2014-09

Improving the analysis capabilities of the Synthetic Theater Operations Research Model (STORM)

Bickel, William G., Jr.

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

IMPROVING THE ANALYSIS CAPABILITIES OF THE SYNTHETIC THEATER OPERATIONS RESEARCH MODEL (STORM)

by

William G. Bickel, Jr.

September 2014

Thesis Advisor: Thomas W. Lucas
Second Reader: Paul J. Sanchez

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2014	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE
IMPROVING THE ANALYSIS CAPABILITIES OF THE SYNTHETIC THEATER OPERATIONS RESEARCH MODEL (STORM)

5. FUNDING NUMBERS

6. AUTHOR(S) William G. Bickel, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943–5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

The Office of the Chief of Naval Operations, Capability Analysis and Assessment Division (OPNAV N81), along with other DOD organizations, utilizes the Synthetic Theater Operations Research Model (STORM) as its primary campaign analysis tool. STORM aids senior-level policymakers in evaluating military strategy and capabilities, force structure, and operational effectiveness. This is a proof-of-concept thesis that determines the feasibility of implementing a simple design of experiments within the complicated framework of STORM. Such a capability will enable quicker and more robust estimates of proposed force structure trade-offs. After utilizing various methods and statistical techniques, this thesis concludes that it is possible to implement small designs within STORM that could offer useful insights to OPNAV N81 analysts. However, the steps needed to successfully complete a design are far from automated and fairly complex. Currently, they require a great deal of time to manually apply. As a pilot study, these results pave the way for future researchers to apply our results to a real-world, classified scenario.

14. SUBJECT TERMS: Synthetic Theater Operations Research Model (STORM), OPNAV N81, campaign analysis, design of experiments, stochastic simulation

15. NUMBER OF PAGES 93

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540–01–280–5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

IMPROVING THE ANALYSIS CAPABILITIES OF THE SYNTHETIC THEATER OPERATIONS RESEARCH MODEL (STORM)

William G. Bickel, Jr.
Lieutenant, United States Navy
B.S., United States Naval Academy, 2007

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2014

Author: William G. Bickel, Jr.

Approved by: Thomas W. Lucas
Thesis Advisor

Paul J. Sanchez
Second Reader

Robert F. Dell
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Office of the Chief of Naval Operations, Capability Analysis and Assessment Division (OPNAV N81), along with other DOD organizations, utilizes the Synthetic Theater Operations Research Model (STORM) as its primary campaign analysis tool. STORM aids senior-level policymakers in evaluating military strategy and capabilities, force structure, and operational effectiveness. This is a proof-of-concept thesis that determines the feasibility of implementing a simple design of experiments within the complicated framework of STORM. Such a capability will enable quicker and more robust estimates of proposed force structure trade-offs. After utilizing various methods and statistical techniques, this thesis concludes that it is possible to implement small designs within STORM that could offer useful insights to OPNAV N81 analysts. However, the steps needed to successfully complete a design are far from automated and fairly complex. Currently, they require a great deal of time to manually apply. As a pilot study, these results pave the way for future researchers to apply our results to a real-world, classified scenario.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION....1
A. LITERATURE REVIEW....2
B. RESEARCH QUESTIONS AND GOALS....3
C. METHODOLOGY....4
D. BENEFITS OF RESEARCH....4

II. STORM OVERVIEW....7
A. MODELING AND SIMULATION....7
B. ADOPTION OF STORM MODEL....7
C. IMPLEMENTATION OF MARITIME COMPONENT IN STORM....8
D. BENEFITS OF STOCHASTIC SIMULATION....8
E. STORM AS A STOCHASTIC MODEL....9
F. STORM−A CAMPAIGN ANALYSIS TOOL....11
G. STORM–A DATA-DRIVEN DESIGN....12
H. REPRESENTATIONS IN STORM....13
1. Storm Assets....13
2. Storm Environment....14
I. STORM–THE USER....14
1. Map Tool....15
2. Graph Tool....16
3. Report Tool....17
J. STORM–PUNIC 21....17
1. Current Situation and Battle Phases....17
2. Order of Battle for Blue and Red Forces....18

III. IMPLEMENTING A DESIGN OF EXPERIMENTS IN STORM....21
A. DOE TERMINOLOGY....22
1. Types of Variables....22
2. Output Metrics–"What It Takes To Win" Metrics....23
3. Designs In General....23
4. 2^k Factorial Design....24
a. Main Effects....24
b. Interactions....25
B. CRITICAL LIMITATIONS IDENTIFIED....25
C. IMPLEMENTING A DESIGN WITH STORM: A FOUR-STEP PROCESS....26
D. CORRELATION AMONG FACTORS....27
E. BUILDING THE DESIGN IN STORM....31
F. CREATION OF INPUT FILES....33
G. RUNNING A SIMULATION WITH CUSTOM INPUT FILES....35
1. STORMMiner Software and Data Collection....36

IV. ANALYSIS OF DESIGN POINTS....37
A. RANDOM NUMBER GENERATION IN STORM....38

B. BLUE FORCE LOSSES–CARRIER....39
C. BLUE FORCE LOSSES–BFNMF....42
D. TIME AT WHICH BLUE FORCES ACHIEVE AIR SUPREMACY....45
E. RED FORCE SAM SITES DESTROYED....48
F. STATISTICAL DIFFERENCES IN DESIGN POINTS....52
G. BFNMF SPEED PROFILE SENSITIVITY ANALYSIS....61
V. CONCLUSION AND RECOMMENDATIONS....65
A. DOE DESIGN AND METHODOLOGY....65
B. BENEFITS OF EXPERIMENTAL DESIGN IN STORM....65
C. RECOMMENDATIONS....66
APPENDIX. R-STUDIO CODE....67
LIST OF REFERENCES....69
INITIAL DISTRIBUTION LIST....71

LIST OF FIGURES

Figure 1. An example input data file for maritime surface ship damage functions categorized by weapon type 10
Figure 2. Diverse activities associated with a STORM campaign (from Group W, 2012a) 12
Figure 3. STORM's GUI (from Group W, 2012c) 15
Figure 4. STORM's map tool interface (from Group W, 2012c) 16
Figure 5. AOI for STORM's PUNIC 21 scenario 18
Figure 6. STORM front input data files 26
Figure 7. RStudio generated pairs plots of nine WITTW campaign metrics 29
Figure 8. Pairs plot of blue force carriers killed and BFNMF killed 30
Figure 9. Snapshot of the *typeaa.dat* file in STORM Front 32
Figure 10. Snapshot of STORM GUI and how it should look after setting up eight separate design point study directories 34
Figure 11. Snapshot of STORM GUI and how to put all three custom files into a local setting. In order to get to this option, right click on *typeaa.dat, sideC2.dat,* and *transaction.dat* files and select "Make Local" 35
Figure 12. Plot showing the average number of blue carriers with a 95% CI over each design point and base simulation mean 40
Figure 13. The number of BFNMN losses over design points one through four 43
Figure 14. The number of BFNMN losses over design points five through eight 45
Figure 15. The time in which blue forces achieve air supremacy over design points one through four 46
Figure 16. The time in which blue forces achieve air supremacy over design points five through eight 47
Figure 17. Plot showing the average number of red force's SAM sites destroyed with a 95% confidence interval over eight design points and baseline simulation mean 49
Figure 18. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the average number of blue carrier losses 54
Figure 19. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the average number of BFNMF losses 55
Figure 20. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the time it takes blue forces to achieve air supremacy 56
Figure 21. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the average number of red SAM sites destroyed 57
Figure 22. Results from Pearson's chi-squared statistic test on whether or not the blue forces achieve air supremacy before the simulation terminates 61
Figure 23. Snapshot of JMP data file that includes BFNMF factor settings and output data associated with all eight design point 62
Figure 24. Partition tree method in JMP for the number of BFNMF losses. The left-most branch corresponds to the lowest number of losses (highlighted in red) 64

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. Table of available representations in STORM....13
Table 2. A list of blue and red maritime assets for the PUNIC 21 scenario....19
Table 3. A list of blue and red air assets for the PUNIC 21 scenario....19
Table 4. Table of the baseline intercept speeds profile for the BFNMF....22
Table 5. Design matrix for a 2^3 factorial design (from Law, 2007)....24
Table 6. Results for PCC, Spearman's, and Kendall's correlation test....31
Table 7. Table of the 2^3 factorial design matrix that modifies the intercept speed levels associated with the BFNMF....33
Table 8. Table showing random number generation in R-Studio for the normal distribution with default and explicit random number seeding....38
Table 9. Table providing the summary statistics for blue force carrier losses for each design point....41
Table 10. Table showing the proportions of the number of replications blue forces achieved air supremacy out of 25 replications....47
Table 11. Table providing the summary statistics for red SAM sites destroyed metric....51
Table 12. Table showing the results of ANOVA test for four WITTW responses: blue carrier and BFNMF losses, time blue forces achieve air supremacy, and the number of red force SAM sites destroyed....53
Table 13. Table of *HSD.test()* results for each response that include grouping identifications, design points, and associated means....59
Table 14. Matrix generated in R-Studio of the number of times the blue force achieves and does not achieve air supremacy over the 25 replications for design points one through eight....60
Table 15. Summary *p-value* and R2 results for the JMP generated full factorial linear regression models....63

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ANOVA	analysis of variance
AOI	area of interest
AOR	area of responsibility
ARC	Anglo Republic and Carthage
ARG	amphibious readiness group
BFNMF	blue future naval multi-role fighter
BCA	Budget Control Act
C2	command and control
CLF	combat logistics force
CNA	Center for Naval Analysis
COA	course of action
CSG	carrier strike group
DMSO	defense modeling and simulation
DOD	Department of Defense
DOE	design of experiments
ESG	expeditionary strike group
GUI	graphical user interface
HQ	headquarters
HSD	honest significant difference
HVU	high value unit
ISR	information, surveillance, and reconnaissance
ITEM	Integrated Theater Engagement Model
M&S	modeling and simulation
MRF	mobile riverine force
NPS	Naval Postgraduate School
OOA	out of action
OPNAV	The Office of the Chief of Naval Operations
QDR	Quadrennial Defense Review
SAM	surface-to-air missile
SCALA	scalable language

SE	Swiss Empire
SEED	simulations, experiments, and efficient design
SQL	structured query language
STORM	Synthetic Theater Operations Research Model
VV&A	verification, validation, and accreditation
WITTW	what it takes to win

EXECUTIVE SUMMARY

The Synthetic Theater Operations Research Model (STORM) is a state-of-the-art, computer simulation that is specifically designed to offer Department of Defense (DOD) organizations key strategic insights into military force structure, capabilities, and overall operational effectiveness. In 2010, OPNAV N81, the U.S. Navy's Assessment Division, adopted STORM as its primary campaign analysis tool due to its stochastic nature, which provides analysts the ability to model the inherent variability of combat through random number generation. N81 utilizes STORM in order to perform quick turn-around analysis on classified scenarios that are developed years in advance. However, STORM is hindered when supporting this type of analysis due to extremely long run times and extensive output. Moreover, it is difficult for analysts to answer questions that pertain to how modifications in force-structure will affect the overall scenario outcome without making major changes to input data files, which consumes a great deal of time and money. In today's budget-stricken military, N81 must look to broaden the scope of its analysis while limiting associated costs and manpower-intensive post-processing data analysis. This proof-of-concept thesis explores the feasibility of implementing a design of experiments (DOE) within the complicated framework of STORM.

Given that STORM is extremely complex and its results can be difficult to interpret, it is often the case that analysts' level of knowledge extends only to the collection and post-processing analysis of output data files. Therefore, it is imperative that a basic working knowledge of the model be obtained prior to attempting a STORM analysis utilizing a design of experiments. For this reason, this thesis provides a condensed version of three developer-written manuals and information papers that highlight STORM's development, basic structure, characteristics, and capabilities. Armed with this knowledge, analysts will possess the tools necessary in order to follow the four-step methodology outlined in this thesis, which is specifically developed for ensuring a successful DOE implementation that is as simple as possible.

The first step in the process is critical—identifying potential correlation relationships among controllable input factors by analyzing output data extracted from a

newly developed software program called STORMMiner. For this research, Pearson's correlation coefficient, Spearman's rho, and Kendall's tau equations reveal a strong positive correlation between the average number of blue force naval multi-role fighters (BFNMF) lost and blue carriers killed. Given the significance of this relationship, a 2^k factorial design is built that includes three intercept speed profiles specific to the BFNMF. In the design, intercept speeds are modified by increasing and decreasing their values by 10%. Given the unique input file format used by STORM, custom files that include this design are generated and incorporated into eight separate sets (called design points) of runs for the PUNIC 21 scenario. Each design point run corresponds to a unique combination of intercept speed profiles, which produces results in the form of exclusive sets of output data that are analyzed in order to determine what effect changing the BFNMF intercept speeds have on the outcomes of the scenario.

The detailed analysis examines each design point as it pertains to four output metrics, which are treated as separate responses: the average number of blue force carrier and BFNMF losses, the average time it takes blue forces to achieve air supremacy, and the average number of red force surface to air missile (SAM) sites that are destroyed. Through the utilization of summary statistics and graphical examination, it initially seems that the variation exhibited in all responses is unique from one design point to the next. This implies changes in BFNMF intercept speeds have a noticeable effect on the outcome of the PUNIC 21 scenario. However, analysis of variance, Tukey's honest significant difference, and Pearson's chi-squared statistical tests reveal that only one metric—the elapsed time until blue forces achieve air supremacy—contains design points that are statistically different from one another. Further investigation of this metric reveals that in design point eight, which includes higher BFNMF intercept speeds, blue forces only took 13.71 days on average to achieve air supremacy—while some other design points failed to achieve air supremacy within 20 days nearly 40 percent of the time. It is interesting to note that design points seven and eight resulted in blue forces not only achieving air supremacy 100% of the time (across 25 replications), but did so earlier in the scenario. The other design points achieved air supremacy an average of 18.5 out of 25 or only 74% of the time. Further analysis was performed, in the form of linear regression and partition

tree models, in order to identify whether any of the intercept speed factors are significant in terms of predicting a response and to detect interactions that may have been contributing to the response output. As a whole, we find that the variability in outcomes inherent in STORM dominates the effects of changing BFNMF intercepts speeds by plus and minus 10%. This is not surprising, as STORM contains many thousands of input factors that could conceivably affect output measures.

The primary purpose of this pilot thesis is not to determine which intercept speed factors of the BFNMF are most significant in determining the average number of blue carrier losses, the average time to blue air supremacy, or any other metric. The goal is ultimately to determine whether or not it is feasible to successfully implement DOE within STORM and, if it is, to develop a sound methodology for doing so. We conclude that despite the lack of statistical differences among the design points for three out of the four metrics, a DOE can be implemented and may be extremely useful to N81 analysts if the methodology used in this thesis can be automated. However, now that we know design is a viable option, additional methods into building software tools to automate this process must be explored, which will allow for larger and more effective designs to be implemented within STORM.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

The past two years at NPS have been one of the most academically challenging experiences I have had. I cannot deny that my experience has been equally rewarding.

I would first like to express my sincere appreciation for all the help my advisor, Dr. Tom Lucas, provided me throughout three classes and especially during the process of writing my thesis. Professor, your expertise, guidance, support, and most importantly, patience has been invaluable to the successful completion of my thesis and in the OR program. To the other members of the SEED center, especially Dr. Susan Sanchez, and my second reader, Dr. Paul Sanchez, thank you so much for taking the time to sit down and meet with me on multiple occasions and providing the type of feedback I needed. To Stephen Upton, thank you for everything! Without you, I wouldn't have been able to successfully run a design in STORM; a task that, at first, seemed very daunting.

I believe the sole reason for my motivation to succeed has been my family; especially, my father, Bill, who has spent a lifetime positively influencing others and getting them to recognize their potential. His dedication to family, friends, and his country has been such an inspiration. Thank you for always motivating me to be the best officer I can be.

The most important person I would like to thank is my wife, Krysta. Despite my many long hours spent away from home, her love and support was unwavering. Sweetheart, you are such an amazing person and you certainly deserve as much credit for my success as I do.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

According to current strategic guidance, the Department of Defense (DOD) must reduce future defense expenditures to $487 billion over the next decade due to caps instituted by the Budget Control Act (BCA) of 2011. A sequestration mechanism was also instituted by the BCA requiring $50 billion in cuts annually (Hagel, 2014). In 2016, unless Congress agrees on a budget, the DOD may be facing additional sequestration cuts. The austerity of such budgetary restrictions will force current and future decisions on matters of national defense to be heavily scrutinized for validity. The Office of the Chief of Naval Operations, Capability Analysis and Assessment Division (OPNAV N81), provides senior DOD officials this validity through crosscutting analysis of U.S. Naval warfare and force-level capabilities. Personnel at N81 utilize many techniques and models to perform their assessments, but the primary modeling environment used for campaign analysis is the Synthetic Theater Operations Research Model (STORM).

STORM is a state-of-the-art, multi-sided, closed-form, stochastic computer simulation designed to provide insight into military strategy, capabilities, force structure, and operational effectiveness in a joint warfighting context (Group W, 2012c). Originally developed for HQ/U.S. Air Force Studies and Analysis and Assessments and Lessons Learned (HQ/USAF A9), STORM offers unique campaign analysis capabilities and was adopted as N81's primary assessment tool once a maritime component was incorporated in 2006 (Sweeney, Hamman, & Biemer, 2011).

The current version, STORM 2.3, allows analysts to input a multitude of factors (often referred to as variables) in a single simulated campaign covering the air, space, land, and maritime domains (Group W, 2012a). Consequently, a single instantiation results in gigabytes of output data that analysts must examine. To adequately characterize the breadth of objects being simulated in a campaign, many thousands of input variables must be specified by the user, as well as entity capabilities, behaviors, and interactions with each other and the environment. Given STORM's stochastic nature, typically 25 to 50 replications are generated for each configuration, which has proven to generate stable results with sufficiently narrow 95% confidence intervals, thereby allowing analysts to

better understand key output measures. As beneficial as having a stochastic model may seem, it exponentially increases the workload of N81 personnel and their turnaround times. The enormous volume of output data impedes fast and efficient use of STORM and requires more than 24 hours of manpower-intensive post-processing before recommendations can be forwarded to senior-level decision makers.

In an effort to improve N81's post-processing capabilities, the Naval Postgraduate School's SEED Center (Simulation, Experiments, and Efficient Design, see http://harvest.nps.edu) initiated a project to increase the overall speed and efficiency of STORM analysis by developing a post-processing tool that extracts scenario relevant metrics. N81 often receives tasking to estimate how perturbations in force structure, platform types, and capabilities might impact the overall effects on a specific campaign. Given that scenario development usually takes a year or more and output responses are based on large input data files, modifying individual factors for the purpose of sensitivity analysis is extremely difficult. Therefore, harnessing the power of experimental design offers a potential solution.

Experimental designs developed specifically for computer models have made it possible for experimenters to explore many more input variables than was feasible only a few years ago (Hernandez, Lucas, & Carlyle, 2012). Additionally, the designs indicate how to efficiently vary the settings of factors to see whether and how they affect outputs. This provides insights that cannot be gleaned from trial-and-error approaches or by sampling factors one at a time (Sanchez, 2007). Providing N81 with the ability to automate running STORM experiments according to a specified experimental design will save valuable time and allow the Navy to operate under a strict budgetary constraint by quickly identifying the dominant factors within a specified campaign.

A. LITERATURE REVIEW

Documentation relevant to STORM 2.3 software is limited to developer-written manuals and serves as reference documents to all users. The User's Manual is written for the end user and provides basic operating instructions, input and output tools, and terminology associated with the graphical user interface (GUI) within which users

interact with the system (Group W, 2012c). The current version of STORM offers many enhancements to its predecessors, which were sponsored by the U.S. Navy and Marine Corps in an effort to integrate campaign-level expeditionary warfare into STORM. These upgrades are explained in the "What's New in Version 2.3" document provided as an add-on to the User's Manual (Group W, 2012d). The Analyst's Manual is intended to promote a level of understanding and skill with STORM on the part of the campaign analyst. It is designed for individuals, with a range of experience levels, who are concerned with employing the simulation as a campaign-level tool to produce credible results for the decision maker (Group W, 2012a). The Programmer's Manual is more technical in nature, and provides guidelines for STORM development. It is intended for the programmers and designers of STORM to use as a guide to develop source code at both the Group W Inc. facility and remotely (Group W, 2012b). Additionally, appendices contain information useful to those who utilize all of the manuals associated with STORM software (Group W, 2012a). These documents serve as the primary resources that provide the background information on STORM for this thesis.

Experimental design has a rich history, with many theoretical developments and practical applications in a variety of fields (Kleijnen, Sanchez, Lucas, & Cioppa, 2005). The implementation of a design of experiments (DOE) within the framework of STORM is primarily based on the work of Professors Thomas Lucas and Susan Sanchez. Their work has influenced more than a dozen DOD modeling environments and countless thesis projects related to the subject. Their research (see http://harvest.nps.edu) is used extensively in this proof-of-concept thesis. Additionally, works by Averill M. Law, who was previously a Professor of Decision Sciences at the University of Arizona, and is now President of Averill M. Law and Associates, are utilized to help set up the initial design in STORM (Law, 2007).

B. RESEARCH QUESTIONS AND GOALS

The primary goal of this research is to determine whether the implementation of a modern-day design of experiments is feasible within the complicated framework of STORM. This is a proof-of-concept thesis applies a recently developed post-processing

tool, known as STORMMiner, to the unclassified, pre-installed PUNIC 21 scenario. STORMMiner allows for the manipulation of specific and carefully chosen factors. As a result, this research is guided by the following questions:

1. Does STORM's complexity allow for the implementation of a design of experiments? If so, what is the most efficient execution of such a design?
2. Should all input variables be considered as significant factors with regards to model output? If not, which ones should be and how can they be determined?
3. Will a single proof-of-concept demonstration be sufficient in determining the analytical potential of the new capabilities?

C. METHODOLOGY

This thesis explains specific details regarding STORM, including some of its input variables and responses, to provide the reader with a basic understanding of this complex campaign analysis tool. Output data analysis is performed in order to initially determine significant factors that may be appropriate to use in a small design of experiments. After a designed experiment is successfully run, analysis is performed to demonstrate the benefits of using experimental design. Following the implementation of the newly developed technique, the results will be provided to N81, extending its current analysis capabilities by enabling statistical insights to be gleaned through the designed modification and experimentation of input variables.

D. BENEFITS OF RESEARCH

This research will assist N81 analysts in capturing the full potential of STORM and provide insights into how modifying model inputs may affect model outputs. This will be critical in verifying and validating (V&V) new scenarios, helping N81 quickly gain confidence in their reliability. Given the substantial time it takes for the development of a single scenario, the implementation of experimental design offers the potential to provide answers to difficult questions in a fraction of the time now required.

In order to provide substantial benefits to N81, the experimental design process developed in this thesis must be automated. This capability has not yet been achieved. In that regard, this pilot study creates a foundation for follow-on research pertaining to STORM. The design and analysis performed in this thesis is a first step in a long journey that will ultimately enhance the utility of STORM to the Navy.

THIS PAGE INTENTIONALLY LEFT BLANK

II. STORM OVERVIEW

STORM is a multi-sided, stochastic, simulation of air, space, ground, and maritime planning and execution. Its framework is extremely complex and often difficult to interpret even for an experienced user. From the analysts' perspective, STORM is a means to an end. For N81 analysts specifically, this means their focus may be on output data and not necessarily on how the data flows through STORM. Developers and DOD civilian contractors are the STORM experts and work closely with N81 if technical questions arise. Therefore, the main purpose of this chapter is to provide the reader and follow-on researcher with a broad overview of STORM, to include its development, basic structure, characteristics, and capabilities from an analyst's perspective.

A. MODELING AND SIMULATION

The Defense Modeling and Simulation Office (DMSO) was established in 1991 following a policy study set forth by the Department of Defense (DOD) on Defense Modeling and Simulation (M&S). This was the precursor to the DOD directive, signed in 1994, which requires each of the military services to adopt their own verification, validation, and accreditation (VV&A) process for M&S (Nunn & Heimerman, 2003).

B. ADOPTION OF STORM MODEL

Following an in-depth review by the Center for Naval Analysis (CNA), the Integrated Theater Engagement Model (ITEM) was adopted by N81 as the U.S. Navy's primary assessment tool in 2003. ITEM provided integrated air, land, and naval warfare engagement models permitting a realistic representation of capabilities utilizing a deterministic method to represent uncertainty in outcomes (Sweeney et al., 2011). At that time, a deterministic approach was preferred over a stochastic one, which requires multiple simulation runs to produce a distribution of outcomes for a single set of inputs (Sweeney et al., 2011). Given significant advancements in technology and computing power over the past few decades, the higher number of required model runs has become less of an issue, making stochastically driven models more appealing due to their ability

to provide a solution space (i.e., distribution of potential outcomes) rather than a point estimate based on assumed probabilities. This was the principal reason why N81 looked to adopt STORM as its new assessment tool.

C. IMPLEMENTATION OF MARITIME COMPONENT IN STORM

STORM was first developed and used by the U.S. Air Force and is managed by the U.S. Air Force Air Staff's Studies and Analysis Directorate (A9). It replaced the theater-level tactical air warfare model known as THUNDER as the Air Force's primary campaign analysis tool in 2004. In 2006, N81 partnered with A9, under the project name STORM+, in an effort to determine the feasibility of adding a maritime component to a predominately air-warfare model (Sweeney et al., 2011). Verification and testing efforts were broken into three distinct phases, each building on the previous phase, with the ultimate goal being the successful implementation of a maritime operational command and control (C2) component similar to the ground and air C2 components that previously existed in the Air Force's STORM model. In July 2010, STORM+ efforts resulted in STORM Version 2.0, which was utilized by N81 as their primary campaign analysis tool until it was superseded by version 2.3 in early 2014.

D. BENEFITS OF STOCHASTIC SIMULATION

Although computing technology has made rapid advancements in the past two decades, it is impossible to model every outcome military combat could generate, especially considering war itself is inherently chaotic, intrinsically unpredictable, and characterized by a great deal of uncertainty (Vinyard and Lucas, 2002). Efforts to close the gap between simulation and reality can best be made by the implementation of a stochastic model, which introduces one or more random variables as inputs to represent uncertainty. This produces an outcome, result or value that depends on chance (Lucas, 2000). Additionally, results are provided as a distribution of outcomes rather than a simple point estimate, which is particularly beneficial to a military analyst because it allows them to identify the entire range of possibilities, or variation, within a specified campaign. Only with a thorough investigation of all associated uncertainty will the

decision maker be allowed to interpret the results in an informed way and make risk assessments (Committee on National Statistics and Committee on Applied and Theoretical Statistics, 1994).

E. STORM AS A STOCHASTIC MODEL

STORM is a stochastic simulation, therefore requiring multiple runs in order to achieve a desired level of confidence in outputs. For highly aggregated data, relatively few replications may be needed. However, for rare events, such as the loss of a carrier, a significantly higher number of runs may be required (Group W, 2012a). As a baseline, N81 analysts typically use 30 runs on scenarios−regardless of complexity. Moreover, as a stochastic model, STORM input data is generated from twelve available random number distributions (e.g., binomial, gamma, uniform, triangular, Weibull), that are pre-set or user-defined. For example, Figure 1 is an input file that defines damage functions for specific maritime surface ships by weapon class.

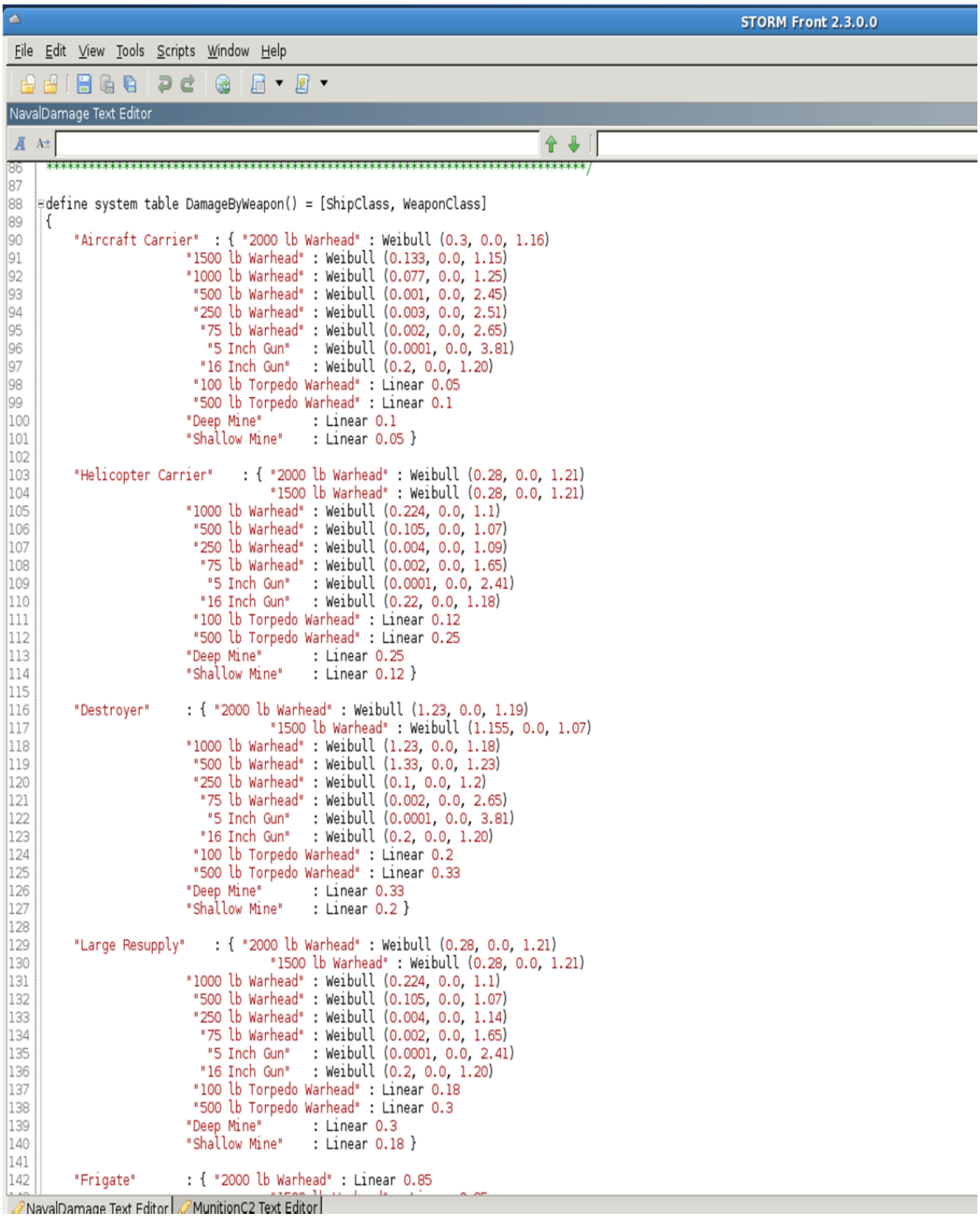


Figure 1. An example input data file for maritime surface ship damage functions categorized by weapon type

F. STORM−A CAMPAIGN ANALYSIS TOOL

The shaping of military strategy involves three key elements: development of an overall objective or end-state; a ways (courses of action); and a means (available resources). STORM is a campaign analysis tool that aids the decision maker in developing and evaluating the above-mentioned characteristics for an improved understanding of policy, acquisition, and operational issues that may arise (Group W, 2012c). Figure 2 illustrates the way STORM captures the overall impact, known as the campaign analysis thread. It is this traceable process—which links systems represented in STORM and their unique capabilities—that provides an outcome of adjudications over a simulated period of time. The four pillars that make up the campaign analysis thread serve as guidelines for a balanced simulation (Group W, 2012c). Systems are the real-world objects (air, land and sea platforms) and their supporting subsystems, which accurately represent the multiple players normally associated with a military campaign. Capabilities are the characteristics of each specific system. For example, destroyers may carry torpedoes with heavier payloads than a hunter submarine or offer different intelligence, surveillance, and reconnaissance (ISR) capabilities. Planning refers to the courses of action (COA) that are carefully planned out in advance by individuals relevant to a specific campaign (such as N81). For example, a COA may include sending a carrier strike group (CSG) in the western Mediterranean for sea denial operations. These COA's are implemented through a set of executable data files and control how assets or groups of assets move during a scenario similar to pieces on a chessboard. Analysts, following execution, will be able to determine the overall impact of their generated scenario, which is the beauty of STORM.

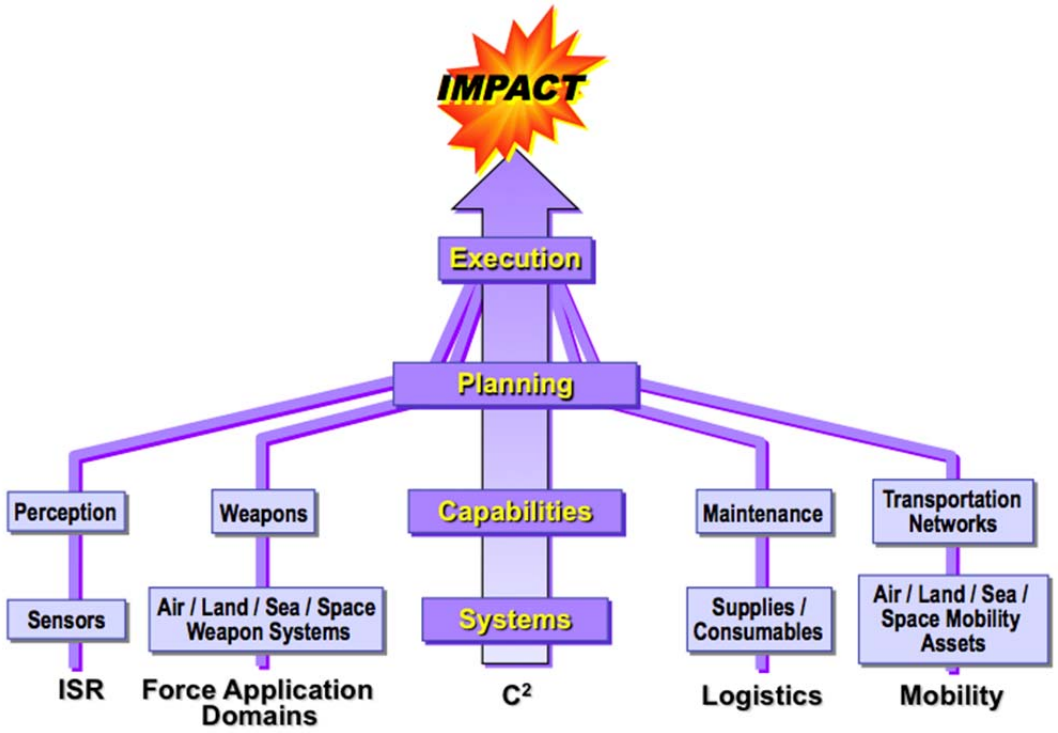


Figure 2. Diverse activities associated with a STORM campaign (from Group W, 2012a)

G. STORM–A DATA-DRIVEN DESIGN

STORM exhibits a data-driven design characteristic, which purposefully avoids the hard wiring of data (Group W, 2012b). N81 is often tasked with running unique classified scenarios that are either relevant to present-day military operations or future campaigns that may fall within the U.S. Navy's area of interest (AOI). Either way, hard wiring data contributes to making modifications complex, extremely costly, and hinders the ability to build scenarios in an appropriate amount of time. Currently, N81 analysts and STORM developers can build a complete real-world scenario in about a year or so, depending on the level of complexity. Without STORM's unique capability, the time required would dramatically increase.

H. REPRESENTATIONS IN STORM

STORM models real-world combat operations through the use of input data files that are organized into five classes or representations: Command and Control (C2), Assets, Intelligence Manager, Interaction Manager, and Environment. The classes interact nearly simultaneously during a simulation to form STORM's conceptual model. Each representation contains sub-classes, which continuously send and receive state-condition reports to the C2, intelligence, and interaction managers. Upon execution of a scenario, the C2 manager will issue initial orders and requirements to the assets and intelligence manager, respectively. As the campaign progresses, assets will send status reports on their current condition, whether fully mission capable, degraded, or out of action (OOA). Those updates are then interpreted and new orders issued by the C2 manager. This cycle continues until the simulation is terminated by a pre-set run time or the opposing force is unable to continue its mission (as determined by a user specified stopping rule). Shown in Table 1 are the five representations and associated sub-classes, including the most relevant classes for this thesis, assets, and environment.

ASSETS	ENVIRONMENT	INTERACTIONS	COMMAND & CONTROL	INTELLIGENCE
Surface	Terrain	Ground-to-Ground	Ground	Sources
Air	Weather	Air-to-Surface	Logistics	Perception-Based Planning
Orbital	Surface Transportation Network	Surface-to-Air	Mobility	
	Geopolitical Boundaries	Air-to-Air	Air	
		Counter-Space	Naval	
		Chemical and Biological Weapons	Space	
		Systemic and Special Effects	ISR	
		ISR Sensor-to-Surface		
		Naval Detection and Damage		

Table 1. Table of available representations in STORM

1. Storm Assets

Real-world physical entities are represented in STORM by assets that move, attack, conduct surveillance, consume resources, and execute orders similar to the way military operations are conducted at the theater level (Group W, 2012c). As in real combat, assets can experience a reduction in capabilities or be taken out of the fight

altogether. There are three categories of assets: Surface, air, and orbital. Surface assets are a representation of naval surface and subsurface platforms (e.g., carriers, cruisers, destroyers, amphibious crafts, submarines), shore installations (e.g., naval and air bases), and ground units (e.g., armored divisions). Air assets represent individual airframes (e.g., strike fighters and reconnaissance aircraft) or squadrons. Surface and air platforms have unique capabilities (e.g., munitions, surface search radar, sonar), which themselves contain distinctive characteristics (e.g., payload, max/min ranges, max/min speeds). Satellites and space-based platforms are representations of orbital assets relating mostly to how effectively surface and air assets communicate (Group W, 2012c).

2. Storm Environment

Environmental conditions are key elements that can impose significant limitations on a campaign. Therefore, the environment class provides unique capabilities to the user that allow them to enforce specific conditions on the AOI, such as terrain type, geospatial location, cloud density, darkness, and time (day/night).

I. STORM–THE USER

STORM offers many tools the end user can utilize for analysis. These tools are broken into three functional areas: Input, execution, and output. Input refers to all input data files, segregated from the model itself, that pertain to a user-generated scenario, referred to as a study. Each study houses relevant data files that are fed into STORM at the execution of a simulated run and can be easily accessed through STORM's GUI under the study manager tab (see Figure 3).

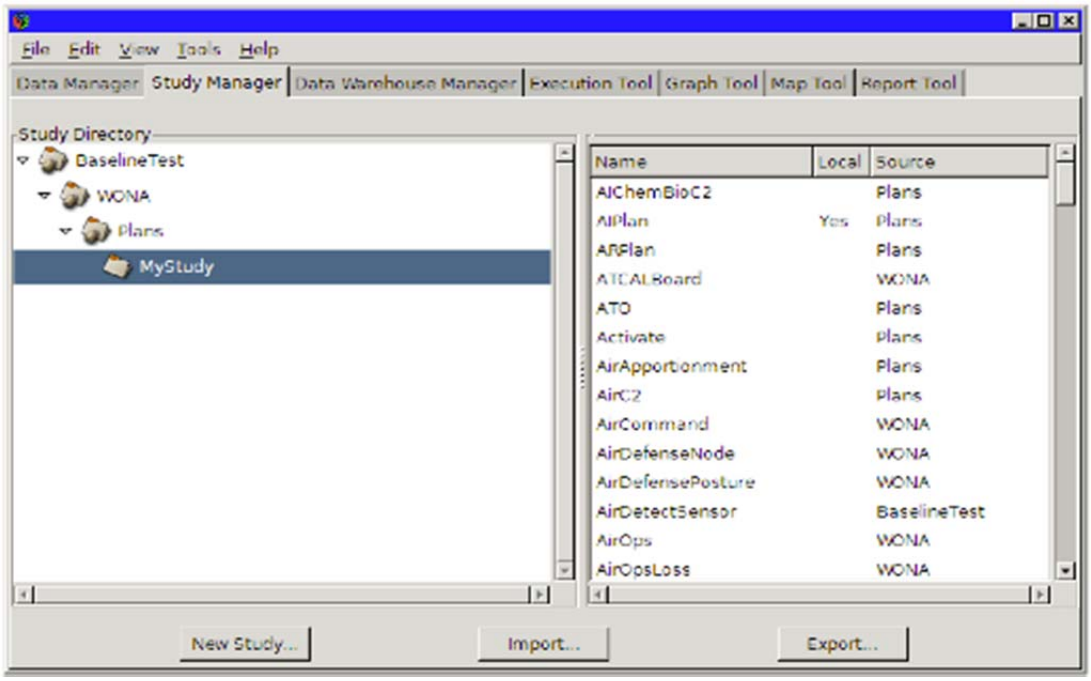


Figure 3. STORM's GUI (from Group W, 2012c)

Additionally, the execution and output functional areas can be easily accessed from this interface to complete each run and perform post-run analysis, respectively. The execution functional area offers many options to customize a particular run configuration, such as the number of runs, specification of a random number stream (1–10), or whether to compress the output file. The output functional area is broken up into three post-run analysis tools: Map tool; Graph tool; and Report tool.

1. Map Tool

The map tool is an interactive application that provides a geographical representation of a previously completed simulation run (see Figure 4). STORM users can specify which assets to view through field filters, fast-forward to a certain point, or zoom in to analyze a particular AOI. Unfortunately, STORM's map tool cannot be utilized in real-time. Each run must first be completed in order for this option to be used.

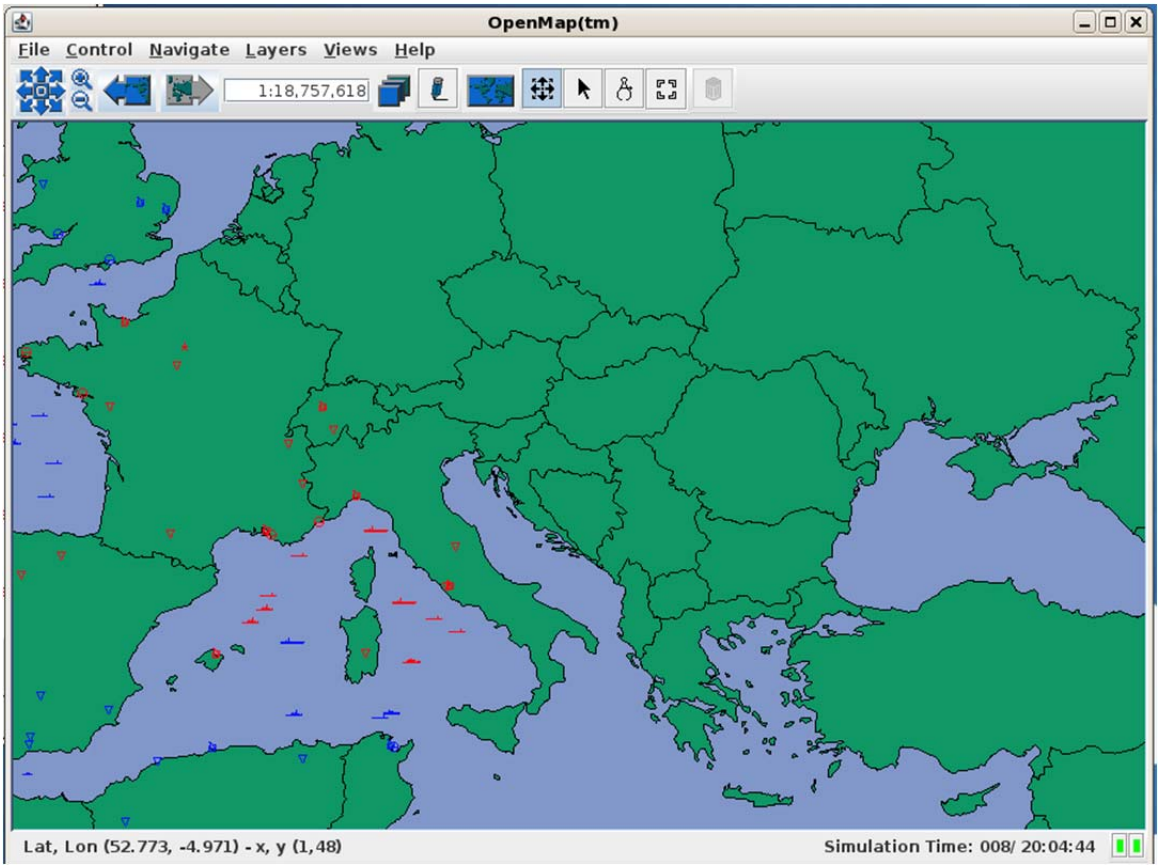


Figure 4. STORM's map tool interface (from Group W, 2012c)

2. Graph Tool

The graph tool is an application that allows the user to view model results in either graph or data table format (Group W, 2012c). These data tables are transferable into external statistical applications (e.g., R, JMP) or spreadsheet programs (e.g., Excel) that give the user even more options with which to analyze the output. Although extremely beneficial, graphs and data table options are limited in STORM to only certain output metrics. For metrics not included, users must utilize the report tool application.

3. Report Tool

The report tool application provides output data in tabular form as either HTML or comma separated file (.csv) files. Unlike the graph tool, it offers specifics on every aspect of the simulation. For example, an analyst can see which maritime assets were killed, who killed them, and with what type of weapon. Although this is useful information, a single run can generates hundreds of files. This makes it particularly difficult for even an experienced STORM user to identify useful information.

J. STORM–PUNIC 21

The current version of STORM includes two unclassified test scenarios; WONA and PUNIC 21. Both were specifically designed to provide users the ability to experience many aspects of STORM's functionality. Due to its maritime aspects, relatively small input data set, and short run time (20 simulated days), PUNIC 21 was selected as the test scenario for this thesis. The purpose of this section is to provide a brief overview so readers may familiarize themselves with PUNIC 21.

1. Current Situation and Battle Phases

PUNIC 21 is predominately a naval battle between blue forces of the allied nations of Anglo Republic and Carthage (ARC), and red forces of the Swiss Empire (SE). Tensions between these nations are on the rise and the area has become increasingly unstable due to the Swiss Empire's determination to seize control of the entire Iberian Peninsula. Figure 5 is a geographic snapshot of the AOI and occupied territories for PUNIC 21. The scenario is broken up into four major phases: Battle of the Atlantic; Battle of the Mediterranean; Fight for Spain; and Fight for Italy. Each phase incorporates surface, sub-surface, air, and ground engagements taking place over a period of twenty days, which was arbitrarily chosen as the battle duration.

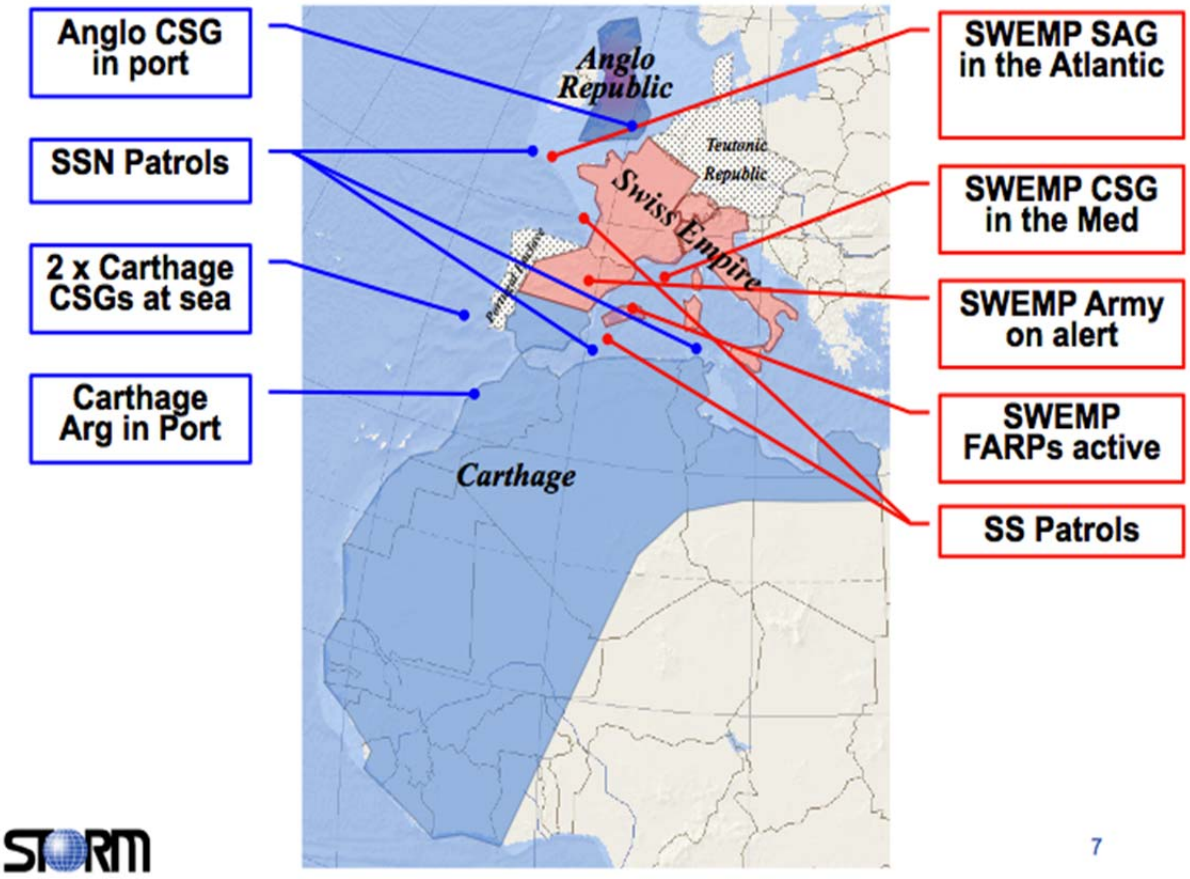


Figure 5. AOI for STORM's PUNIC 21 scenario

2. Order of Battle for Blue and Red Forces

Both ARC and SE possess a particular level of military force strength. However, both are relatively at parity and contain maritime, air, land, and logistical elements. The maritime order of battle (OOB) for blue and red forces is shown below in Table 2. Darker colored rows indicate totals for combat vessels, combat logistic forces (CLF), and mobile riverine forces (MRF).

The air OOB shown in Table 3 is similar to the maritime OOB, that is, darker rows indicate totals for combat aircraft, non-combat aircraft, and missiles (for red only). For PUNIC 21, red forces possess 180 surface-to-surface missiles, 40 short-range ballistic missiles, and 24 intermediate-range ballistic missiles. Although this is a capability the blue forces lack, it does not provide a significant advantage to the SE.

Blue Navy	Quantity	Red Navy	Quantity
CV (Carrier)	3	CV	1
LHD (Amphibious Assualt)	3	LHD	0
CG (Guided Missile Cruiser)	8	CG	12
DDG (Guided Missile Destroyer)	24	DDG	27
MIW (Counter-Mine)	2	MCM	0
SSN (Attack Submarine)	10	SSN	11
SSGN (Guided Missile Submarine)	1	SSGN	0
Combat Vessels	51	Combat Vessels	50
CLF (Combat Logistic Force)	11	CLF	2
CLF Oiler	6	CLF Oiler	
MRF-N (Mobile Riverine Forces)	120	MRF-N	100
MRF-M	40	Total MRF	100
MRF-EW	15	AEW	3
MRF-Tanker	15	MPA	8
Total MRF	190	Vertical Assault	0
AEW (Airborne Early Warning)	9		
MPA (Maritime Patrol)	12		
Vertical Assault	40		

Table 2. A list of blue and red maritime assets for the PUNIC 21 scenario

Blue Air	Quantity	Red Air	Quantity
MRF	138	MRF	144
MRF-EW	12	MRF-EW	10
FTR	70	FTR	64
BOMBER	32	BOMBER	32
Combat Aircraft	252	Combat Aircraft	250
Tanker	36	Tanker	0
AEW	12	AEW	10
HVA (ISR)	8	HVA (ISR)	8
UAV (ISR)	16	UAV (ISR)	16
AIRLIFT	24	AIRLIFT	24
Total Aircraft	348	Total Aircraft	308
		SSM	180
		SRBM	40
		IRBM	24
		Total Missiles	244

Table 3. A list of blue and red air assets for the PUNIC 21 scenario

THIS PAGE INTENTIONALLY LEFT BLANK

III. IMPLEMENTING A DESIGN OF EXPERIMENTS IN STORM

This chapter provides the foundation for successfully implementing a design of experiments within STORM. As discussed in Chapter II, STORM is a complex combat simulation model that is data-driven rather than "hard-coded." This allows N81 analysts to build campaign scenarios without having to possess an intimate programmers-level knowledge of STORM. However, associated with the multifaceted STORM modeling environment are thousands of input variables spread across hundreds of input files, a vast number of which may be significant in determining model output. Only through a great deal of post-processing analysis, which is time and manpower intensive, can correlations be identified and insights gained. This severely impacts N81's ability to accomplish quick turnaround tasking, especially if questions arise that pertain to how perturbations in force structure, platform types, and capabilities impact a particular campaign.

The benefits of a well-designed experiment can provide invaluable insights to analysts who seek to identify the most significant input variables or how making modifications to them could impact a combat scenario. For example, senior military decision makers are often interested in looking at utilizing different surface or air assets to perform a mission. Likewise, they may want to identify the overall impact of decreasing submarine presence in an AOI. These questions are exceptionally difficult to answer in a relatively short period of time because STORM does not currently support quick-turn analysis due to long simulation run times and the enormous output generated. Hindering this further are the vast numbers of input files that are associated with a single scenario, such as PUNIC 21. For real-world classified scenarios this number increases significantly, placing an even bigger burden on N81 analysts who seek to answer difficult questions in a timely manner. Therefore, this chapter explains the terminology associated with a design of experiments (DOE), the step-by-step methodology that is required in order to implement a design in STORM, and issues that arose during the process of doing this proof-of-concept demonstration.

A. DOE TERMINOLOGY

This section covers important terminology associated with a DOE, bearing in mind it may be a foreign concept to some. In DOE terms, experimental designs specify how to vary a set of input variables in order to identify whether and how they affect a particular response or responses (Sanchez, 2007). For example, this could include, but is not limited to, changes in overall force composition or enhancing a system's capabilities.

1. Types of Variables

Variables are classified as either quantitative or qualitative. Quantitative variables are those that take on a numerical value, such as the maximum speed of a blue force destroyer, an aircraft's minimum engagement range, or the initial force level. Qualitative variables are not measured by a numerical value; they are categorical and may have no natural sense of ordering, such as different types of undersea warfare offensive weapons (e.g., Mark 46 torpedo or Mark 48 torpedo). Experiments may contain both types of variables. For this research specifically, quantitative variables (intercept speed profiles) were chosen that relate to the blue future naval multi-role fighter (BFNMF). These variables were not chosen arbitrarily. Reasons for their selection are discussed later in this chapter. The baseline intercept speeds are illustrated in Table 4. Intercept speeds are significant in terms of reinforcing friendly aircraft for supporting missions and intercepting hostile aircraft and surface vessels that may pose a threat to blue forces in general or their high value units (HVU), such as a carrier. Aircrafts with vastly superior speed profiles are likely to have substantial advantages in combat.

Friendly Intercept Speed (NM/HR)	Hostile Low Intercept Speed (NM/HR)	Hostile High Intercept Speed (NM/HR)
600	540	540

Table 4. Table of the baseline intercept speeds profile for the BFNMF

2. Output Metrics–"What It Takes To Win" Metrics

Output metric is another term associated with DOE. Given STORM's inherent stochasticity, output metrics of interest are often a particular measureable outcome averaged across a set of replications, such as the number of blue forces remaining or the number of sorties flown by an aircraft. Key measures that enable blue forces to achieve victory are defined by N81 as "what it takes to win" (WITTW) campaign metrics. These are interesting outcomes or events that either may or may not take place given certain input parameters or that are triggered by a preceding event like red forces entering the Mediterranean Sea. Given that PUNIC 21 is a terminating simulation, certain events or goals may not take place or be achieved due to some element or other variable that is inherently invisible to analysts at first glance. Further exploration by N81 of the output data may reveal dependencies or correlations among controllable input factors. However, all analysis is constrained by the initial factor settings. For example, a quadratic effect can only be estimated if at least three distinct values of an input variable are used in the experimentation. Building a DOE and incorporating factors that exhibit interesting relationships will reveal a much broader scope of possible cause-and-effect associations.

3. Designs In General

The term *design matrix* refers to a matrix where columns correspond to factors and the entries within each column are the settings for that factor. Each row in the matrix is a *design point* that specifies all of the factor settings for that simulation run–where each factor setting is varied at the researcher's discretion. Factor levels are often characterized by high and low settings from the baseline. For example, an analyst may want to vary a factor by plus and minus 20% of a typical setting. If the baseline is 100 units of some measure, the high value would be 120 units, and the low value 80 units. Lastly, the response is a metric that is being explored and will likely vary over each design point run. A word of caution: High and low settings should be set to provide a somewhat realistic interpretation. That is, some settings may go beyond a factor's physical capability and thus DOE results would be worthless in determining true effects on an outcome.

4. 2^k Factorial Design

There are many approaches to designing experiments, such as sophisticated nearly orthogonal Latin hypercube (NOLH) designs (see Cioppa & Lucas, 2007). For this research, which is the first that the authors know of utilizing a DOE on STORM, we use a relatively simple 2^k factorial design in order to determine the effects of factors on the response. A 2^k was chosen for its simplicity and also because with it we can measure and examine interactions (Law, 2007)–such interactions can be critical in combat. Table 5 shows a 2^3 factorial design in matrix format, also known as a design matrix. That is, three factors at two levels each, and an associated response (R_j) for each factor setting over eight design points. Setting up a matrix facilitates calculations of the factor effects and interactions once the design is implemented. A plus indicates that the factor is set at its high setting for that run. Likewise, a minus indicates that factor is set at its low setting for that design point.

Factor Combination (Design Point)	Factor 1	Factor 2	Factor 3	Response (R_j)
1	–	–	–	R_1
2	+	–	–	R_2
3	–	+	–	R_3
4	+	+	–	R_4
5	–	–	+	R_5
6	+	–	+	R_6
7	–	+	+	R_7
8	+	+	+	R_8

Table 5. Design matrix for a 2^3 factorial design (from Law, 2007)

a. Main Effects

Often examined are the *main effects* (e_j) of each factor—which is the average change in response due to moving the factor from low to high levels while holding all

other factors fixed (Law, 2007). This is done with all eight design points over all possible combinations of the other k–1 factors (for a total of 2^{k-1} differences).

b. Interactions

Interactions describe the case where two or more factors behave synergistically, i.e., the effect of one of the factors upon the outcome is altered by the settings of other factors. We may be unable to detect interactions or separate them from main effects unless the experiment is carefully designed. Since interactions may or may not be present in STORM scenarios, it is imperative to use designs that would permit us to identify their presence or absence. Statistical packages such as R or JMP can perform the calculations needed to estimate main effects and interactions as long as the data have been created with a suitable design.

B. CRITICAL LIMITATIONS IDENTIFIED

Given STORM's complexity, there are critical limitations that must be taken under consideration before applying a DOE. For smaller models, a design matrix is usually generated in a spreadsheet program, such as Microsoft Excel (.xls or .csv format). Once the design is complete, it is fed into a simulation model and looped over each factor setting, creating a single output file containing data for analysis. Unfortunately, STORM's input files are *.dat* or data files and using Excel generated designs are not possible without first converting them into a data file, which takes time and requires analysts with coding experience. The best approach to implement a DOE in STORM is to maintain the design in a data file format. Additionally, STORM has 148 input data files that represent various categories of data including platform performance, geographic locale, operational planning, tactical planning, and inventory (Group W, 2012a). Figure 6 is a snapshot of STORM Front and all of the input files associated with a study. On the left side is the expanded view of the highlighted Naval Asset file along with 147 additional files relevant to PUNIC 21, such as Naval C2, Naval Unit, and Naval Tactical C2. Each file contains the specific information relevant to aspects of the scenario, like an asset's ID (e.g., Anglo Republic Carrier Strike Group Medium-Range Fighter Squadron

A) or Asset Type (e.g., Blue Advanced Destroyer). Some asset details are referenced in only three input files, but others are referenced in more than a dozen. Most files call on others to perform specific tasks throughout a scenario. This dependency adds to the difficulty in terms of DOE implementation and severely limits which factors can be chosen for a design.

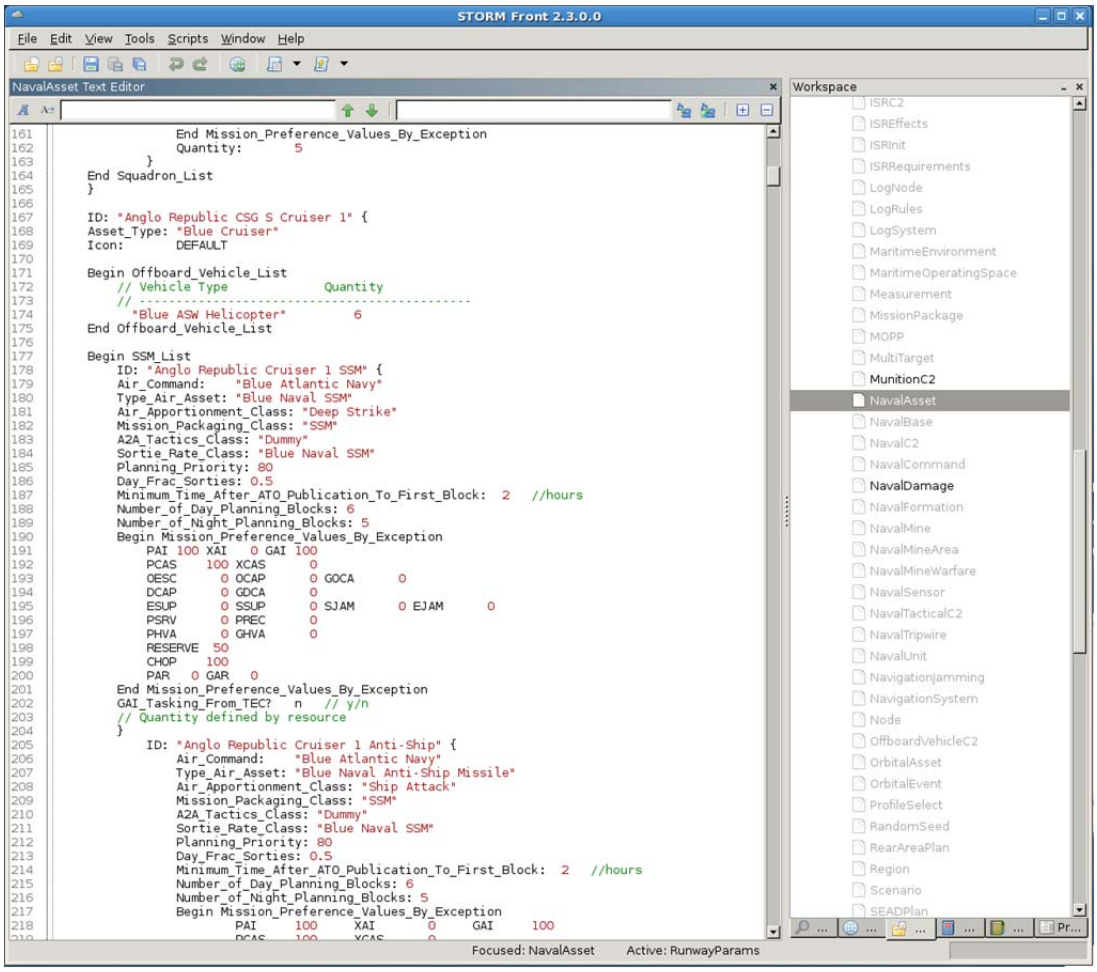


Figure 6. STORM front input data files

C. IMPLEMENTING A DESIGN WITH STORM: A FOUR-STEP PROCESS

Taking into consideration the critical limitations discussed above, building and successfully implementing a DOE in the complex STORM environment can be done with four relatively straightforward steps:

1) The analyst should examine scenario-specific output from a baseline experiment and identify correlations and patterns among responses of interest. Preferably, moderate to high correlation is observed using Pearson's Product-momentum correlation, Spearman's correlation criteria, and/or Kendall's tau. As we shall see, the relationships among responses help identify input factors to explore.

2) Given the first step, carefully choose the input factor(s) that likely will impact the responses. The analyst will then build a DOE, in our case a 2^k factorial design with levels set at their discretion. For this study, the levels were set to +/- 10% of the baseline value, but the analyst may increase those levels depending on the factors themselves and scenario. As this capability expands, more sophisticated designs (see Kleijnen et al., 2005) can be used.

3) The analyst must create separate study folders in the STORM GUI, then create custom input files that contain each design point setting and save each file in their study directory. Ultimately, this process needs to be automated.

4) Once each study folder (one study folder per design point) contains all custom generated files, the analyst may perform replications for each design point. With access to a computing cluster, these can be run in parallel, dramatically decreasing the clock time required to execute the STORM experiments. Please note that analysts must ensure those files are run locally—details will be discussed later in this chapter.

D. CORRELATION AMONG FACTORS

In order to identify significant relationships among responses, output data from 30 replications of PUNIC 21 were initially examined and graphically analyzed in R-Studio using Pearson's product-momentum correlation or Pearson's correlation coefficient (PCC). PCC identifies linear associations between two variables by assigning a value between -1 and +1. A perfect negative correlation, or a coefficient of -1, indicates a linear relationship: if one variable increases, the other decreases by a proportionate amount. Conversely, a coefficient of +1 indicates that the two variables are perfectly positively correlated, so as one variable increases, the other increases linearly by a proportionate amount (Field, Miles, & Field, 2012). These values are found by estimating the covariance between two variables and then dividing it by the product of their standard deviations, see Equation 3.1.

$$r = \frac{\mathrm{COV}\, xy}{Sx * Sy} \tag{3.1}$$

Since there are 2,589 variables associated with the PUNIC 21 output file, attempting to identify relationships among all of them is a tedious task–and should be automated as much as possible. Therefore, the scope of variables tested in this research was significantly narrowed to the WITTW campaign metrics. Figure 7 is a pairs plot of nine WITTW metrics that included (from top left to bottom right): losses to blue force carriers; advance destroyers; hunter submarines; advanced multi-role fighters; amphibious assault ships; boomer submarines; cruisers; destroyers; and future naval multi-role fighters. Illustrated on the diagonal in light blue are histograms representing the number of losses for each metric at the end of the 20-day terminating simulation for 30 replications of the PUNIC 21 scenario, which provide insights to the possible normality (or other distribution) of the data. The lower left-half boxes are scatter-plots that include a red trend-line to help visually determine what type of relationship exists between the two variables. Finally, the upper-right boxes include *p-values* (*p*) and PCC values (*r*), which provide numerical information on the relationship indicated by each scatter-plot under an assumption of normality. Additionally, output data for 30 replications were chosen and used as a baseline for significant variable identification. Due to the reasons noted in the critical limitations section, only 25 replications were allowed per design point. Therefore, it was essential to this research to maintain a similar number of replications for the eight design points as the baseline.

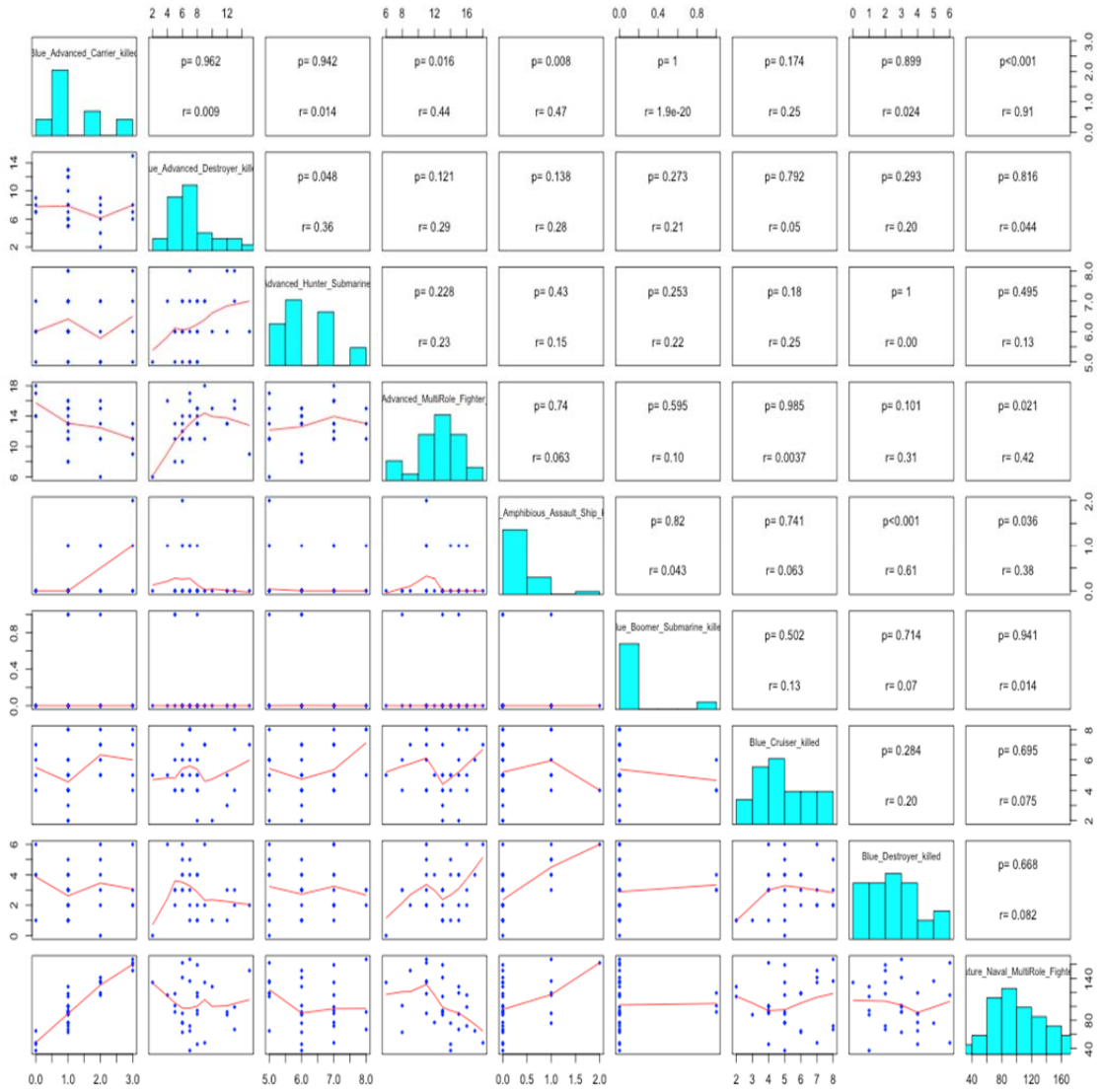


Figure 7. RStudio generated pairs plots of nine WITTW campaign metrics

Figure 7 reveals a strong positively correlated relationship ($r = 0.91$) between blue carrier losses and the number of BFNMF killed. That is, the higher number of BFNMF losses, the greater number of carriers blue forces can expect to lose. Moreover, the *p-value* is less than 0.001; meaning the probability of getting a coefficient this big if the null hypothesis were true is very low. Therefore, there is a high level of confidence that this relationship is genuine (Field et al., 2012). Additional relationships are identified

with having moderate positive correlations when blue forces lose BFNMF aircraft, such as the number of blue advanced hunter submarines (r = 0.13) and blue amphibious assault ships (r = 0.38). Figure 8 is a magnified view of the relationship between blue carriers and BFNMF killed.

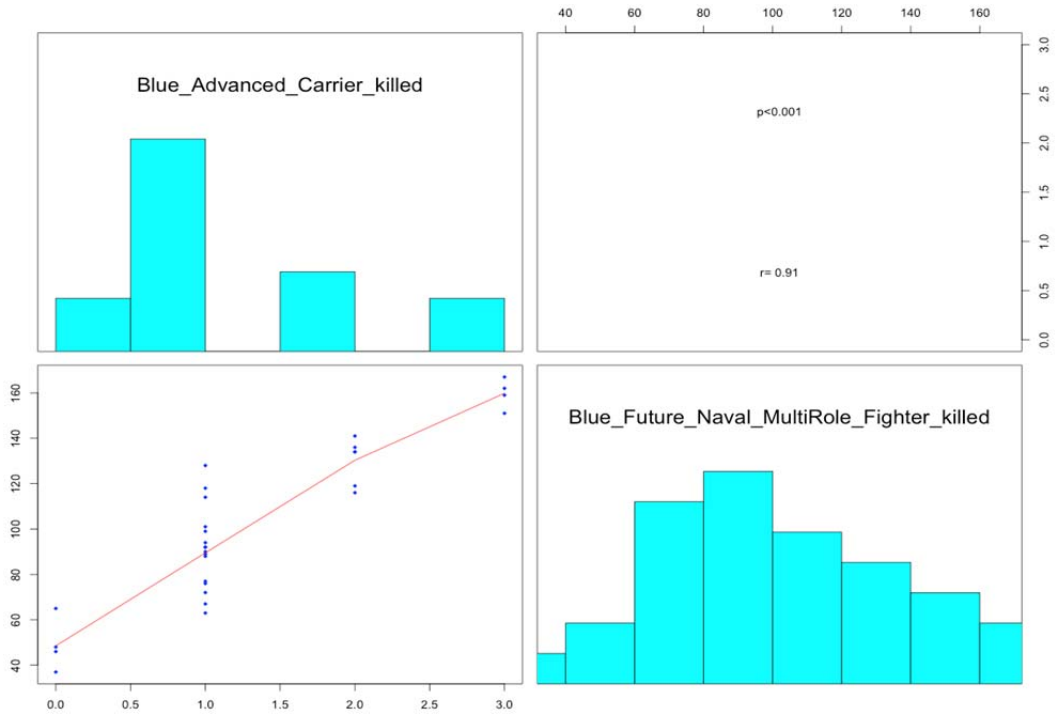


Figure 8. Pairs plot of blue force carriers killed and BFNMF killed

Although the data for the number of BFNMF killed appears relatively normal, PCC does not require normality. Two additional correlation tests are used in order to verify the robustness of the PCC previously generated. Both Spearman's rho (r_s) and Kendall's tau (τ) are non-parametric statistics that can be used to quantify the association between two variables (Field et al., 2012). Spearman's test first ranks each variable in the data and then applies Pearson's equation to each rank. Conversely, Kendall's test is commonly used when the data set is small and contains a large number of ranks of similar value. Each test was performed on the blue carriers and BFNMF pair of variables; see Table 6.

Correlation Test / Relationship	Pearson's Correlation Coefficient	Spearman's rho Test (non-parametric)	Kendall's tau Test (non-parametric)
Blue Carriers Killed & BFNMF Killed	0.9123	0.8938	0.7928
95% Confidence Interval	[.8224, .9578]	N/A	N/A
P-Value	2.24E-12	2.90E-11	4.03E-08

Table 6. Results for PCC, Spearman's, and Kendall's correlation test

The Spearman's rho and Kendall's tau tests in R-Studio do not generate confidence intervals. Perhaps by using a bootstrap method this range can be identified. However, given the very low *p-values* for each test, bootstrapping was not necessary. Although coefficient results decreased roughly 13% from the PCC test to the Kendall test, the relationship between blue carriers and BFNMF killed is still highly statistically significant, with the *p-values* all being less than .0000001.

E. BUILDING THE DESIGN IN STORM

Following the identification of significant relationships within the output data set with regards to blue carriers, the next step is to build an actual DOE that varies input variables. However, two additional tasks must be performed beforehand: First, identify the exact number of input files that are associated with the chosen metric; and Second, the analyst must decide on what metric aspect to modify. For this research, the BFNMF is specifically referenced in one input file, *typeaa.dat* (type air asset) (see Figure 9).

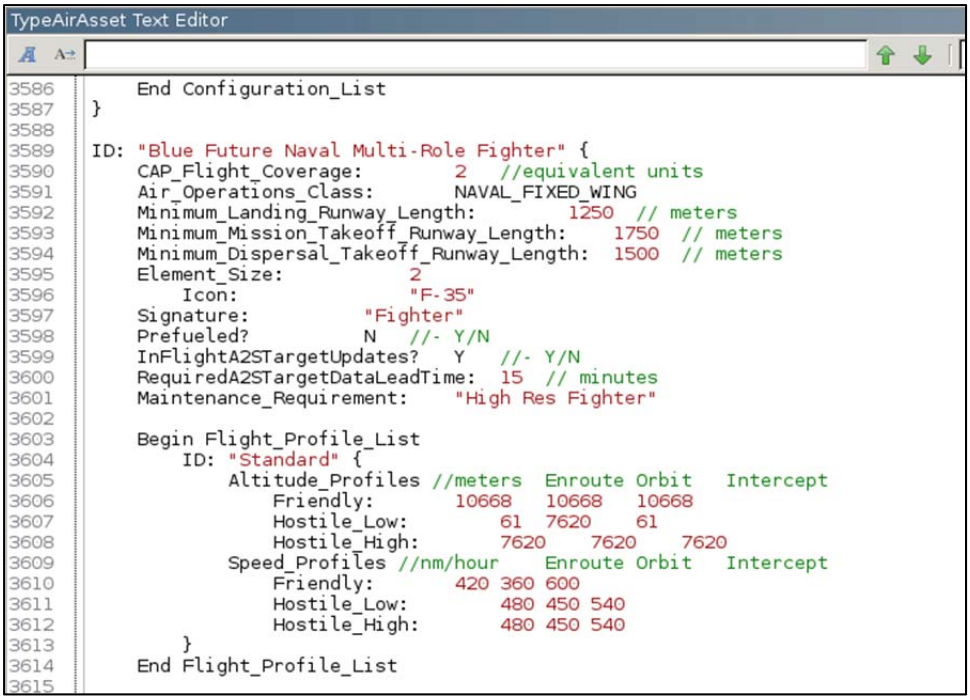


Figure 9. Snapshot of the *typeaa.dat* file in STORM Front

Identifying the number of input files that reference the BFNMF is essential because each file requires modification. Finding significant metrics that are in relatively few input files is key to successfully implementing a DOE in STORM. The second task, identifying which metric to modify for the design, should only be performed after careful considerations and discussions with various experts who have extensive knowledge pertaining to that metric. After collaborating with various pilots, it was determined that the speed profiles of the BFNMF were believed more vital to accomplishing a mission than operating at a higher altitude or making changes to its weapons load-out. Therefore, the intercept speed attributes were implemented into a design. Table 7 below is the design matrix that was built and implemented with STORM. There are three variables that are explored: friendly intercept speed (nautical miles per hour); hostile intercept low speed; and hostile intercept high speed. Each variable is assigned two levels, for a total of eight design points. Each design point includes a combination of variable levels and a specified response (R).

Factor Combination (Design Point)	Friendly Spd (NM/HR)	Hostile Low Spd (NM/HR)	Hostile High Spd (NM/HR)	Response (Blue Carriers Lost)
1	540	486	486	*R1*
2	660	486	486	*R2*
3	540	594	486	*R3*
4	660	594	486	*R4*
5	540	486	594	*R5*
6	660	486	594	*R6*
7	540	594	594	*R7*
8	660	594	594	*R8*

Table 7. Table of the 2^3 factorial design matrix that modifies the intercept speed levels associated with the BFNMF

The response in Table 7 indicates a focus on blue carriers lost. However, responses analyzed in Chapter IV are not limited to this single metric. Additional WITTW metrics are included, such as additional naval platforms (e.g., submarines, amphibious ships), and studied in order to determine if there are any significant changes to the PUNIC 21 scenario output and, more specifically, if intercept speed profiles for the BFNMF are in fact significant.

For this research, each speed profile, or factor level was set to +/- 10% of the base value, see Table 4. Modifications of +20% would have pushed the aircraft's speed past its physical capability. Therefore, in an effort to make the change more "real-world," levels were set to the values indicated in Table 7.

F. CREATION OF INPUT FILES

After identifying specific factors to explore and building a 2^k factorial design, the next step is to generate separate input files for each design point. To start, the analyst must create separate "design point" study folders in STORM, as seen in Figure 10. These folders represent individual study directories that house all relevant input data for the PUNIC 21 scenario. Additionally, all output data files associated with each design point run will be located in the data warehouse specific to each folder. This allows the analyst to examine each design output separately from the others. For further DOE analysis,

analysts must currently concatenate output files manually. Further information on how to accomplish this step is provided in Chapter IV. Eventually, to be practical, this needs to be automated as much as possible.

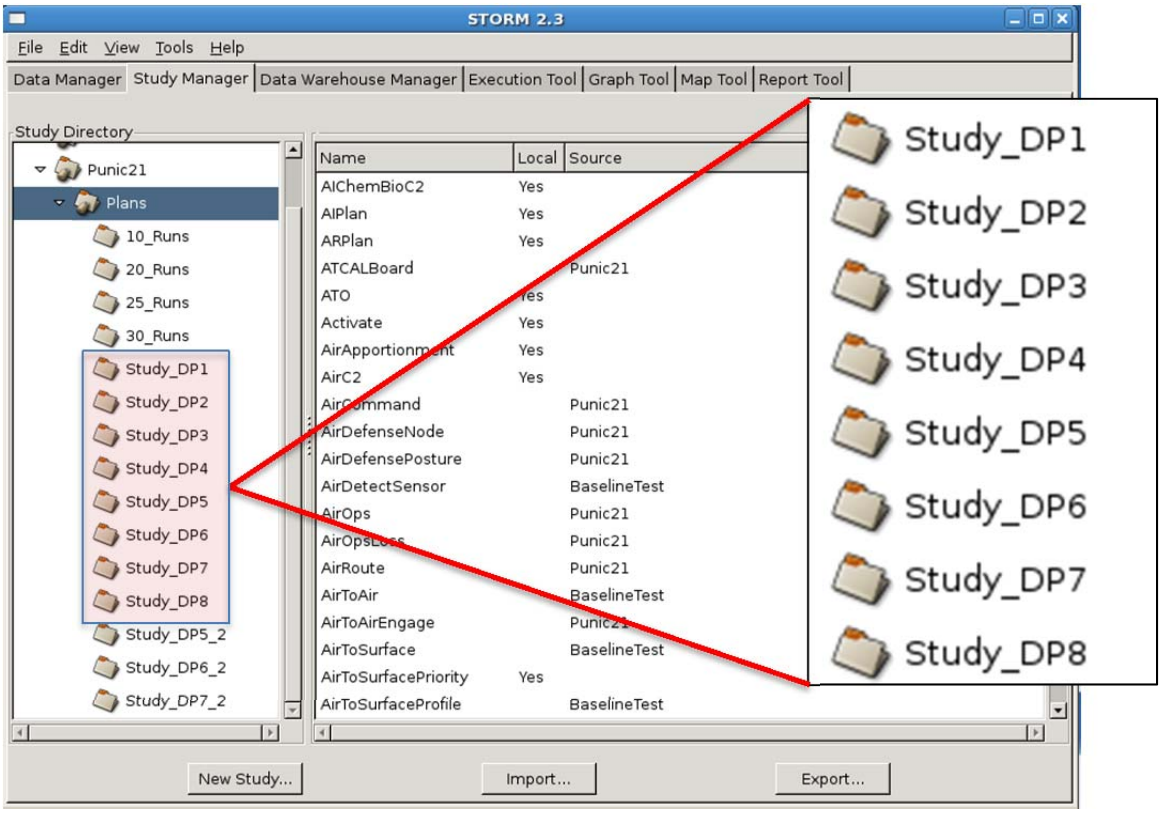


Figure 10. Snapshot of STORM GUI and how it should look after setting up eight separate design point study directories

Once each study folder is set up, the next step is to create custom input files that contain all information relevant to the design. For this research, a program written in the SCALA programming language is used in order to generate each file. The programming code is used to read in a templated version of each input file. Each file template contains the location pertaining to each speed profile we are looking to modify. Placeholders are then generated that match where the individual factor level settings would be. The design matrix from Table 7 is then read in concurrently with the template file, which generates a custom *typeaa.dat* file for all eight design points. This process is then repeated for two additional input files that are needed that pertain to command & control (*sideC2.dat*) and

interactions (*transaction.dat*) files. These files ensure that the custom input file appropriately interacts within STORM's logistical framework when a scenario is run.

After generating each file, it is placed in its respective study directory. That is, the *typeaa.dat* file for design point one must be placed in the first study directory, along with *sideC2.dat* and *transaction.dat* files, which is found in the STORM home folder. After this occurs the analyst should ensure each of the three files is turned on as a "local" file. This step guarantees that when a simulation is run, STORM reads in the custom files and not the default files that contain the original input parameters (see Figure 11). If successfully made local, the word "Yes" will appear next to that specific input file.

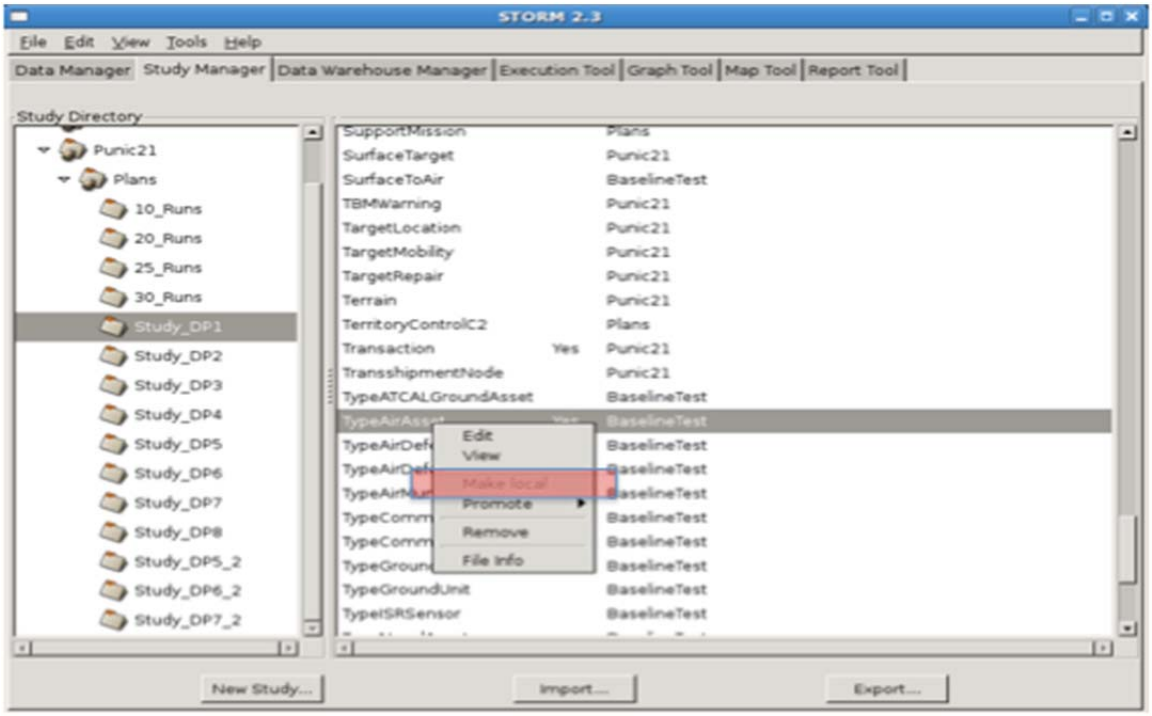


Figure 11. Snapshot of STORM GUI and how to put all three custom files into a local setting. In order to get to this option, right click on *typeaa.dat, sideC2.dat,* and *transaction.dat* files and select "Make Local"

G. RUNNING A SIMULATION WITH CUSTOM INPUT FILES

The last step in implementing a DOE in STORM is running each design point simulation individually over a number of replications and collecting the output data.

Setting the number of replications is at the analyst's discretion, but they must maintain cognizance of the available base memory of their operating system. For this research, 25 PUNIC 21 replications of each design point were performed through a virtual machine (VM) called Oracle VirtualBox. The VM is a Red Hat 64-bit operating environment with a Mac OSX (16 GB, 1333 MHz DDR3 Memory) host system. In order to have sufficient space, the VM's base memory was set to 9,046 MB prior to performing a simulation run. However, only 25 replications could be performed given certain limitations on the host operating system.

1. STORMMiner Software and Data Collection

Once each PUNIC 21 scenario is successfully run for each of the eight design points, output data files are collected from each study directory for analysis and passed through a software program called STORMMiner. This newly developed program is designed to parse the output database in order to quickly obtain specific metrics (WITTW) that are important to N81. This is accomplished through multiple MySQL queries, which identify those metrics and dump them into an Excel file for analysis.

MySQL is an open source SQL (Structured Query Language) database server. MySQL allows a program or user to store, manipulate, and retrieve data in table structures. This server was chosen because it was much faster than other open-source relational databases at the time when STORM was first developed. Since there are large amounts of output data associated with STORM, MySQL's performance efficiency was necessary in order to pull out the WITTW campaign metrics (Group W, 2012b). The data tables that are generated from STORMMiner enable N81 to narrow the focus of their post-processing analysis efforts, thereby significantly reducing turn-around times, and allowing for further exploration of the output data. This thesis is a test platform for STORMMiner, which was used extensively in order to analyze the effect of modifying each of the three speed profiles. Without its capabilities, this research would not be possible.

IV. ANALYSIS OF DESIGN POINTS

This chapter focuses on analyzing the output data of each design point as it pertains to four WITTW campaign metrics, which are treated as separate response variables. This analysis uses an unclassified data set and is intended primarily to illustrate the possibilities and potential of using DOE with STORM. As discussed in the previous chapter, only 25 replications were performed per design point due to host machine limitations. This resulted in 200 total runs. Output data was then directly extracted from each study directory, run through STORMMiner, and exported to an Excel file to be analyzed. Three software programs were used to conduct analysis on each metric; Excel, R, and JMP.

The following list of responses are analyzed in this chapter:

- Blue force losses—specifically carrier and BFNMF losses
- The time at which blue forces achieve air supremacy.
- The number of the red force's advance surface-to-air missile (SAM) sites destroyed.

Summary statistics are analyzed for each metric and comparisons are made over each design point in order to illustrate the variability in responses due to changing the BFNMF intercept speed profiles. Statistics are then tested using analysis of variance (ANOVA) and Tukey's honest significant difference (HSD) tests in order to determine whether there is in fact a statistically significant difference between the design points. Further analysis is conducted to determine the significance of each modified speed profile—using regression analysis and partition tree models, determine which, if any, speed profiles have a significant effect on the response. The ultimate goal, however, is to show how impactful a DOE can be when introduced into STORM and why N81 analysts should incorporate these methods for scenario analysis.

A. RANDOM NUMBER GENERATION IN STORM

Prior to analyzing design point output data, it is important to understand the use of STORM's random number generation as it pertains to scenario replication. As discussed in Chapter II, an important benefit of STORM is its stochastic nature. The stochasticity produces variability in output data, which provides N81 analysts with a range of uncertainty or risk—as is often found in combat situations. This is accomplished through a random number seed that determines the specific sequence of random numbers that are used within the model. Within the STORM GUI, analysts are able to set their own random number seed for a set of replications. If the same random number seeds are chosen for multiple sets of replications, then the output data for each set will be identical. For example, Table 8 is a set of four separate runs produced in RStudio. The top function, *rnorm(1)*, generates one random number from the normal distribution using default (randomized) seeding. Over four runs, the number produced changes each time. The runs in the lower row use identical seeding, and produce identical values.

R-Code	Run 1	Run 2	Run 3	Run 4
`rnorm(1)`	0.1836	-0.8356	1.5953	0.3295
`set.seed(1); rnorm (1)`	-0.6264	-0.6264	-0.6264	-0.6264

Table 8. Table showing random number generation in R-Studio for the normal distribution with default and explicit random number seeding

STORM determines the random number seed for a replication by computing a function of replication number (r), error retry number (n), and stream (s)−or $f(r,n,s)$. The stream parameter allows the user to change the third argument for a set of replications (W, STORM User's Manual Verson 2.3, 2012). In order to identify whether changing the BFNMF speed profiles make a difference in scenario output, the same random number seed was chosen for each set of replications.

B. BLUE FORCE LOSSES–CARRIER

One of the most important output metrics to examine in this campaign is the surface ship force level prior to and at the conclusion of a simulation. In terms of specific, theater-level area of responsibility (AOR), the proportion of force degradation provides insights to initial force composition and the size required to achieve success. For N81, an HVU, such as a carrier, is the most important surface ship metric focused on during post-process analysis. Figure 12 is a bar-plot of the average carrier losses over each design point. Additionally, each light blue colored bar contains a vertical black line that represents the 95% confidence interval for that specific set of runs. For comparison, the dark-blue horizontal dashed line represents the average number of carrier losses seen in the baseline scenario of PUNIC 21 over 30 replications. An examination of the plot reveals a moderate amount of variability in the carrier data as we move across each design point. Design points one, three, and six are all below the baseline average. Visually, design point four has the highest average number of blue carrier losses. Recall Table 7 from Chapter III, which indicated design point four as having higher friendly and hostile-low intercept speeds, but lower hostile-high intercept speeds. Summary statistics for this metric are found in Table 9.

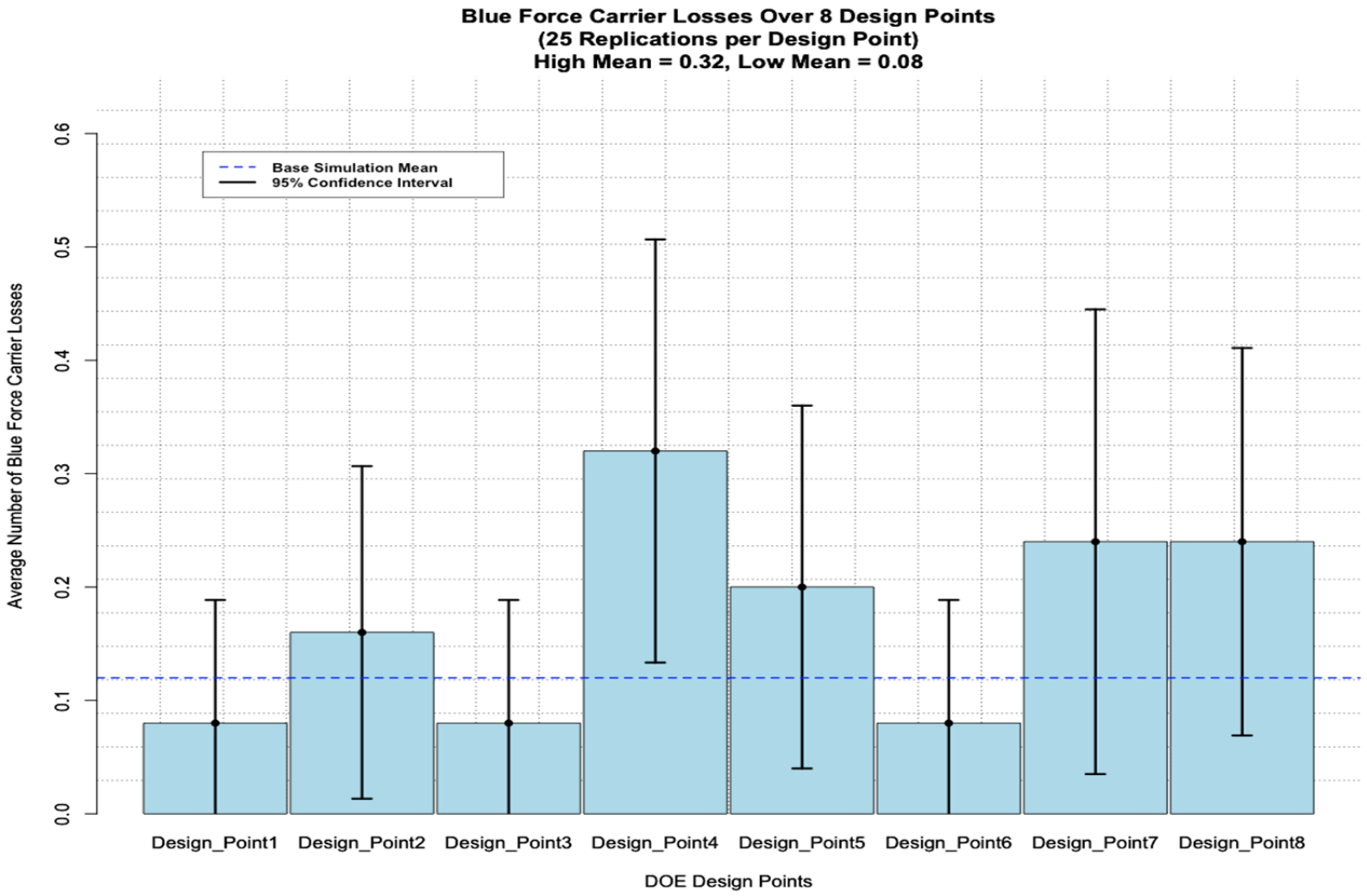


Figure 12. Plot showing the average number of blue carriers with a 95% CI over each design point and base simulation mean

Statistic	DP 1	DP 2	DP 3	DP 4	DP 5	DP 6	DP 7	DP 8
Mean	0.08	0.16	0.08	0.32	0.2	0.08	0.24	0.24
Standard Deviation	0.28	0.37	0.27	0.48	0.41	0.26	0.52	0.44
95% Confidence Interval	[0,0.19]	[0.005,0.31]	[0,0.19]	[0.12,0.52]	[0.03,0.37]	[0,0.18]	[0.02,0.46]	[0.06,0.42]
Min	0	0	0	0	0	0	0	0
Max	1	1	1	1	1	1	2	1

Table 9. Table providing the summary statistics for blue force carrier losses for each design point

Across all design points, the maximum number of blue carriers killed before the PUNIC 21 scenario terminates is at least one-third of their entire carrier force inventory. The exception is design point seven, where roughly 67% are destroyed in at least one replication. Given that this design point exhibits such a wide confidence interval and the largest standard deviation, seeing such a rare event is possible. An additional rare case is design point four, where only one carrier was killed in any one set of replications, but that event occurred most frequently at 32% of the time.

Given each design point run was performed using the same random number seed, it can be concluded that the variation over each set of replications is either due to modifying the BFNMF intercept speeds or, which is more likely the case, by the ordering at which certain events occur during the scenario. Although seemingly insignificant in terms of changes in average carrier losses across each design point because of the minuscule differences among the means, blue forces only possesses three carriers in the PUNIC 21 scenario. Therefore, small changes could have a relatively large impact on the overall campaign. Furthermore, included in the majority of input files that pertain to blue force platform specifics are prioritization criteria for protecting such HVU platforms. Since carriers are the pillar for power projection, it has a protection priority class of "1." This means that losing a carrier is a rare event and should not often occur. As such, an analyst may want to either further investigate the details surrounding both design points four and seven to get a better understanding of why there was such a large proportion of carrier losses, or take more replications and compare the output.

C. BLUE FORCE LOSSES–BFNMF

Since variables chosen for the DOE pertained specifically to the BFNMF, it makes sense to examine the total losses this air platform suffers in the PUNIC 21 scenario. Figure 13 is a histogram of the first four design points illustrating the frequency at which red forces kill the BFNMF during the 20 days of simulated battle.

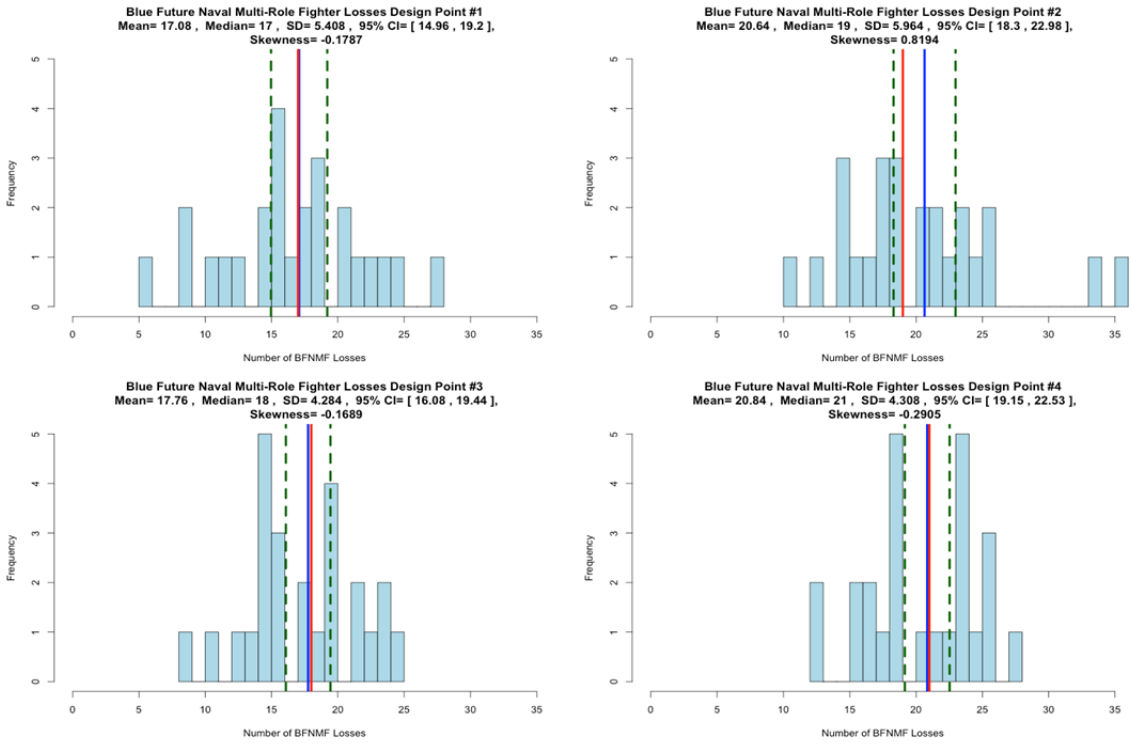


Figure 13. The number of BFNMN losses over design points one through four

In each plot, the mean, median, standard deviation, 95% confidence interval, and skewness are included under the plot title. Skewness, which is a measure of asymmetry, is equal to zero for a symmetric distribution such as the normal (Law, 2007). In Figure 13, the solid dark blue vertical line indicates the mean, the solid red vertical line represents the median, and the dashed dark green lines signify the lower and upper confidence interval limits on the mean.

Compared to blue force carriers, there is much more variability in this output because speed changes directly affect the BFNMF and its capabilities. Giving the BFNMF a higher speed profile may directly correlate to a lower number of killed platforms. The lowest mean from the first four design points is 17.08 (design point one). With the exception of design point two, which exhibits a slight right tail distribution, all means are relatively close to the median and are distributed somewhat equally on either side. Recall from Chapter III that each combination of intercept speeds had a lower hostile-high speed profile set to 486 nm/hr and resulted in two design points with means

exceeding 20. This is important to an analyst because it clearly identifies a threshold speed below which the BFNMF should not drop when intercepting a hostile. If that speed should exceed the current physical capability of an assigned airframe, it would be beneficial to consider an aircraft that has the ability to intercept enemy forces at higher speeds.

Figure 14 includes the same type of plots, but illustrating design points five through eight. Again, there is above average variation in this response, but contrary to the first four plots each group of factor settings included higher hostile-high intercept speeds set at 594 nm/hr. This resulted in 16.92 BFNMF's killed (design point six), which was the lowest mean out of all eight design points, but also exhibits a wide confidence interval. The highest mean, 18.28, was seen in design point five, which also had the smallest difference in median and mean.

Despite only a 12.6% decrease in the average number of BFNMF lost from design point four to design point five, its significance should not be overlooked considering it could mean the difference between mission failure or success. Moreover, as the number of aircraft losses goes up, the blue forces' ability to gain air supremacy (discussed in the following section) goes down. Therefore, insights gained by Figures 13 and 14 provide analysts with invaluable knowledge that they would otherwise not have by running a scenario with a single set of input metrics.

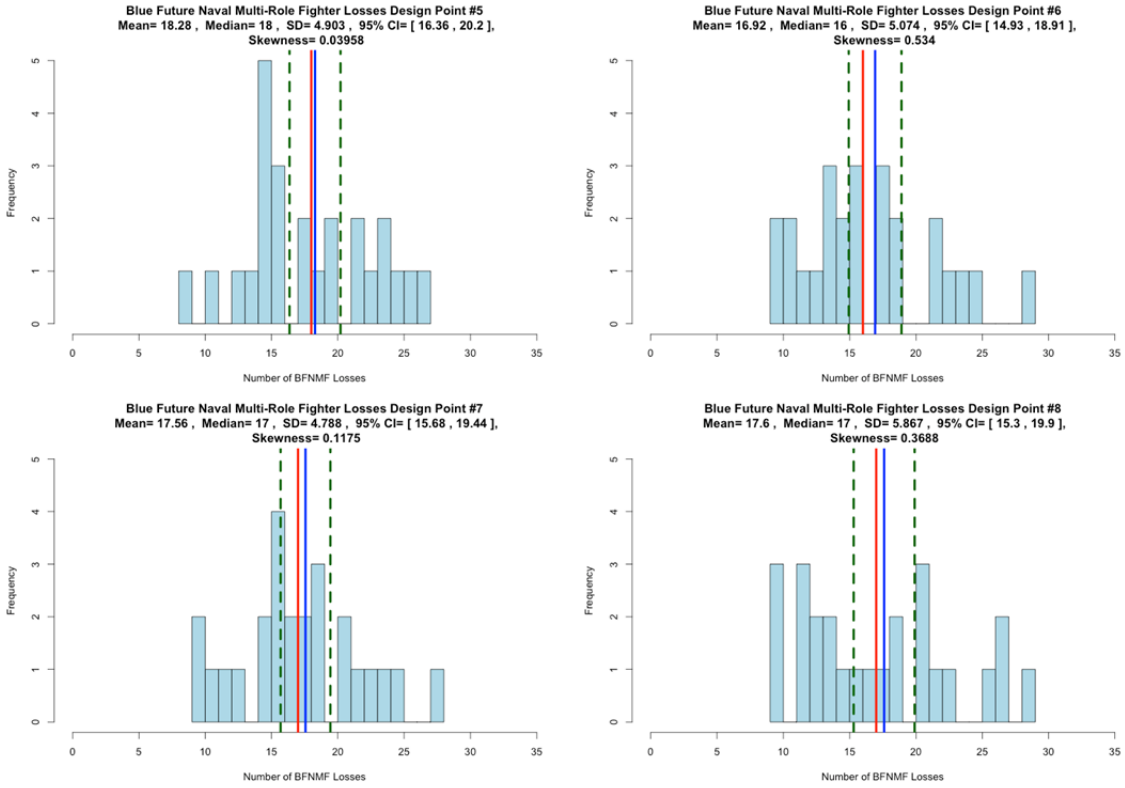


Figure 14. The number of BFNMN losses over design points five through eight

D. TIME AT WHICH BLUE FORCES ACHIEVE AIR SUPREMACY

This section examines whether the blue forces achieve air supremacy and, if so, the time at which it occurs. For a senior-level decision maker, this particular WITTW campaign metric is of significant interest because many military operations require control over the AOR air space. If air control is not obtained, it may severely limit the choices combatant commanders have in order to achieve a strategic objective. In the first 20 days of the PUNIC 21 scenario, blue forces achieved air supremacy ranging from a minimum of 64% of the time (design points two, three, and five) to a maximum of 100% of the time out of 25 replications (design points seven and eight); see Table 10. Design point four had the second largest proportion at 88%. Overall, the average number of times air supremacy was achieved is 20.13, which is roughly 81% of the time. The highest frequency of supremacy occurred most often towards the simulation's terminating point. Figures 15 and 16 are cumulative plots of the time air supremacy is gained across design points one through eight.

The output data (from STORMMiner) used for the generation of each plot breaks each day into quarter segments (e.g., 0.25, 0.50, 0.75) and identifies the specific time during PUNIC 21 when air supremacy was achieved by placing a "1" next to that value and a "0" next to all other times for when supremacy was not achieved. For example, in the first replication of design point one, air supremacy was gained at day 19.75, but not achieved for the second replication. Therefore, their values would be 1 and 0, respectively. Additionally, it is important to note that once air supremacy is gained in PUNCI 21, it is maintained; meaning all follow-on values are "1", which is represented graphically by the light blue bars reaching an apex towards the end of the scenario. Of course, it must be determined whether the difference in design points is simply the result of random variation.

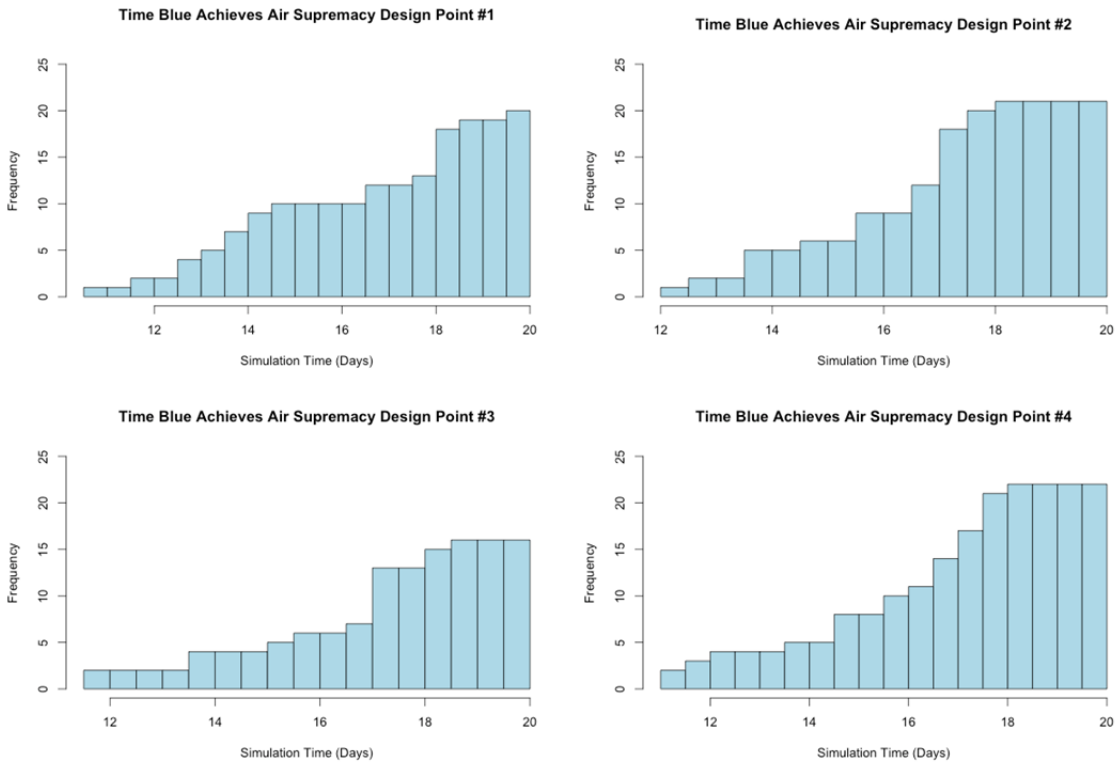


Figure 15. The time in which blue forces achieve air supremacy over design points one through four

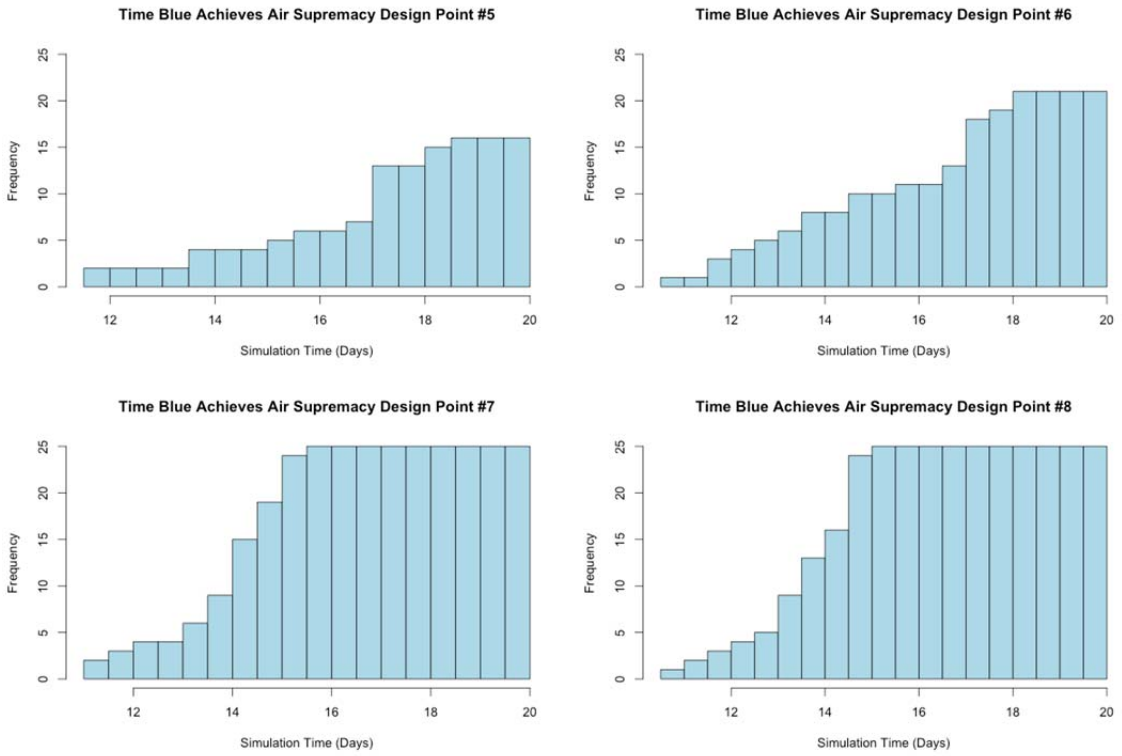


Figure 16. The time in which blue forces achieve air supremacy over design points five through eight

	Design Point 1	Design Point 2	Design Point 3	Design Point 4	Design Point 5	Design Point 6	Design Point 7	Design Point 8
Proportion Of Blue Air Supremacy Over 25 Replications (%)	0.8	0.64	0.64	0.88	0.64	0.84	1	1

Table 10. Table showing the proportions of the number of replications blue forces achieved air supremacy out of 25 replications

It is critical for an analyst to identify the total time in which it takes to gain control over the AOR air space because the longer this objective is left unaccomplished the more opportunities red forces have to diminish blue force levels (i.e., kill more ships, submarines, and aircraft). Therefore, the ultimate objective is to reduce the overall mean number of days to alleviate unnecessary blue force losses. Figures 15 and 16, along with Table 10, provide an excellent representation of how impactful the DOE (i.e., altering the BFNMF speed profiles) is, especially design points seven and eight. Both not only exhibit blue forces achieving air supremacy for all 25 replications, but they are doing so earlier on in the PUNIC 21 scenario.

E. RED FORCE SAM SITES DESTROYED

As discussed in the previous section, gaining air supremacy is an integral part of successful military operations. Directly correlated to this metric, however, is the important objective of destroying an enemy's SAM sites. N81 is particularly interested in this metric because it often accounts for a large portion of the aircraft sorties flown in a combat situation. A greater number of sites destroyed directly enhances the friendly forces' ability to achieve air supremacy and ultimately lower the inherent risk to blue force aircraft.

Figure 17 is a barplot (similar to Figure 12) of the average number of destroyed red force SAM sites over the eight design points. Each red bar contains a vertical black line that represents the 95% confidence interval that reveals a narrow window of uncertainty for each set of replications. As in the BFNMF losses and blue air supremacy case, the benefit of higher intercept speed capability can be seen across all design points. This is particularly noticeable in design points six and eight, which account for the highest mean number of SAM sites destroyed (29.20), while design point one accounts for the lowest (23.08).

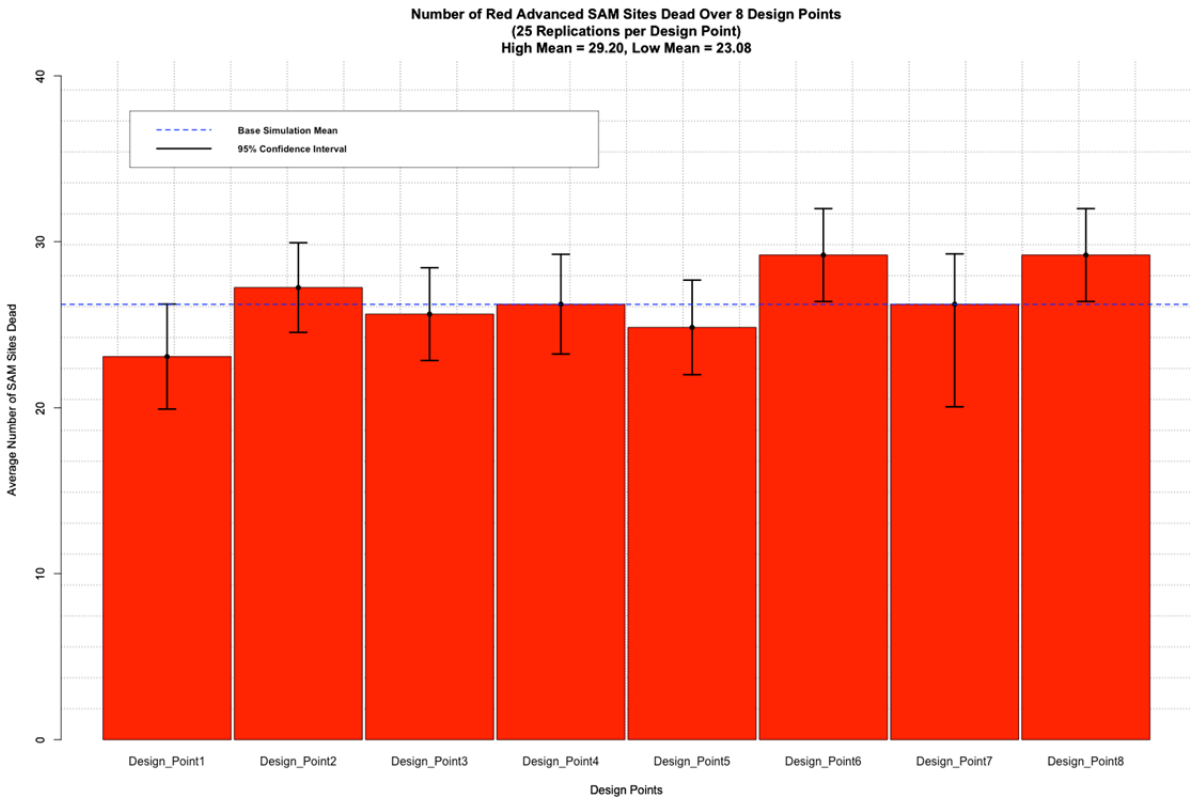


Figure 17. Plot showing the average number of red force's SAM sites destroyed with a 95% confidence interval over eight design points and baseline simulation mean

Recall that design points one through four have BFNMF hostile-high intercept speeds set to 486 nm/hr and five through eight is set to 594 nm/hr, therefore, the latter design points should be exhibiting an overall greater number of average SAM sites destroyed. In this particular case, the exception is design point two. Even with lower hostile-high intercept speeds, blue forces were still able to take out on average 15.3% more sites than design point one. This number could be a false representation of the true value, considering only 25 replications were taken, and may reflect the inherent variability associated with combat. If more runs of PUNIC 21 were performed, that average number may possibly decrease. Either way, this is a specific case analysts may want to examine further in order to try and identify additional events that took place throughout that specific set of runs which might have caused a higher number of sites to be destroyed.

Table 11 displays the summary statistics of red force SAM sites destroyed over all eight design points. As previously noted in this section, the precisions in the estimated means are moderately consistent across all design points.

Statistic	DP 1	DP 2	DP 3	DP 4	DP 5	DP 6	DP 7	DP 8
Mean	23.08	27.24	25.64	26.24	24.84	29.2	26.24	29.2
Median	24	29	27	28	27	30	27	30
Standard Deviation	8.07	6.89	7.13	7.67	7.3	7.14	7.73	7.13
95% Confidence	[19.92,26.24]	[24.54,29.94]	[22.85,28.43]	[23.24,29.25]	[21.99,27.70]	[26.41,32]	[20.01,29.23]	[26.3,31.8]
Min	0	5	0	5	1	0	0	0
Max	35	36	35	36	35	36	35	36
Skewness	-1.14	-1.94	-2.34	-1.6	-1.96	-2.78	-1.95	-2.7

Table 11. Table providing the summary statistics for red SAM sites destroyed metric

F. STATISTICAL DIFFERENCES IN DESIGN POINTS

This section focuses on determining whether the design points are statistically different from each other in order to validate the resulting variability seen in design points for each response examined earlier in this chapter. For this analysis, ANOVA and Tukey's HSD tests are performed using the following hypotheses:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0 \text{ for at least one } i \ , \quad i = \{1,2,3,4,5,6,7,8\}$$

The null hypothesis (H_0) states that the means of all design points (β_i) are statistically equal to each other, whereas the alternative hypothesis (H_a), states that there is at least one design point that is statistically different from the rest. In hypothesis testing, the null hypothesis prevails unless sufficient evidence to the contrary is produced, in which case the null hypothesis is rejected in favor of the alternative hypothesis. If the ANOVA results suggest a rejection of the null hypothesis (*p-value* less than .05), Tukey's HSD test will be performed to analyze where specific differences may lie.

ANOVA is a statistical test used to analyze whether the means of a several groups are equal (i.e., our eight design points). For this case, a generalized linear regression model was fit using each response and all associated design points. The results shown in Table 12 indicate retention of the null hypothesis (*p-value* = .2348) for blue force carrier losses and the number of destroyed red force SAM sites (*p-value* = .058), which is highlighted in red. This means there is no statistical difference in any of the eight design points for those specific responses. Conversely, the results for BFNMF losses and the time blue forces achieve air supremacy both qualify for rejection of the null hypothesis with *p-values* of .0301 and less than .001, respectively (highlighted in green). Since both *p-values* are less than .05, at least one design point is statistically different from the rest. In the time to air supremacy case, however, this conclusion could be misleading, as it is a result of only accounting for those instances where supremacy was actually achieved. Further analysis to validate the ANOVA results for this specific case is performed later in this section. In order to determine where the specific differences lie, Tukey's HSD test results are presented in the form of a plot comparing each combination of design point

pairs. See Figures 18, 19, 20, and 21. Additionally, an alternative way of examining each pair to evaluate differences is accomplished by grouping each design point utilizing the *HSD.test()* function in R-Studio, which is found in the *agricolae* package (see Table 12).

Response: Carrier Losses	**Degrees of Freedom**	**Sum of Squares**	**F-Value**	**P-Value**
Design Point	7	1.435	1.337	0.2348
Residuals	192	29.44		
Response: BFNMF Losses	**Degrees of Freedom**	**Sum of Squares**	**F-Value**	**P-Value**
Design Point	7	416	2.276	0.0301
Residuals	192	5012.6		
Response: Air Supremacy	**Degrees of Freedom**	**Sum of Squares**	**F-Value**	**P-Value**
Design Point	7	152.78	5.238	2.24E-05
Residuals	153	637.5		
Response: Red SAM Sites	**Degrees of Freedom**	**Sum of Squares**	**F-Value**	**P-Value**
Design Point	7	761	1.993	0.058
Residuals	192	10475		

Table 12. Table showing the results of ANOVA test for four WITTW responses: blue carrier and BFNMF losses, time blue forces achieve air supremacy, and the number of red force SAM sites destroyed

Figures 18, 19, 20, and 21 are graphs of the 95% family-wise confidence levels for each response. Along the y-axis is a comparison of each combination of design points, and along the x-axis is the value that corresponds to the differences in mean levels of each response as it pertains to each design point combination. If a pair of design points are statistically similar the confidence interval includes zero, which is indicated by the red vertical line. A clear indication that design point pairs are different is when a particular confidence interval is either entirely to the left or right side of the red vertical line. For the average number of carrier losses and red SAM sites destroyed cases (Figures 18 and 21), Tukey's HSD test confirms the ANOVA results that there are no statistically different pairs of design points. Contrary to the ANOVA results, however, is the average number of BFNMF losses case. Despite a *p-value* that is less than .05,

Tukey's HSD results (Figure 19) show that all confidence intervals include zero. Borderline cases include design point pairs six-four and six-two.

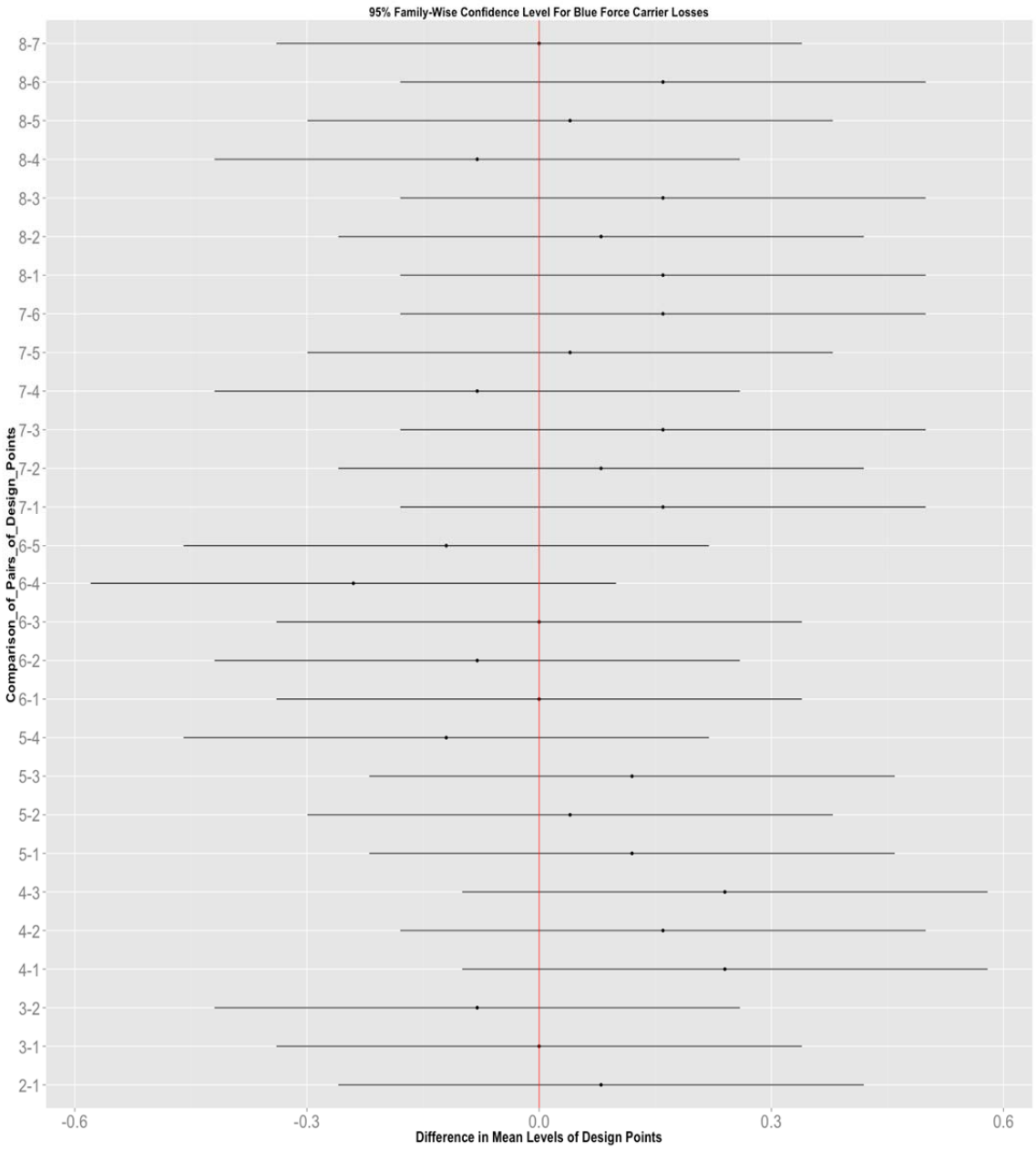


Figure 18. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the average number of blue carrier losses

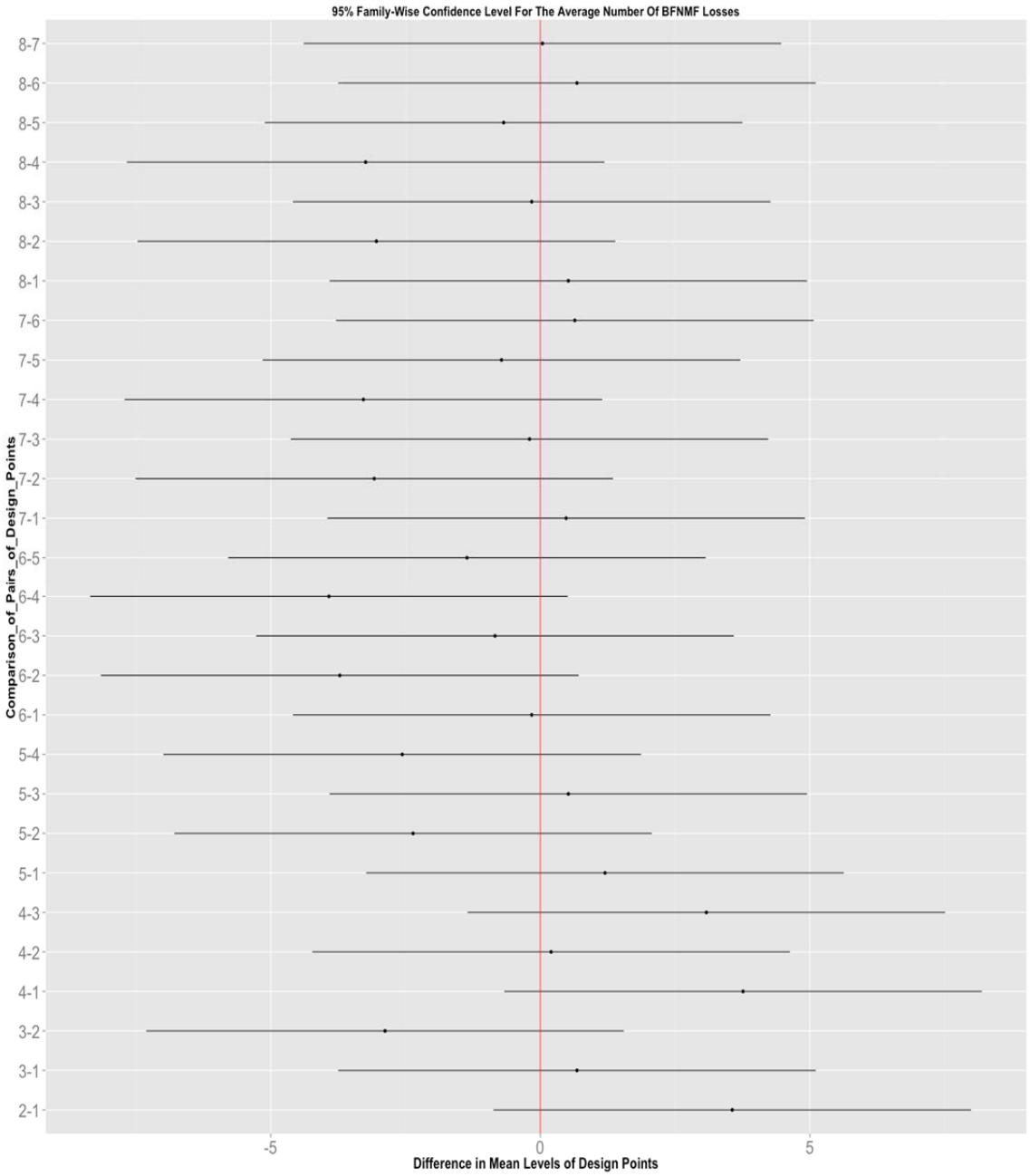


Figure 19. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the average number of BFNMF losses

Figure 20. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the time it takes blue forces to achieve air supremacy

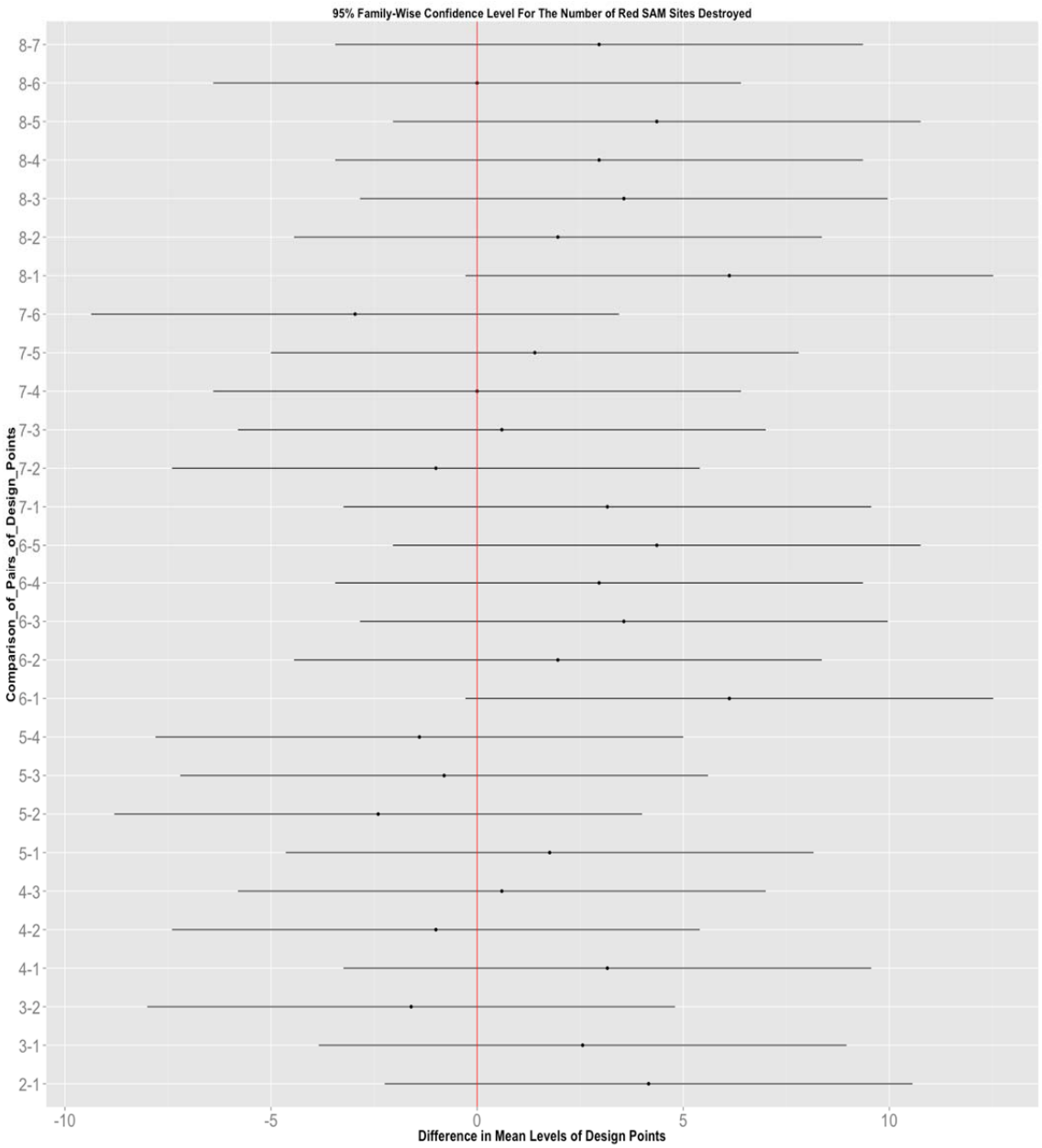


Figure 21. Graph of Tukey's HSD test results in the form of 95% confidence intervals for the average number of red SAM sites destroyed

The only unique case identified is the time it takes blue forces to achieve air supremacy. The ANOVA results concur with the graph generated by Tukey's HSD test (Figure 20). We can conclude that design point pairs seven-two, seven-three, seven-five, eight-one, eight-two, eight-three, eight-four, and eight-five, are statistically different from the remaining pairs. Therefore, according to the ANOVA and Tukey's HSD test, the only response that is statistically different in this case is the time it takes blue forces to achieve air supremacy—given that air supremacy is achieved.

Since the BFNMF losses case exhibited a discrepancy between ANOVA and Tukey's HSD test results, further analysis is conducted on all responses by examining the *HSD.test()* results. This test makes multiple comparisons of design points and provides each one with a grouping identification. The results in R-Studio provide summary statistics on each metric over each design point, as discussed earlier in this chapter, and output that is similar to Table 13. The results of this test confirm that the time at which blue forces achieve air supremacy is the only response that includes statistically different design points. The results have grouping identifications "a", "ab", "abc", "bc", and "c". This means that design points one and four are considered to be a single group ("ab"). Design points, six, seven, and eight are all unique design point groups. For the three remaining responses, all design points are considered to be a part of the same group, which is denoted by grouping identification "a".

Carrier Losses		
Groups	**Design Point**	**Means**
a	4	0.32
a	7	0.24
a	8	0.24
a	5	0.2
a	2	0.16
a	1	0.08
a	3	0.08
a	6	0.08
BFNMF Losses		
Groups	**Design Point**	**Means**
a	4	20.84
a	2	20.64
a	5	18.28
a	3	17.76
a	8	17.6
a	7	17.56
a	1	17.08
a	6	16.92
Air Supremacy		
Groups	**Design Point**	**Means**
a	2	16.56
a	3	16.09
a	5	16.09
ab	1	15.65
ab	4	15.61
abc	6	15.2
bc	7	14.03
c	8	13.71
Red SAM Sites		
Groups	**Design Point**	**Means**
a	4	29.2
a	7	29.2
a	8	27.24
a	5	26.24
a	2	26.24
a	1	25.64
a	3	24.84
a	6	23.08

Table 13. Table of *HSD.test()* results for each response that include grouping identifications, design points, and associated means

Although both tests indicate that there are statistical differences among design points for the time blue forces achieve air supremacy response, it is important to note that these results could be misleading, as mentioned earlier. Since only those replications where air supremacy was actually gained were used in the regression model, the results from the ANOVA and Tukey's HSD tests could be providing a false representation that design points are in fact unique—especially given that design points one through eight varied in the number of times blue forces achieved air supremacy out of 25 replications (see Table 10). For this reason, Pearson's chi-squared statistic test is generated in R-Studio to test if there is a difference in the proportions in which air supremacy is achieved within 20 days across the eight design points. The hypothesis test used in this case is:

H_0: Design Points are independent from the probability of achieving air supremacy

H_a: Design Points are not independent

In preparation for the test, a matrix is generated that contains the number of times the blue force achieves or does not achieve air supremacy in the 25 replications for each design point (see Table 14). The data are then compared using Pearson's chi-squared statistical test.

	Design Point 1	Design Point 2	Design Point 3	Design Point 4	Design Point 5	Design Point 6	Design Point 7	Design Point 8
Observed Time (in Days) to Air Supremacy	20	16	16	22	16	21	25	25
Difference From 25 Replications	5	9	9	3	9	4	0	0

Table 14. Matrix generated in R-Studio of the number of times the blue force achieves and does not achieve air supremacy over the 25 replications for design points one through eight

Since Figure 22 reveals a *p-value* that is significantly lower than .05, we reject the null hypothesis that states the design points are independent from the probability of achieving air supremacy. Therefore, we can ultimately conclude that the results show a statistical difference in design points as it pertains to this response. We conclude that BFNMF intercept speeds impact when air supremacy is achieved, and that faster intercept speeds are better—as one would expect.

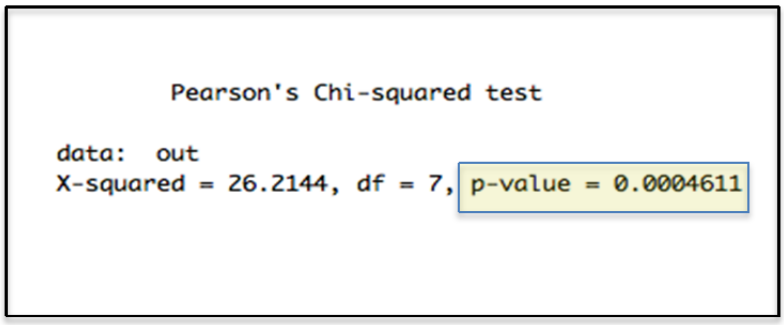


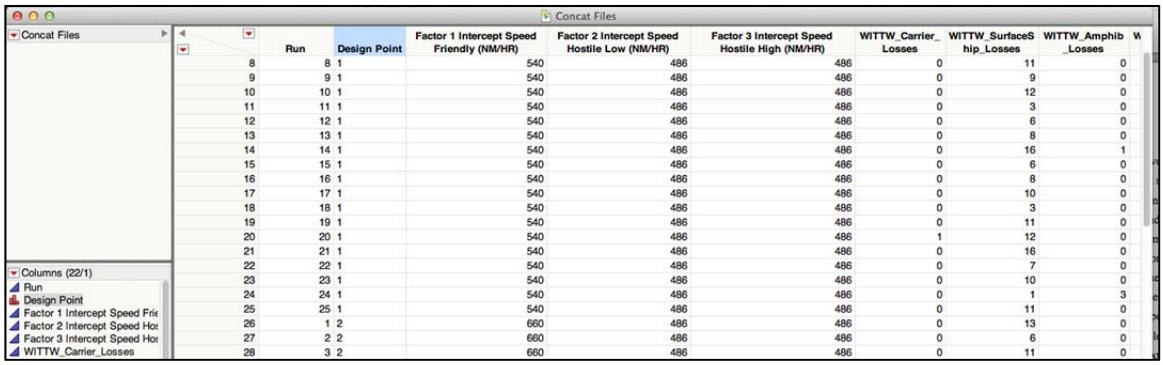
Figure 22. Results from Pearson's chi-squared statistic test on whether or not the blue forces achieve air supremacy before the simulation terminates

Although there are no significant statistical differences among design points as it pertains to blue carrier and BFNMF losses or red force SAM sites destroyed, this does not indicate that the implemented DOE was not successful. It emphasizes the point that small changes in a couple of factor levels (i.e., +/- 10%) may not have been sufficient to affect the specific campaign measures used for the PUNIC 21 scenario. If we varied more factors, or used wider ranges, we would have the potential to see greater impacts.

G. BFNMF SPEED PROFILE SENSITIVITY ANALYSIS

This section focuses on examining the output data over each design point as it pertains explicitly to each factor setting (i.e., changes in speed profiles of the BFNMF). Specifically, we seek to identify main effects and any interactions between factors by performing regression analysis to determine factor effects on a specified response. This process is known as factor screening or sensitivity analysis (Law, 2007). Data manipulation and preparation is key to the analysis performed in this section and must be completed initially on each design point data set. First, data files for design points one through eight are opened individually in R-Studio. Since each file contains significantly more variables than just the WITTW campaign metrics, each data set is narrowed by extracting 19 variables and placing them into a separate data frame. Once completed for all design points, the results are saved as an Excel document and imported into JMP. Once in JMP, its custom design application (under heading DOE) is used to incorporate the design matrix as it corresponded to each set of simulation runs (25 replications) and

the associated response output. The final step is to concatenate all eight design point files into a single data frame that can be utilized for study. Figure 23 is a snapshot of the generated data frame necessary to conduct analysis on each intercept speed variable.



Run	Design Point	Factor 1 Intercept Speed Friendly (NM/HR)	Factor 2 Intercept Speed Hostile Low (NM/HR)	Factor 3 Intercept Speed Hostile High (NM/HR)	WITTW_Carrier_ Losses	WITTW_SurfaceS hip_Losses	WITTW_Amphib _Losses	W
8	1	540	486	486	0	11	0	
9	1	540	486	486	0	9	0	
10	1	540	486	486	0	12	0	
11	1	540	486	486	0	3	0	
12	1	540	486	486	0	6	0	
13	1	540	486	486	0	8	0	
14	1	540	486	486	0	16	1	
15	1	540	486	486	0	6	0	
16	1	540	486	486	0	8	0	
17	1	540	486	486	0	10	0	
18	1	540	486	486	0	3	0	
19	1	540	486	486	0	11	0	
20	1	540	486	486	1	12	0	
21	1	540	486	486	0	16	0	
22	1	540	486	486	0	7	0	
23	1	540	486	486	0	10	0	
24	1	540	486	486	0	1	3	
25	1	540	486	486	0	11	0	
26	2	660	486	486	0	13	0	
27	2	660	486	486	0	6	0	
28	2	660	486	486	0	11	0	

Figure 23. Snapshot of JMP data file that includes BFNMF factor settings and output data associated with all eight design point

Once the concatenated file is set up correctly in JMP, as indicated by Figure 23, full-factorial linear regression models were estimated with each BFNMF intercept speed factor (friendly, hostile-low, and hostile-high) as it pertains to each of the four responses discussed throughout this section. A condensed version of the model results is shown in Figure 24 for all responses, which provides R-Squared (R^2), ANOVA, individual factor, and interaction *p-values*. Values other than R^2 are identified as either being significant (highlighted in green) or insignificant (highlighted in red). The first model uses blue carrier losses as the response and reveals that each factor exhibits neither individual significance nor two- or three-way interactions. The results are similar for the time it takes blue forces to achieve air supremacy and red force SAM sites destroyed, where only a three-way interaction and friendly intercept speed, respectively, are found to be significant. As previously noted, the values pertaining to the air supremacy response may be misleading because only those instances where supremacy was actually achieved are accounted for. However, in the model for BFNMF losses we find that friendly and hostile-high intercepts speeds, along with their two-way interaction, are statistically significant—additionally indicated by the ANOVA results (*p-value* = .03).

Source	Blue Carrier Losses	BFNMF Losses	Time To Blue Air Supremacy	Red SAM Sites Dead
R-Square	0.00825	0.0766	0.0564	0.0412
ANOVA *p-value*	0.9782	0.03	0.126	0.3176
Friendly Intercept Speed (*p-value)*	0.8207	0.0044	0.3038	0.0444
Hostile-Low Intercept Speed (p-value)	0.8207	0.0672	0.1179	1
Hostile-High Intercept Speed (p-value)	0.4968	0.0406	0.3064	0.1516
Friendly Intercept Speed * Hostile-Low Intercept Speed *(Interaction p-value)*	0.8207	0.7506	0.4711	1
Friendly Intercept Speed * Hostile-High Intercept Speed (Interaction p-value)	0.4968	0.0065	0.8597	0.1516
Hostile-Low Intercept Speed * Hostile-High Intercept Speed (Interaction p-value)	0.8207	0.7506	0.2934	1
Friendly Intercept Speed * Hostile-Low Intercept Speed * Hostile-High Intercept Speed (Interaction p-value)	0.4968	0.5162	0.0231	1

Table 15. Summary *p-value* and R2 results for the JMP generated full factorial linear regression models

The partition tree method gives an additional way of examining and visualizing these results. JMP provides an interactive tool for partitioning which factor groupings affect each response metric. Each partition that is made improves coverage of the data and reveals the factor levels that correspond to a specific response value. For more complicated designs, the partition tree method provides a unique way for an analyst to examine output data and is extremely easy to follow. Figure 24 is an example of the partition tree method used for the number of BFNMF losses response. According to the partition tree, the lowest number of losses occur when hostile-high and friendly intercept speeds are greater than or equal to 594 nm/hr and 660 nm/hr respectively, and hostile-low intercept speeds are strictly less than 594 nm/hr. These specific factor settings result in a mean loss of only 16.92 BFNMF's (highlighted in red).

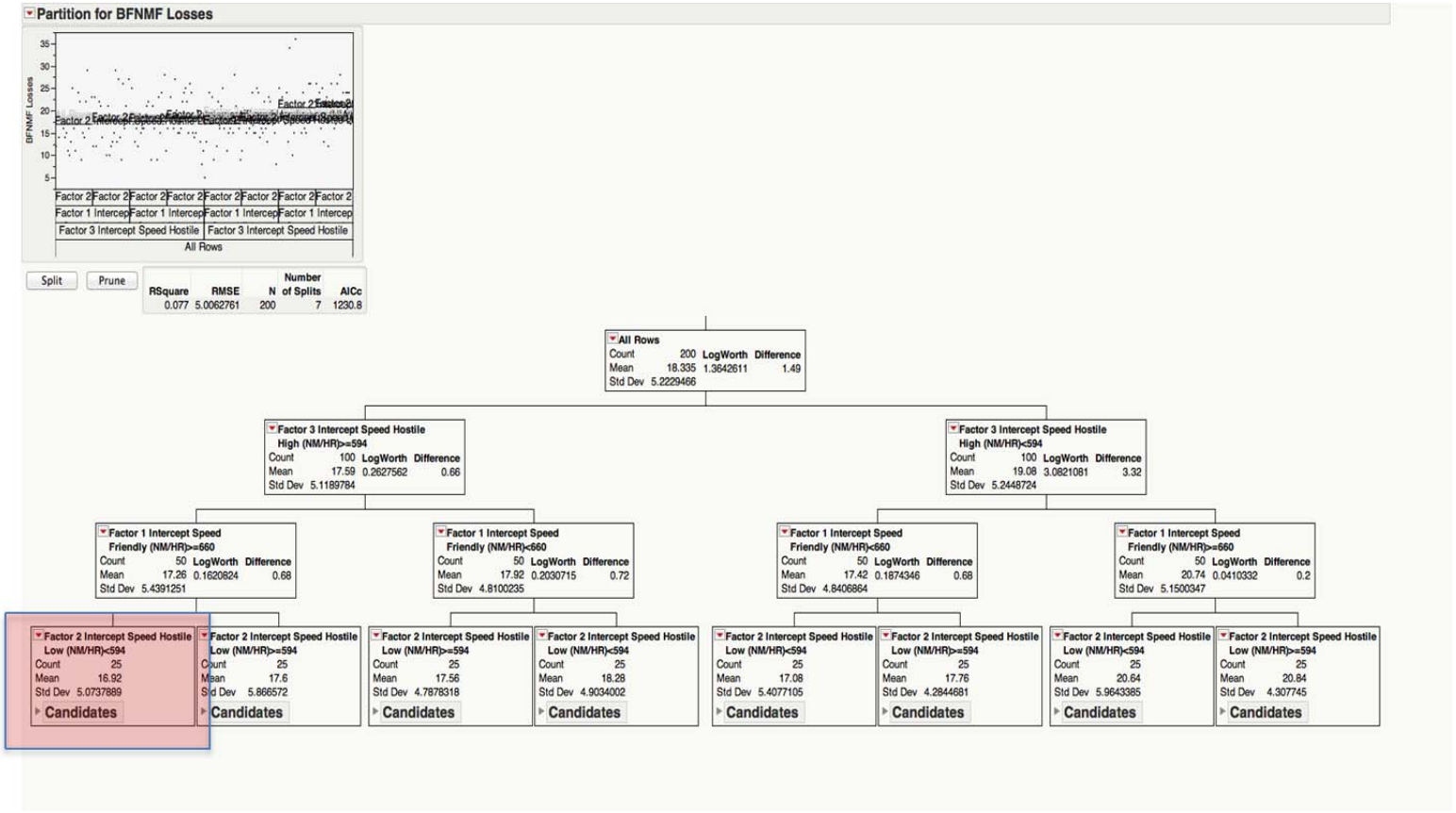


Figure 24. Partition tree method in JMP for the number of BFNMF losses. The left-most branch corresponds to the lowest number of losses (highlighted in red)

V. CONCLUSION AND RECOMMENDATIONS

This research is a first step in a long journey that will ultimately enhance the overall analysis capabilities of STORM and provide N81 analysts with the capability to enhance their level of knowledge and an ability to provide leadership with quick turnaround analysis. Although the results from Chapter IV indicate that the BFNMF factors are only moderately significant for BFNMF losses in this proof-of-concept analysis, it can be concluded that a small design of experiments, like the one generated in Chapter III, can be implemented into STORM if the four-step methodology is followed and limitations are carefully considered. In addition, it is important for the user to have a broad understanding of STORM, as discussed in Chapter II, before an experiment is attempted—it is an extremely complicated simulation environment. As a pilot study, this thesis provides the foundation for future researchers to further explore either automating the approach, or applying it to a real-world classified scenario.

A. DOE DESIGN AND METHODOLOGY

Chapter III discusses the methodology behind choosing significant factors from a baseline output data set, incorporating those factors into a 2^k factorial design, and implementing that design into STORM. Initially, it is important for the user to become familiar with STORM, its GUI, where to locate input and output data files, and the various tools that can be utilized for analysis. Additionally, understanding and adhering to the critical limitations section is extremely important for successful DOE implementation. Choosing a specific factor that is referenced multiple times throughout STORM's input data files requires significant coding efforts. Unfortunately, this currently limits the overall effectiveness of this process. However, significant insights can potentially be gained even from small designs, as seen from the analysis in Chapter IV.

B. BENEFITS OF EXPERIMENTAL DESIGN IN STORM

The main purpose of Chapter IV was to analyze output data as it pertained to four WITTW metrics in order to demonstrate that N81 analysts can gain significant insights

from a successfully implemented design. Summary statistics, histograms, and barplots were utilized to gain intuition about how changing the intercept speed profiles for the BFNMF was affecting the PUNIC 21 scenario outcomes. Since variations were exhibited throughout all metrics and all design points, ANOVA and Tukey's HSD tests were run in order to validate whether this variation resulted in statistically different design points. It is important for an analyst to analytically confirm the variation in separate responses. Indeed, statistical tests may reveal that differences that appear to be significant via plots or summary data may in fact be the result of random variation. For this thesis, the results revealed that only one response, the time it takes blue forces to achieve air supremacy has statistical significance across different design points. Moreover, it can be concluded that this difference resulted from making a slight change of only 10% to a very small number of factors in a hypothetical scenario. If the methodologies discussed in this thesis are applied to a real-world scenario, the results could prove to be even more informative.

The ultimate goal for this research was to test the feasibility of implementing a DOE within STORM, knowing that how each factor affects the response in the PUNIC 21 scenario, or any scenario for that matter, will provide valuable insights to N81 analysts. Armed with the specific tools used in Chapter IV, they will be able to significantly improve their overall ability to analyze scenarios and provide decision makers with options that are backed by statistical evidence.

C. RECOMMENDATIONS

Although the benefits of a DOE within STORM are seen throughout this research, the methodology discussed in Chapter IV is far from being an automated process. When implemented manually it requires a significant amount of time and effort, which is not conducive for current N81 operations. Additionally, a moderate level of coding experience is required in order to create custom input files for design implementation. Therefore, follow-on research should explore additional methods such as automating these procedures to extend the reach of the work presented in this thesis. The ultimate goal is to incorporate a DOE in a real-world scenario so N81 can gain significant insights similar to what was presented in this research.

APPENDIX. R-STUDIO CODE

This Appendix contains the R code that was used to conduct the statistical analysis on each design point, as discussed in Chapter IV. Upon reading in the desired scenario output data file (retrieved using STORMMiner software), the code generates summary statistics, histograms, barplots, and statistical-difference tests (ANOVA and Tukey's HSD) of each user-specified metric.

```
##Calculates the summary statistic, bar-plots, and histograms for each design point
##LT William Bickel
##September 2014

#Reads in specific file that is to be examined
Output_File<-read.csv(file.choose())

#Creates a data frame with only the metric of interest. Must be performed for design
#points one through eight
n<-nrow(Output_File$value)
average<-mean(Output_File$value)
std_dev<-sd(Output_File$value)
error<-qnorm(0.975)*(std_dev)/sqrt(n)
lower<-average-error
upper<-average+error

#Install R Package e1071 for skewness of data
skew<-skewness(Output_File$value)

#brings all eight design point summary statistics together
metric_file_avg<-rbind(average1, average2, …, average8)
metric_file_lower<-rbind(lower1, lower2, …, lower8)
metric_file_upper<-rbind(upper1, upper2, …, upper8)


##Building the bar-plot for examining averages over each design point
plot<-barplot(metric_file_avg, col="light blue",ylab="Y-axis label", main="Main Title")
errbar(plot[,1], metric_file_avg, metric_file_upper, metric_file_lower, add=T, xlab="")
legend(locator(1),lty=1,lwd=c(2,3),col= 'black', legend=c("95% Confidence Interval"))

## Building a 2 by 4 histogram with mean, median, and 95% Confidence Intervals
par(mfrow=c(2,4)
```

```
hist(Output_File$value,breaks=20,col="lightblue",main=paste("MainTitle”,
Mean=",signif(average,digits=4),",Median=",signif(median(Output_File$value),digits=4)
,",SD=",signif(sd4,digits=4),",95%CI=
[",signif(metric_file_lower,digits=4),",",signif(metric_file_upper,digits=4),"],\nSkewness
=",signif(skewness(Output_File$value),digits=4)),xlab="X-axis label")
abline(v=average,col="blue",lwd=4)
abline(v= metric_file_lower,col="dark green",lty=2,lwd=4)
abline(v= metric_file_upper,col="dark green",lty=2,lwd=4)
abline(v=median(Output_File$value),col="red",lwd=4)

##Creating ANOVA table, Tukey’s HSD test results and graph

#Create a linear regression model over each design point
File_lm<-lm(Output_File$(“WITTWMetric”)~Output_File$Design.Point,
data=Output_File)
summary(File_lm)
anova(File_lm)
Output<-aov(Output_File$(“WITTWMetric”)~Output_File$Design.Point,
data=Output_File)
posthoc<-TukeyHSD(x=Output,'Design.Point',cof.level=.95)
plot(posthoc)
HSD.test(File_lm,'Design.Point',console=TRUE)
Output.hsd <-data.frame(TukeyHSD(Output, which = "Design.Point")$Design.Point)
Output.hsd$Comparison_of_Pairs_of_Design_Points<-row.names(Output.hsd)
ggplot(Output.hsd, aes(Comparison_of_Pairs_of_Design_Points, y = diff, ymin = lwr,
ymax = upr))+geom_pointrange() + ylab("Difference in Mean Levels of Design
Points")+coord_flip()+ggtitle("MainTitle")+theme(plot.title = element_text(lineheight=2,
face="bold"))+geom_hline(yintercept=c(0,0),color="red")+theme(axis.text=
element_text(size=20),axis.title=element_text(size=16,face="bold")
```

LIST OF REFERENCES

Cioppa, T. M., & Lucas. T. W. (2007). Efficient nearly orthogonal and space-filling Latin hypercubes. *Technometrics,* 49(1), 45–55.

Committee on National Statistics and Committee on Applied and Theoretical Statistics, N. R. (1994). *Statistical issues in defense analysis and testing.* (J. E. Rolph, & D. L. Steffey, Eds.) Retrieved May 28, 2014, from The National Acadamies Press: http://www.nap.edu/openbook.php?record_id=9686

David M., & Pugh, C. U. (2000, March 13). *A validation assessment of the STORM air-to-air prototype algorithm*. Wright-Patterson Air Force Base, OH: United States of America: Air Force Institute of Technology.

Department of Defense. (2014, 15 June). *Department of Defense dictionary of military and associated terms* (JP 1-02.). Retrieved from http://www.dtic.mil/doctrine/new_pubs/jp1_02.pdf

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* Thousand Oaks, CA: United States of America: SAGE Publications.

Group W. (2012a). *STORM:Analyst's manual version 2.3*. Fairfax: Group W

Group W. (2012b). *STORM: Programmer's manual version 2.3*. Fairfax: Group W

Group W. (2012c). *STORM: User's manual version 2.3*. Fairfax: Group W

Group W. (2012d). *What's new in STORM version 2.3.* Fairfax: Group W

Hagel, C. (2014). *Quadrennial defense review.* The White House, Department of Defense, Washington, D.C.

Kleijnen, J. P., Sanchez, S. M., Lucas, T. W., & Cioppa, T. M. (2005). A user's guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing,* 17(3), 263–289.

Hernandez, A. S., Lucas, T. W., & Carlyle, M. (2012). Constructing nearly orthogonal Latin hypercubes for any nonsaturated run-variable combination. *ACM Transactions on Modeling and Computer Simulation*, 22(4), 20:1–20:17.

Law, A. M. (2007). *Simulation modeling & analysis* (4th Edition ed.). New York, New York, United States of America: McGraw-Hill.

Lucas, T. W. (2013, October 10). *Stochastic Lanchester models.* Unpublished class notes, Department of Operations Research, Naval Postgraduate School, Monterey, California.

Lucas, T.W. (2000). The stochastic versus deterministic argument for combat simulations: Tales of when the average won't do. *Military Operations Research*, *5*(3), 9–28.

Nunn, W. R., & Heimerman, K. T. (2003, October). *Review of the integrated theater engagement model (ITEM)*. Alexandria, Virginia, United States of America: Center for Naval Analyses.

Sanchez, S. M. (2007). Work smarter not harder: guidelines for designing simulation experiments. *Proceedings of the 2007 Winter Simulation Conference,* 84–94.

Sanchez, S. M., Lucas, T. W., Sanchez, P. J., Nannini, C. J., & Wan, H. (2012). Designs for large-scale simulation experiments, with applications to defense and homeland security. In K. Hinkelmann, *Design and Analysis of Experiments* (Vol. 1, p. 600). Hoboken, NJ: John Wiley & Sons.

Sweeney, R. L., Hamman, J. P., & Biemer, S. M. (2011). The application of systems engineering to software development: A case study. *John's Hopkins APL Technical Digest,* 29 (4), p. 11.

Vinyard, W., and Lucas, T.W. (2002), Exploring Combat Models for Non-monotonicities and Remedies. *PHALANX,* 35(1), 19, 36–38.

Wackerly, D., Mendenhall III, W. & Scheaffer, R. (2008). *Mathematical statistics with applications.* Belmont, CA: Brooks/Cole.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California