Uncertainty Analysis Demonstration: *A Missile Case Study*

Navreeta Singh and Jeremy Werner, DOT&E
November 2022

Background

Validation is a key component of modeling and simulation (M&S). Rigorous validation, in particular, requires quantification of the uncertainty between live data and simulation output. The following demonstration supposes that we have data from a live-fire missile explosion test and simulated explosions.[1] A statistical analysis determines the extent to which the data and simulation agree.

Primer on Uncertainty

Uncertainty quantification estimates the extent to which a quantity, as measured, may differ from its actual value. The uncertainties themselves arise from limitations in measurements or M&S and can be categorized as *statistical* or *systematic*. Figure 1 below highlights the differences between the two.

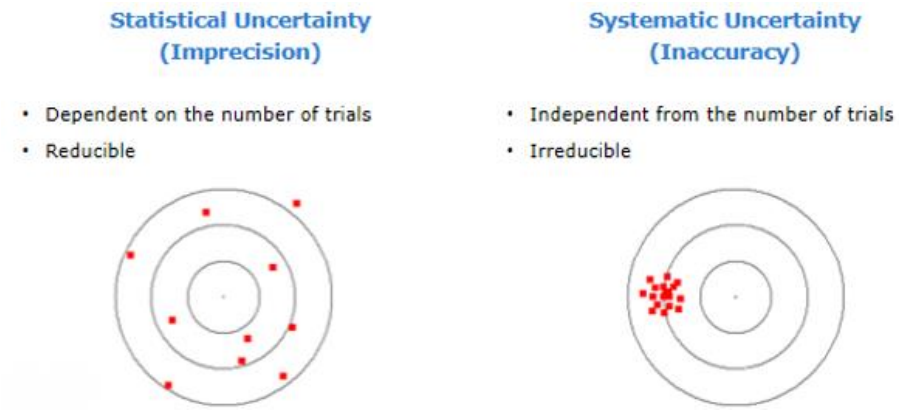

Figure 1: Visualization of the difference between Statistical and Systematic Uncertainty.

Statistical uncertainty arises from stochastic effects (probabilistic effects that occur by chance) in a measurement process and is an estimate of imprecision. As the cause is random, statistical uncertainty may be reduced by accumulating more samples, and it approaches zero as the number of samples goes to infinity. Take determining the mean weight of a basketball approved for NBA games as an example. Weighing 1,000 different balls then calculating the mean would yield a measurement with a much smaller statistical uncertainty than weighing only 10 balls.

Systematic uncertainty, on the other hand, is due to unknown but constant errors in measurement or M&S, which makes it independent of the number of samples. A systematic uncertainty estimates inaccuracy, and calibration error is a common source.

[1] This study builds upon and uses input from IDA's 2018 "Comparing M&S Output to Live Test Data: A Missile System Case Study" by D. Thomas and K. Avery, to include the visualization shown in the left side of Figure 2.

If the scale used to measure the weight of the basketballs was calibrated only to a tolerance of 10 grams, then the systematic error in the mean weight of the basketballs will always be 10 grams, no matter the number of samples.

A Case Study[2]

A model that characterizes a missile's impact on the target can significantly aid design and testing of that missile. By depicting the number of fragments that perforate the target at a given distance from missile burst, the model can predict the amount of damage caused. Model output can then be used, for example, to inform a proximity sensor on the missile and help maximize area coverage on its target.

Figure 2 below shows an experimental set-up for testing these parameters. The missile explodes in the center of the range, surrounded by "witness panels" at various distances from it. Few fragments impact the panels closest to the burst, as the explosion has not yet spread out. The panels farthest away also register few fragments, as the force of the explosion has dissipated. This leaves a mid-range "sweet spot," where the number of perforations is highest.

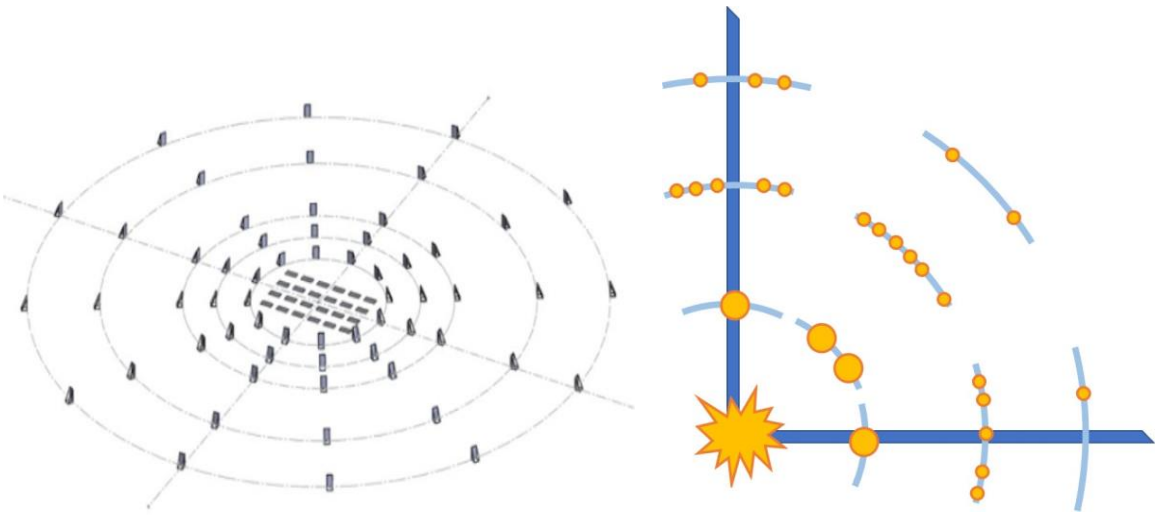

Figure 2: Experimental set-up and perforation dependency tied to radius.

Method

The live data are first fit to a regression model. Because the perforation data are count data (i.e., discrete rather than continuous), we considered the following two main model options: Poisson and Negative Binomial. The Poisson model typically is used to describe random events that occur over time or space, such as the number of car accidents per month or the number of pieces of gum on a sidewalk square. Poisson models assume that each event is independent and that the mean of the data equals the variance. However, in practice, events often are correlated, which typically causes the variance to become larger than the mean. This larger variance,

[2] The analysis software needed to automatically reproduce all the findings in this study is publicly available at http:FIXME.

or overdispersion, is indeed observed in the missile data, owing to the fact that the events – the fragment bursts from a single mission explosion – are not independent.

The following two models can account for overdispersion: the quasi-Poisson and the Negative Binomial. The quasi-Poisson model includes an extra dispersion parameter to estimate how many times larger the data variance is than the mean. The Negative Binomial model considers the distribution parameter itself as a random variable whose variation accounts for the overdispersion. Although these two models were considered in addition to Poisson, the quasi-Poisson model (as implemented in the base library of the R Project statistical computing language[3]) is functionally identical to the Poisson model in terms of the regression it produces. This analysis discusses only the Poisson and Negative Binomial results.[4]

After fitting the live data with Poisson and Negative Binomial distributions, the team computed the R^2 values (the proportions of variance in the data that are captured by the fitted regressions, also known as the coefficient of determination) and the probability values (p-values) of the χ^2 "goodness of fit" statistics, which estimate how well the models characterize the underlying data. Figure 3 below shows values computed from the average of 100 simulation runs, as well as from live data. The R^2 values for the Negative Binomial and Poisson fits are comparable, as the regression curves capture a similar proportion of the variance present in the data. However, the fits' χ^2 p-values reveal the difference between the two: The standard deviation of the Negative Binomial distribution is much wider, as this model accounts for overdispersion and, accordingly, yields a higher χ^2 p-value. Conversely, the Poisson fit doesn't account for overdispersion; thus its χ^2 p-value is almost negligible. The Negative Binomial fit therefore is the better choice and is used for the remainder of this analysis.

[3] https://www.r-project.org/

[4] When using statistical software, quasi models like quasi-Poisson also have limitations due to the fact that they do not produce exact likelihood. Several statistical tests and fit measures are unavailable.

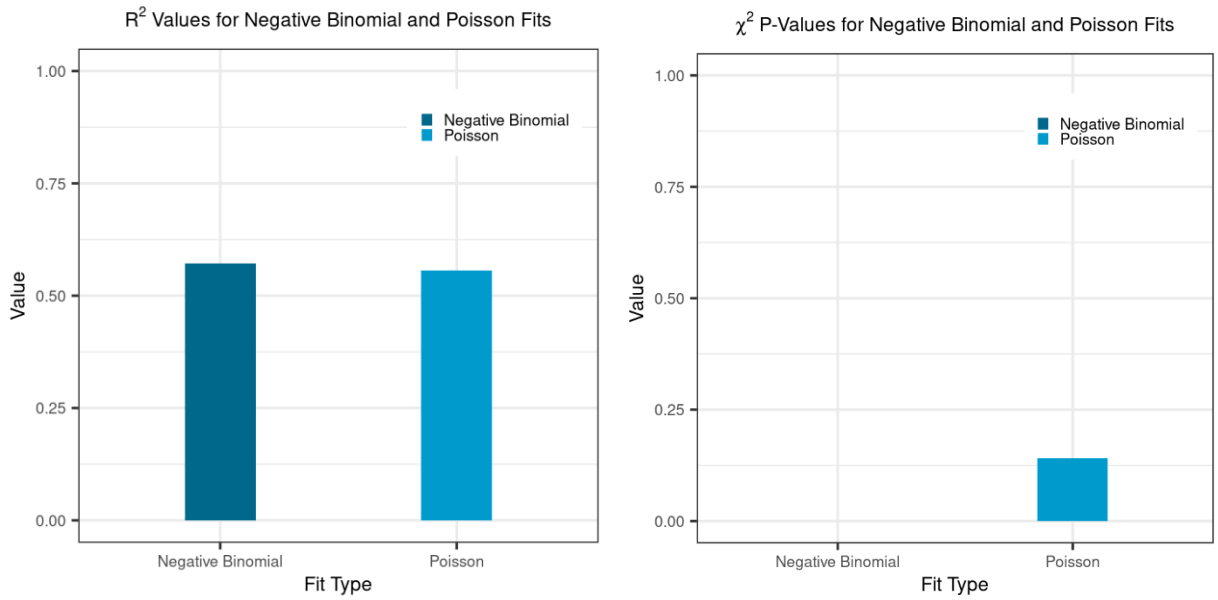

Figure 3: R^2 and χ^2 p-values for Negative Binomial and Poisson Fits.

Next, we compared the live data and simulation (see Figure 4 below). The 95% confidence band on the fit to the data, which is shown in gray, indicates that, for any given value on the horizontal axis, we are 95% sure that the mean of the parent distribution from which the data were sampled falls within these bounds. Shifting this interpretation, we can claim that, to be considered consistent with the data, the mean of the simulation itself (teal curve) must fall within the confidence band – and so we correspondingly changed the color of the gray confidence band to teal.

We then drew points from the simulation, while assigning each point an error bar that represents the width of the confidence band, as any given simulated point could be drawn from the lowest or highest end of the band and remain within bounds. In doing so, we reinterpreted the *statistical uncertainty* that is latent in the fit to the data (due to the limited number of samples from which it was generated) as a *systematic uncertainty* in the simulation; the mean of the simulation could fall anywhere within the gray band and still be considered consistent with the data.

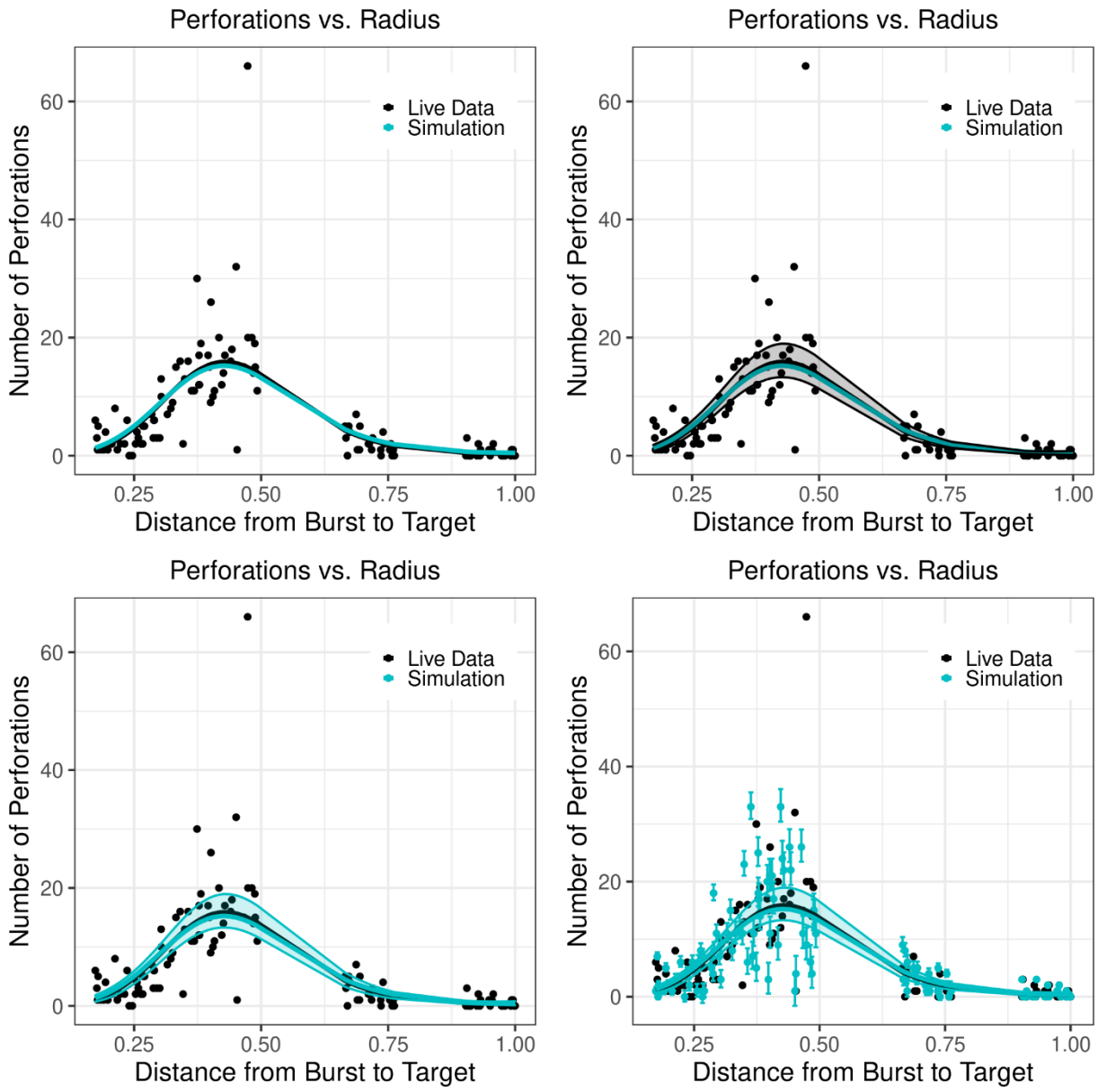

Figure 4: Live data and model curve, with 95% confidence bands and simulated values.

Analysis

Choosing the right hypothesis test is key. The left graph in Figure 5 below more clearly visualizes the extent of any disagreement between the two distributions, plotting the difference between the live data and simulation (black curve) with the 95% confidence band of that difference (gray band). The confidence band encompasses the horizontal axis (i.e., Live – Simulation = 0) for the entire range, demonstrating that the distributions are consistent.

Hypothesis testing allows us to quantify the extent to which the live data and simulation agree or disagree. Gaussian distributed data can be analyzed with a

student's t-test, which compares the mean of the simulation to that of the live data. But, because the data aren't Gaussian distributed, we instead used nonparametric tests that don't presume any distribution: the Kolmogorov-Smirnov (KS) test and the Mann-Whitney U (MW) test. The KS test compares the shapes of distributions through their empirical cumulative distribution functions (ECDFs), a plot of which is included on the right in Figure 5 below. The Mann-Whitney test compares the medians of the distributions in either the horizontal or vertical axes. By using a KS test, along with horizontal and vertical MW tests, we obtained a holistic comparison of the distributions.

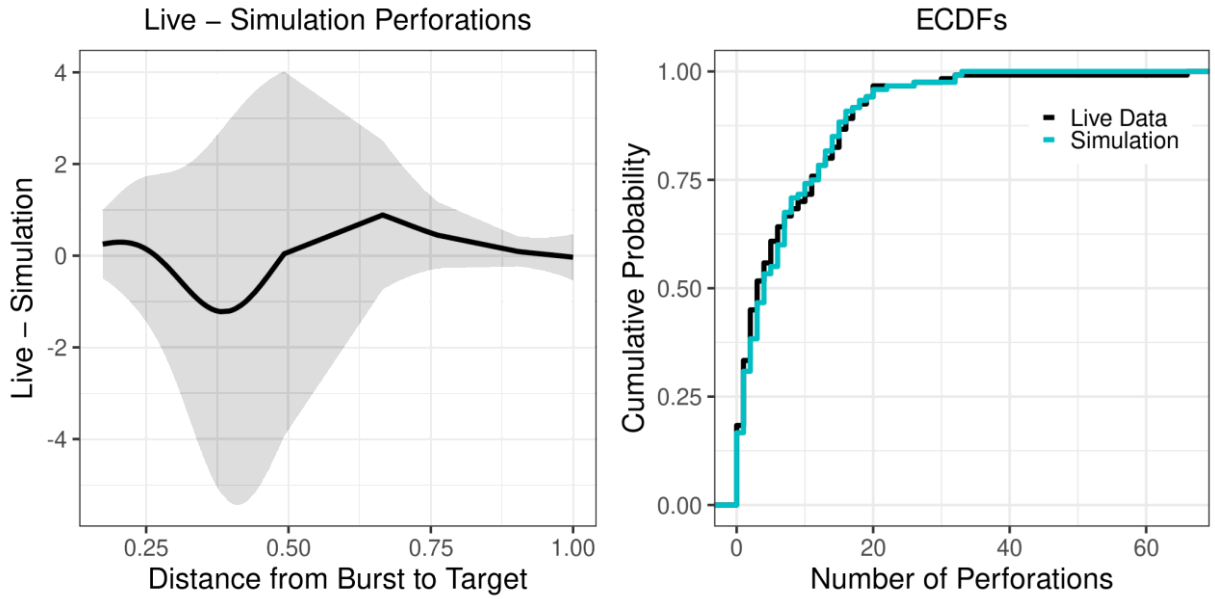

Figure 5: Live – Simulation Plot (left) and ECDFs (right).

Figure 6 exemplifies the importance of using multiple hypothesis tests. Each simulation was run 100 times, and the means of the resulting p-values plotted. When we shifted the simulation to the right, both the KS and vertical MW test p-values remain unchanged. An analyst looking at these values might presume — incorrectly — that the live data and model are in agreement, when they clearly are not. However, the horizontal MW test, which accounts for the direction in which the simulation is shifting, exhibits a steep downward trend as the simulation moves farther away from the live data. It intersects the horizontal p = 0.05 line at a shift of about 0.05 (arbitrary units), at which the probability of the data and model being in agreement is 5%. The corresponding shift is exhibited in the Perforations vs. Radius plot below (Figure 6). This is the point at which the data and model no longer exhibit the same distribution at the 95% confidence level.

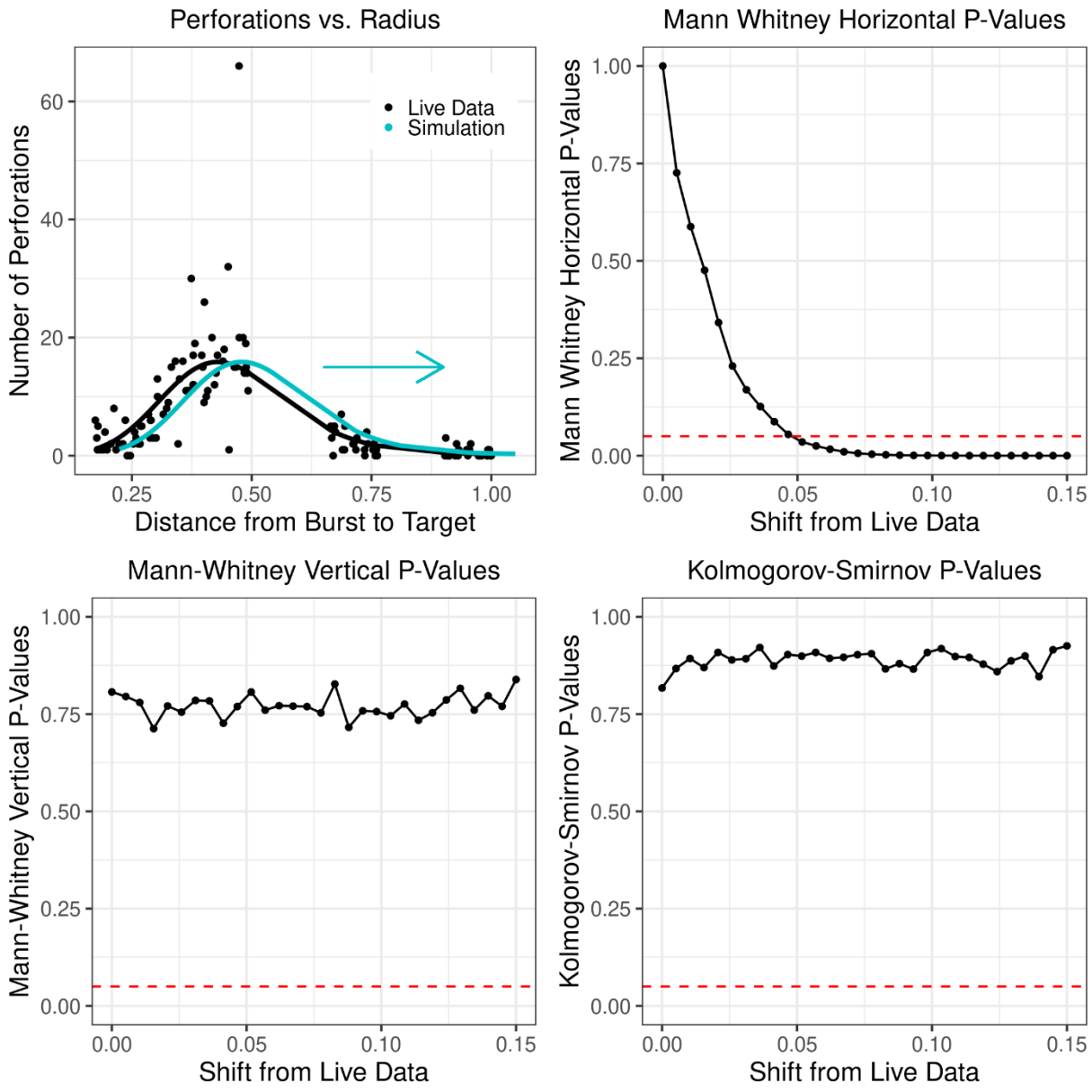

Figure 6: Horizontal and Vertical Mann-Whitney and Kolmogorov-Smirnov P-values.

Similarly, the KS test is sensitive to vertical shifts and changes in the shape of the distribution. We demonstrate this in Figure 7 below by comparing the live data against its regression line, shifted vertically by some value between 0 and 1. The KS p-values exhibit a sharp drop around a vertical shift of 0.4 (number of perforations). They cross the p = 0.05 line at a shift of only 0.55, demonstrating the sensitivity of the KS test to changes in the distribution.

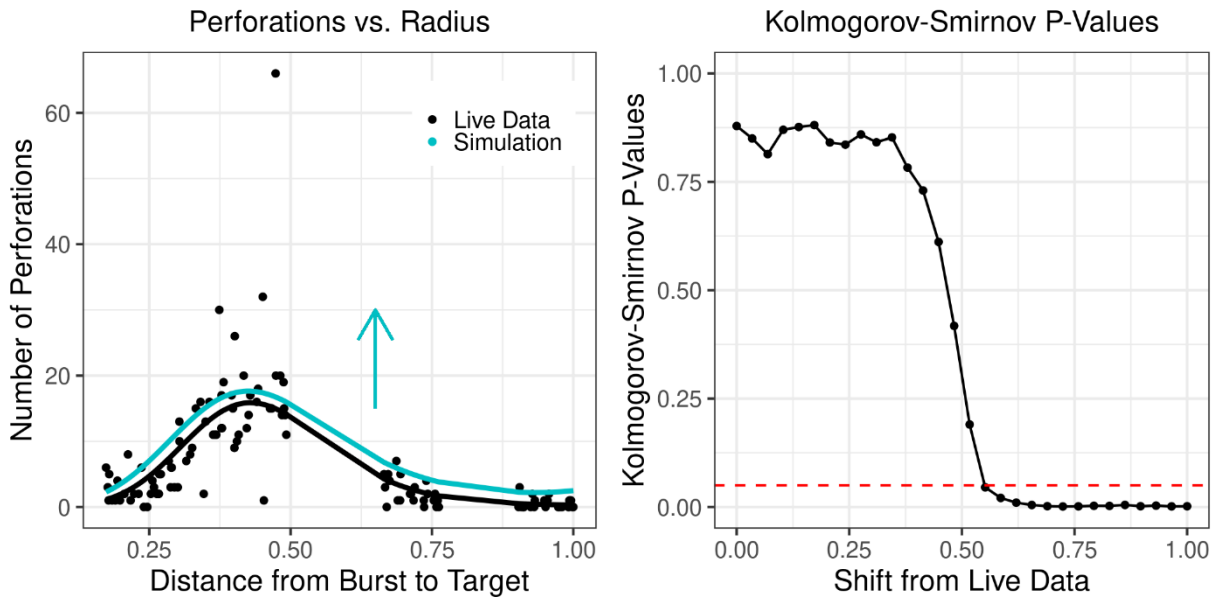

Figure 7: KS Response to Vertical Shift.

What Does It Mean?

This analysis centered around uncertainty quantification, which moves beyond asking whether the data and simulation agree to determining the extent to which they may, in fact, vary but still be considered consistent. In particular, we established that the simulation could fall anywhere within the 95% confidence band and still be considered consistent with live data. This allowed us to reinterpret the statistical uncertainty latent in the fit to the data – due to the limited number of samples from which it was generated – as a systematic uncertainty in the simulation.

As U.S. capabilities, the operating environment, and potential threats become more complex and challenging, operational test and evaluation will rely more and more heavily on M&S. Just as with live data, quantifying the uncertainties that occur in M&S assessments of system performance is critical. Uncertainty quantification conveys the accuracy and precision of M&S results, helps to ensure those results' reliability and reproducibility, and allows testers and the intended user to have greater confidence in the predicted outcome. That, in turn, is critical to executing credible and adequate operational test and evaluation that provides decision makers and warfighters information they can trust. The straightforward method presented here for deriving systematic uncertainty will serve as a crucial tool in validating M&S venues – and setting the foundation to earn that trust.

Navreeta Singh *is a graduate student at Princeton University, pursuing a master's degree in mechanical and aerospace engineering. She earned a Bachelor of Science in in the same field from Princeton, along with certificates in materials science engineering*

and history, and the practice of diplomacy. Navreeta most recently worked in the Office of the Director, Operational Test and Evaluation (DOT&E) at the Pentagon, which sparked her interest in T&E; she previously conducted research at the Air Force Research Laboratory as an air armament scholar.

Jeremy Werner, PhD, Senior Scientific Professional (ST) *was appointed DOT&E's chief science advisor / chief scientist in December 2021 after starting at DOT&E as an action officer in the Naval Warfare Division in August 2021. Before then, Jeremy was at Johns Hopkins University Applied Physics Laboratory, where he founded a data science-oriented military operations research team that transformed the analytics of an ongoing military mission. Jeremy previously served as a research staff member at the Institute for Defense Analyses, where he supported DOT&E in the rigorous assessment of a variety of systems and platforms. Jeremy received a PhD in physics from Princeton University, where he was an integral contributor to the Compact Muon Solenoid collaboration's experimental discovery of the Higgs boson at the Large Hadron Collider at CERN, the European Organization for Nuclear Research in Geneva, Switzerland. Jeremy is a native Californian and received a bachelor's degree in physics from the University of California, Los Angeles, where he received the E. Lee Kinsey Prize (most outstanding graduating senior in physics).*