fundamental question then becomes how accurately the validation method classifies a given test data set as either in-family (consistent with the comparison data) or out-of-family (inconsistent with the comparison data)?

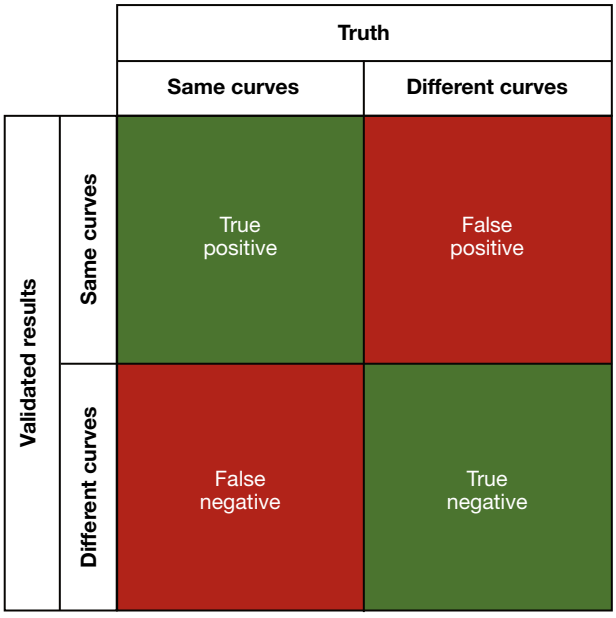

Figure 1. A contingency matrix for evaluating a method's performance. The sensitivity (true positive) and specificity (true negative) are representative of the correctness of methods implemented.

To evaluate the accuracy of the validation methods used to assess a simulation, we can compare simulation data with other simulation data. Because the true family of the simulation data is always known, it is straightforward to verify whether the classification of the validation method was correct or not. The result may then be classified as either a true positive (TP) or true negative (TN) (the test data set was correctly classified as in-family or out-of-family) or false positive (FP) or false negative (FN) (the test data set was incorrectly classified as in-family or out-of-family). For classification problems, these four measurements are combined to produce a contingency matrix, an example of which is shown in Figure 1. It is desirable to maximize the true positive (sensitivity) and true negative (specificity) rates. We define the validation method performance in Eq. 1.[3] This matrix can be used to assess the validation method's performance across various metrics in simulation models.

$$\text{Performance accuracy} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{TN}} + N_{\text{FP}} + N_{\text{FN}}} \quad (1)$$

To test the performance of the validation methods, time-series data were generated for several scenarios using a 6-degree-of-freedom (DoF) missile system simulation. For each scenario, N (270–300) trial data sets were generated, and each scenario tracks the values of four metrics (ground range, ground range velocity, height, and height velocity) over time. Comparing trials from a given data set with other trials from the same data set produces the true positive and false negative rates in the left column of the contingency matrix. Comparing the data from a given trial with the trials in a different data set produces the true negative and false positive rates in the right side of the contingency matrix.

To perform the within-scenario comparison, a trial from one scenario data set is selected and treated as the test data, while the remaining $N - 1$ trials are treated as the comparison data set. A validation method is then used to compare the "test data" with the comparison data and classify it as either in-family or out-of-family. This process is repeated for each trial in the scenario data set to acquire the true positive and false negative rates. To perform the between-scenario comparison, a trial from a scenario data set is used as test data and compared with the trials from a different scenario data set as the comparison data. Here, no trials are excluded from the comparison data since the test data trial is not a part of it. This process is performed for each metric. This analysis approach enables the identification of how different two curves are, given a metric to evaluate, which can help in sensitivity analysis.

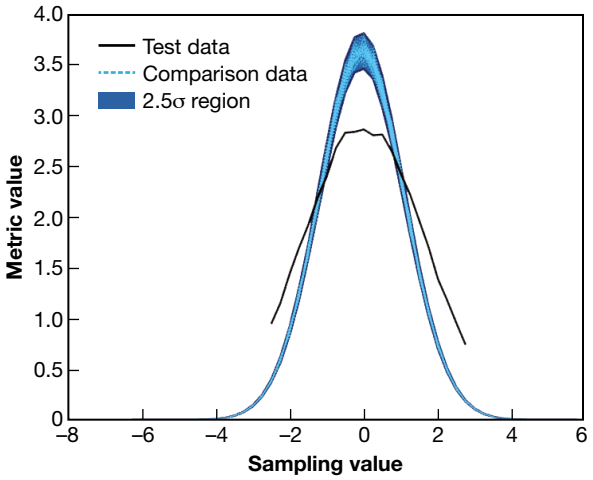

Figure 2. An example of a comparison of test data with an out-of-family comparison data set.

STATISTICAL METHODS FOR VALIDATION

The missile system model is a physics-based simulation that emulates the behavior of missile system flyouts. The missile system model used a two-sided one-sample hypothesis test[4] on four metrics (ground range, ground range velocity, height, and height velocity) to perform validation. To compare the performance of the selected validation methods, they were applied to data from two different scenarios for all four of the previously employed metrics. The scenarios were selected because they exhibit similar metric data distributions. A good statistical method should be able to distinguish between different scenarios that have similar behavior. In the analysis of the 6-DoF missile system model, for the test data to be considered in-family, all metrics must meet the in-family criteria. We evaluated four validation methods: two-sided hypothesis, area-outside-threshold hypothesis, extrema hypothesis, and extrema *p*-value hypothesis.

Two-Sided Hypothesis

The two-sided hypothesis is the current approved validation method. This method takes all the trials in the comparison data set and calculates the mean and standard deviation for each time sample. Then, a region is defined bounded by $\mu_n \pm 2.5\sigma_n$, where μ_n and σ_n are the mean and the standard deviation for nth time sample, respectively. A test data metric is classified as in-family if more than 90% of its time samples are within this region. Figure 2 shows an example of test data (black line) compared with an out-of-family comparison data set (blue) and failing the two-sided hypothesis test.

The two-sided hypothesis test has some potential limitations. If there are few time samples, even a small number of samples going out of threshold can cause test data to be rejected as out-of-family. Additionally, test data that runs along the edge of the threshold, such as that depicted in Figure 3a, can result in many time samples going out of threshold. Both these scenarios can potentially produce false negatives.

Area-Outside-Threshold Hypothesis

The area-outside-threshold hypothesis attempts to improve performance of the two-sided hypothesis by considering instead the area between time samples that are out of threshold and the threshold. This "area outside" is depicted for an out-of-family comparison case in Figure 3b. Test data are judged by calculating the total area outside of threshold and scaling it by the time range over which the data are defined. To judge this value, the comparison data are compared with themselves. The

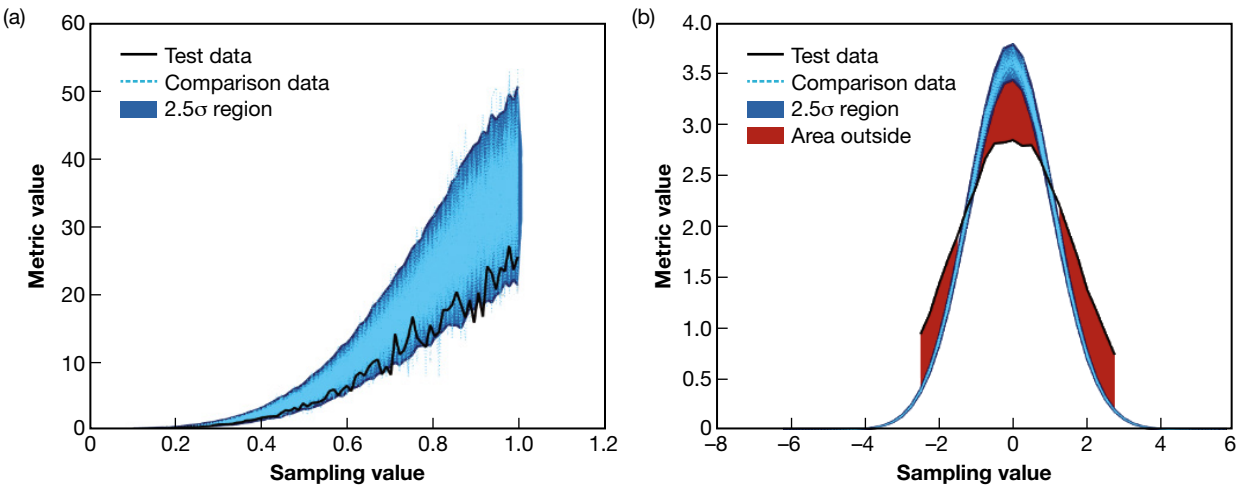

Figure 3. Area-outside-threshold hypothesis. (a) Sometimes test data may run along the edge of the two-sided threshold region. This can result in too many time samples being out of threshold. (b) A depiction of the area-outside-threshold test. The red region represents the area outside of the two-sided threshold.

area outside of threshold is calculated for each trial of a data set relative to the other trials. The results from each trial then form a distribution from which the xth percentile is calculated for use as a threshold. Any test data with a total time-scaled area outside the threshold less than the xth percentile value are classified as in-family. Otherwise, they are rejected as out-of-family.

Extrema Hypothesis

The extrema hypothesis is a simplified form of the Rosenblatt process modified to accommodate the missile system time-series data.[5] It breaks up the N trials into g groups of size $N_g = N / g$ trials per group. Time samples in the test data are then compared with the comparison data in each group individually. For a given time sample, if the test data are higher than the maximum or lower than the minimum compared with any of the trials in a group, the test data are an extrema. For each group, the total number of extrema (N_{ex}) in the test data relative to the group in question is counted and scaled by the number of time samples (N_s). This fraction, N_{ex} / N_s, is then compared with a threshold of $2 / (N_g + 1)$. If the scaled number of extrema for *any group* exceeds this threshold, the test data are rejected as out-of-family. Otherwise, they are considered in-family. Figure 4 shows an example of the extrema hypothesis applied for an out-of-family comparison. The red dots indicate test data time samples that are extrema.

Extrema *p*-Value Hypothesis

The extrema p-value test was developed as an attempt to implement portions of the extrema hypothesis but "soften" the in-family threshold. Assume that a trial is composed of data with a Gaussian distribution around some truth function in a statistically independent fashion. If this is true, then the likelihood of a given test data time sample being an extremum is given by $p = 2 / (N + 1)$, where N is the number of trials. This is derived from the fact that for any given collection of time samples, there will be one maximum and one minimum. It then follows that the number of observed extrema in the test data will follow a binomial distribution with a "probability of success" equal to p and a "number of trials" equal to the number of time samples.

To determine whether test data are in-family, a similar process to the extrema hypothesis is performed. The minimum and maximum of each time sample are calculated, and if the test data are greater than the maximum or less than the minimum, they become extrema. The total number of extrema N_{ex} is then calculated.

Now, consider the null hypothesis $\mathcal{H}_0$ that the test data are in-family and the alternative hypothesis $\mathcal{H}_1$ that the test data are out-of-family. These hypotheses can be differentiated by defining a test metric $\lambda = N_{ex}$. Since this value follows a binomial distribution, it is possible to calculate the p-value for likelihood of seeing N_{ex} extrema or more. If the p-value is below the standard threshold of $\alpha = 0.05$, then the test data are rejected as out-of-family. Otherwise, they are accepted as in-family.

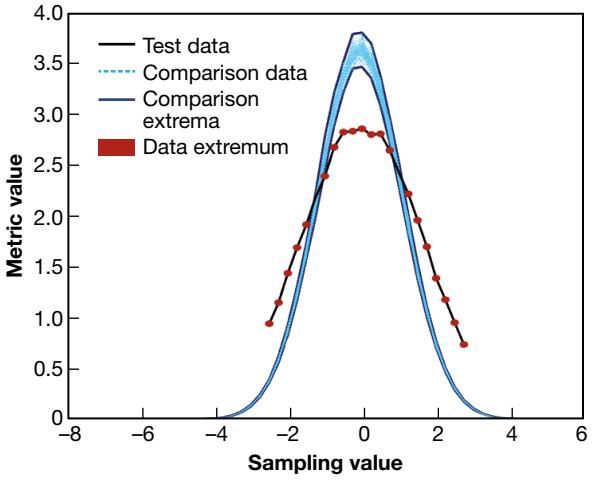

Figure 4. An example of the extrema hypothesis for an out-of-family comparison. Each red dot represents a time sample in the test data that is either a maximum or minimum relative to the comparison data set.

VALIDATION METHOD ASSESSMENT FRAMEWORK

To evaluate the performance of the statistical methods proposed to validate the simulation model, we created the validation method assessment framework (VMAF). VMAF is a flexible, extensible code base for implementing and running validation methods written in MATLAB. It is designed to standardize as much of the validation process as possible and render it easy to run without any underlying knowledge of the methods or code that make up the framework. Additionally, it is designed to be portable to enable easy sharing across organizations. The VMAF is split into three parts: data preprocessing, analysis preprocessing, and finally the validation method itself.

Data Preprocessing

The data preprocessing step collates the raw data files. The preprocessor allows for only the specific metric data needed to be drawn into the framework, where the data from across all trials are compiled in a standardized way for all scenarios and then saved to disk to allow for faster subsequent processing. The resultant output, a compiled data object, can serve as an input to any validation method. Additionally, the data preprocessor can perform limited transformations of metric data during the collation process.

Analysis Preprocessing

From the compiled data object, an analysis preprocessor generates the data the corresponding validation method needs to compare against the test data. Since it is impossible to know what an arbitrary validation method may need a priori, an analysis preprocessor is defined for each validation method. In pursuit of the goal of standardizing the process as much as possible, all the background machinery, such as extracting data from the compiled data object and creating self-comparison data (the "test data" trial), was carried out before the analysis preprocessing. The only step that must be specifically defined for a given validation method is what data to generate. Again, a standardized object is generated and output.

Validation Method

Lastly, the validation method is defined. It takes as input test data, comparison data, and the analysis preprocessing output for the comparison data. From this, it will classify the trials present in the test data as in-family or out-of-family. As with the analysis preprocessing step, only the actual classification logic needs to be defined when implementing new methods. All the other backend steps are automated and consistent across all validation methods.

Control Scripts

Each stage of the VMAF is run via a control script. A control script is a basic script that has fields for the user to define the required input (such as which files to collate and which metric data to load for the data preprocessor or the size of the thresholds for the two-sided test). It includes significant documentation to describe what is needed and examples to assist the user. Once provided with the necessary input, it can simply be run, and the rest of the process is handled automatically.

RESULTS AND DISCUSSION

The VMAF was employed to test two scenarios of 6-DoF missile system data using the two-sided, area-outside-threshold, extrema hypothesis, and extrema *p*-value hypothesis tests for the four metrics described previously. The values defining the contingency matrix for all metrics are given in Figures 5–8, respectively, for each validation method. The ground range and ground range velocity metrics exhibited very similar forms between the two scenarios and proved difficult to differentiate. The height and height velocity metrics presented sufficiently distinctive forms, so they were reliably differentiable.

Two-Sided Hypothesis

The two-sided hypothesis test performed well. As shown in Figure 5, it reliably produced near-100% true positive rates as well as 100% true negative rates for height and height velocity, resulting in near-100% performance accuracy for these areas. True negative rates for the ground range and ground range velocity were poorer, yielding an overall averaged accuracy of 76.1% across all metrics. However, since rejecting one metric rejects the entire test data, this method is still able to consistently separate the test data from the out-of-family comparison data.

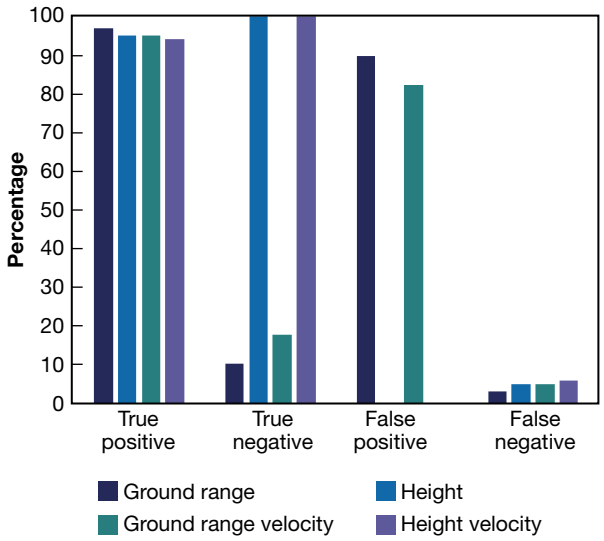

Figure 5. The results of the two-sided hypothesis test. True positive rates are extremely good, and at least some true negative rates are extremely reliable.

Area-Outside-Threshold Hypothesis

Figure 6 shows the performance of the area-outside-threshold hypothesis with 90th percentile and 50th percentile thresholds. Using the loose 90th percentile threshold, the area-outside-threshold hypothesis test was able to improve on the true positive rate compared with the two-sided threshold hypothesis test. However, it was considerably less effective at identifying true negatives. The method presented a metric-averaged accuracy of 62.9%. Using a stricter 50th percentile threshold, the area-outside-threshold test was able to achieve similar true positive and true negative rates to the two-sided threshold test, but no percentile resulted in improvement. Additionally, the need to calculate the percentile threshold uniquely for each comparison data set is not ideal. However, the tested implementation was fairly simple, so while it presently offers no benefit over the two-sided hypothesis and thus is not preferred in its current form, there may be room to improve the area-outside-threshold hypothesis in the future.

Extrema Hypothesis

The extrema hypothesis test was performed splitting a total of 300 trials into N_g = 10 and N_g = 30 groups,

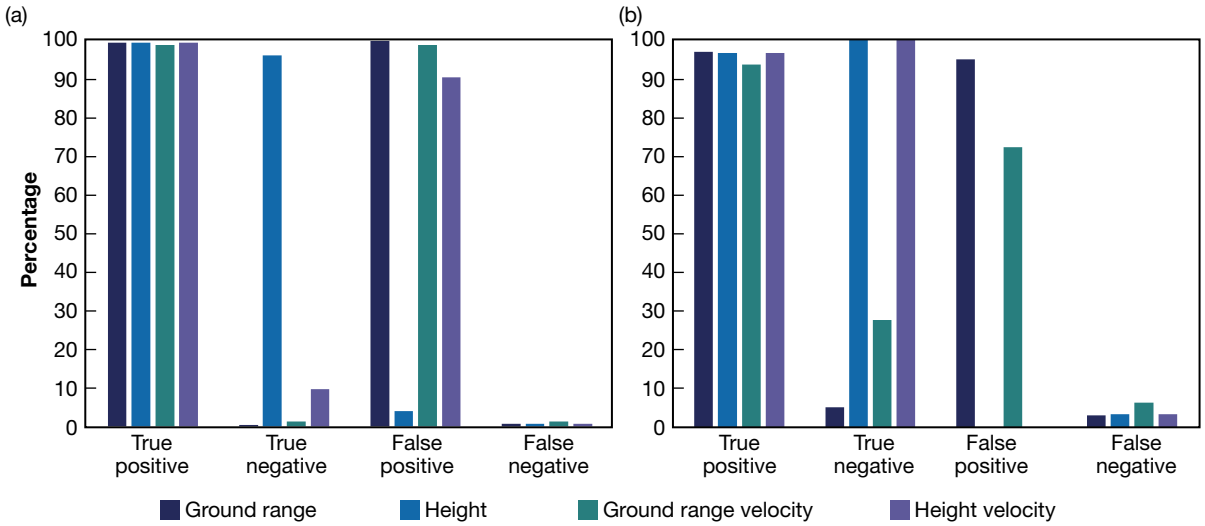

Figure 6. The results of the area-outside-threshold hypothesis test. (a) The results with a 90th percentile threshold. The true positive rate is improved over the two-sided hypothesis, but at the cost of worse true negative rates. (b) The results with a 50th percentile threshold. The true negative rates are now in line with the two-sided hypothesis, but so are the true positive rates.

representing 30 and 10 trials per group, respectively. The method's metric-averaged accuracies were 74.5% and 55.2%. The strict threshold on the number of extrema resulted in a very good true negative rate. However, this came at the cost of rejecting a large percentage of in-family comparisons and thus producing a notably lower true positive rate. Figure 7 shows the results for each number of groups.

The performance of the extrema hypothesis test fairly strongly correlates with the number of trials in groups and time samples. Increasing the number of trials will generally increase the comparison data minimum and maximum, and thus reduce the likelihood of seeing an extremum in the test data. This effect can be observed by the improved performance of the smaller grouping size—fewer groups means more trials per group. This could additionally be improved by simply including more trials overall. Meanwhile, including more samples increases the likelihood of seeing more extrema. Conversely, it also reduces the "cost" of each extremum observed, since the total is scaled by the total number of time samples. Balancing these properties could improve the performance of the extrema hypothesis, but in general it is not suitable for the missile system simulation.

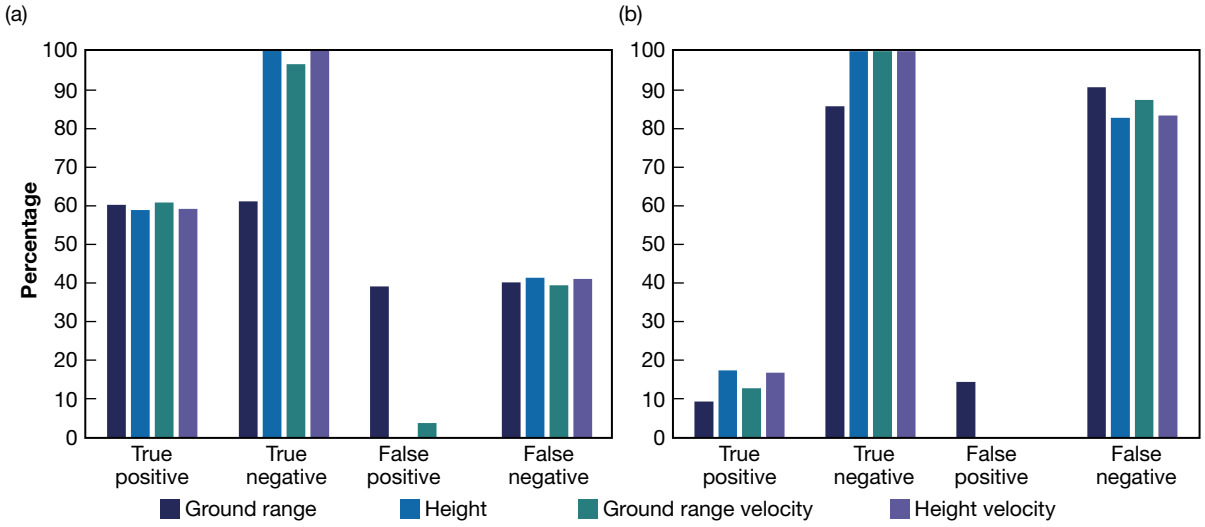

Figure 7. The results of the extrema hypothesis test. (a) The results for 10 groups. The true negative rate is very good, but it comes at the cost of a low true positive rate. (b) The results for 30 groups. The true negative rate is further improved, but the true positive rate is further reduced.

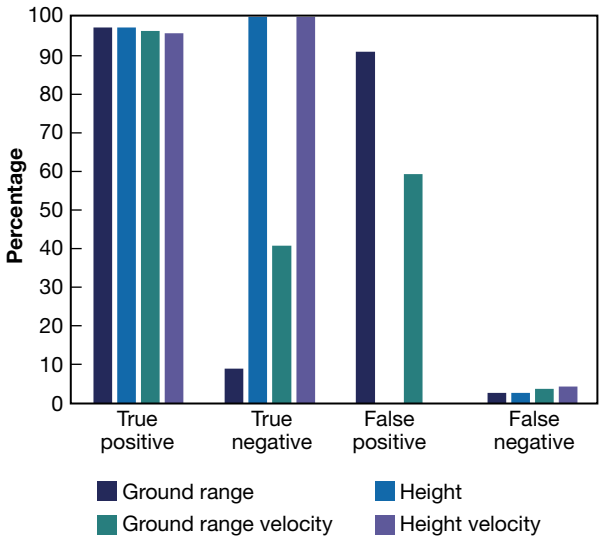

Figure 8. The results for the extrema *p*-value hypothesis test. It performs on par with the two-sided threshold hypothesis in most areas but actually outperforms it in others (such as the ground range velocity true negative rate).

Extrema *p*-Value Hypothesis

The extrema *p*-value test performed exceptionally well. In terms of true positive rate, it matched the two-sided threshold hypothesis, and in terms of true negative rate, it matched or exceeded the two-sided threshold hypothesis. This is notable in the ground range velocity true negative rate, which is 30% higher for the extrema *p*-value hypothesis.

Since it is able to match or exceed the two-sided threshold hypothesis across all metrics, the extrema *p*-value test is a viable contender for the preferred validation method for the 6-DoF missile system model.

FUTURE PROSPECTS

The VMAF provides the capability to consistently evaluate the performance of validation methods for many different data sources. Further, it can help alleviate the difficulty of sharing evaluation methodologies across organizations by providing a single unclassified, portable, and easy-to-use code base for implementing and running said methods. Each organization can test and improve the validation methods with their own models.

The two-sided threshold, area-outside-threshold, and extrema validation methods have been presented by the authors to external organizations. The authors will continue forward to present the validation method performance results from VMAF. Additionally, external organizations have shown interest in applying the VMAF to a time-series sensor model to determine whether the framework is extensible. Through this analysis and potentially others, the authors aim to address the need for strong statistical validation methods for all combat system models.

REFERENCES

[1]"VV&A recommended practices guide," Modeling and Simulation Enterprise, last modified May 18, 2011, https://vva.msco.mil/default.htm?Key/default.htm

[2]"DoD modeling and simulation (M&S) verification, validation, and accreditation (VV&A)," Department of Defense, *DoD Instruction 5000.61* (Dec. 2009), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500061p.pdf.

[3]M. Gulati, "How to choose evaluation metrics for classification models," last modified Oct. 11, 2020, https://www.analyticsvidhya.com/blog/2020/10/how-to-choose-evaluation-metrics-for-classification-model/.

[4]"NIST/SEMATECH e-handbook of statistical methods," National Institute of Standards and Technology, last modified Apr. 2012, https://www.itl.nist.gov/div898/handbook/prc/section2/prc22.htm.

[5]P. A. Jacobs, "An approach to the validation of a computer model with time series output," working paper, Naval Postgraduate School, Monterey, CA, 2022, https://faculty.nps.edu/pajacobs/.

Stephanie Y. Su, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Stephanie Y. Su is a systems engineer and Discovery Program member in APL's Research and Exploratory Development Department. She has a BS in physics from Pennsylvania State University and a PhD in physics from the University of Michigan. With expertise in particle physics research and data acquisition system, Stephanie has supported projects across multiple sectors involving operations analysis, system requirements analysis, and development for a space objects tracking system. During her rotation in APL's Air and Missile Defense Sector, she took on the effort, detailed in this article, to develop the evaluation framework to assess the performance of applying various validation methods on system models. Her email address is stephanie.su@jhuapl.edu.

Samantha K. McCarty, Air and Missile Defense Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Samantha K. McCarty is an analyst and software developer in APL's Air and Missile Defense Sector. She has a BS in physics and computer science from Gettysburg College and a PhD in physics from the University of New Hampshire with an emphasis on experimental particle physics and dark matter research. Samantha has over 10 years of experience in software engineering, primarily in a scientific environment. She has played lead roles in designing and maintaining calorimeter simulation software packages, optimizing and designing calorimeter data selection algorithms, performing statistical signal processing of physics data, and developing and improving model validation methodologies. Her email address is samantha.mccarty@jhuapl.edu.

Joseph D. Warfield Jr., Force Projection Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Joseph D. Warfield is a principal statistician and chief scientist of the System Performance Analysis Group in APL's Force Projection Sector. He received a BS in mathematics from Loyola University, an MS in statistics from Virginia Polytechnic Institute, and a PhD in statistics from University of Maryland, Baltimore County. His group focuses on testing and evaluation for Navy and Air Force programs. In addition, the group provides statistical consulting to programs across the Laboratory and to various government and defense agencies in the areas of experimental design, response surface methodology, nonlinear regression analysis, predictive analytics, and reliability-related methods. Dr. Warfield's current research areas focus on optimal design approaches for generalized linear model applications, Bayesian estimation of system reliability and risk quantification, and predictive analytics using structured and unstructured data sources. He has taught courses on design of experiments and regression analysis through APL's Strategic Education program. His email address is joseph.warfield@jhuapl.edu.

Eric J. Uthoff, Air and Missile Defense Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Eric J. Uthoff is a section supervisor in APL's Air and Missile Defense Sector. He has a BS in electrical engineering from the University of Texas and an MS in electrical engineering from the University of Michigan. Eric has 8 years of work experience, specializing in systems and signal processing, with 5 of those years in missile systems engineering. He specializes in digital signal processing, in particular estimation, detection, filtering, and machine learning. He has more than 3 years of simulation development experience and has demonstrated a strong technical proficiency in areas related to radio frequency sensors and missile sensor systems, as well as modeling of related missile system components. In addition, he is equipped with working knowledge of the various 6-degree-of-freedom (DoF) simulations and has a strong understanding of the Standard Missile-2 6-DoF simulation. His email address is eric.uthoff@jhuapl.edu.

Simone M. Youngblood, Air and Missile Defense Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Simone M. Youngblood is a member of the APL's Principal Professional Staff in the Air and Missile Defense Sector. She has a BA in mathematics and a BS in computer science from Fitchburg State University, as well as an MS in computer science from Johns Hopkins University. Leveraging an extensive background in simulation development and credibility assessment, Simone has served as a Department of Defense (DOD) verification, validation, and accreditation (VV&A) focal point for the past 28 years. She was the editor of the *DOD VV&A Recommended Practices Guide* and has chaired the development of several VV&A-related standards, including IEEE Standards 1278.4, 1516.4, and 1730.2 as well as MIL-STD-3022. She has served as the V&V and/or accreditation agent for numerous modeling and simulation efforts that span a broad organizational spectrum including Program Executive Office Integrated Warfare Systems 1, the Defense Threat Reduction Agency, the Department of Homeland Security, the US Naval Air Systems Command, and the US Army Medical Research and Development Command. Her email address is simone.youngblood@jhuapl.edu.