Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

CASE STUDY IN MODELING AND SIMULATION VALIDATION METHODOLOGY

Scott D. Simpkins

Program Analysis and Evaluation Directorate
U.S. Army Recruiting Command
Fort Knox, KY 40121, U.S.A.

Eugene P. Paulo

Director
TRADOC Analysis Center
Monterey, CA 93943, U.S.A.

Lyn R. Whitaker

Operations Research Department
Naval Postgraduate School
Monterey, CA 93943-5221, U.S.A.

ABSTRACT

The military develops simulations to analyze nearly every aspect of defense. How accurate are these simulations and to what extent do they produce dependable results? Most guidance available to DoD analysts provides broad recommendations geared towards management and coordination of the validation processes. Here, we focus on practical validation from the analyst's perspective in the form of a case study. The platform used is the theater missile defense (TMD) aspects of Extended Air Defense Simulation (EADSIM) and a new simulation called Wargame 2000. The focus is not to validate Wargame 2000 but to develop real, usable tools for analysis. Measures of effectiveness include defense battery search, engagement and intercept times against threat missiles. Insight is provided into developmental and data production issues making the validation process more effective and meaningful.

1 INTRODUCTION

Simulation effectively analyzed and supported can save money on acquisition and reduce more costly live-fire testing to verify results. Indeed there are military applications where simulation may be the only method of quantitative analysis when enemy equipment or technology specifications are required for real-world tests. Validation of a simulation makes its results more acceptable to analysts and decision makers. However, the process can be very intricate and extremely cumbersome depending on the level of accuracy and detail of expected results. Additionally, while numerous publications address need for validation, we know of none that deal directly with analysis of model output in specific formats, or that even present methodologies, metrics or programs for validating simulation output.

This study addresses the simulation of defending against a ballistic missile attack. National missile defense (NMD) and theater missile defense are the pillars of the United States' ballistic missile defense program. NMD was intended to provide a shield of protection across the territory of the United States to intercept long-range missiles. TMD has grown out of the NMD efforts. Theoretically, TMD establishes an umbrella of protection for a theater of operations much smaller than the United States. Each of these defenses relies on integration of detection and intercept systems to engage and destroy inbound ballistic missiles.

Presented here is a case study intended to develop and illustrate a bottoms-up approach to simulation validation by using specific measures of effectiveness (MOE) and fairly simple graphical and statistical methods. The secondary objective is to begin a body of practical case studies that can be used to support a move toward validation commonality among Department of Defense modeling and simulation.

2 SIMULATION OF THEATER BALLISTIC MISSILE DEFENSE

This study centered on the analysis of two combat simulations. EADSIM was considered the baseline system, while a new simulation called Wargame 2000 (WG2K) was evaluated and its output compared to EADSIM output.

2.1 EADSIM

EADSIM is a workstation-hosted, system-level simulation that is used to assess the effectiveness of theater missile defense and air defense systems against the full spectrum of extended air threats. EADSIM provides a many-on-many theater-level simulation of air and missile warfare, an integrated analysis tool to support joint and combined force

operations and a tool to provide realistic air defense training to maneuver force exercises. EADSIM models fixed- and rotary-wing aircraft, tactical ballistic missiles, cruise missiles, infrared and radar sensors, satellites, command and control structures and fire support in a dynamic environment which includes the effects of terrain and attrition on the outcome of the battle.

2.2 Wargame 2000

The department of defense is developing a software simulation for ballistic missile defense that can be used for command and control analysis, provide insight into technology development and provide a training platform for system operators/users. Wargame 2000 is a virtual, real-time, discrete event, command and control missile defense simulation used to investigate human interactions. It is the successor to the Advanced Real-time Gaming Universal Simulation (ARGUS) that has been used for years. WG2K is intended to provide a simulated combat environment that allows war-fighting commanders, their staffs and the acquisition community to examine missile and air defense concepts of operation. This is accomplished through the use of human-in-the-loop experiments and other events.

WG2K has been under development since 1997 and is half way through its developmental lifecycle. It is prudent at this point in development to assess the simulation's accuracy and ability to perform required tasks. Primary attention has been paid to NMD in the past and now development is shifting focus to include TMD (Deis, 2000).

3 MODEL VALIDATION

3.1 Purpose and Necessity

The policy of the Department of Defense is that all its components establish VV&A policies and procedures for modeling and simulation projects they develop and/or manage. Thus, the DoD stresses an effort to establish standards and guidelines promoting VV&A procedural commonality. Furthermore, VV&A is required to be part of all modeling and simulation developments (DoD 5000.61 2000).

3.2 Selection of Measures

The obvious MOE to use in a timeline comparison is time. Clearly, there are multiple physics based characteristics of missile performance not associated with time (e.g. thrust or radar sensitivity). However, only critical times and associated ranges were produced as output during this early stage of development. Typically, time-to-go until some critical event, such as impact, is useful for comparison – as opposed to time after some critical event, such as launch. The percentage difference between the two simulations is more meaningful and the accuracy demand increases as the missile progresses. For instance, 10 percent of the time-to-go-to-impact gets smaller as impact approaches, demanding a better match, while 10 percent of time-after-launch becomes large. This is to say that accuracy near impact is more important than accuracy near launch if the objective is to avoid impact. However, impact time is not one of the reported fields for the baseline EADSIM data and time since launch was used under the assumption that detection and intercept occur early enough in flight for time-since-launch to be equally significant as time-until-impact.

Six MOE's for three batteries against six threats were adopted: detection time, detection range, 1st launch time, 1st intercept time, last launch time, last intercept time (see Figure 1). This leads to potentially 3·6·6=108 comparisons of output variables between EADSIM and WG2K. There are those cases among the batteries and threats where no engagement occurs because the threat was beyond the capability or range of the battery and some cases where no values were reported; this reduced the data set and analysis. Ultimately only 76 output variables were collected.

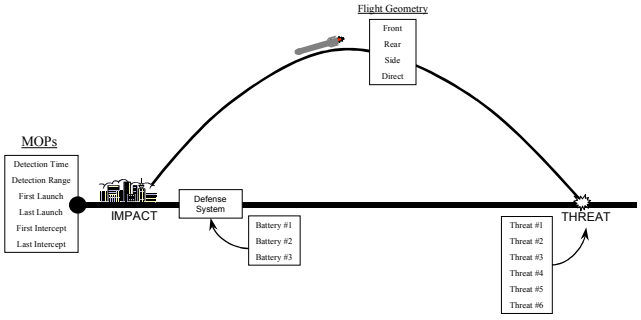

Figure 1: Input Parameter and Output MOE Combinations.

3.3 Planning Data Collection

Threat data is based on current intelligence estimates of ballistic missiles and is considered to be valid for the intended use. However, few of these missiles have ever been engaged, monitored or possibly even launched in any setting other than fielding tests. Thus, the baseline data for combat is, in most cases, extrapolated from small samples collected under artificial conditions.

National and theater missile defense have very different parameters and expected engagement scenarios. BMDO must ensure the simulation accurately portrays a United States defensive response to ballistic missiles within theater by using output from EADSIM. If there is no significant difference between WG2K when compared to EADSIM, WG2K may be validated with respect to TMD. Results are taken from each model and compared using statistical analysis. Also, this work provides insight into developmental and data production issues which make the validation process more effective and meaningful.

4 DATA

4.1 Summary of Data

Data was collected in two phases. The Applied Physics Laboratory at The Johns Hopkins University collected baseline data from EADSIM. For each of six different scenarios (with different input parameters), one hundred replications were recorded where detection time varied randomly. The Joint National Test Facility ran Wargame 2000 stochastically but collected only one set of output data for each scenario due to fiscal constraints.

The primary focus for initial simulation runs is high fidelity range and time performance behavior for the sensor and interceptor. This is to imply that interceptor probability of kill P_k is not important in the consideration of physics based flight. The data is structured so each battery will engage each threat missile separately. Additionally, each threat could assume any of the different geometries: front, back, side and direct. That would be a maximum of $6 \cdot 4 \cdot 3 = 72$ total runs of the system producing multiple output values each to get all possible combinations for the three batteries. The objective is to demonstrate single missile, detailed timeline performance and behaviors in an uncluttered situation (Deis, 2000). All data collected was for one-on-one tests where a single threat missile was detected, tracked and engaged by a single battery. Each of the interceptor systems was positioned in a location determined to represent a valid real-world location. From this location, the defensive radar scanned for threat missiles. In some cases the defense system was placed such that it would not engage threats in order to examine detection only.

Data was collected and recorded for flight times in seconds. This data, representing flight time of both missiles and interceptors, reflects the actual performance characteristics of the modeled entities. As such, the recorded times are classified and are not available in this document. Normalized data representing percentages of the total threat missile flight time is presented here to mask identifiable characteristics defense batteries.

4.2 Derivation of the Database

EADSIM and WG2K were loaded with identical threat characteristics, threat positions, interceptor performance parameters and defensive battery locations. Thus, the radar cross-section (RCS), trajectory and geographic reference remain constant for each threat or battery across all runs involving that system. For these runs, detection and engagement (if any) occur at the same location. Detection data is reported for missiles not engaged.

The development of baseline or control data must not be taken for granted. This data set is considered ground-truth for the systems and attention must be given to its verification. The principle concern with baseline data is that it accurately reflects the input parameters that will be provided to the perspective new simulation. Initially, a face validation or common sense check must be conducted of the baseline output to ensure the performance characteristics represent the expected performance of the actual system and that critical functions of the real system are modeled and reported by the simulation. One of the first baseline runs produced detection and engagement output shown in Figure 2.

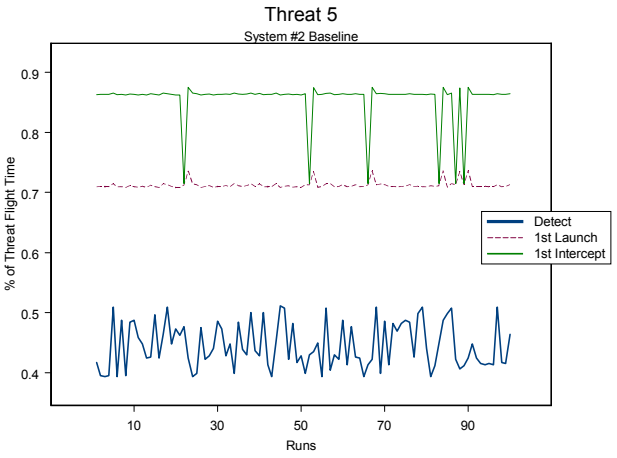

Figure 2: Initial Output from System #2

The few instances where launch and intercept are the same in Figure 2 standout as incorrect. But, it is not clear what the problem is or how to correct it from an initial look. This example is easily identified by inspection; most are not this obvious. Subtle discrepancies can pass undetected through the control data set and derail an otherwise solid validation effort.

Defensive batteries analyzed are identified as systems #1, #2 and #3 for easy reference. The actual identification of defense and threat systems has been classified by the source and has no bearing on the comparative statistics used or methodology recommended here. Defending batteries were taken from a set of allied systems and six threat missiles were selected from a data set of all known threat ballistic missiles developed by the special projects center at the JNTF and approved by the Defense Intelligence Agency (DIA). Thus, no discussion or explanation of threat characteristics such as RCS, velocity or range is included here. The modeling of threat systems was done by the JNTF and provided as input for both simulations.

4.3 Description of EADSIM Output

Each system was placed in a location near an area of interest designated as the impact point. Batteries were assigned the mission of searching a sector of sky from which a known threat would appear. The radar's measure of effec-

tiveness is not whether it found the threat but when and with how much delay it identified and began tracking.

The defense was established as one-on-one; even though a threat was detected by radar and tracked, it is possible the threat was not engaged if the flight path toward the impact point was beyond the defended area of the battery. An interceptor missile was launched when a detected threat entered the defended area or when a threat already in the defended area was detected based on the limitations of the individual system.

Characteristic output between systems is shown below. Figure 3 shows invariability of interceptor launch and engagement times that imply the interceptor is limiting the system (i.e. the interceptor's region of coverage is smaller than the radar's area of coverage). Figure 4 demonstrates high variance in both launch and engagement that indicates it is limited by radar performance.

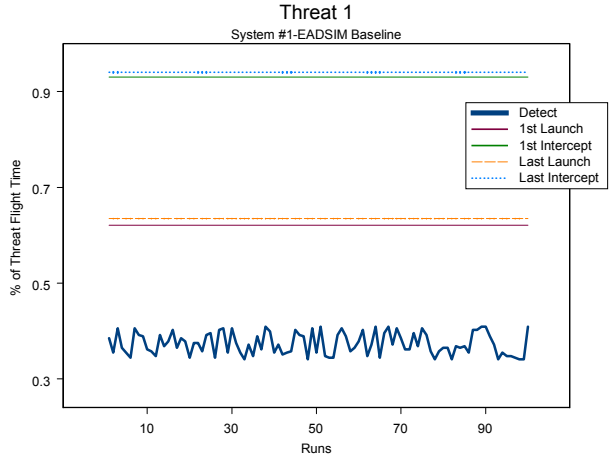

Figure 3: Interceptor Limitations Lead to Small Variance in Launch and Intercept

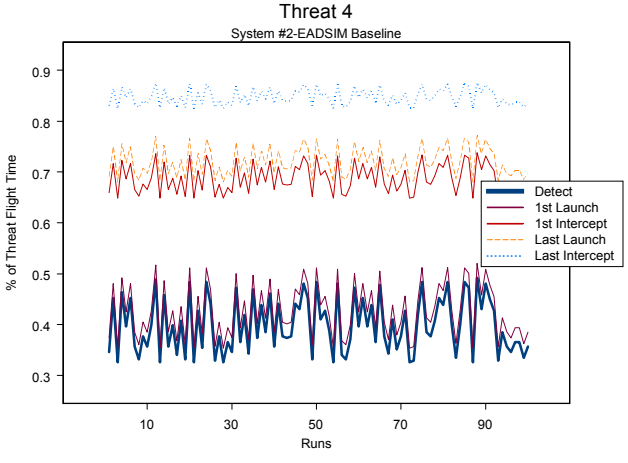

Figure 4: Radar Limitations Produce Immediate Launch and High Variance

4.4 Scope and Limitations

Only one-on-one scenarios are considered at this stage of WG2K development. There is no hand-off between radar systems as would happen if threats were detected early by a system and then 'passed' to another system with higher P_k. Therefore, no interaction between threats or interceptors is found. It is anticipated that there will be dependence on threat and interceptor type when there are multiple systems.

WG2K essentially has one random variable: stochastic radar detection using a Normal distribution with almost no variability. EADSIM uses two stochastic variables, radar detection which is Normal and sensor frame time, which is Uniform, producing detection times distributed with a convolution of the two input distributions. In other cases only one stochastic variable is used, sensor frame time which is uniform. Interceptor flyout is always deterministic in both simulations.

5 RESULTS

5.1 Exploratory Data Analysis

We assumed that the 100 runs for each scenario are in fact independent. As a quick check for independence of EADSIM output, a test was used to see if MOE times from EADSIM were truly independent. Here, a 'runs test' for above and below the median is used to test the null hypothesis that the sequence of detection times for the 100 iterations of the simulation is indeed independent. Specifically a sequence of 100 binary variables is constructed where the i^{th} variable takes value 1 if the MOE time of the corresponding i^{th} simulation run is above the median, and 0 otherwise. The number of runs above the median m, below the median n and the total number of runs $R = m + n$ are computed. In the example of Figure 5 there are three runs of ones (of lengths 2,3,3) and two runs of zeros.

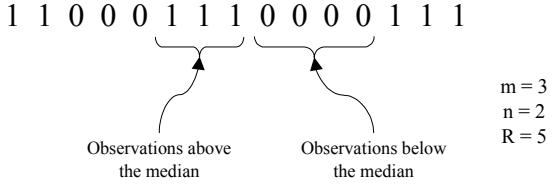

Figure 5: Example of MOE Times Indicating Distribution Above and Below the Median

For large samples, the test statistic

$$\frac{R - \frac{2m}{(1+\gamma)}}{\sqrt{\frac{4\gamma m}{(1+\gamma)^3}}}$$

where $\gamma = m/n$, has a standard Normal null distribution (Lehaman, D'Abrera, 1975). At 5% level of significance none of the 76 sets of simulation runs failed the test for randomness.

As illustrated in Figures 6 and 7, the mean and variance of detection times differed significantly between systems. This may be attributed to specific limitations of the batteries. The distribution of detection times also differed from system to system.

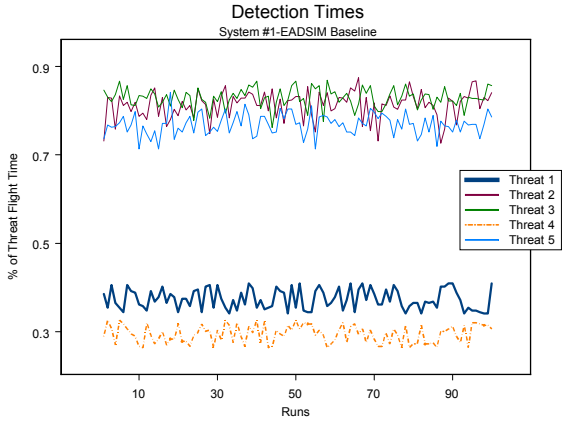

Figure 6: Detection Times for System #1

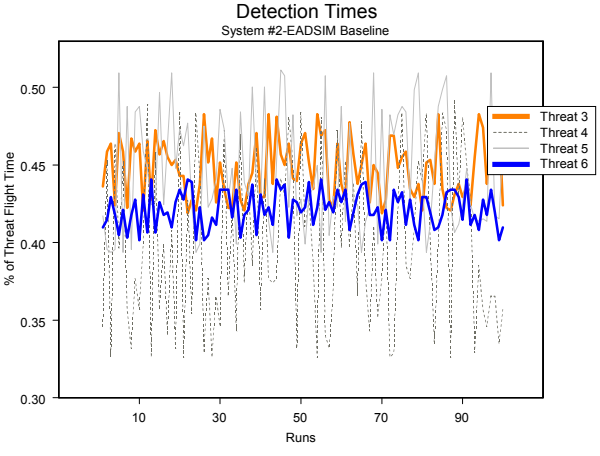

Figure 7: Detection Times for System #2

There is a clear difference between these two systems depicted in Figures 6 and 7. Both batteries are positioned in the same place and threats follow the same flight path against them. Figure 6 shows an average near 80% of threat flight time for three missiles while Figure 7 shows none of the detection times above 50%.Variance for Threat 4 is three times higher in System #2 than System #1. Variance in detection times is acceptable. However, critical to analysis by comparison is that WG2K demonstrate similar behavior when modeling the same combinations. Graphic analysis can quickly identify areas of interest or anomalies within the data. But, when provided with a single output value, graphic analysis is limited in comparing the two simulations. An interesting observation taken from the Figures 6 and 7 is that for most threats variance in detection is the same for each system even though System #1 takes twice as long to detect on average.

Stochastic models can be viewed in two distinct classes. The first class involves sampling from a probability distribution of inputs such that, once a sample of inputs is generated, the model is deterministic. In this situation, no random events occur within the simulation. The second class contains those models in which events during the simulated course of battle are affected by the results of dynamically generated random numbers. Of course, models can contain both a random sample of inputs and have internal random events (Lucas, 2000).

Simulations modeling the TMD environment contain both stochastic classes. The outcome of many simulation runs will vary across a spectrum of values, even when the model is provided deterministic input. It is expected that a valid model will produce output MOE's that tend to cover the entire range of the model's capabilities but are concentrated near the expected value of the functions producing them. Knowing the distribution functions of the model is important when determining how close the simulation has come to producing the expected outcome.

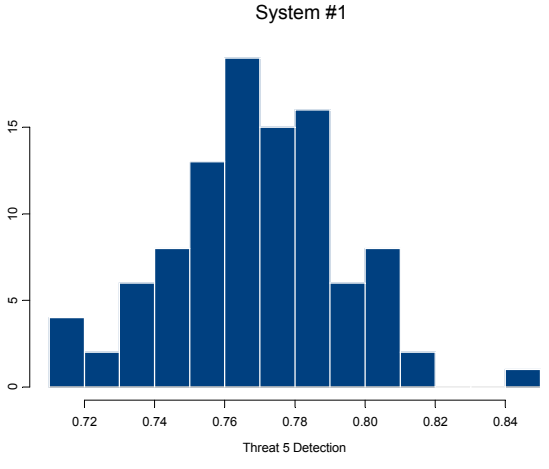

Figure 8: Distribution of Random Detection Times for System #1 Against Threat 5

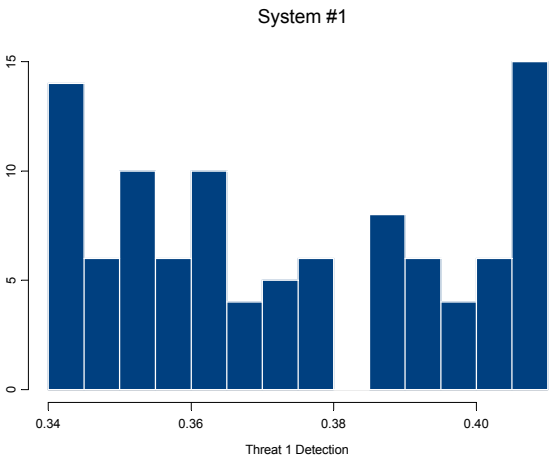

Figure 9: Distribution of Random Detection Times for System #1 Against Threat 1

In the case of WG2K, only one run of the simulation is provided so a complete analysis of the output cannot be compared to a distribution function. An analysis of the distributions resulting from EADSIM provides some insight into the expected behavior of WG2K. Histograms in Figures 8 and 9 contrast detection time distributions from one system against two separate threats.

However, comparing Figure 8 and 9, it is clear detection times do not come from the same distribution. What can be causing such a dramatic shift in detection time distribution for the same system against similar threats? The battery modeled here has a search pattern that makes its detection time highly dependent on threat flight trajectory. Threat missiles launched from far away are detected normally as they enter the top of the radar coverage (Figure 8) while threats launched from close range are detected uniformly as they enter the bottom of the radar coverage (Figure 9). It is sufficient to say that the distribution of detection times cannot be assumed normal and that no regularly used parametric distribution captures all of the detection time distributions. This type of disparity among time parameter distributions relating MOE's is common to many systems/threat combinations.

5.2 Comparing Wargame 2000 with EADSIM

5.2.1 Graphic Analysis

The spread of times can be further broken down into quartiles separating the 100 observations into groups of 25 separated in sequence by the 1st, 2nd (the median) and 3rd quartiles. One expects the result provided by WG2K to fall between the 1st and 3rd quartile of EADSIM implying the value is relatively close to that expected for validation.

A box plot places a box around the middle 50% of the data, with the upper edge at the 3rd quartile and lower edge at the 1st quartile (Devore, 1995). The whiskers in box plots for all MOE's extend from the box up to the largest observation and down to the smallest observation. In general, extreme observations are reported as points beyond the whiskers, no such extreme values were observed in this study. The most visual feature is the box that shows the limits of the middle half of the data. Box plots not only show the location and spread of data but indicate skewness, as well. Box plots for EADSIM parameters were compared to WG2K values as shown in Figure 10. Each parameter that had output was compared by plotting the WG2K result on the horizontal axis against the EADSIM spread on the vertical axis. Ideally, a line with 45° slope will intersect the box plot at the mean indicating a one-to-one match between the WG2K result and the average EADSIM result.

Figure 10 displays several comparisons where EADSIM and WG2K outputs agree and are statistically similar. Even System #1's first launch times compare nicely although EADSIM exhibits zero variance as represented by the flat line in-lieu of a box plot. Small deviation, as shown in System #3's first intercept times by the intersection of the box and line, is acceptable. In general, Figure 10 is good news for the new simulation. A preliminary look at these comparisons indicates WG2K is producing detection, launch and intercept times very close to EADSIM.

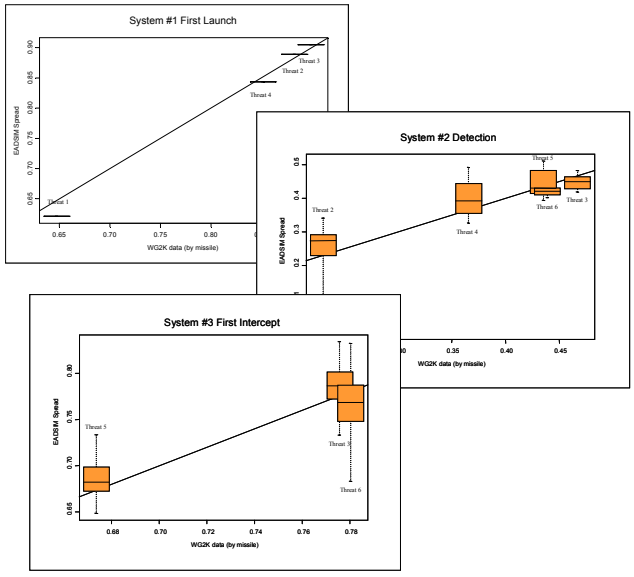

Figure 10: Box Plots Reveal Trends when Compared

Further inspection of the data reveals system/threat combinations with larger variability however. Figure 11 indicates two detection times far outside the baseline distribution for two of the threats. One of the threats was not detected by WG2K leaving only five; this further con-

founds the results and implies simulation issues larger than interceptor flight time such as sensor detection modeling or sensitivity parameters.

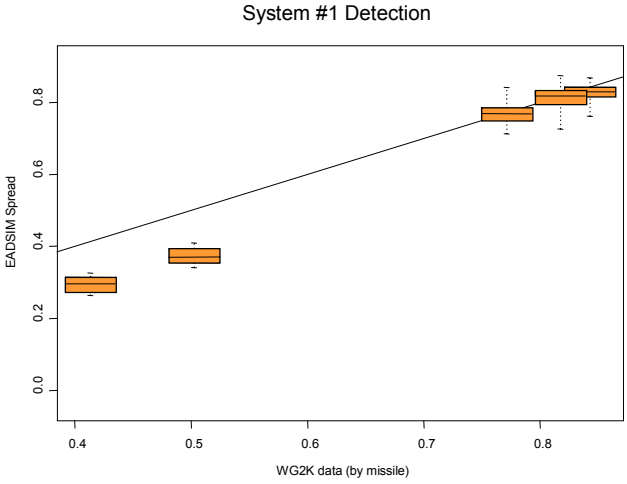

Figure 11: Box Plot Showing Detection Delays in WG2K for a Specific System

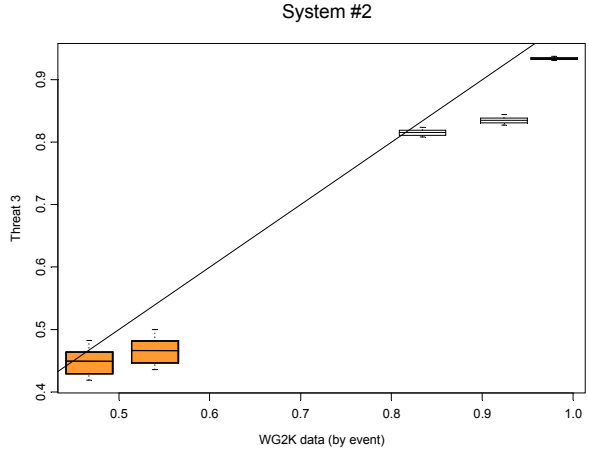

Figure 12: Box Plot Showing System Delays in WG2K for a Specific Threat

It should be noted that poor performance was not consistent with any system or threat combination. The overriding consistency was in WG2K's inability to identically engage threats EADSIM engaged. Furthermore, there were two cases where WG2K engaged threats that EADSIM never engaged.

Time for single systems can be compared as in Figure 12, which shows results across all measures of effectiveness for System #2. The boxes represent the individual MOE's beginning with detection in the lower left corner and moving forward in time through last intercept in the upper right. Each of these measures exceeds the baseline. All WG2K events occur much later than expected for this system/threat combination.

5.2.2 Inference

Although each scenario was replicated 100 times for EADSIM, because WG2K is run in real-time, only one realization of WG2K is available for each scenario.

Often, there is the temptation to treat the output of such a run as an expected value (i.e. to treat a detection time as the expected detection time). This seems reasonable because WG2K has little to no variability among input variables as discussed. However, for each run the output MOE's of WG2K are certainly non-linear functions of the input. In general, the expected value of a non-linear function is not equal to the function of the expected value of those random variables. In practical terms this means that the average detection time over many replications of WG2K will not be the same as one detection time with average input. Thus, the output of one run of WG2K and indeed one run of any model run in small input variance should not be considered to be the expected output.

Another approach, and the one adopted here, is to treat WG2K output as one realization of WG2K had it been run in true stochastic mode. This is not an entirely appropriate model since some inputs are constant rather than random. However, it will provide a more realistic comparison.

5.2.3 Re-Sampling from the Baseline

Wargame 2000 is striving to match the fidelity of EADSIM. Thus, if WG2K is in fact reproducing EADSIM the distributions of output parameters (under the same input conditions) should be the same. This implies that the mean and variance of output should also be the same for both simulations. The difficulty with comparing the mean and variance is that WG2K was only run one time in each scenario. A bootstrapping approach is used to test the null hypothesis that WG2K has the same output distribution as EADSIM. The empirical distribution of the 100 values from EADSIM is used as an estimate of the null distribution for both EADSIM and WG2K. With bootstrapping we can approximate the null distribution of a test statistic using Monte Carlo simulation. In this case, the simulation involves repeatedly "drawing" a sample of 101 from EADSIM. These repeated draws of 101 observations do not require re-running EADSIM or WG2K. They are independent 101 psuedo-random detection times generated from the empirical distribution of the 100 actual EADSIM detection times. Note that generating the WG2K psuedo-detection time from the empirical distribution of EADSIM detection times is consistent with the null hypothesis that WG2K and EADSIM have the same output distributions.

The percentage difference in mean detection times for each scenario for both models can be computed to compare WG2K with EADSIM. In particular, for a scenario let $\bar{Y}_E$ and Y_W represent the average detection time from 100 runs

of EADSIM and the detection time from one run of WG2K respectively. They define the test statistic:

$$T_{stat} = \frac{\overline{Y}_E - Y_W}{\overline{Y}_E}$$

The test statistic (T_{stat}) represents the percent difference between EADSIM's percent of total flight time for each MOE and WG2K's.

Bootstrapping is used to estimate the sampling distribution of T_{stat} under the null hypothesis. Sampling from the empirical distribution of EADSIM detection times is equivalent to draws with replacement from the 100 actual EADSIM values. In total $1000 \cdot (100 + 1)$ draws with replacement were made from the 100 EADSIM values. The results are depicted in Figure 13; $\overline{X}_{Ei}$ i = 1,2,…, 1000, are the 1000 averages of 100 draws each while X_{Wi} consists of individual 1000 draws.

From the $\overline{X}_{Ei}$, X_i, the bootstrapped value of the test statistic T_i is computed as in Figure 13. Figure 14 depicts this process; shaded areas near the tails represent the proportion of observations more extreme than T_{stat}. This area is one half of the p-values for a two-sided test.

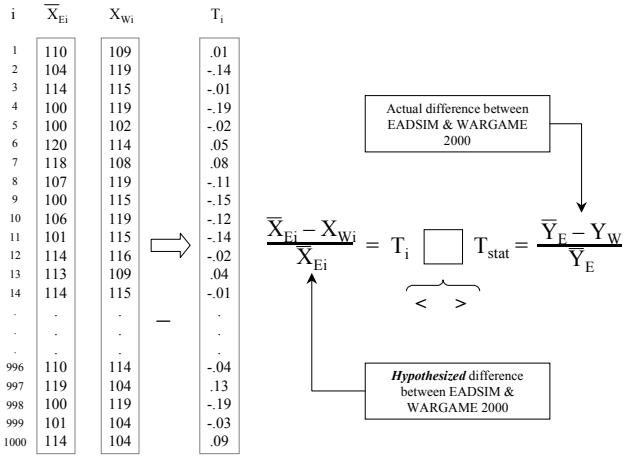

Figure 13: Re-Sampling for P-Values

The 1000 bootstrapped values of T_i i = 1,2,…,1000, and the actual value of the test statistic T_{stat} are used to approximate a p-value for the test where F_W and F_E are the MOE distributions for WG2K and EADSIM respectively. The p-value is approximated as the proportion of bootstrapped T_i that are more extreme than the actual observed T_{stat}. These p-values indicate the strength of evidence against the null hypothesis that WG2K's result represents a possible result from EADSIM. There are examples supporting both acceptance and rejection of the null hypothesis. High p-values for system #3 accept the null hypothesis in all cases where data is recorded for both simulations. However, there are very low p-values for system #1, indicating rejection for nearly every MOE clearly based on the invariability of EADSIM. System #2 showed mixed results depending on the MOE.

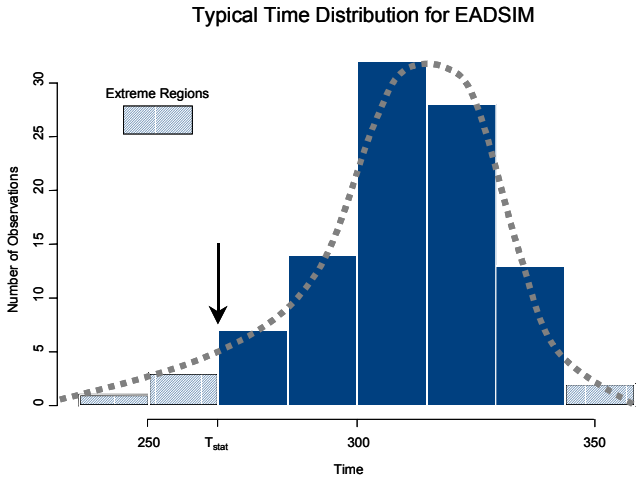

Figure 14: Using P-Values and Test Statistics to Identify Extreme Points.

There are several cases where EADSIM exhibits zero variance. This is expected as most of the many random variables available in EADSIM are set for zero variance. In these cases it is difficult to determine whether WG2K is accurately reflecting EADSIM or not. The simulation outputs differ very little and it may be that the developer considers a zero variance in WG2K at its current value to be within an acceptable range.

It is important to note that in general, the test statistic chosen here, the fact that only one realization of WG2K is used and the non-parametric nature of the bootstrapping procedure all contribute to a testing procedure that is not powerful against all alternatives to the null hypothesis. Thus, this procedure will not be able to detect all types of differences between WG2K and EADSIM. In particular, because only one realization of WG2K is available, there is no way to tell if the variability in times simulated by WG2K will be the same as the variability of those simulated by EADSIM.

6 CONCLUSIONS

WG2K has many capabilities and accepts diverse input from which it builds a TMD scenario. The challenge of validating a simulation model early in its development comes from ensuring that a small set of capabilities and data produced fit into the big picture of the model using all capabilities; this is not unique to WG2K. Very little is available, published in the open literature or from DoD, that gives specific guidance for how to validate a simulation model based on the limited data usually available in such attempts. Formulating such an

approach applicable to the majority of validation attempts would be extremely difficult. Thus, although case studies are situation dependent, they are vital for providing practical guidance for validation.

6.1 Summary of Results

Presented here is a straightforward approach to validation that uses various methods to examine simulation output. This begins by choosing a small number of appropriate MOE's. Limiting the MOE's examined can ignore significant ranges of output but allows the analyst to focus on those most relevant to simulation performance. A graphic, explanatory analysis is used which is supported by subsequent inferential statistics. The challenge in this particular case study is that WG2K scenarios are replicated only one time.

The importance of graphic analysis should not be overlooked. Graphics allow quick, accurate analysis as long as appropriate comparisons are depicted. The boxplots used in this study clearly show differences between EADSIM and WG2K. The informed eye can discriminate very small differences. These must be confirmed but a clearly thought-out, accurate graphic representation of the data can narrow one's focus by identifying those areas requiring effort.

A non-parametric bootstrap is used to test the null hypothesis that the distributions of specific MOE's are the same for both models. The results of this inference confirm the graphical analysis. Note though that this procedure is conservative in that rejecting the null hypothesis (i.e. finding differences between the models) does provide evidence that the models differ while failing to reject does not provide evidence that they are the same. With only one run per scenario from WG2K it is not possible to draw the conclusion that the distributions of MOE's are the same for WG2K and EADSIM.

6.2 Recommendations

The most important recommendation is that a series of case studies, such as this one, need to be compiled and made available to those analysts actually doing validation. These case studies need to show by example practical but simple approaches to sorting through and making sense of complex simulation output. It is clear that a single document that tries to take a top down approach and that encompasses all types of validation and possible output is impractical. This approach leads to volumes of general guidance but nothing specific enough to be useful to the analyst in practice. A case study approach tackles the problem of practical analytical guidance from the bottom up. With a number of such case studies, available in one place, patterns of what approaches prove most useful should emerge rather quickly.

REFERENCES

Deis, F., *Wargame 2000 White Paper vs. EADSIM Anchoring Effort Planning Proposal*, (2 Nov 2000).

Lucas, T. W., "The Stochastic Versus Deterministic Argument for Combat Simulations: Tales of when the Average Won't Do", *Military Operations Research*, vol. 5, no. 3, pp. 9-28 (2000).

Department of Defense, *DoD 5000.61: DoD Modeling and Simulation (M&S Verification, Validation, and Accreditation (VV&A)*, 2000. Available: http://www.ailtso.com/simval/Documents/5000.61/dod5000.61.htm

Department of Defense Research and Engineering, *Verification, Validation and Accreditation (VV&A) Recommended Practices Guide*, 1996.

Hodges, J., Dewar, J. *Is It You or Your Model Talking? A Framework for Model Validation*. Santa Monica: RAND, 1992.

Devore, J. *Probability and Statistics for Engineering and the Sciences*. Pacific Grove: Duxbury Press, 1995.

Lehaman, E., D'Abrera, H. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day, Inc, 1975.

Mitchell, B. "Underwater Launch Technology Sustainment," briefing on Trident Missile Validation (3 Oct 2000).

AUTHOR BIOGRAPHIES

SCOTT D. SIMPKINS is a Captain in the United States Army. He is an analyst in the Program Analysis and Evaluation Directorate of the US Army Recruiting Command. He received his Masters Degree from the Naval Postgraduate School in 2001. His research interests are in entity level simulation and data analysis. His email address is `scott.simpkins@usarec.army.mil`.

EUGENE P. PAULO is a Lieutenant Colonel in the United States Army. He is the Director of TRAC-Monterey and an Assistant Professor in the Department of Operations Research at the Naval Postgraduate School. His email address is `paulog@trac.nps.navy.mil`.

LYN R. WHITAKER is an Associate Professor in the Department of Operations Research at the Naval Postgraduate School. Her research interests are in reliability, statistical analysis of simulation and combat simulation, and in categorical data analysis. Her email address is `lwhitaker@nps.navy.mil`.