2015

Developing Discrete Empirical Distributions for Tractable Stochastic Programming Problems with Application for U.S. Army Force Sizing

Checco, John

http://hdl.handle.net/1920/9880

a service of Mason Publishing
DEVELOPING DISCRETE EMPIRICAL DISTRIBUTIONS FOR TRACTABLE STOCHASTIC PROGRAMMING PROBLEMS WITH APPLICATION FOR U.S. ARMY FORCE SIZING

by

John C. Checco
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Systems Engineering and Operations Research

Committee:

Dr. Bjorn Berg, Dissertation Director
Dr. Daniel B. Carr, Committee Member
Dr. Karla L. Hoffman, Committee Member
Dr. Andrew G. Loerch, Committee Member
Dr. Ariela Sofer, Department Chair
Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering

Date: Summer Semester 2015
George Mason University
Fairfax, VA
Developing Discrete Empirical Distributions for Tractable Stochastic Programming Problems With Application for U.S. Army Force Sizing

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

John C. Checco
Master of Science
Kansas State University, 2007
Master of Science
Missouri University of Science and Technology, 2001
Bachelor of Science
Lehigh University, 1996

Director: Bjorn Berg, Professor
Department of Systems Engineering and Operations Research

Summer Semester 2015
George Mason University
Fairfax, VA
This work was neither directed by nor endorsed by U.S. Army or any other government agency. Any opinions reflected in this thesis are the author’s.
DEDICATION

To my family who supported me throughout this endeavor.
ACKNOWLEDGEMENTS

I would like to thank my dissertation director, Dr. Bjorn Berg for his guidance, patience, and willingness to learn about Army force structure in order to support this research. His instruction on stochastic optimization inspired me to apply that approach to an Army problem that I had encountered. He kept my efforts focused in a productive direction while his questions, comments, and revisions greatly improved the quality of this product.

I would also like to thank my committee members, Dr. Andrew Loerch, Dr. Karla Hoffman, and Dr. Daniel Carr. Their insightful comments broadened my perspective on this problem, and lead me in interesting directions that I did not anticipate.

I want to acknowledge the institutions and professionals who aided in this research without compensation. Their dedication to supporting and improving the U.S. Armed Forces is commendable, and much appreciated. The Center for Army Analysis hosted several meetings for me and provided the historic contingency data set used in this analysis. Colonel Scott Nestler, Mr. Tucker Hughes, Mr. Jason Southerland, Mr. Jack Zeto, Mr. Troy Wilke, Lieutenant Colonel Richard Green, Major Jonathon Welborn, and Lieutenant Colonel Josh Helms all provided valuable insight to current force structure analysis and research. The Army Training and Doctrine Command Analysis Center provided unclassified planning scenarios used for force demands; thanks in particular to Mr. Cody Beck and Lieutenant Colonel Wayne Nitzschner. Dr. Nancy Huff, Dr. Colin Doyle, and Mr. John Brinkerhoff at The Institute for Defense Analysis provided an overview of their model as well the notional force structure and force requirements used in this research. Dr. Frank Camm and Dr. Wade Markel from the RAND Corporation provided valuable insight regarding institutional force requirements and relationships within the Army. Ms. Anne Morton at the Army Program Analysis and Evaluation Division provided technical support and access to the Army’s budgetary database.
Furthermore, I would like to thank the Army Senior Leaders who took the time to discuss and help shape my research topic, Lieutenant General (Retired) Joe Martz, Major General John Ferrari, Mr. Roy Wallace, and Dr. Steve Stoddard.

I am indebted to Dr. Kaye McKinzie and Dr. Ilene Keltz for recognizing my potential and encouraging me to pursue a doctorate degree.

I am thankful to Raylene for reviewing my grammar and identifying many mistakes and incomplete thoughts. Any mistakes remaining are my own.

Finally, I am eternally grateful to my wife for supporting me and making my accomplishments possible, my children for the numerous sacrifices they made on my behalf, and my parents for their encouragement.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>xiv</td>
</tr>
<tr>
<td>Abstract</td>
<td>xvi</td>
</tr>
<tr>
<td>Chapter 1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2 Background</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Army Operating Environment</td>
<td>6</td>
</tr>
<tr>
<td>2.1.1 Demand Uncertainty</td>
<td>7</td>
</tr>
<tr>
<td>2.1.2 Dynamic Environment</td>
<td>7</td>
</tr>
<tr>
<td>2.1.3 Decision Timing</td>
<td>8</td>
</tr>
<tr>
<td>2.1.4 Costs</td>
<td>8</td>
</tr>
<tr>
<td>2.1.5 Manpower</td>
<td>9</td>
</tr>
<tr>
<td>2.1.6 Army Organization and Classification</td>
<td>15</td>
</tr>
<tr>
<td>2.1.7 Adjusting Force Size</td>
<td>21</td>
</tr>
<tr>
<td>2.1.8 Deployment Demand</td>
<td>26</td>
</tr>
<tr>
<td>2.2 Army Force Development</td>
<td>33</td>
</tr>
<tr>
<td>2.2.1 DoD Guidance and Constraints</td>
<td>33</td>
</tr>
<tr>
<td>2.2.2 Army Force Structure Development</td>
<td>38</td>
</tr>
<tr>
<td>2.2.3 Historic Context</td>
<td>39</td>
</tr>
<tr>
<td>2.2.4 Total Army Analysis</td>
<td>40</td>
</tr>
<tr>
<td>2.2.5 TAA Based Approaches</td>
<td>44</td>
</tr>
<tr>
<td>2.2.6 Other Stochastic Force Structure Approaches</td>
<td>51</td>
</tr>
<tr>
<td>2.2.7 Analytic Capability Gaps</td>
<td>57</td>
</tr>
<tr>
<td>2.3 Stochastic Programming</td>
<td>61</td>
</tr>
<tr>
<td>2.3.1 Stochastic Programming Overview</td>
<td>61</td>
</tr>
</tbody>
</table>
2.3.2 Notation and Problem Representation ...................................................... 66
2.3.3 Solution Methods ........................................................................................ 69
2.3.4 Other Stochastic Optimization Approaches ............................................... 74
2.3.5 Related Applications ................................................................................... 77
2.4 Research Objectives ........................................................................................... 83
2.4.1 Purpose ....................................................................................................... 83
2.4.2 Study Questions .......................................................................................... 84
2.4.3 Scope ........................................................................................................... 89
Chapter 3 Stochastic Deployment Demand ............................................................. 93
3.1 Mathematical Model of Deployment Demand .................................................. 93
3.1.1 Event Arrivals .............................................................................................. 94
3.1.2 Event Categorization and Distribution ...................................................... 104
3.1.3 Planning Scenarios .................................................................................... 114
3.2 Stochastic Demand Generation ....................................................................... 123
3.2.1 Simple SDG Example ................................................................................. 125
3.2.2 Stochastic Demand Verification and Validation ....................................... 129
3.3 Application for Stochastic Programming ......................................................... 140
3.3.1 Stochastic Demand Sample Space ............................................................ 140
3.3.2 Approximation Methods ........................................................................... 142
3.3.3 Approximation Method Comparison ........................................................ 162
Chapter 4 Army Force Size Model ............................................................................. 167
4.1 Key Relationships ............................................................................................. 167
4.2 System Definition ............................................................................................. 172
4.2.1 System Diagram ........................................................................................ 172
4.2.2 Temporal Model of Periods and Decisions ............................................... 177
4.3 Mathematical Model ........................................................................................ 178
4.3.1 Variable Definition .................................................................................... 178
4.3.2 Objective Function .................................................................................... 189
4.3.3 Personnel Constraints ............................................................................... 191
4.3.4 Unit Constraints ....................................................................................... 199
4.3.5 Constraints Linking Personnel and Units .................................................. 203
5.8 Results Summary ................................................................. 284
Chapter 6 Conclusions ................................................................. 287
  6.1 Contributions ........................................................................ 288
  6.2 Future Research ................................................................... 292
Appendix A Notes On Model Tractability ........................................ 294
  A.1 Experimental Conditions ..................................................... 294
    A.1.1 Design Limitations ......................................................... 294
  A.2 Default Solution Times ......................................................... 295
    A.2.1 Expected Value Problem ................................................ 295
    A.2.2 Method Z ...................................................................... 295
  A.3 Optimizer Tuning ................................................................. 296
    A.3.1 Optimization Algorithm (Method) ................................. 296
    A.3.2 Other Optimization Parameters .................................... 298
    A.3.3 Barrier Method Crossover to Simplex ............................ 301
    A.3.4 Solution Verification ..................................................... 303
  A.4 Practical Model Size Limits ................................................ 304
Appendix B Probability Distribution for 2-stage Approximation ........ 306
  B.1 Method E – Subset of SDG simulation scenarios .................... 306
  B.2 Approximation Method Comparison ..................................... 310
References .................................................................................. 314
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1 Stochastic programming notation.</td>
<td>67</td>
</tr>
<tr>
<td>Table 3.1 Summary of data screening for historic contingency operations.</td>
<td>100</td>
</tr>
<tr>
<td>Table 3.2 Rate of MCOs over various time intervals in U.S. history.</td>
<td>110</td>
</tr>
<tr>
<td>Table 3.3 Empirical distribution of contingency events with modified MCO sample.</td>
<td>114</td>
</tr>
<tr>
<td>Table 3.4 Contingency force demands and likelihood probabilities.</td>
<td>119</td>
</tr>
<tr>
<td>Table 3.5 Fixed ongoing demands for U.S. Army forces.</td>
<td>122</td>
</tr>
<tr>
<td>Table 3.6 Summary of force demand data for fixed and ongoing operations.</td>
<td>122</td>
</tr>
<tr>
<td>Table 3.7 Example scenarios with deployment demands.</td>
<td>125</td>
</tr>
<tr>
<td>Table 3.8 Example scenario probabilities.</td>
<td>126</td>
</tr>
<tr>
<td>Table 3.9 Example period 1 demand from scenarios A &amp; B.</td>
<td>126</td>
</tr>
<tr>
<td>Table 3.10 Total example demand from the first two periods.</td>
<td>127</td>
</tr>
<tr>
<td>Table 3.11 Total demand truncated to three years.</td>
<td>127</td>
</tr>
<tr>
<td>Table 3.12 Simulated stochastic demand for a 10-year period.</td>
<td>128</td>
</tr>
<tr>
<td>Table 3.13 Simulated contingency event list.</td>
<td>129</td>
</tr>
<tr>
<td>Table 3.14 Correlation of total demand between periods.</td>
<td>138</td>
</tr>
<tr>
<td>Table 3.15 Method A empirical distribution intervals with total demand values.</td>
<td>146</td>
</tr>
<tr>
<td>Table 3.16 Method B discrete empirical distribution levels.</td>
<td>151</td>
</tr>
<tr>
<td>Table 3.17 Method B empirical distribution partitioning for each demand level.</td>
<td>151</td>
</tr>
<tr>
<td>Table 3.18 Method C Markovian transition probabilities.</td>
<td>155</td>
</tr>
<tr>
<td>Table 3.19 Summary of demand approximation methods.</td>
<td>162</td>
</tr>
<tr>
<td>Table 4.1 Summary of variable index definitions.</td>
<td>179</td>
</tr>
<tr>
<td>Table 4.2 Summary of general constraint parameter definitions.</td>
<td>184</td>
</tr>
<tr>
<td>Table 4.3 Summary of other parameter definitions.</td>
<td>186</td>
</tr>
<tr>
<td>Table 4.4 Summary of variable definitions.</td>
<td>188</td>
</tr>
<tr>
<td>Table 5.1 Alternate Method B demand distribution designs.</td>
<td>243</td>
</tr>
<tr>
<td>Table 5.2 Alternative Method D demand design levels.</td>
<td>253</td>
</tr>
<tr>
<td>Table 5.3 Discount factors.</td>
<td>266</td>
</tr>
<tr>
<td>Table 5.4 Value of stochastic solutions.</td>
<td>283</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1 Hierarchical organization of an Infantry Division (simplified) ........................................ 19
Figure 2.2 ARFORGEN force pools and readiness progression .......................................................... 29
Figure 2.3 Example integrated security construct (ISC) .................................................................. 36
Figure 2.4 Three stage scenario tree for flipping a coin ................................................................. 64
Figure 3.1 Histogram of contingency event inter-arrival days ....................................................... 96
Figure 3.2 Quantile-Quantile exponential plot of contingency event inter-arrival times. .................. 97
Figure 3.3 Historic number of contingency events per year ........................................................... 101
Figure 3.4 Histogram of screened contingency event inter-arrival periods .................................... 102
Figure 3.5 Quantile-Quantile exponential plot of screened contingency event inter-arrival times ......................... 103
Figure 3.6 Observed contingency events by category ..................................................................... 105
Figure 3.7 Observations of each contingency type over time .......................................................... 106
Figure 3.8 Timeline of U.S. Major Conflicts .................................................................................... 109
Figure 3.9 The Q-Q exponential plot of MCO inter-arrival times from 1898 to 2014 .................... 111
Figure 3.10 The inter-arrival times of major U.S. conflicts from 1898 to 2014 ............................. 111
Figure 3.11 The Q-Q normal plot of MCO inter-arrival times from 1898 to 2014 ....................... 112
Figure 3.12 Simulation process diagram to generate stochastic demand ...................................... 124
Figure 3.13 Distribution of average total deployment demand ....................................................... 132
Figure 3.14 Distribution of peak total deployment demand ............................................................. 133
Figure 3.15 Stochastic demand profiles by percentile ................................................................. 134
Figure 3.16 Histogram of average total demand observations ....................................................... 135
Figure 3.17 Stochastic demand profiles by percentile of total average demand ......................... 136
Figure 3.18 Total demand levels using skewed distribution ............................................................ 137
Figure 3.19 Comparison of historic to simulated troop demands .................................................. 139
Figure 3.20 Discrete total demand levels by year for Method A .................................................... 148
Figure 3.21 Deployment demand profiles generated using Method A ............................................ 149
Figure 3.22 Empirical distribution values for method B ................................................................. 152
Figure 3.23 Example demands generated by method B ................................................................. 153
Figure 3.24 Example demands generated by Method C ............................................................... 157
Figure 3.25 Histogram of Methods C scenario probabilities .......................................................... 158
Figure 3.26 Cumulative distribution function for Method C .......................................................... 158
Figure 3.27 Cumulative distribution function from method D ....................................................... 160
Figure 3.28 Example demands generated by Method D ............................................................... 161
Figure 5.25 Comparison of discount factors ................................................................. 267
Figure 5.26 Comparison of penalty factor levels using costs and unmet demand. ...... 269
Figure 5.27 Examples of policy recommendations under varying penalty factors. ...... 270
Figure 5.28 Cumulative distribution function of unmet demand by scenario probability.
    Left graph shows full function with inset area detail provided on right. .......... 271
Figure 5.29 Five-year expected cost of policies organized by 95% confidence index of
    unmet demand. .................................................................................................. 273
Figure 5.30 Active Component force levels from models with varying penalty factors. 275
Figure 5.31 Selected policy metrics comparing resource constrained solutions. Graphs
    on left show full scale for perspective, while graphs on right are re-scaled to
    provide greater detail. ........................................................................................................ 277
Figure 5.32 Unmet demand and total cost for resource constrained solutions. ........... 279
Figure 5.33 Comparison of resource constrained solutions with differing penalty factors.
    Graphs on left show full scale for perspective, while graphs on right are re-scaled
    to provide greater detail. .................................................................................... 280
Figure 5.34 Comparison of the expected value solution to the recourse problem
    solution. ....................................................................................................................... 284
Figure A.1 Anecdotal solution times for increasing numbers of periods. ...................... 296
Figure A.2 Anecdotal solution times for different optimization algorithm ................. 297
Figure A.3 Anecdotal solution times comparing dual problem pre-solve settings. ...... 299
Figure A.4 Anecdotal solution times for various settings of the prepass parameter..... 300
Figure A.5 Anecdotal solution times for the scaling parameter .................................. 301
Figure A.6 Anecdotal solution times for the barrier crossover method parameter ..... 302
Figure A.7 Anecdotal solution times with increasing numbers of periods............... 303
Figure B.1 Cumulative distribution function from method E. ................................. 309
Figure B.2 Average total demand of Method E scenarios ....................................... 309
Figure B.3 Example demands generated by method E ............................................ 310
Figure B.4 Average total demand comparison of demand approximation methods...... 311
Figure B.5 Standard deviation comparison of demand approximation methods ........ 312
Figure B.6 Correlation comparison of demand approximation methods ................. 312
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCT</td>
<td>Armor Brigade Combat Team</td>
</tr>
<tr>
<td>AC</td>
<td>Active Component</td>
</tr>
<tr>
<td>ACM</td>
<td>Army Contingency Costing Model</td>
</tr>
<tr>
<td>ADP</td>
<td>Approximate Dynamic Programming</td>
</tr>
<tr>
<td>AFSP</td>
<td>Army Force Size Problem</td>
</tr>
<tr>
<td>AGR</td>
<td>Active Guard Reserve</td>
</tr>
<tr>
<td>AMCOS</td>
<td>Army Military-Civilian Costing System</td>
</tr>
<tr>
<td>ARFORGEN</td>
<td>Army Force Generation Model</td>
</tr>
<tr>
<td>BCT</td>
<td>Brigade Combat Team</td>
</tr>
<tr>
<td>CAA</td>
<td>Center for Army Analysis</td>
</tr>
<tr>
<td>CAB</td>
<td>Combat Aviation Brigade</td>
</tr>
<tr>
<td>CAPE</td>
<td>Cost Assessment and Performance Evaluation</td>
</tr>
<tr>
<td>CBP</td>
<td>Capabilities Based Planning</td>
</tr>
<tr>
<td>CONOPS</td>
<td>Concept of Operations</td>
</tr>
<tr>
<td>DAC</td>
<td>Department of the Army Civilian</td>
</tr>
<tr>
<td>DAS</td>
<td>Defense Acquisition System</td>
</tr>
<tr>
<td>DoD</td>
<td>Department of Defense</td>
</tr>
<tr>
<td>DPS</td>
<td>Defense Planning Scenario</td>
</tr>
<tr>
<td>EEV</td>
<td>Expectation of the Expected Value Problem</td>
</tr>
<tr>
<td>EV</td>
<td>Expected Value Problem</td>
</tr>
<tr>
<td>FMCA</td>
<td>Force Mix and Composition Analysis</td>
</tr>
<tr>
<td>FOB</td>
<td>Forward Operating Base</td>
</tr>
<tr>
<td>FYDP</td>
<td>Future Years Defense Program</td>
</tr>
<tr>
<td>IBCT</td>
<td>Infantry Brigade Combat Team</td>
</tr>
<tr>
<td>IDA</td>
<td>Institute for Defense Analysis</td>
</tr>
<tr>
<td>ISC</td>
<td>Integrated Security Construct</td>
</tr>
<tr>
<td>JCIDS</td>
<td>Joint Capabilities Integration and Development System</td>
</tr>
<tr>
<td>JICM</td>
<td>Joint Integrated Contingency Model</td>
</tr>
<tr>
<td>LN</td>
<td>Local National</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Program</td>
</tr>
<tr>
<td>MCO</td>
<td>Major Contingency Operation</td>
</tr>
<tr>
<td>MIP</td>
<td>Mixed Integer Program</td>
</tr>
<tr>
<td>MOS</td>
<td>Military Occupational Specialty</td>
</tr>
<tr>
<td>MSFD</td>
<td>Multi-Service Force Deployment</td>
</tr>
</tbody>
</table>
MSSP ................................................................. Multistage Stochastic Program
MTOF ........................................................................... Mission Task Organized Force
NCO ............................................................................. Noncommissioned Officer
OCO ............................................................................ Oversees Contingency Operations
OCS ................................................................................ Officer Candidate School
OOTW ........................................................................ Operations Other Than War
OPTEMPO ................................................................................ Operations Tempo
OSD ........................................................................... Office of the Secretary of Defense
PPBE ............................................................................... Planning, Programming, Budgeting and Execution
RANGER IPOD ........................................................................................................

........Randomly Generated Requirements Informed by Past Operational Deployments
RC ........................................................ Reserve Component
RHS ......................................................................................... Right Hand Side
ROTC ........................................................ Reserve Officer Training Corps
RP .................................................................................. Recourse Problem
SADE ........................................................ Stochastic Analysis for Deployments and Excursions
SARA ........................................................ Stochastic Active-Reserve Assessment
SARDE .................................................. Stochastic Analysis of Resources for Deployments and Excursions
SBCT .......................................................... Stryker Brigade Combat Team
SDG .......................................................... Stochastic Demand Generator
SP ................................................................................ Stochastic Program
SSA ........................................................ Support to Strategic Analysis
SSC .......................................................... Small Scale Contingency
SSSP ........................................................ Steady State Security Posture
TAA ........................................................ Total Army Analysis
TOE ........................................................ Table of Organization and Equipment
TPU ........................................................ Troop Program Unit
TRAC ........................................................ Training and Doctrine Command Analysis Center
TTHS ............................................................ Trainees, Transients, Holdees, and Students
VSS ........................................................................ Value of the Stochastic Solution
WS ................................................................. Wait and See Problem
WWII .............................................................. World War II
ABSTRACT

DEVELOPING DISCRETE EMPIRICAL DISTRIBUTIONS FOR TRACTABLE STOCHASTIC PROGRAMMING PROBLEMS WITH APPLICATION FOR U.S. ARMY FORCE SIZING

John C. Checco, Ph.D.
George Mason University, 2015
Dissertation Director: Dr. Bjorn Berg

Providing for the common defense of the United States and its interests is a vital, though expensive, endeavor. To that end, the U.S. Army is expected to grow and shrink as necessary to conform to changing circumstances. Current analytic methods focus on operational (combat) requirements in developing resource constrained operational force designs. These tend to rely on simulation of static operational force structures with fixed readiness policies. Additionally, the Army’s primary strategic force structure analysis models have no direct linkage to cost, instead relying on manpower limits. This type of approach is ill-suited to consider the cost vs. benefit tradeoff of the Army, or the overall size and composition of the Army, and how it should adapt over time to changing conditions.
This thesis develops a multi-stage optimization model to determine dynamic force size over a number of years while accounting for uncertainty of future demand for forces. We begin by providing a method to approximate force demand as a discrete empirical distribution amenable to a stochastic optimization framework. Next, we develop a strategic level stochastic linear program of the Army to include the total workforce, operational and institutional forces, deployment policies, and costs. Solution methods are evaluated and presented as a computational study. Finally, we offer insights for further development and potential use of this approach to augment future force structure analysis.
CHAPTER 1 INTRODUCTION

The military provides for the common defense of the United States, a critical function of the federal government. It is very costly, however, budgeted for $615B in 2014, or 3.6% of the gross domestic product [1]. That is 16% of the total U.S. budget (which is operating at a deficit), or 50% of discretionary spending (excluding Social Security, Medicaid, Medicare and other mandatory programs). Therefore, while the U.S. must provide adequate military capability to ensure security, allocating too much to National defense may invite economic ruin.

The questions surrounding military capability requirements and how much funding to allocate toward defense are inherently difficult to answer. Predicting the future state of the world with a multitude of actors and dynamic environmental factors is near impossible, making future mission requirements uncertain by nature. Even with the world’s largest economy [2] and most powerful military [3], America may still find it has limited military options in many cases. Furthermore, it is difficult to quantify the benefit of military capability in terms of dollars. For example, not only does a strong military provide security, it can also assure regional or global influence. Both politics and economics conspire to shape requirements and limitations, respectively. Despite
bearing resemblance to a philosophical question, “How much is enough?” requires a concrete answer, regardless of the underlying complexity.

Current planning procedures require significant simplifications and assumptions. Specified planning scenarios are used as the basis of all current military analysis, reducing an infinite number of future possibilities to a small discrete set. In many cases, stochastic aspects are ignored in favor of simpler deterministic methods. Arbitrary funding and manpower constraints are imposed in order to determine a feasible military force. Force structures and policy decisions are typically held constant over an analysis period of years rather than allowing them to shift dynamically as conditions change. Various portions of the process are performed sequentially or nearly independently, and therefore not optimized or improved as a cohesive system. Several important factors are not currently considered inside the main optimization system causing them to often be ignored. Each of these issues will, in general, lead to sub-optimal results; seeking good feasible solutions rather than the best solution.

This research focuses on the Army’s force size. The Army, more so than the other military branches, is expected to grow and contract as needs wax and wane. This is possible because the Army is personnel centric, while the Navy and Air Force are more equipment centric. New soldiers can be recruited and trained much more quickly than new ships and aircraft can be designed and built. The Marine Corps is very similar to the Army, but only a fraction of the size, and resourced through the Department of the Navy
which can buffer budget fluctuation as well as provide a consistent presence mission for the Corps.

The Army’s budget for 2014 is $130 Billion, for a force of 520,000 active duty and 560,000 reserve (part time) soldiers [1]. The Army divides its resources between three general categories: force structure, readiness, and modernization. Force structure refers to the number and type of each Army unit along with the detailed personnel and equipment requirements for each unit type. Readiness includes training, maintenance, transportation, facilities and infrastructure. Modernization deals mainly with weapons and equipment, encompassing research, development, tests, procurement, and upgrades. Each of these must be balanced in order for the Army to operate efficiently. Of these, force structure has primacy. Both readiness and modernization are scaled by force structure, so funding is only relevant within the context of a force structure. We therefore limit our current focus to force structure. Explicit treatment of readiness and modernization may be the subject of future work. Furthermore, we use the term force size here rather than force structure to emphasize analysis at an aggregated, strategic level, whereas Army force structure analysis is typically much more detailed by specific unit type.

This research addresses the fundamental question of how large the Army should be. This thesis introduces a novel approach to model the total Army workforce at a strategic level which incorporates future uncertainty. Policy implications resulting from the model will be analyzed as well as the trade-off between cost and the likelihood of
meeting the deployment demand for soldiers. This approach may be used to augment current systems and yield better planning solutions as well as better understanding of Army resourcing, or funding distribution.

The remainder of this thesis is organized as follows. Chapter 2 provides background for this research. It begins with an introduction to general Army concepts and force management issues in particular before providing a review of current and existing analytic methods. Next we provide a stochastic programming overview and literature review of relevant work, and justification of the multistage stochastic programming approach. Finally, we state the study questions addressed in this thesis and place it in the context of the available literature.

Chapter 3 develops a model of stochastic deployment demand based on current and published practices and implemented as a discrete event simulation. This simulation is used to study the characteristics of the deployment demand distribution, which is then used to develop a probability distribution of demand. Several refinements on the demand distribution are presented and compared, resulting in a discrete distribution derived empirically from a stochastic demand simulation. These demand distributions are then tested and compared in Chapter 5.

Chapter 4 describes the concepts necessary for an accurate representation of Army force management systems. It provides a detailed description of the Army force size problem (AFSP) model which was created through this research. Unlike previous works, the AFSP model accounts for the entire Army force structure and all segments of
the workforce while providing for the dynamic aspect of adjusting force size and composition over time as new information becomes available. A special adaptation of non-anticipative constraints are used to generate a multi-year force plan that accurately manages information available at each decision point, and allows for the flexibility to adjust plans in future years.

Chapter 5 discusses model implementation and data for the AFSP model. It continues to outline solution methods and model verification methods. Next we conduct and analyze a series of experiments to evaluate the various deployment demand probability distributions developed in Chapter 3. This was used to select the distribution used in subsequent research. Additional experimentation allowed us to set remaining model parameters such as penalty factors by benchmarking to current conditions. We also define a relationship between force cost and the likelihood of meeting demand by solving the AFSP model for a series of penalty values. The addition of budget and manpower end strength constraints then allowed us to study their effect on recommended force levels. Finally, we show through the calculation of the value of the stochastic solution that accounting for the uncertain deployment demand is an important aspect of the Army force size problem.

Chapter 6 concludes this thesis by summarizing the most important contributions provided by this research and identifying areas for future related research.
CHAPTER 2  BACKGROUND

This chapter provides background information germane to this thesis. It begins with a general overview of Army concepts which may be unfamiliar to those outside the Army. Next we review the Army force development process in detail with focus on analytic methods both within and external to the Army. Following that, we cover stochastic programming with a brief overview and discussion of relevant methods and related applications. This chapter concluded by formalizing the objectives set forth by this thesis within the context of the existing literature.

2.1  Army Operating Environment
Managing the U.S. Army is accomplished through a complex federation of systems and procedures that are themselves ever evolving in a quest for greater efficiency. Many portions of the overall Army management system are modeled, or otherwise supported by quantitative analysis. Often, however, portions are considered independently or updated sequentially leading to overly constrained decisions that may not be well integrated.

The section that follows introduces the key concepts and phenomena that are important to understanding the models being proposed and discussed in this thesis. There are many complex and inter-related items presented for the benefit of non-
military readers, and to provide context for the sections that follow. These concepts are intentionally simplified here in the interest of brevity and conciseness.

2.1.1 Demand Uncertainty
The Army leadership must make decisions to shape the future force without knowing the future demands that will be placed on that force. The primary mission of the Army is to produce combat ready units available for deployment and use by a Combatant Commander (military leader of a joint, regional command), as needed. The future demand for combat units is the greatest source of uncertainty in the Army force management system. Explicitly modeling and integrating that uncertainty is one of the novel aspects of our approach.

2.1.2 Dynamic Environment
The Army operates in a dynamic environment such that a steady state equilibrium is potentially never reached. Ideally, feedback loops between inter-related processes would influence management decisions in a timely manner. In reality, the complexities of the processes involved often only allow updates once per year. Although most major processes are synchronized to allow one decision or update to inform the next, an unexpected problem or new information can upset this process and force decisions to be made without the benefit of current or updated information. Setting one subsystem while assuming others are fixed or operating at a constant rate of change may be sub-optimal, but typical in practice.
2.1.3 Decision Timing
Key force structure and funding decisions in the Army are synchronized in a yearly cyclic process that ultimately provides a budget request to Congress. The long development and approval time of the budget necessitates that funding and force structure decisions being considered usually pertain to a time period two years in the future. Each yearly budget is incorporated in a five year plan in order to prevent erratic behavior and myopic decisions. This implies that the Army needs to look about seven years into the future for budget programming. For long range strategic planning, the Army considers 20 years or more.

2.1.4 Costs
Base costs refer to the Army’s peacetime operating budget. They include the routine maintenance, upkeep, and missions of the Army. Examples include the salary and benefits of the workforce, training, supplies, repair parts, service contracts, utilities, leases, construction, repairs, weapons research, and equipment purchases. Base funding reached a height of $144 billion during the Iraq and Afghanistan wars and is currently set at $1 billion for 2014 [1]. Base funds are planned for in detail, and requested each year from Congress.

Supplemental costs are also referred to as overseas contingency operations (OCO). They are related to deployments or contingency operations, and can be thought of as wartime costs in laymen’s terms. Examples of supplemental costs include hazardous duty and separation pay, related medical expenses, transportation, equipment repair and replacement, forward operating bases, service contracts, fuel,
ammunition, and emergency weapon and equipment procurement. Supplemental funding reached a height of $121 billion during the Iraq and Afghanistan wars and is currently set at $130 billion for 2014 [1]. Supplemental funding is typically not planned for in advance and only requested when needed. Supplemental funding requests tend to receive less Congressional scrutiny.

2.1.5 Manpower
The Army work force is commonly referred to as manpower, though it can also mean soldiers specifically, as in military manpower. Almost every aspect of the Army is directly tied to manpower. The number of combat units, amount of training, and required equipment are all directly related to the size of the work force. Likewise, the systems that man, train, organize, equip, lead, and support those units are also directly tied to personnel and force structure, or the number and types of units.

Manpower Types
Manpower can be classified as soldiers, civilians, or contractors. The most recognizable type of manpower in the Army is the military component, or Soldiers. All combat units are composed of soldiers and many of the supporting functions require military manpower, at least in part. A soldier’s hierarchical position, or rank, is determined mainly by seniority, performance, and experience. They start at a low entry level rank and work their way up. Promoting from within results in a closed, rigid personnel system. The Army needs more lower rank troops and first line supervisors than middle and upper level management, which necessitates turn-over and attrition.
Soldiers are relatively expensive because they require significant training (initially and recurring), experience high turn-over, and are paid a premium to compensate for potential deployment and associated hazards and hardships of the job. Soldiers are cliff vested in retirement benefits at 20 years of service, which strongly influences behavior at the mid-career point and beyond. Service is limited by a mandatory retirement date which varies based on rank and the needs of the Army.

The Department of the Army also employs a substantial civilian workforce, or DA Civilians. They fill many supporting and administrative functions where deployment or frequent relocation is not required, therefore avoiding the premium cost of soldiers. Although they too work within a similar hierarchical framework as soldiers, they may enter the workforce at any level as needed (provided appropriate skills and experience), which provides some flexibility to the DA Civilian personnel system. The civilian workforce provides an element of stability and institutional knowledge to Army organizations as there are no mandatory re-assignments or retirement dates. Civilians can become niche specialists, operating in the same job or field for decades. By contrast, soldiers are expected to be generalists, adaptable to any number of related jobs, and required to change jobs and locations periodically. It has become more acceptable, in recent years, to deploy DA Civilians to perform support and training functions in conjunction with combat units oversees. This is usually on a voluntary and individual basis, and in relatively small numbers compared with soldiers. This provides a means to reduce stress on the military component and deliver a skill or capability that may not exist within the
military component. The down side, however, is that deployed civilians cost significantly more than their soldier equivalents, require pre-deployment training, leave their organization short-handed, and are not permitted to bear arms, thereby limiting their role in a conflict zone.

Contractors, though not directly employed by the Army, have become essential to Army operations. Contract labor and services can provide value for the Army in several ways. A significant benefit of using contractors is the flexibility that it affords. When workload is uneven or uncertain, contractors may augment the workforce for the specific period required without the need to recruit, hire, train, and then continue to employ or terminate, as is the case for soldiers or civilians. Contractors can also provide specific expertise and niche skills that are not cost effective to maintain within the Army. Sometimes contractors can provide services more cheaply or efficiently than the Army could provide internally, either due to economies of scale or core competencies in a particular area. One should expect a for-profit company to charge a premium to provide a flexible labor pool, but those costs can be offset or even surpassed by possible savings, as described above. However, there are positions where it would be inappropriate or illegal to use a contractor, such as making decisions representing the government, accounting and oversight of funds and contracts, and potential conflicts with other commercial endeavors. In such cases DA Civilians or Military personnel would be required.
Like civilians, the role of contractors on the battlefield has been expanding over the last several decades. Armed contractors providing security are a small minority. Limited to defensive or protective roles, they are typically used by the Department of State or other agencies that do not have a robust armed force at their disposal. Increasingly, contractors have deployed to provide technical expertise and maintenance for new, rapidly fielded equipment or systems for which support infrastructure has not yet been integrated into the Army force structure. Providing translators has become a critical contract function recently, often embedding in combat units. It has become common to have a substantial contractor presence on established forward operating bases (FOBs). Established FOBs are usually well protected and relatively secure. Contractors from the U.S and closely allied countries routinely provide infrastructure services, such as providing electrical power and food services. The indigenous population, or local nationals, are often contracted to provide janitorial or labor services, as well as supplying goods and materials. These examples illustrate how deployed contractors are used to either provide a capability that does not exist within the Army force structure, or allows a soldier to fight in a combat unit who might otherwise be needed in a support role. Deployed contractors can have exorbitant costs, however, and there are security concerns when employing contractors on the battlefield, especially local nationals.
**Army Components**

The Army’s military manpower can be classified as active component (full time) or reserve component (part time). Active component soldiers provide higher readiness and greater flexibility for deployments and reassignment (relocation). Reserves can be activated or mobilized to active duty status when needed. Non-mobilized reservists are often referred to as an M-day soldiers or TPUs (for Troop Program Unit). The reserve component also has a modest contingent of full time soldiers (still designated reserve component) to provide organizational and administrative functions, referred to as full time support or AGR (Active Guard Reserve). Soldiers can transfer between active and reserve components, but only on a limited basis. Soldiers that choose to separate from active duty prior to retirement are often encouraged to consider the reserve component, which affords the Army to continue benefitting from their training and experience. During times of need, programs exist that allow reserve component soldiers to transfer to the active component, though this process is slow, difficult, and limited in scope.

**Soldier Types**

The primary distinction between soldiers is that of Officer or Enlisted. Officers provide leadership, planning, and management for Army units, and are typically college educated. There is also a small class of Officers, called Warrant Officers, which have greater specific technical skills than Enlisted (such as helicopter pilots and head mechanics), but less responsibility than Commissioned Officers. Enlisted soldiers make up the workforce of the Army, and as they rise in rank, become Non-Commissioned
Officers (NCOs), which are still enlisted ranks. NCOs are the supervisors or foremen in charge of executing tasks and managing daily operations. Enlisted soldiers usually have high school diplomas, and frequently have associates’ or bachelors’ degrees, especially as NCOs. Officers cost more to train and employ than enlisted, and it is more difficult to expand or contract the officer pool than the enlisted force. Officers are managed within a fairly rigid system, taking 10 years to grow a mid-level officer, and 20 years or more for a senior level officer. By contrast, enlisted soldiers are promoted based on requirements, which allows for more rapid expansion, though quality may suffer. Whether Officer or Enlisted, the soldier begins at entry level within their class, and works up the ranks through seniority, performance and experience; hence the closed personnel system previously noted. Enlisted soldiers (typically young NCOs) can apply to become either Warrant Officers or Commissioned Officers though only a fraction of enlisted soldiers make this transition.

Rank or pay grade is also used to classify Soldiers. There are nine pay grades for enlisted soldiers (E1-E9), five for Warrant Officers (W1-W5), and ten for commissioned officers (O1-O10). Segregating by rank is important for detailed personnel planning, but beyond the scope of this project. Other personnel models exist for this purpose.

Military Occupational Specialty (MOS) is also commonly used to differentiate soldier type. With hundreds of job specialties available, this aspect is also important for personnel management, but beyond the scope of this project. Again, other models exist for this purpose.
**Force Caps**

The number of soldiers in each component is set by Congress. This number can be negotiated up or down by reducing or increasing the average cost per soldier. For example, two higher level positions may be eliminated to allow three lower level positions to be added, thereby increasing the force cap by one [4]. Therefore the force cap functions like a constraint on total personnel costs, by component. Salary levels for each grade are specified by U.S. Code, and cannot be negotiated individually by the services [5].

Civilians do not have a force cap in the same manner that soldiers do. With recent budget reductions, however, the Department of Defense has imposed mandatory civilian reductions on the services to reduce cost.

Contractors are not managed as individual employees. Although contract manpower equivalents (the number of billable hours to equal one man-year) seem as though they would be a useful metric, they are not reported accurately. Currently, the most effective way to monitor and limit contract labor is through accounting and budget constraints.

2.1.6 **Army Organization and Classification**

There are several important criteria, which can be considered as multi-dimensional factors, to understand the basic structures and functions of the Army.

**Operational & Institutional Forces**

The Army can be partitioned as operational or institutional. The Operational Army consists of the units designed and organized for deployment. The Institutional
Army is everything else, providing the systems and infrastructure to support the operational force. The Operational Army cannot exist and function effectively without the support of the Institutional Army, while the Institutional Army does not have a purpose to exist without the Operational Army. These entities are clearly related, but that relationship is complicated and difficult to define due to the many diverse services the institutional force provides for the operational force.

The Army’s mission is to provide trained, deployable forces to conduct full spectrum operations globally. These deployable combat units are the operational force. Operational units are made up almost entirely of soldiers, either from the active or reserve component. Each is organized with a complete set of deployable equipment, and conducts unit and multi-echelon training in order to achieve deployment readiness.

Institutional forces provide the personnel, individual training, equipment, facilities, and management functions necessary for the operational force to accomplish their mission. Institutional activities such as recruiting and training directly impact how quickly the Army can expand or replace losses. Other institutional functions include: materiel support (supplies and equipment), installation activities (housing, services, and real property at forts and bases), and functional or special commands (such as medical, legal, or human resources) that provide service to the entire force.

The required capacity for each of these functions is related to the size of the operational force, though not all are easily adjustable or proportional. A certain level of fixed overhead is required no matter what size the operational force is. For example,
writing doctrine for 10 Brigades (of the same type) is not inherently more difficult than for 1 Brigade. Also, adjusting institutional capacity can be much less flexible than changing the operational force size; reducing training areas and closing bases takes years, and most likely can never be reversed.

Soldiers receive the same basic pay and benefits whether they are in an operational or institutional unit. Institutional organizations typically do not have full sets of combat equipment or the requirement to conduct training to maintain deployment readiness. This tends to make institutional soldiers less expensive than operational soldiers, when the additional equipment and readiness costs are apportioned.

**Army Component Considerations for Units**

Like soldiers, Army units and organizations can be classified as active component or reserve component. The reserve component is further sub-divided as National Guard or U.S. Army Reserve. National Guard units tend to be the combat units (infantry, tanks, artillery, etc.), while Army Reserve units tend to be service and support (logistical) units. National Guard units are aligned to a state or territory and under the control of the governor until federally activated, whereas Army Reserve units remain under federal control whether activated or not.

Active component operational units continuously cycle through reset, train, and ready phases to maintain a high level of readiness and availability for deployments. This also makes them expensive to maintain. Reserve component units progress through the same phases, but more slowly and only achieve the highest levels of readiness once
mobilized and designated for a deployment. This makes them less expensive to maintain than active component units when not deploying, but they also require more time and resources to prepare for a deployment than active component units. Reserve component units are also available for deployment less often than their active component counterparts.

Most deployable units are purely active or reserve component, but it is not unusual for institutional organizations to have a mix of both types of soldiers in addition to civilians.

**Echelon**

Army units and organizations operate within a hierarchical command structure. A sample of some common unit echelons from the operational force is shown below in Figure 2.1 [6]. Higher echelons are formed by combining several of the lower echelon units along with headquarters and additional support elements. Organization within the institutional force is often unique and functionally based, though they still have a hierarchical command structure.
Figure 2.1 Hierarchical organization of an Infantry Division (simplified). The units highlighted are composed primarily of infantryman.

**Unit Type**
Types of combat units are differentiated by their branch, key characteristics, and echelon (above). Branch describes the primary function of the unit, and often reflects the primary type of soldier assigned to that unit. For example, infantry units consist mainly of infantrymen and field artillery (fires) units consist mainly ofartillerymen. Lower echelon units are more homogeneous with a narrow mission set, while higher echelons are more diverse in personnel with a wide set of capabilities. This is reflected by the organizational chart in Figure 2.1.
Additional identifying characteristics can vary, but a specific weapon type, vehicle, or capability are commonly used. An infantry unit may be light (no armored vehicles), mechanized (with armored vehicles), or Stryker (named for their medium weight armored Stryker vehicles). Furthermore, each of the above could also have a scout designation based on their battlefield function. Light infantry may also be airborne (paratroopers) or air assault (helicopter mobile). Field Artillery units are described by their howitzers or launchers. These examples illustrate the myriad of unit types; there are thousands of unique unit types in the Army inventory.

*Combat arms* units are the units directly responsible for engaging enemy forces, such as infantry. This is in contrast to the *combat service* and *combat service support* unit, also referred to as combat *enablers*. Examples of enablers include military intelligence, communications, engineering, and supply units.

The Brigade Combat Team, or BCT is now considered the Army’s basic self-sustaining combat arms building block for deployment. It is built on a brigade structure with additional support elements to enable it to be self-sufficient. BCTs are Infantry (IBCT), Stryker (SBCT), or Armored (ABCT). Each BCT has approximately 4,500 soldiers, varying slightly by type. Two other important combat arms units for this model are the fires brigade, which provides artillery and rocket support, and the combat aviation brigade (CAB) providing attack, reconnaissance, and cargo helicopters.
2.1.7 Adjusting Force Size

The Army governs force size by controlling inputs to the system through recruiting and losses from the system through separations. The Army can also manipulate the composition of various segments of the force by managing transfers between segments as well as the recruiting and separation rates noted above.

Although conscription is still technically an option, the Army does not consider it viable in the foreseeable future. A draft would run counter to the all-volunteer policies currently in place.

Recruiting & Training

The Army recruits new active and reserve soldiers to replace losses and grow its ranks when necessary. Recruiting stations are positioned throughout the country and U.S. territories for this purpose. New enlisted recruits must go through basic training (general) as well as initial skills training (job/MOS specific). The time from recruitment until arrival at their first duty assignment is typically one to two years. Officer recruits are sent to The U.S. Military Academy at West Point (fixed number), or attain a civilian degree through the Reserve Officer Training Corps (ROTC), which provides some flexibility. The third avenue for creating officers is through Officer Candidate School (OCS), which transitions enlisted soldiers to commissioned officers. This program offers the most flexibility and is regulated by the needs of the Army. OCS graduates must then attain a college degree over the next several years. All officers attend common core training as well as branch/MOS specific training. The time from selection until arrival at a duty station is usually three to six years, accounting for schooling. The
capacity and funding for all these functions within the institutional force limits the rate at which the Army can grow, assuming there is sufficient equipment available.

DA Civilian hiring is similar to corporate practices. There can be an associated internship or on the job training period preceding full employment. Training for civilians is much less structured and organized than for soldiers. It is normal for a year to elapse between a position vacancy and having a new replacement. It can take an additional year if a new position must be created first.

Contract labor can be expanded quickly. A new contract can take up to a year to be processed and approved, but expanding an existing contract can be accomplished within a few months. Contractors are usually required to arrive adequately trained. Again, contractors provide the most flexibility within the labor force.

Separations
A term of enlistment is usually from two to five years, depending on the soldier’s preference as well as the specialty and bonuses involved. At the end of a term, a soldier may re-enlist if permitted by current Army requirements. Acceptance of additional training or a new duty assignment may require re-enlistment, or an additional service obligation. At the 10 year point, enlisted soldiers are given an indefinite status, and must request to resign if they choose to separate before their mandatory retirement or the Army selects them for separation. Once soldiers reach 18 years of service, they are protected until they reach 20 years, and are eligible for retirement benefits.
Officers need to request resignation in order to leave the Army before mandatory retirement or involuntary separation. They also incur service obligations from their initial schooling and subsequent training and moves. Once the service obligations are fulfilled, the officer is eligible to resign, at the convenience of the Army.

The 20 year vestment period for retirement affects soldier’s separation behavior. There is typically a large turnover in personnel after the initial service period, in the early years of an Army career. Once a soldier passes 10 to 12 years, however, they typically try to earn retirement benefits, unless they are stressed by frequent deployment or other personal considerations.

The Army can influence separation behavior to help shape the workforce. Separations can be decreased by offering incentives and bonuses to soldiers who re-enlist. The Army can also institute a stop-loss policy in extreme situations which can prevent most separations and involuntarily extend soldiers for the duration of the crisis. There is a financial and psychological cost to the stop-loss, however. Soldiers may be compensated with an additional allowance, which is usually modest in comparison to their regular pay and benefits. More importantly, morale of the soldiers affected and public perception of the Army can suffer if the stop-loss is lengthy or perceived as unjustified, reducing combat effectiveness. A stop-loss can also make it more costly or impossible to meet recruiting goals in order to expand the force, further stressing the available force.
Separations can also be increased when a force reduction is necessary. Separation bonuses and early retirement packages can entice additional soldiers to leave the service, but voluntary separations may not be sufficient or in the proper proportions of specific grades and specialties required. Then involuntary separations can be used. In order to reduce the force, targets are set for each grade and MOS. A centralized board is convened to review the records of each soldier under consideration for separation. Once the board has concluded and the results certified, the soldiers are then notified, and given adequate time and assistance to transition from military life. It takes from one to two years from the decision to conduct a separation board until the selected soldiers are discharged from service.

Soldiers may also be separated for personal hardship, medical, or disciplinary reasons, but these cases are negligible at this scale. Potential combat losses should be accounted for, though fortunately the U.S. has not experienced a significant number of casualties recently which would necessitate special treatment separate from normal separations.

DA civilians are managed similar to a corporate workforce. There is no mandatory service period, and the retirement plan is yearly matched 401k contributions. Civilians are not typically offered bonuses to stay longer or to leave earlier. There is no stop-loss for civilians, but they may be fired or laid off as need to meet force and budget requirements. Although there is not a centralized board process as with soldiers, it can take up to a year to set target reductions among the commands,
who must then select which jobs to eliminate and select the employees to let go. The DA civilians are also given ample time and support to transition from Army service.

Reducing contractor support is even simpler. By not renewing contracts or renegotiating at lower levels of support, adjustments can typically be made to the contractor work force in under a year.

Transfers & Moves
Changing job assignments and duty stations is routine for active duty soldiers. By policy, active soldiers usually move every two to four years, and moves are centrally managed by human resources command. For units in the operational force, moves are synchronized with their readiness cycle. This routine movement makes it easy to move soldiers from the operational force to the institutional force and vice versa.

Army Reserve soldiers who are full time (AGR) are managed similar to the active duty soldiers above. National Guard soldiers and TPU soldiers from the Army Reserve relocate geographically only when they choose to, usually for a new position that allows for their promotion, or for personal reasons. Although they may change jobs over time, these soldiers can elect to stay in the same units for years, providing less opportunity to reshape the force composition.

DA Civilians are stable, consistent workforce compared with soldiers. They have no requirement to move, but like RC soldiers, may choose to for promotion or personal reasons. Since most civilians are in the institutional force, move infrequently, and hired
for their particular skills, there is limited opportunity to reshape force composition through their moves.

Contractors are not managed as individual employees as are soldiers or civilians. Any necessary adjustment in force composition or labor support can be made by adjusting contracts or letting new contracts as necessary.

**Trainees, Transients, Holdees, and Students (TTHS)**
Soldiers, who are not assigned to a unit, whether operational or institutional, are accounted for through the TTHS account. This includes soldiers in training, moving between assignments, hospitalized, in detention, and the Military Academy cadets. This accounting practice prevents units from having to keep soldiers who are no longer available for duty on their roles, which would prevent the units from receiving timely replacements.

2.1.8 **Deployment Demand**
The uncertainty of future deployment requirements is central to the question of military capacity, or in this case Army force size. Current practices give little consideration to this aspect, however, opting for deterministic methods.

**Demand Type**
The requirement for deployed forces may be considered contingency operations or enduring operations. Military operations are typically precipitated by an event such as the aggression of a nation-state or other emerging threat. An Army response, if chosen, usually includes a rapid build-up of forces and decisive operations followed by
stability operations and transition. Contingency operations (in colloquial use, not USC 10 definition) refer to the initial and decisive period of response to the actor, usually marked by high intensity and limited duration. Thus the initiation of contingency operations is inherently unpredictable. Enduring operations are usually peacekeeping or stability operations, which may include counterinsurgency. Whether following a contingency operation, as above, or as a stand-alone requirement, enduring operations are more predictable. They also likely take place over a longer period of time, necessitating the periodic rotation of forces to provide continuous support. Enduring operations should end with a successful transition of control to local or allied forces.

From the cold war until recent times, National Military Strategy has focused the Army on large scale contingency operations. Other types of operations, such as peacekeeping or stability operations, were considered “lesser included” operations. It was assumed that if units were prepared for a high intensity, large scale conflict, they would inherently be ready to respond to other types of conflict as needed. Enduring operations such as those in Bosnia or Sinai were relatively small. Recently, however, the Global War on Terrorism challenged the assumption of enduring operations as “lesser included” operations. Sustaining a continuous rotation of forces to the Middle East became the primary mission. This strained Army systems and forced the Army to restructure how it supplies forces.
Force Generation

Force generation is used to describe the policies and systems governing the supply of available forces for deployment. Because deployment demand can fluctuate more rapidly than the force can expand and contract, force generation policies seek to provide an adaptive supply of units to best meet demands with the force on hand [7].

Previous policies include individual replacement and tiered readiness. Under individual replacement, once a unit is deployed it remains for the duration of the conflict, being sustained by rotating individual unit members into and out of the theater. This creates a continuous state of personnel turnover, reducing readiness and lacking the unit cohesion of current policies. Tiered readiness maintains units at different readiness levels determined by their planned contingency deployment timeline, with the first deployers at the highest readiness and later deployers lower readiness. Tiered readiness worked well to supply deploying units for contingency operations at low cost, but could not provide the continuous supply needed for recent enduring operations in the Middle East. Furthermore, tiered readiness created a culture of “haves” and “have-nots” with early deploying units receiving the lion’s share of resources.

The current readiness policy is called ARFORGEN, for Army force generation. It prescribes that operational units progress through three phases: reset, train/ready, and available; repeating to form a cycle, illustrated in Figure 2.2. Units in reset undergo high personnel turnover, supplies and equipment are reconstituted, and many personnel attend offsite professional
development training. Throughout the train/ready stage, units conduct individual and collective training to increase their mission proficiency and deployment readiness. Units may be deployed or mobilized at this stage if necessary. Finally, units enter the available stage at their highest level of readiness; either deploying as scheduled, or serving as a contingency force should the need arise. ARFORGEN is further described in Army Regulation 525-29 [8].

![Unit Progression Diagram]

Figure 2.2 ARFORGEN force pools and readiness progression. Timeline depicted is exemplary only; each phase as well as total cycle time is adjustable.

ARFORGEN provides the flexibility to increase or decrease the supply of available units as necessary by adjusting the duration of each phase. Shortening reset reduces non-available time, but increases pressure on personnel, equipping, and training systems resulting in some combination of increased cost, reduced standards, and more stress on the force. Lengthening the reset phase has the opposite effects, but an overly
long reset is wasteful, providing no increase in readiness while incurring costs; and can result in reduced morale through lack of mission and increased busy-work.

Likewise, decreasing the train/ready phase increases the proportion of available time at the cost of additional pressure on the units, support systems, and training systems. This also may result in increased cost, reduced standards, and more stress on the force. An overly long train/ready phase will also increase costs both for payroll and repeating training to prevent skill atrophy. Soldiers may also grow complacent or apathetic if the training schedule is not challenging.

Decreasing the available phase directly reduces the proportion of time each unit is available, thereby reducing the number of units available for a given force pool. Lengthening the available phase makes more units available for the same force pool. Newly deployed units tend to be energetic and alert, but also unfamiliar with their new surroundings and potentially inexperienced. After several months of deployment, units are at the peak of their effectiveness. Toward the end of a long deployment, soldiers become overly complacent and more accident prone. Deployment length, set by the available phase, therefore, has a direct impact on casualties.

Finally, the total cycle length is determined by the sum of the phases above. Each time a unit deploys, it takes time and resources to transport the personnel and equipment to the battlefield. There is an overlap period while the new unit is being acclimated and assuming responsibility from the departing unit. The unit undergoes another transition and transport period to re-deployment to home station. The
transportation and transition periods count against deployment available time, even when the unit is not actively fulfilling a deployment requirement. Therefore, the longer the cycle time, the greater the portion of units that can fulfill deployment requirements.

This ARFORGEN description applies to both active and reserve units, but there are additional considerations for the reserve component units. Reserve component units are on longer cycle times than active component; most of the reset and training phase is accomplished part time. Then reserve units are mobilized or activated to active duty status to complete collective training and prepare for deployment. After returning from deployment, the reserve unit must demobilize by receiving the returning equipment and completing personnel actions before returning to part time status. The mobilization and demobilization time counts against the reserve units’ available period, effectively adding overhead cost and time to their ARFORGEN cycle.

In order to simplify discussion of various ARFORGEN cycles, it is common to specify the ratio of available to not available time as well as the total cycle time. For example, a 1:3 ratio in a 36 month cycle means that the unit is available to fulfill a deployment requirement for 9 months every 3 years.

The ARFORGEN cycle is meant to be adapted to fluctuating deployment demands. Current policy prescribes a steady state rotation, a surge rotation, and a full surge. Steady state rotation is the preferred operating mode, used when supply meets or exceeds deployment demand. The Army can sustain a steady state rotation indefinitely. A surge rotation is used to meet increased deployment demand by
increasing the deployment availability ratio. This places additional strain on Army personnel and systems while also requiring additional resources. A full surge would be used when deployment demand exceeds rotational capability. In such a case, resourcing and deployment policies would be re-evaluated and adapted to address the situation [8].

**Fulfilling Deployment Demand**

Once tasked to provide forces to a Combatant Commander, the Army determines its ability to supply the required forces, and selects which units and personnel to provide. If adequate forces are not available, additional military and non-military options may be generated, including increasing resources, accepting additional risk, or modifying policies.

It is within the Army’s purview to designate which units to deploy. It may be necessary to substitute an available unit type for one that is not, such as an ABCT instead of a SBCT if there is a shortage of Stryker units. Also, the component (active or reserve) must be selected, along with civilian and contractor augmentation. There are various constraints on the number of each type selected, beyond the inventory level available. Although active Army units should be the least expensive option, not all skills and unit types are resident in the active Army. In addition to reducing operational stress on the active component, reserve units are often used both for political reasons as well as to maintain deployment and warfighting skills. It can also be beneficial to use civilian and contract labor in supporting roles in order to augment and reduce stress on military
support units. This can keep more military units available to operate in more dangerous areas, respond to future contingencies, or reduce the number of support units, effectively shifting more resources to combat units. The selection of the deployment force is, therefore, a non-trivial task.

It is reasonable to consider that not all deployment demands will be satisfied. This occurred recently during surge operations in Iraq and Afghanistan. Although commanders were requesting additional forces, the costs and detrimental effect of further stressing Army forces did not warrant increasing the deployment levels. There also seems to be a fundamental distinction between force requirements for operations where the U.S. is already militarily engaged versus emerging contingency operations where military options might be useful. Ongoing operations tend to take precedence, and national leaders may be forced to address the emerging situations with non-military or limited military options. This helps to effectively limit the potential peak demand of multiple contingency operations coinciding in time.

2.2 Army Force Development
This section provides both an overview of the Army’s force development process within the context of the greater national defense framework, and a synopsis of research topics related to Army force structure analysis.

2.2.1 DoD Guidance and Constraints
Army force structure design is based on requirements ultimately derived from national strategy to include The National Security Strategy [9], The National Defense
strategy [10], and the national military strategy [11]. DoD uses national strategy as an
outline to develop guidance and directives to coordinate its own staff as well as the
military services. This section provides context for the force development process, as
well as a description of current Army practices.

The three main decision making systems in DoD are the Joint Capabilities
Integration and Development System (JCIDS); the Planning, Programming, Budgeting,
and Execution System (PPBE); and the Defense Acquisition System (DAS) [12]. JCIDS
establishes joint capability requirements, identifies capability gaps between current
capabilities and required capabilities, and provides non-material (not procurement
related) solutions or mitigation. For example, during the post-cold war transition of the
1990’s, the Army identified a lack of capability to provide a highly mobile armored force
that could be rapidly deployed by aircraft. The result was the Stryker vehicle and related
formations (though this was in large part a material solution) which saw widespread use
in Iraq. The purpose of this capability based planning (CBP) is to develop a force capable
of responding to a broad range of missions [13]. PPBE develops a short to medium range
plan, covering a window two to six years in the future, allocating limited funds in order
to provide the necessary capabilities. This resource allocation plan is called the future
years defense program (FYDP). The Defense Acquisition System selects, develops, and
procures weapon systems and equipment to address validated capability gaps. Each of
these systems must work in concert to provide sound decisions and produce effective
military capabilities.
The Support for Strategic Analysis (SSA) produces approved planning scenarios and related items to provide a common analytic basis or starting point for DoD studies and analysis. Defense planning scenarios are developed by OSD from selected Joint operations concepts [12]. These scenarios represent potential future threats and military requirements in general terms including a high-level summary of the conflict, strategy, objectives, and forces involved. Only scenarios approved by the Office of the Secretary of Defense (OSD), called Defense Planning Scenarios (DPS), are permitted for use to justify future force requirements. The individual military services, left to their own devices, would be apt to choose or generate scenarios that challenge their particular service, thereby making it possible to justify a larger force, new equipment, or greater resources. OSD, therefore, selects and develops scenarios that are consistent with the National Strategy documents.

The current set of scenarios includes 5 surge (major contingency operations) and about 100 small scale, or steady state events. The Joint Staff further refines the scenarios to produce a multi-service force deployment (MSFD) or concept of operations (CONOPS) which includes forces necessary for a successful conflict resolution. Further analysis by OSD Cost Assessment & Performance Evaluation (CAPE) produces baseline data for the scenario which contains the logistically feasible force flow, or a list of forces required over time. These products serve to coordinate analytic efforts across DoD systems [12].
Integrated Security Constructs (ISC) are assembled from a selection of both steady state and surge scenarios arrayed over a set time period, as shown in Figure 2.3. This provides military planners with a time indexed list of force requirements, facilitating analysis within a global context allowing for multiple overlapping scenarios [12]. The scenarios or ISC provide the analytic foundation for determining capability requirements and ultimately force structure. It is worth noting the importance of the assumption that chosen scenarios are representative of the probable future. If the scenarios omit or over-represent a potential threat or operation type the resulting analysis may be skewed.

![Figure 2.3 Example integrated security construct (ISC). The diagram shows a smooth transition from current operations to foundational activities. Several contingency operations are layered to produce a realistic potential future force requirement.](image)

The Joint Staff analyzes ISC to determine initial force structure requirements or mitigation strategies to address any shortfalls. They assess whether the services have shortage or excess of capability (force structure), and consider options, such as substituting a Marine or Allied Brigade for an Army Brigade, or adjusting missions by
lengthening air strikes to delay a ground offensive. The force analysis, among other factors, also influences funding levels determined through the PPBE process. Although accepted practices continuously evolve, currently it is normal for OSD to direct the number of major combat units in the Army force structure, such as brigade combat teams (BCTs) or headquarters. The detailed planning of the force structure is then left to the Army.

Through analyzing the scenarios or ISCs, an optimal force can be found to respond to the specified scenario. The military services are rarely provided the resources to build such a force. Therefore, they ascertain the best force possible within resource limitations. The use of scenarios is a simplifying assumption to make the planning and staff work manageable while also making analysis feasible. The inherent assumption of this analysis, however, is that the future which will come to pass is similar to the analyzed scenario. This is akin to selecting several arbitrary reference scenarios from infinite possibilities and optimizing the system for that scenario set. In reality, the military tends to remain prepared for the last war it fought rather than adapting for the future.

It is clear that the Army receives direction in a variety of venues. DoD provides: military strategy and general guidance to prioritize and focus missions, validated scenarios and CONPLANs to use for analysis, a directed portion of the force structure, and a tentative funding limit or budget. Congress also specifies the total force size by
component, as discussed in the previous section. Each of these directly bears on the Army force structure problem.

2.2.2 Army Force Structure Development

Force structure development takes place within the context of the broader Army force development process.

The Force Development Process is used to identify requirements, build organizational models, define the total force structure required to meet the National Military Strategy, and document authorizations. It consists of defining military capabilities, designing force structures to provide these capabilities, and translating organizational concepts based on doctrine, technologies, materiel [military hardware and supplies], manpower requirements, and limited resources into a trained and ready Army. [14]

The force development process begins with identifying necessary Army capabilities and designing or modifying organizations to provide those capabilities which aligns with the JCIDS process above. The proper structure, equipment, and personnel are determined as well as the supported/supporting relationship to other units. Total Army Analysis (TAA) is the analytic process to determine the number of each type of unit which will constitute the force structure. Unit designs are considered fixed inputs to the analysis. The force development process concludes with documentation of the force and update of databases [14]. This thesis pertains to the analytic determination of force size, which is related to the TAA process.
2.2.3 Historic Context

The Army (and the rest of the DoD) use defined planning scenarios to consider possible future conflicts because planning for infinite possibilities is impossible. Using approved planning scenarios also provides a common analytic basis for all the military services. The importance of considering multiple possible scenarios has long been recognized in order to design a robust force, which is capable of responding to various contingencies without foreknowledge of the mission type, size, or timing of the requirements. A stochastic process can be defined to select which individual scenarios occur, arrayed in time, in order to create possible futures for planning and analysis. This is predominantly done through discrete event simulation. This section provides an overview of Army force structure techniques developed to address uncertainty in the post-Cold War era.

During the Cold War, the U.S. military was designed to counter a singular and well known Soviet threat. The U.S. had a large Army, and assumed that this Army, designed to defend Europe against the Soviet threat would also be capable of responding to any smaller conflicts or emergencies as the needed. This belief was supported by operational successes such as Operation Just Cause, a small contingency operation in Panama, and Operation Desert Storm, a major contingency operation (MCO) in Kuwait and Iraq.

After the Cold War and dissolution of the Soviet Union, the U.S. Army (and rest of the U.S. military) was reduced in size. The Army remained focused on fighting major
theater wars, though this period was marked by a dramatic increase in the number of smaller operations [15]. These operations were referred to as operations other than war (OOTW) or more recently, small scale contingencies (SSCs) or foundational activities. These operations can range from humanitarian aid missions and peacekeeping operations, to military strikes and interventions, though clearly smaller than a major theater war.

Many of these SSCs required different proportions of unit types than the MCOs for which the Army was designed. The frequent SSCs and smaller overall force required some unit types such as Civil Affairs, Special Forces, and Military Police to deploy repeatedly. The assumption that capability to respond to SSCs was inherent in the force designed for major combat operations was invalid [15].

An early approach to addressing the SSC requirements was conducted by the Joint Staff via a series of war games in the mid-1990s [16]. The war games considered a possible future created by translating the occurrence of previous events onto a future timeline, preserving their order and spacing. The participants then planned what forces would be required to respond to each of the events in the notional future scenario. Though crude, this provided a baseline of SSC activities to be considered in addition to any planned MCO requirements.

2.2.4 Total Army Analysis
TAA is a rigorous analytic process that develops a force structure solution to provide the necessary forces within the given constraints. This force structure, once
approved, is used in the Army PPBE process, which allocates the funding for all Army programs. TAA analytic procedures are improved each year; current practices as well as new initiatives are summarized below.

*TAA determines the requirements (number and type of units) ... in the active Army, ARNG, and USAR components.* ... Through TAA, the Army provides the combatant commanders with the best force structure capabilities within allocated resources to execute the National Military Strategy and defense planning guidance tasks. TAA takes into account force guidance and resource availability to produce a balanced and affordable force structure. [AR 71-32]

The process comprises several steps, beginning with campaign analysis. The OSD support to strategic analysis (SSA) products provide the context for analysis. The CONOPS are refined to add necessary detail about Army missions and forces. A combat simulation is used to evaluate the campaign, and ensure combat forces are adequate to resolve the conflict. A deterministic, theater-level simulation is used for this purpose, named the Joint Integrated Contingency Model (JICM). Simulation results may be used to select the most favorable military course of action, and update force requirements as necessary for an iterative solution. The output is a detailed operational plan including forces and supplies required, outcomes, and expected losses [17].

Enabler force analysis builds on the results of the campaign analysis. Enablers are the support forces, such as supply, engineer, military intelligence, and medical units that facilitate the main combat forces (BCTs), in successfully completing their mission.
These types of units are not considered in the campaign analysis. This step determines the quantity and type of enabler units needed for the campaign, based on rules of allocation. Given the combat force (directed force) from the campaign analysis, the enablers are added based on the existence of combat units (such as one support brigade for every three BCTs) or workload (such as truck companies based on the amount of ammunition and supplies needed and distance). Enabling forces may even be added manually when the existence or workload rules are insufficient. The force generator model, or FORGE, is a deterministic theater-level model that reaches a solution iteratively. Once enablers are added, they generate their own support and consumption requirements, which in turn call for additional enablers, continuing until the solutions converge [17].

Next, rotational analysis applies a specified force generation policy (in this case, ARFORGEN) to determine the forces necessary to meet operational demand. For a system at steady state with evenly distributed forces, this is a relatively simple calculation, but dynamically representing the rotational forces over time is accomplished by the MARATHON simulation. The key model inputs are: demand (an integrated security construct (ISC) generated from the scenarios analyzed above), policy rules for force generation (fixed for period of analysis), and the supply of units (specified force structure). This process may be used for two different purposes. First, if a force structure is not fully specified, the model will dynamically create units as they are required. This provides the unconstrained force structure requirement. With the force
caps in place, the model will stop generating units once the end strength constraints are reached and record missed demand. The constrained force results are analyzed to compile candidate feasible forces. Second, the model can be used to evaluate a specified force and determine the portion of demand satisfied, as well as the nature of the demand that was missed. The candidate forces can then be evaluated and compared based on their simulation performance to choose the best feasible force structure [17].

The deterministic model will always yield the same answer if given the same input, so the scenario or ISC is varied in order to provide a measure of robustness to the solution. For each ISC, two surge scenarios are randomly selected with random start times, such that the entire event will fit within the analysis period of 15 years or more. These are overlaid on a fixed series of foundational activities. The foundational activities are a selected subset of the small scale scenarios, arranged across the analysis period. By varying the surge scenarios selected, as well as their timing (so they may overlap), many possible ISCs are generated. Subsequently, many candidate force structures are generated, one from each ISC. The candidate force structures are then evaluated using a common testing set of ISCs for comparison.

A substantial portion of the Army force structure is determined outside the analytic process described. The majority of combat forces (BCTs, etc.) are directed by OSD, although the Army is a participant in this process. The Army staff determines force levels for the number of soldiers in schools (not assigned to a unit), providing support
for other services, and serving in the Army generating force. These levels are set primarily by adjusting from past levels, based on need and constraints. This is the practice because reliable models that adequately incorporate each of these aspects have not been developed. That leaves less than 10% of the force to be determined through modeling.

2.2.5 TAA Based Approaches
This section describes research efforts to extend the TAA force structure analytic capabilities. Many have already been incorporated in the current methodology.

SADE - Stochastic Analysis for Deployments and Excursions
Dubois and Kastner from the Center for Army Analysis developed a method to forecast future SSC events based on historic data [2]. Their method, named “Stochastic Analysis for Deployments and Excursions (SADE)”, is based on queuing theory. In the model, the occurrence of an event requiring military response is the calling population, and Army response is the service mechanism. The inter-arrival time of events and service time are modeled from the historic data. The study used data from the post-Cold War period from 1990 to 1997.

The SADE method defines five steps. First, the historic contingency data is collected, reviewed, and categorized. In this case, the analysts collected the start and end dates of each event, geographic region, and maximum level of soldiers deployed for that event (because more detailed troop levels and unit types are not typically
available). The events were categorized by mission type and size to balance the desire to accurately model distinct SSC types with the need for significant sample sizes.

The second step develops models based on the data using goodness of fit tests. An exponential model of inter-arrival times was shown to fit the data well. The duration of SSCs by type was also modeled. End dates for ongoing operations were estimated given the current elapsed operational time. Exponential, Weibull, gamma, and log-normal distributions were used to model duration for specific category types. The frequency of event types is also estimated from the data set. SADE used a log-linear model to account for the multi-factor categorical data of region, contingency type, duration, and interactions. The model developed incorporated the primary factors for region, mission type, and duration in addition to a region-mission type interaction term. This approach is not necessary for single factor data or if interactions can be ignored.

The third step builds the simulation model. A discrete event simulation was constructed using a commercial simulation software package. Events are generated with an exponential inter-arrival time, categorized by an empirical frequency distribution, and duration is determined by the type-specific distribution fit in step 2. This provides the capability to quickly generate many probabilistic futures for analysis.

The fourth step was to run the simulation, verify, and validate the model. Simulation output was compared to historic events and found to have similar behavior. Additionally, the SADE study conducted a separate time series analysis using integrated
auto-regression and moving averages. The time series analysis yielded similar results, providing additional verification of the simulation model.

The final step was output analysis, dependent on the study issues. The SADE simulated SSC output was used by both the Army and OSD in the planning of force requirements. When the simulated demand is used to derive force requirements, as in the TAA process, the risk of the design may be defined as \((1 - p)\), where \(p\) is the percentile of the simulated future demand used to generate the force requirements ranked from the fewest to most SSCs.

This model was validated by the Army and used to forecast the frequency and type of SSCs used for force planning. It was quickly expanded upon to not only generate possible future event timelines, but also translate them into force demand levels. That extension is the subject of the “Stochastic Analysis of Resources for Deployments and Excursions (SARDE)” by DuBois at CAA [18].

**SARDE - Stochastic Analysis of Resources for Deployments and Excursions**

The SARDE methodology follows the same procedure as SADE, with several sub-steps added in order to translate the SADE event list into a force demand by unit type, or SRC. During the second step, while fitting the probability distributions, it is also necessary to determine the forces necessary to respond to each category defined. These notional deployment forces are referred to as the mission task organized force (MTOF). For the initial implementation, MTOFs were not available for each SSC category, so the analysts substituted historical deployment data where necessary. The historic force data
was less desirable than the planned MTOFs because the historic deployments may not be representative of future requirements or the category in general.

The multi-factor log-linear categorical model used in SADE was not carried over to the SARDE study. Instead, the SSCs were classified by mission type and sub-divided by size. Category likelihood was then determined by simple proportion of observations within that category.

The simulation development from step 3 must also incorporate the MTOF data so that when an event occurs and is designated by category type, the proper force demand is collected among the statistical data generated.

The SARDE study was used to inform the TAA process, and showed that significant shortfalls were expected in some unit types, predominantly support units. Also, the SARDE risk metric is extended to address specific unit types and their likelihood of meeting deployment demand.

There are several considerations for using this type of approach [13]. It assumes that the occurrence of future contingencies will be similar to the past, but historic factors may not be applicable for the future. The accuracy of data will affect results, and it is difficult to obtain complete data as well as determine which data should be included. Design decisions such as category definitions are subjective, and will also affect outcomes. The forces used in past conflicts may not accurately represent future requirements. The data does not typically include instances when a requirement was missed or avoided for lack of forces or other reasons.
The SADE and SARDE methodologies are outlined in greater detail in chapter 4 of [13]. These studies from CAA provide the basis for generating stochastic deployment demands for SSCs still used in TAA today, and which we will draw upon for this research.

**Incorporating OOTW in Force Structure Analysis**

A stochastic programming based methodology was developed by Loerch and Coblentz [15]. They present an analytic tool to aid the process of developing a feasible force structure within manpower constraints given a force structure requirement that exceeds manpower limits, as in the second phase of the TAA process.

This approach assumes that the capability to respond to two major theater wars is inherent, and therefore a fixed requirement. It stochastically generates SSC requirements via the SARDE method. A two stage stochastic program determines forces to satisfy the fixed major theater war requirements in the first stage, and the stochastic SSC requirements in the second stage. The objective is to minimize the penalty costs for missed demand. The penalty costs are specific to each unit type and conflict/mission to reflect a weighting priority, which inherently requires an element of subjectivity.

The model would provide a measure of which missions were well covered, and which assumed risk, as well as indicate which unit types had shortage or surplus in inventory. This method was not fully implemented.

**RANGER IPOD – Randomly Generated Requirements Informed by Past Operational Deployments**

Despite the ability to produce stochastic future SSC deployment demands, force analysis continued to use only a single fixed profile for computational reasons. The
variation of SSCs was considered less important than major contingency operations, so the type and initiation time of MCOs or surge events was varied arbitrarily, while demand profile for foundational activities was fixed. Helms and Stoll, also from CAA, introduced “Randomly Generated Requirements Informed by Past Operational Deployments (RANGER IPOD)” to incorporate stochastic SSC demands into the TAA process [19].

The RANGER IPOD method generates stochastic future SSC demands in a similar manner to SARDE. Instead of using an arbitrary categorization and deriving mission task organized forces (MTOFs) for each category, the process is more formalized through the use of Steady State Security Posture (SSSP) vignettes, or approved defense planning scenarios for SSCs. Event frequencies are still based on historic conflict data, but the analyst must decide which SSSPs to include, and match the categorized historic conflict data to the categorized SSSP set.

The RANGER IPOD methodology was coupled with a separately developed simulation methodology to generate future demand profiles, generate the rotational force structure requirements to meet those demands, and evaluate the alternative force structures. Each of the force demand profiles generated by RANGER IPOD would be used to develop constrained force structures using rotational force simulation, as in TAA. Each of the force structures would then be used as input for capacity analysis, again using TAA methodology, against a common suite of demand profiles. The force designs that met a higher percentage of the overall demand across the set of possible
futures were considered more robust, or better able to meet the future uncertain force demands.

In order to investigate the importance of stochastically generated SSCs, surge events or MCOs were held constant. The historic data was partitioned into a post war reduction period, peaceful steady state, and growth during persistent conflict. Therefore, the nature of the future demand could be selected by choosing the most appropriate historic data as input. The RANGER IPOD study considered three cases. In the first, the set of SSC events is set, but the timing and sequence is varied. For the second, the duration of SSC events was varied in addition to timing. In the third case, the frequency, size, and type of SSCs were derived from the historic conflict dataset, similar to the SARDE procedure. Each of these cases was compared to the base case, the force structure developed with a single fixed demand for foundational activities.

The RANGER IPOD analysis showed that the array of SSCs used to develop the current force structure was not representative of historic data in terms of event types, duration, and size. It also showed that each of the three alternatives outperformed the base case where the SSC events were predetermined. Varying the event duration and timing did not show any improvement over varying timing alone, but the third case where frequency was also treated stochastically yielded the greatest improvement.

**FMCA- Force Mix and Composition Analysis**

Force Mix and Composition Analysis (FMCA) is an ongoing CAA effort to provide analytic decision tools to the TAA resourcing panels which are part of the more
qualitative second stage of the process [20]. Its objective is to provide risk-informed force level recommendations to the panels. The resourcing panels, in turn, recommend to Army leadership which of the force requirements should be funded given manpower constraints. FMCA employs the RANGER IPOD methodology for generating stochastic foundational activities, along with an enduring counter-insurgency conflict and randomly timed MCO and homeland defense events. Each of the demand profiles is used to generate a force structure, and the resulting force structures are then evaluated for performance across a common set of possible future demands, similar to RANGER IPOD, above.

The analysis considers the variability in demand of each unit type, and the ability of the force to satisfy that demand both overall, and at peak demand levels. Recommended floor and ceiling fielding levels are created for each unit type based on the observed simulated demands to establish acceptable risk levels across each of the unit types. Units with a recommended band above the current design levels should then be considered for expansion, while units with a recommended ceiling below the current design levels should be considered for reduction in order to create a robust force structure, capable of responding to a wide variety of possible future demands.

2.2.6 Other Stochastic Force Structure Approaches

**RAND SLAM**

The RAND SLAM program was developed by RAND National Defense Research Institute for the Office of the Secretary of Defense (OSD) [21]. It is a tool for analyzing
the performance of both active and reserve component rotational forces in a stochastic environment. In contrast to the simulation based approaches above, this method is optimization based.

The program requires user inputs including force structure, scenarios, deployment rules, and costs. Multiple deployment contingencies may be defined, designating states (such as peace, war, and transition) with Markov transition probabilities. Deployment rules govern the selection and behavior of units for deployment. Unit operating costs are also needed for each system state.

The program calculates arbitrary “disutility” factors for each force type, deployment status, and period combination; these factors are consistent with the specified deployment policy. The objective of the program is then to minimize the disutility of deployed units, thereby developing a deployment plan that best meets the guidance and rules provided.

The program may be solved as a linear program or as a mixed integer program. In either case, a heuristic is applied sequentially to a sliding window of periods to create a deterministic deployment demand. The first period of the analysis window is determined stochastically from the preceding period state using the transition matrix. Subsequent periods in the analysis window are then determined by a user selected heuristic. The first is most likely, where the transition probabilities are calculated for the next period, and the most likely (highest probability) state is selected, for each period in the analysis window. The other heuristic is tolerance, where the most demanding
scenario from within the most likely quantile specified is selected. Deploying units are selected, inventories and states are updated, and the program advances to the next period.

Average cost per period, unmet demand, and unit dwell time provide useful metrics for comparing force designs. The illustrative example included in the report shows that the results from a stochastic treatment of the deployment demand can be dramatically different from a deterministic or steady state analysis which may unduly favor active forces. Under uncertainty, reserve units may provide additional value with cost-effective additional capacity to deal with contingencies.

The RAND SLAM program requires a large amount of user defined input without offering a methodology for developing that input, namely the force structures to be considered, the contingencies to use, and the transition matrices to govern the stochastic states. It does provide an example of optimization in force structure analysis as well as a potentially useful heuristic for reducing the dimensionality of stochastic problems.

**GTO – Generator to Operator Model**

A growing interest in functions and resourcing of the Institutional Army has led to series of reports and projects commissioned by the Army and provided by the RAND Corporation [17],[22],[23]. The latest of these produced economic input-output model, treating the Institutional Army as an economy, and tracing the inputs required to the outputs produced in terms of manpower equivalents. These outputs are required by the
Operational Army, and entities external to the Army, such as families, veterans, and Joint organizations.

Every unit in the Institutional Army was consolidated into one of about 180 activities, each with a distinct purpose. An un-capacitated network flow model was constructed with the activities as nodes and supporting relationships as arcs. Input manpower adds to the flow (sources). The value of manpower (flow) is passed between activities through the supported/supporting relationships defined by the network arcs, where cycles are not permitted. There are about 110 outputs modeling the various commodities and services provided by the Institutional Army. These accumulate the manpower value flow in the model (sinks).

The GTO model used Army force structure data for all personnel authorizations in the Institutional Force. GTO incorporates Active and Reserve Soldiers, DA Civilians, and Contractors by converting all authorized billets to Active Duty E-4 (specialist rank) equivalents by using Army personnel cost factors. These manpower equivalents provide values for the model, and provide a means for determining the division of support between agencies; based on proportion of manpower within the agency. The GTO authors recognize that this apportionment of support is not always accurate, and posit that it is not unreasonable for the Army to require units to report their division of support to other units and output functions, thereby providing better data for the model.
The GTO model offers a way to analytically determine how that Institutional Army should change in order to meet changing demands of the Operational Army and external requirements. This is in contrast to the current decentralized and non-objective method of units self-reporting their requirements. The authors show how the institutional changes recommended by the model may be compared with those implied by the President’s Budget. Any significant differences indicate opportunities to improve the model, improve the budget, or both, thus focusing the budget analysts’ efforts.

Although the GTO model accounts for all types of manpower, it does not distinguish between them once aggregated; all recommended changes are proportional to the original activity composition. The model can evaluate and recommend changes in institutional force structure based on changing requirements and resourcing, but it is a steady state model of the Army, and does not account for dynamic changes. Furthermore, the GTO model is deterministic, and does not inherently consider uncertainty, though input values may be manipulated to consider various cases.

Although considered for incorporation into the TAA process, those efforts have been suspended due to lack of reliable data. Instead, CAA is pursuing a linear regression model of institutional to operational relationships.

**SARA – Stochastic Active-Reserve Assessment Model**

The Stochastic Active-Reserve Assessment (SARA) Model was developed by the Institute for Defense Analysis (IDA) for the Office of the Secretary of Defense (OSD) to study cost and capability implications of force mix (active vs. reserve component
composition) and readiness policy [24]. The SARA model uses simulation to evaluate force structures with a stochastic deployment demand. It allows analysts to consider active and reserve unit force mix, force composition by unit type, force size, and readiness policies. Although originally formulated for Army force structure, additional services are being incorporated.

The SARA model requires input data for the contingencies to be used, the force structure to be evaluated, readiness and deployment policies, and unit operating costs. The study suggests a SARDE-like approach to developing scenario data, though the model will accept any user specified data that is properly formatted. The force structure to be evaluated must be specified along with costs. The authors suggest using a brigade-level notional force structure, where the Army force structure is simplified to 20 different brigade-level aggregated unit types. This approach provides a useful compromise between considering aggregate troop levels or brigade combat teams only and analyzing the entire Army force structure with thousands of individual unit types. The force generation and deployment rules control how units are cycled through readiness phases and selected for deployment. Unit operating costs were derived from official Army costing models as well as in-house models.

The model uses simulation to generate stochastic requirements and explicitly manages the readiness progression and deployment rotation of units to determine cost, utilization rates, and the ability of a force structure to meet aggregate deployment demands.
The authors use an illustrative analysis to demonstrate the model’s utility. They show that unit utilization is not uniform for the current force structure; some units were considerably stressed while others were not. Their analysis also indicated that the active/reserve force mix has similar overall costs across a broad range of values. Finally, there is a potential for savings by using reserve units in a strategic role rather than in a cyclical readiness policy like the current ARFORGEN policy.

The SARA model provides a useful approach to consider aggregated notional unit types without representing thousands of unique elements. The authors also use a Pareto frontier to express the relationship between cost and risk, which is helpful to avoid subjectively assessing a penalty or cost of missing demand.

2.2.7 Analytic Capability Gaps
The models and methods outlined here provide a fair range of analytic capability to address force structure issues. We can generate stochastic deployment demands using defined scenarios with force requirements based on historic event frequencies. Those demand profiles can be used to generate rotational force requirements as well as assess the ability of a specified force structure to meet a fixed or stochastic demand and assess the risk, or probability of not meeting demand. We have seen examples of both simulation and stochastic programming to aid in translating force requirements into a feasible manpower constrained force structure. Institutional support requirements can be derived from the Operational force structure. We can assess the cost and
performance of specified force structures in order to improve the active / reserve component mix, reduce cost, reduce risk, or modify readiness and deployment policies.

It is worth noting that the models developed by RAND and IDA, both external to the Army, incorporate cost, while all of the TAA based approaches deal with manpower constraints and no explicit cost representation.

Despite the current capabilities, there remain important areas that have yet to be adequately addressed. There are two significant gaps in the current Army force structure analytic capability that we address with this research: the ability to analyze changing force size and readiness policy dynamically over time, and the ability to consider the entire Army workforce simultaneously.

In an ever changing world with developing technologies, evolving threats, competing demands for resources, and short term political goals, we should expect that the Army, as a system, will never reach equilibrium. In fact it is a requirement that the Army be capable of expanding and contracting to meet the nation’s requirements. Furthermore, manipulating the force readiness and deployment policies is the only way to adjust unit availability in the near term and provide flexibility to the force generation process [7]. Each of the analytic methods described here, however, only considers fixed force structures and readiness/deployment policies. Army personnel models can estimate the feasibility and cost of expanding or contracting the force size at a particular rate, but that is currently external to rather than integrated in force structure analysis.
The capability to analyze a dynamically changing force size is important in order to determine how quickly to grow and by how much in times of war as well as how quickly to reduce force size and by how much in times of peace. Accounting for flexible readiness and deployment policies is also important, otherwise, analysis will tend to yield overly conservative (or expensive) force structures in order to meet peak deployment demands.

The second gap is the need to address the entire Army workforce. The Operational Force comprises the deployable combat units, while the Institutional Force handles the training, equipping, administrative, support, and management functions of the Army. The two are inextricably linked; the Institutional Force creates and supports the soldiers and systems for the Operational Force, which in turn provides the requirement for the Institutional Force. It would be imprudent to change the Operational Force size without considering the capacity of the Institutional Force.

The Army employs soldiers, both Active and Reserve, as well as civilians and contractors in order to meet its labor requirements. Although the vast majority of positions in the operating force require soldiers, a great many functions in the institutional force could be filled by a variety of personnel types. Soldiers may be reassigned from the institutional force to the operating force, and vice versa. Contractor support has played a critical role in recent conflicts, with contractors on the battlefield outnumbering deployed soldiers at times[25]. It is inefficient and limiting to consider
any one labor type in isolation or to analyze the operational force separately from the institutional force.

The GTO model provides a linkage between the Institutional and Operational Force. It accounts for all personnel types, but aggregates them, thereby offering no advice on how the distribution of the work force should change. Each of the other analytic methods reviewed above address only active and reserve component units of the operational force. The ongoing regression analysis at CAA will provide a useful model of institutional requirements, but again, not necessarily how to balance the composition of the institutional force. No model or tool that we are aware of considers both the operational and institutional force together with each personnel type in the context of force size or force structure.

Modeling the entire workforce makes a meaningful analysis of the total force size and composition possible. Furthermore, it is necessary in order to provide a reasonable treatment of a dynamically changing force size. These capabilities will permit the investigation of the cost and capability of a force, as well as how the size and composition of that force should change over time to efficiently provide necessary defense forces. This will enable us to address our central study questions: “What size should the Army be?”, “What composition should the Army have?”, and “How should the Army size and composition change over time?” This approach will also enable us to consider both the Army end strength constraint for soldiers and the budget constraint
for funds to determine which is more restrictive, and how the Army may want to relax that more restrictive constraint.

2.3 **Stochastic Programming**

“Stochastic programs involve an artful blend of traditional (deterministic) mathematical programs and stochastic models.” [26]

This section provides a brief overview of stochastic programming, introduces the terminology and notation used throughout the rest of this report, and reviews related research topics. An introductory tutorial on stochastic programming is available from Higle [26], and a thorough treatment of the subject is given by Birge and Louveaux [27]. This thesis investigates the application of stochastic programming to address the Army force size problem (AFSP), and therefore focuses on multistage stochastic programs (MSSP).

2.3.1 **Stochastic Programming Overview**

Stochastic programming has been closely tied to linear programming and other deterministic forms of mathematical programming since its introduction by Dantzig in 1955 [28]. It is distinguished, however, by its ability to extend deterministic mathematical programming models to explicitly address uncertainty in the form of random variables with a probability distribution.

Stochastic programming also differentiates between decisions that must be made under uncertainty, and decisions that can benefit by waiting until new information resolves some or all of the uncertainty. This temporal aspect of making
decisions, re-evaluating, and making further decisions with new information is referred to as recourse in the model. The recourse model formulation provides the best immediate decisions required under uncertainty while accounting for possible future decisions as well. This is a distinct improvement over myopic models that may maximize immediate benefit, but produce poor policies in the long run [26].

The basic stochastic program with recourse is a two stage model. It pertains to a problem where a decision must be made under uncertainty, additional information is revealed, and then an additional recourse decision is made. A classic example of this is the news vendor problem. The news vendor must order newspapers from the distributor prior to knowing how many customers will want a paper that day. Over the course of the day, the demand for newspapers is revealed. At the end of the day, the vendor may return unsold papers for a salvage price. The order for newspapers is the first stage variable. The daily demand for newspapers is the unknown random variable. Finally, the number of papers to return for salvage price is the second stage or recourse variable.

The two stage model can be extended by adding additional decision stages to become a multistage model. A common example of a multistage problem is investment, or portfolio, management. In this problem a manager must choose between a set of investment options based on risk and expected return. Only after each investment period does the manager discover his gains or losses. The manager is faced with another set of investment choices for the next period, which may depend on the previous choices as well as the stochastic market outcome. Note that the decision stages need
not include similar decisions or be periodic as in the example above; periodic models are a special case of the multistage class, which will include our problem. Although this is a simple extension of the two stage problem, multistage problems can be considerably more difficult to solve.

An alternate form of stochastic programming uses *chance or probabilistic constraints* which must hold true, but only at a specified proportion of the time. This is typical when dealing with system reliability, or if system failure (infeasibility) does not translate easily into the units of the objective function to add a reasonable penalty [29]. These two approaches are not mutually exclusive, though they are often introduced and employed separately.

In practice, the uncertainty in variables is modeled as a discrete probability distribution, which may be depicted as a *scenario tree*. An example (Figure 2.4) shows the scenario tree for flipping a coin two times with four possible outcomes. Consider that an arbitrary decision, such as calling *heads or tails*, must be made prior to flipping the coin. This tree represents the discrete potential realizations of the random variables at each stage or decision point. In a straight-forward approach, each of these discrete scenarios must be evaluated simultaneously. Continuous random variables are often approximated with a discrete distribution or sampled to create an empirical distribution.
Acceptable solutions must be both *admissible* (capable of yielding feasible solutions for all possible scenarios) and *implementable* (decision variables specified at a decision node are identical for all future scenarios that are not yet distinguished; the model does not take advantage of information that is not yet available). In Figure 2.4 above, the decision at stage 2 for scenarios 3 and 4 must be the same, or it implies foreknowledge of the second coin flip. Implementability is also referred to as *non-anticipativity*, and is an important consideration in multi-stage models. Solutions are typically reported in terms of first stage decision variables only, as subsequent stages are probabilistic and may be too numerous to list.

There are several special classes of recourse problems that have desirable properties. *Fixed recourse* describes problems where the coefficients in the second stage constraints do not vary, simplifying computation. *Simple recourse* arises when the recourse decision can be determined directly from the outcome of the first stage decision, as in the news vendor example. *Complete recourse* requires that second stage
feasibility can always be achieved, such as through a penalty function and slack variable. 

*Relatively complete recourse* requires that a feasible second stage solution exists for every feasible stage one outcome. Simple, complete, and relatively complete recourse guarantee feasibility of the second stage for all feasible stage one decisions which is useful in decomposition methods. This concept extends to the multistage case as well. These cases are described conceptually here, and thorough technical definition is available from [27].

The additional functionality of stochastic programming comes at the price of increased model complexity and the resulting effort necessary to solve the problem. Increasing the number of random variables, the number of possible discrete states, and the number of decision stages causes combinatorial growth, which quickly creates explosive problem sizes.

Solution methods for stochastic programs are based on linear programming and integer programming techniques. Integer formulations can be quite challenging, especially when the integer variables are in the second (or later) stage because feasibility of choices in the previous stage must be considered.

Modest problems can be solved directly by modern commercial mixed integer program (MIP) solvers, solving all scenarios simultaneously via the *extensive form* or deterministic equivalent formulation. Many practical problems require specialized techniques that can exploit the special structure of stochastic programs, including
decomposition and statistical estimation [26]. A more detailed discussion of these techniques follows in Section 2.3.3.

2.3.2 Notation and Problem Representation
Unlike linear programming, there is not a universally recognized method and notation for describing stochastic programs. This thesis will attempt to adhere to the notation used in [27]. An excerpt is included below as Table 2.1. Additional notation will be defined as it is introduced.
Table 2.1 Stochastic programming notation.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>first stage constraint coefficient matrix</td>
</tr>
<tr>
<td>( b )</td>
<td>first stage right hand side</td>
</tr>
<tr>
<td>( c )</td>
<td>first stage cost coefficient</td>
</tr>
<tr>
<td>( E )</td>
<td>expectation operator (expected value)</td>
</tr>
<tr>
<td>( h )</td>
<td>second stage right hand side</td>
</tr>
<tr>
<td>( H )</td>
<td>number stages in multistage problem</td>
</tr>
<tr>
<td>( q )</td>
<td>second stage cost coefficient</td>
</tr>
<tr>
<td>( Q )</td>
<td>second stage value function</td>
</tr>
<tr>
<td>( Q )</td>
<td>second stage recourse function (expected value)</td>
</tr>
<tr>
<td>( t )</td>
<td>stage index (superscript)</td>
</tr>
<tr>
<td>( T )</td>
<td>technology matrix – coefficients of first stage decision variables in the second stage constraints</td>
</tr>
<tr>
<td>( W )</td>
<td>recourse matrix – coefficients of second stage decision variables in the second stage constraints</td>
</tr>
<tr>
<td>( x )</td>
<td>first stage decision variable vector</td>
</tr>
<tr>
<td>( y )</td>
<td>second stage decision variable vector</td>
</tr>
<tr>
<td>( z )</td>
<td>objective function value</td>
</tr>
<tr>
<td>( \xi )</td>
<td>random vector</td>
</tr>
<tr>
<td>( \Xi )</td>
<td>support of the random vector ( \xi )</td>
</tr>
<tr>
<td>( \omega )</td>
<td>random event</td>
</tr>
<tr>
<td>( \Omega )</td>
<td>set of all random events</td>
</tr>
</tbody>
</table>

A two stage stochastic problem with fixed recourse can be represented in classical form:

...
\[
\begin{align*}
\text{min } z &= c^T x + E_\xi \left[ \min_y q(\omega)^T y(\omega) \right] \\
\text{s.t. } &Ax = b \\
&\quad T(\omega) x + W y(\omega) = h(\omega) \\
&\quad x \geq 0, y(\omega) \geq 0,
\end{align*}
\] (2.1)

or implicitly as

\[
\begin{align*}
\text{min } z &= c^T x + Q(x) \\
\text{s.t. } &Ax = b \\
&\quad x \geq 0,
\end{align*}
\] (2.2)

where \( Q(x) \) is the recourse function:

\[
Q(x) = E_\xi Q(x, \xi(\omega))
\] (2.3)

with \( Q \) as the second stage value function and \( \omega \in \Omega \):

\[
\begin{align*}
Q(x, \xi(\omega)) &= \min_y q(\omega)^T y \\
\text{s.t. } &W y = h(\omega) - T(\omega)x \\
&\quad y \geq 0,
\end{align*}
\] (2.4)
The multistage expression adds an additional index, shown as a superscript here, to indicate the problem stage rather than different variables symbols as in the two stage problem. Transpose symbols are omitted to simplify notation.

\[
\begin{align*}
\min z &= c^1 x^1 + E_{\xi^2} \left[ \min c^2 (\omega^2) x^2 (\omega^2) + \cdots + E_{\xi^H} \left[ \min c^H (\omega) x^H (\omega^H) \right] \right] \\
\text{s.t.} \quad W^1 x^1 &= h^1 \\
T^1 (\omega^2) x^1 + W^2 x^2 (\omega^2) &= h^2 (\omega^2) \\
\vdots \\
T^{H-1} (\omega^H) x^{H-1} (\omega^{H-1}) + W^H x^H (\omega^H) &= h^H (\omega^H) \\
x^1 &\geq 0; x^t (\omega^t) \geq 0, \quad \text{for } t = 2, \ldots, H
\end{align*}
\] (2.5)

2.3.3 Solution Methods
Two naïve approaches to solve stochastic programs are *wait and see* (WS) and *expected value* (EV). For the wait and see approach, deterministic methods are used to find the optimal solution for each scenario as if perfect information were available to eliminate uncertainty. The WS value is the average objective function value from all the scenarios, shown in (2.6), where \( \bar{x} (\xi) \) represents an optimal solution under the conditions of \( \xi \). The result is a collection of potential decision vectors that are not implementable, violating non-anticipativity. Selecting any one of the WS solutions
arguments or even averaging them is not assured to yield an admissible (feasible) solution for each possible outcome. The WS objective value, in general, is not attainable, as it requires perfect information.

\[
WS = E_{\xi} \left[ \min_x z(x, \xi) \right] = E_{\xi} z(\bar{x}(\xi), \xi)
\]

(2.6)

The expected value approach assumes the average value of the stochastic terms, and solves the problem deterministically. Again, there is no assurance that this solution is admissible, and the outcome will not be optimal, in general. If feasible, the resulting objective value provides the EV (2.7). The expectation of the expected value problem (EEV) is found by evaluating the EV solution argument for each scenario and averaging the objective function values (2.8), where \( \bar{x}(\xi) \) is the optimal solution arguments using the expected value of \( \xi \).

\[
EV = \min_x z(x, \bar{x}) \quad (2.7)
\]

\[
EEV = E_{\xi} \left( z(\bar{x}(\xi), \xi) \right) \quad (2.8)
\]
The issues with both of these approaches are clear, but they can provide simple useful bounds for the true recourse problem (RP). Solving the RP can be expensive, so it is worthwhile to check the gap between the bounds (2.9) to see if there is sufficient benefit available to justify the cost. Better bounds can be determined by statistical sampling [30] or improving on the classical Jensen’s inequality and Edmundson-Madansky inequality with partitioning [31].

\[
WS \leq RP \leq EEV, \quad (2.9)
\]

Where

\[
RP = \min_x E_{\xi} z(x, \xi). \quad (2.10)
\]

The RP (2.10) can be solved directly using the extensive form [32], provided the capability of the optimization software and system hardware are not exceeded. The high dimensionality of many practical problems and exponential growth of stochastic programming problem size often require more sophisticated techniques. These techniques typically employ decomposition, approximation, or both.
There is a large body of research pertaining to decomposition of stochastic programs, exploiting their structural properties to break the large problems into smaller solvable pieces then recombining them to achieve global optimality.

Van Slyke and Wets [33] developed the L-shaped approach to addressing stochastic two-stage linear programs based on Bender’s decomposition, or outer linearization. This method decomposes the problem by stage, and iteratively adds feasibility and optimality cuts until the optimal solution is obtained. Glassey [34] provides a nested decomposition approach to handling multi-stage problems using Dantzig-Wolf decomposition. This method decomposes by decision branch so that each nested level is one stage shorter than the previous level. Birge [35] presents an alternative nested decomposition method for solving large stochastic multi-stage problems based on an extension of the L-shaped method. This horizontal decomposition breaks the problem along its natural stage boundaries. He shows the advantage of Bender’s decomposition for dealing with large period multi-stage problems. This algorithm is the basis of many current nested decomposition methods. Gassmann [36] made several improvements to nested decomposition including: applying improved subproblem sequencing, incorporating bunching with modifications for various random sources, adapting the multicut generation procedure for multistage problems. Donohue and Birge [37] incorporate sampling procedures in abridged nested decomposition using decomposition and sampling for multistage problems to establish lower and upper bounds.
Progressive hedging, introduced by Rockafellar and Wets [38] provide a robust solution approximation for multistage stochastic problems derived from solving a selection of WS problems, referred to as scenario analysis. This amounts to a Lagrangian relaxation of the non-anticipativity constraints with a penalty for infeasibility. Watson and Woodruff [39] apply progressive hedging as a heuristic for stochastic mixed integer problems, providing an effective way to deal with the issue of integer variables in MSSP. Gade et al [40] recently provided a method to determine the lower bound for the progressive hedging application in MSSPs.

Although there seems to be a prevailing school of thought favoring primal problem formulations and use of Bender’s decomposition, there are many competing examples working with dual formulations and column generation. Mulvey and Ruszczynski [41] provide diagonal quadratic approximation for decomposition of the multistage dual problem which facilitates parallel computing, further developed for the augmented Lagrangian approach by Rosa and Ruszczynski [42]. Higle et al [43] develop the stochastic scenario decomposition algorithm for MSSPs. Their method uses scenario decomposition (column generation) of the dual problem to generate cutting planes, integrated with statistical sampling. They are able to limit the columns of the master problem as sample size grows.

It is apparent that there are a multitude of techniques available to address common problems in stochastic programming, whether the multistage formulation is
simply too large, there are too many scenarios to evaluate, the problem contains integer variables, or the solution space is non-convex.

2.3.4 Other Stochastic Optimization Approaches

There are a variety of other analytic approaches for dealing with decision making under conditions of uncertainty. Stochastic programming, however, appears to be best suited to address the AFSP.

Stochastic programming is a generalization of deterministic mathematical programming, representing uncertain data with probability distributions. Typical models include a limited number of stochastic variables, but potentially large numbers of deterministic variables that can take on many potential values. With a basis in mathematical programming, they use efficient algorithms to exploit the problem structure whenever possible. Stochastic programs employ discrete time periods to represent decision making. The emphasis is on finding a solution after the problem is defined and modeled [27].

Likewise, decision theory seeks the variable levels that produce the best outcome of an experiment. This results in a basic problem similar to stochastic programming. Decision theory works through evaluation of each possible decision, making it applicable when the decision variables produce a small finite number of cases [27]. Decision theory, therefore, is not appropriate for the high dimension design space of this problem.
Dynamic programming focuses on finding optimal actions to take, given a discrete system state, such that the potential resulting system states are favorable. The models have a Markovian structure with regard to state transition. Backward recursion is used to arrive at an optimal decision at each stage [27]. The size of the state space must be carefully limited for dynamic programming, which would require significant simplification of the AFSP. Dynamic programming also requires state transition probabilities that are not easy to obtain.

Approximate dynamic programming (ADP) techniques, although computationally intensive, can deal with large state spaces. ADP requires a significant learning period to evaluate the state space, after which, decisions can be made quickly. Through this learning period, it avoids the need for transition probabilities. ADP is most useful where the time between decisions is on the order of seconds or minutes. For effective learning, stochastic paths are typically simulated, as would be appropriate in a network optimization setting. Although ADP may provide a viable option for the AFSP, it does not fit the general ADP framework of quick successive decisions in a network setting [44]. Therefore, approximate dynamic programming may not be an efficient approach for this application.

Machine learning and online optimization typically involve a changing objective function without foreknowledge of the change. Decisions are made based on past observations, such that the regret of the decision under consideration is minimized [27].
Machine learning and online optimization are not necessary for the AFSP, as it does not have a changing objective function (by the description provided).

Although stochastic statistical control models may not have any clear distinction from stochastic programming, they tend to be of lower dimension, have more restrictive constraint assumptions, and place emphasis on developing controls with specific decision rules [27], which is not consistent with our problem objective.

Stochastic simulation, by itself, is not an optimization method, but it can be used to effectively model very complex stochastic systems, and evaluate alternative designs or policies. It is widely used in science, engineering, military, and industrial applications. Simulation can be coupled with a statistical sampling procedure to explore the design space or a search heuristic to find good solutions. Simulation model design is infinitely flexible, but usually computationally intensive with no exploitation of problem structure. In general, simulation approaches work to improve the upper bound of a minimization problem, but do not provide a lower bound.

Optimization through stochastic simulation offers a reasonable approach for the AFSP, although perhaps not ideal. The simulation approach offers more modeling flexibility (such as non-linear functions and random variables with unknown joint distributions), but given the current level of understanding of system relationships and state of data available, additional precision may not be warranted at this time. More importantly, exploring or searching the high dimensional, continuous design space of this problem would be resource intensive. Stochastic programming offers a
methodology to efficiently find solutions in high dimensional space, albeit with a potentially simplified model. This stochastic programming solution may then be used in subsequent, more detailed simulation based studies.

Of the alternatives outlined above, stochastic programming is the most appropriate to address the AFSP. From the problem description, deployment demand is the primary source of uncertainty with a partially unknown, continuous distribution. There are over one hundred nearly continuous decision variables that must be decided at each time period in order to specify the size and composition of each workforce segment. The objective is to minimize cost while maintaining a specified level of service.

Two potential issues with stochastic programming are developing a model that is realistic, yet simple enough to solve and developing a distribution to represent the unknown demand function. Both of these issues will be addressed by this dissertation. We now consider stochastic programming applications in other areas used to solve related problems.

2.3.5 Related Applications
This section provides a sample of related stochastic programming research from other sectors. It is not meant to imply that stochastic programming is the only appropriate choice for these problems, but rather demonstrate the active and ongoing work in in this field.
Financial planning represents one of the largest application areas of stochastic programming [27]. Many financial management problems have multiple decisions over time as well as inherent risk or uncertainty which lend themselves well to MSSP formulation. Bradley and Crane [45] developed a MSSP to solve a portfolio management problem. They applied column generation to reduce the number of constraints in the master problem to less than 10% of the full set. The relatively easy sub-problems made the decomposed problem manageable. Mulvey and Vladimirou [46] used a stochastic network optimization model for multiperiod portfolio management. The dynamic multiperiod formulation provided superior results than myopic model. The authors employed an approximation technique from Rockafeller and Wets [38] that would come to be known as progressive hedging. Zenios et al [47] demonstrated that their multistage stochastic formulation for portfolio investment provided better solutions than the two-stage model. The multistage solutions were also more robust, providing better outcomes, on average, even when out-of-sample scenarios were evaluated. Frauendorfer & Schürle [48] develop a MSSP to manage banking deposits under variable interest rates that are also correlated with variable volume. The authors established bounds and approximated a solution using barycentric approximation to generate possible scenarios. They showed that despite the increased complexity, dynamic management policies were superior to static strategies.
These financial applications all deal with choosing multiple sequential investments from a portfolio of options in order to maximize benefit (returns). This is similar to the AFSP where there are multiple sequential decisions to allocate funds to different sectors in order to maximize capability.

**Energy**

Stochastic optimization has a long history of application within the energy sector. This area tends to have complex models with significant uncertainty in any number of variables, including: demand, price, cost, competition, and regulation.

Ventosa et al [49] provide a taxonomy of modeling approaches in the energy sector. They describe the optimization category as most applicable for single firm revenue optimization, considering the competition for quantity as well as possibly the effect on price. The authors indicate that explicit treatment of stochastic factors (such as price) through approximate dynamic programming or stochastic programming improves results. Louveaux [50] develops the multistage quadratic stochastic programming algorithm and applies it to solve an energy related capital investment problem. Expansion options via various types of electrical plants are considered in order to meet unknown future demand at the lowest cost with a trade-off between up-front investment costs and ongoing maintenance and production costs. Manne and Richels [51] investigate the cost and potential benefit of reducing carbon dioxide emissions now versus waiting for more information. They develop an optimal hedging strategy based on a scenario tree analysis with uncertainty about the future as well as potential
damage of greenhouse emissions. The authors determine the value of information, and the potential cost of delaying action to support policy making decisions. Mirkhani & Saboohi [52] illustrate the value of a MSSP over a deterministic model. They use a case study with the natural gas fuel price providing uncertainty while making capacity expansion decisions involving timing and energy source.

The energy problems presented pertain to capacity expansion investment decisions with potentially long lead times in a dynamic, inter-related system with uncertainty in multiple variables including future demand. The Army problem also makes investment decisions (in terms of manpower and funding) in order to shape future capacity with a long lead time in a complex inter-related system with uncertain future demand.

**Water**

There has been an interest in applying stochastic models to solve water resource and reservoir management problems since the late 1960’s. Dynamic programming applications, such as the work of Larson and Keckler [53] appeared first, followed by stochastic programming applications by Prékopa et al [54] beginning in the 1970’s.

Recently, a series of studies have focused on MSSP applications in water resource and reservoir management. The multistage models allow for dynamic system representation and multiple sequential decisions to accurately evaluate policies. Li et al [55] integrate inexact mathematical programming with multistage stochastic integer programming in order to deal with uncertain cost coefficients as well as the uncertain
water supply and discrete policy options. Li et al [56] pair stochastic programming with fuzzy mathematical programming and interval programming in a multi-stage formulation in order to account for various types of uncertainty accurately while maintaining tractable problem. Huang et al [57] incorporated a water run-off simulation to form a simulation-based multistage interval-stochastic model. Suo et al [58] combined inventory theory (economic order quantity) with multistage stochastic programming to determine water transfer decisions as well as water allocation policies in a reservoir system with uncertain supply.

Although the water resource problems have their principle uncertainty in supply whereas the AFSP deals with demand, both must adjust operating policies to mitigate shortfalls while balancing cost and benefit of meeting various requirements until fortunes change or additional capacity can be bought.

**Supply Chain**

There have also been a large number of studies using multistage stochastic formulations to address inventory and supply chain management problems. Escudero et al [59] provides an early example of a MSSP in supply chain management. Showing the limitations of deterministic modeling, they implement a scenario tree to represent uncertainty and allow for recourse decisions in a multi-product production and inventory problem with unknown demand. The authors are able to solve relatively large problems for the time by using decomposition to exploit the staircase structure in the problem. Kress et al [60] present a minmax multidimensional knapsack problem to
represent a military ammunition distribution problem. They formulate this as a two period chance constrained stochastic program with recourse, solving through decomposition by period, thus exploiting the special structure of the problem. Avital [61] adapted this approach to a similar problem in a naval setting with the addition of assigning shooters to targets via heuristic. Lejeune and Ruszczynski [62] study a multistage supply chain problem with uncertain demand. They formulate a stochastic program using probability constraints to specify service level (restricting chance of stock-out). They generate p-efficient demand trajectories, and using a subset, employ column generation to solve. The result is a method to attain robust, sustainable plans. Guan and Philpott [63] address a multi-stage stochastic supply chain problem with uncertain supply, contracting, and price-demand curves for the dairy industry. The authors use Bender’s decomposition with sampling approximation in the dynamic outer-approximation sampling algorithm (DOASA). They are able to compare policies, and show improvement over deterministic method.

There are several similarities between the AFSP and supply chain management. They both occur in a dynamic environment with a series of similar decisions occurring over time. Unknown future demand is one of the principle stochastic factors. Production decisions will affect available inventory and system cost, with distribution (deployment) decisions as recourse. In each case cost should be minimized, but a specified level of service must be achieved. The applications surveyed also suggest that probabilistic
constraints may be appropriate to model the Army’s likelihood of meeting uncertain demand.

**Overall Relevance**
In each of the areas and studies described, researchers chose to apply stochastic programming to model and solve their challenging problems. Each segment has similarities with the AFSP of interest to this thesis, as described above. Multistage stochastic formulations allow modeling of complex, dynamic systems with multiple decision points and various sources of uncertainty producing solutions with more accuracy than two stage or deterministic models. Therefore there is evidence to suggest that a MSSP formulation would be appropriate and effective for the problem at hand.

### 2.4 Research Objectives
This section outlines the research described in this thesis in relation to the existing literature. It states the research objectives and identifies areas of contributions.

#### 2.4.1 Purpose
Chapter 1 introduced our fundamental research question: “What size should the Army be?” This is an important and timely topic in light of recent defense budget reductions. We provided an argument in Section 2.2.7 that current analytic methods did not adequately answer this question. They were lacking in two principal areas, namely the ability to conduct analysis in a dynamic setting, and the ability to consider the entire Army with each segment of the Army workforce. Furthermore, in Section 2.3 we made a case for the application of a MSSP to address these deficiencies. The purpose of the
remainder of this thesis is to implement a multistage stochastic programming formulation of the Army force size problem while addressing related challenges in order to provide additional insights for Army and national level policy makers.

2.4.2 Study Questions

We outlined the main focus areas of our research through these study questions.

Deployment Demand Probability Distribution

Most stochastic programming literature deals with solution and approximation methods for challenging problem formulations, or the application of those techniques to new and interesting problem domains. What we found lacking was a satisfactory treatment of representing the stochastic variables as probability distributions for implementation in stochastic programming.

In our case, deployment demand is the stochastic variable in our model. As explained in Section 2.2, in defense resource planning, it is imperative to base analysis on validated scenarios. Therefore, our first study question became “How can we develop a credible probability distribution of deployment demand which is amenable to stochastic programming?”

This question is addressed directly in Chapter 3 where we first study the characteristics of simulated demand distribution, then develop a method to approximate it through a discrete empirical probability distribution. Furthermore, in Chapter 5, numerical experimentation is performed to show that the deployment
demand probability distribution developed in this thesis provided better representation of the demand function than simple, naïve approaches.

**Multi-Period Problem Approximation**

In the AFSP, the decisions made in each period affect the forces and choices available in subsequent periods. This process continues in perpetuity. As a practical matter, we can only model and solve a limited number of decision periods. Early experimentation with test models indicated that the AFSP model did not terminate gracefully, that is, there were obvious anomalies in the ending periods of the solution attributable to the time horizon of the model. Furthermore, our attempts to mitigate this through additional constraints or ending conditions were unsatisfactory. A common modeling solution is to simply extend the model time horizon and ignore the anomalous periods at the end. The issue in stochastic programming is that scenario trees grow exponentially, and those sacrificial periods can dominate the cost of solving the model. This issue provides us with a second study question: “How can we accurately and efficiently approximate a long horizon problem with a finite stage model?”

The demand probability distributions developed in Chapter 3 take the idea of sacrificial ending periods into account. The fidelity of the distribution diminishes as the stages increase in order to limit the growth of the scenario tree, while still providing some stochastic effect in the late model provide quality solutions. The numerical analysis in Chapter 5 compares methods with and without this adaptation, as well as models of varying length to determine the effect on AFSP model solutions. We observed
that at least two sacrificial periods were necessary, but that four was noticeably better while still being affordable.

**Policy Generation**

Multistage stochastic programs are typically used to find optimal solutions for decisions made in the first stage only. Subsequent stage variables are probabilistic in nature and often too numerous to list. For our problem, however, we sought to create a multi-period force plan (or *policy*), based on the decisions from multiple periods of the model.

Simply averaging the decisions at each period, across all scenarios, weighted by probability would yield a single decision value for each period, but this approach is too lenient. It uses decisions based on information that was not available at the beginning of the model when the policy decision needed to be made.

Conversely, creating additional policy variables that would be decided in the first period of the model, then used to restrict the related decisions in subsequent periods was too restrictive. In reality, although a multi-year policy is required, each year the previous policy can be modified, incrementally based on new information. The multi-year policy is not strictly fixed for all years.

This leads to our next study question: “How can we model AFSP force decisions to yield multi-year policies that are admissible, implementable, and consistent with the actual system? Again, we find this subject absent from literature. We develop a variation on basic non-anticipative variables to model separate policy variable behavior
in Chapter 4. This allows us to decide a policy in the first stage of the model about multiple future periods, then, accounting for the information available and progression of each scenario, permit subsequent policies and decisions to deviate from the ancestor policy decision. This is a more accurate representation of the true system, as shown through experimentation in Chapter 5.

**Solution Methods**

Although we set out to create a modestly sized representation of the AFSP, limited to continuous linear relationships, we knew that this model would become large simply through the addition of stages and scenarios. This provided another study question: “How might we optimize a multistage stochastic program formulation of the AFSP?”

This issue is partially addressed through the careful representation of demand probability distributions developed in Chapter 3 and revised in Chapter 5. By limiting the size of the scenario tree at the outset, the problem size is much more manageable.

The AFSP model was implemented, with initial experimental results detailed in Appendix A. There we showed that the problem in its extensive form could be solved directly and efficiently by seeking an approximate interior point solution. We found these solutions were of high quality, and beyond the precision of our capability to estimate model parameters.
**Value of the Proposed Approach**

The research described in this thesis required considerable time and effort. Furthermore, to refine, validate, implement, and maintain the model introduced here would be even more costly. Our next question addresses that subject: “Does a MSSP formulation of the AFSP provide a benefit worth the effort?”

Qualitatively, we answer “yes”. This approach provides the ability to directly address changing force size issues over time, which is not easily derived from current methods. Quantitatively, we show the value of a stochastic solution in Chapter 5, revealing that the difference between a stochastic and deterministic solution of the same model can be 5% or more. Finally, we offer that any analytic tool that can help policy makers better understand and communicate complicated cost – risk relationships is inherently valuable, especially when applied to a $600 billion program.

**Army Policy Insights**

Finally, our initial motivation to study this problem was to provide useful insights to Army policy makers through the development of a novel force size analysis methodology. That desire provided our final study question: “What insights does the AFSP model allow us to draw for Army policy making?”

This question is addressed in Chapter 5 where we recommend force size and composition changes over a five year period and illustrate how both external budget and manpower constraints affect the recommended solution. Furthermore, we developed a cost – risk relationship based on tolerance for unmet demand at a 95%
confidence level and show how risk preference can influence force recommendations under resource constrained conditions.

2.4.3 Scope
This section will describe the bounds of the research project in terms of constraints, limitations, and assumptions [64]. Within the context of this section, constraints are defined as deliberate design choices to define or reduce the design space. Constraints are controlled by the study sponsors, in this case, the authors. Limitations are imposed by a lack of capability or resources to consider study objectives more fully. Limitations are (at least in part) beyond the control of the study designers. Assumptions are unverified statements taken as fact for study purposes in order to proceed efficiently. They often address constraints or limitations.

Constraints
As noted in the introduction, this research focuses on Army size rather than another service or the entire Department of Defense. The Army, with greater size fluctuation, has the most interesting business case. Furthermore, our expertise and frame of reference pertains to the Army, making it a logical choice. Finally, limiting the study to one service helps to reduce the problem to a manageable size.

This research addresses strategic level questions, and is modeled at a highly aggregated level. We take a holistic view of the Army work force (Section 2.1.5) to capture the dynamic interdependencies. This work is not intended to address personnel needs by grade, specialty, or seniority, nor will it address individual unit types to any
specificity. The existing detailed personnel and force structure models do this adequately.

Force size, as a precursor to force structure is at the center of this study. Readiness and modernization levels will be fixed or addressed implicitly through the data available. This is necessary in order to limit the project to a manageable size. Force structure is the logical area to begin study as it shapes requirements for the other areas such as readiness and modernization. It would be worthwhile to include explicit consideration of readiness and modernization in the future.

Modeled uncertainty in this project is limited to deployment demand (Section 2.1.1). Although other sources of variability exist, such as uncertain future funding levels or the number of soldiers who voluntarily leave the Army, addressing the variable deployment demand is our primary concern.

This project is intended to aid in efficient, fiscally sound force size decisions within the context of modern limited warfare to which the U.S. has been accustomed in the late 20th and early 21st centuries. An imminent threat of invasion of the U.S. or a close ally, a full national mobilization, a re-instatement of conscription, or widespread use of weapons of mass destruction (nuclear, biological, or chemical warfare) are beyond the scope of this model. Neither the appropriate data nor behavior rules are anticipated to be incorporated for such events. Much like a full surge under ARFORGEN, it would necessitate a complete re-evaluation of policies, and likely result in a new operational paradigm.
Only rotational force policies are considered in this model, in keeping with current Army policy [Section 2.1.8]. Incorporation of other readiness policies as they are developed or as needs arise should be feasible as future work, but is not warranted at this time.

The model envisioned will not explicitly account for gamesmanship or politics that are not unexpected in policy making at this level.

**Limitations**

Not surprisingly, much of the military’s work concerning requirements and capabilities as well as intelligence, threat assessments, and operational plans is classified. Without ready access to classified information or the means to process classified data, this research is restricted to unclassified information only.

The Army does not currently have a model that addresses the total workforce of both the operational and institutional force. The AFSP model used in this research will be developed for this purpose. As such, it has not been validated by the Army.

Many aspects of Army effectiveness are tied to the human condition. Morale, experience, fatigue, and public perception are just some examples that can influence the Army’s capabilities. These are qualitative or subjective, and do not easily translate into monetary terms. Although some of these aspects will be implicitly included in the manpower model, further research may be warranted in this area.
Accounting for contractors is problematic. The data available for contract man-hours is unreliable. We will use anecdotal data when available, and adjust DA Civilian data as a surrogate where necessary.

Assumptions

We make several assumptions in order to address concerns about the constraints and limitations above. The unclassified planning scenarios provided used in this research represent likely conflicts in the future. We assume that decisions should be transparent and logically based on economic benefit. Sufficient unclassified material exists in order to produce credible results. We expect the model and process used would be readily adaptable to accept classified data if desired in the future.
CHAPTER 3  STOCHASTIC DEPLOYMENT DEMAND

The uncertainty of future deployment demand represents the stochastic element of the AFSP. The sample space for deployment demands is complex, and not well defined. The majority of previous work, including currently validated methods, uses discrete event simulation to produce stochastic deployment demands.

This chapter develops a model of deployment demand, and implements that model using Monte Carlo simulation to generate stochastic demand. The simulation output is analyzed and approximated to provide a distribution suitable for a stochastic programming approach to our Army Force Size Problem (AFSP).

3.1 Mathematical Model of Deployment Demand

This section details the development and design choices of our model for deployment demand based on historic data. Future deployment demand is uncertain for our AFSP, and represents the stochastic element for our problem. A probability distribution of the deployment demand is necessary for stochastic programming. This analysis permits the development of the demand distribution.

The approach here is similar to the process outlined by RANGER IPOD [19] where the frequency and type of SSC are varied, but here stochastic MCOs are incorporated as well. However, approved planning scenarios and their resulting data are classified.
Therefore, the model developed here is limited by using unclassified, surrogate data from credible sources. The force mix and composition analysis (FMCA) project [20] has a working, validated program to generate stochastic demand using the approved, classified planning scenarios. If ever desired by the Army, our method could be repeated with the classified scenarios as input. Alternatively, another validated simulation may be used to generate stochastic demand data for further analysis, as described in this chapter.

3.1.1 Event Arrivals

**Historic Contingency Data**

The occurrence of contingency events is modeled after historic contingency data. This analysis used the data set originally compiled for the SADE study [16], and later expanded and re-categorized for RANGER IPOD [19]. The data covers the period from December, 1989 through April, 2011, or 21$\frac{1}{3}$ years. The data set contains 574 records of events involving U.S. military personnel during that period. Each record has fields for event type (categorized by the RANGER IPOD study), regional command responsible for the operation, operation name (if assigned), location (country or state), a yes/no indicator for Army involvement, start date, end date, and maximum Army participation level. From the dates the event duration and inter-arrival times are also calculated. Not all records were complete however; many were missing the maximum Army participation level.
We chose not to update the event data because we do not have the means or access to complete the period to present time without introducing an additional bias factor in the data. We accepted the existing record of events as sufficient for our research.

Poisson Process

Previous studies such as [16], [18], [15], [19], [20], and [24], have modeled the inter-arrival times of events as an exponential distribution, an assumption which we wanted to validate. The exponential distribution assumes that events occur one at a time, and independently. We intuitively expect that events are not truly independent; a peacekeeping operation is likely to follow an armed conflict or a natural disaster is likely to affect areas within the same region. Despite violating the independence assumption, exponentially distributed inter-arrival times are often used quite effectively to model arrival processes, such as customer arrivals.

There is a strong precedent for using an exponential distribution to model event arrivals for military contingencies which is further supported by general practices in discrete event simulation [65], [66]. We hypothesize that our contingency inter-arrival times are from the exponential distribution, with the alternative that they are not, shown in (3.1).

\[ H_0 : \text{event inter-arrivals are from exponential distribution.} \]
\[ H_1 : \text{event inter-arrivals are not from exponential distribution.} \]  

(3.1)
Three techniques were used to validate the assumption of exponential inter-arrival times. First, a histogram of the inter-arrival times, shown in Figure 3.1, follows the general shape of the exponential distribution. Second, a Q-Q plot of the data using an exponential distribution, shown in Figure 3.2, indicates a good fit with some minor irregularities at the tail of the distribution. Finally, a chi-squared ($X^2$) goodness of fit test with 33 degrees of freedom yields a p-value of 0.9, while p-values ≤ 0.05 would indicate strong evidence against $H_0$. We fail to reject the null hypothesis: that event inter-arrivals are from the exponential distribution.

Figure 3.1 Histogram of contingency event inter-arrival days. This is the general shape of the exponential distribution.
Figure 3.2 Quantile-Quantile exponential plot of contingency event inter-arrival times. The exponential distribution fits the data very well for the majority of the smaller 99% of the data values. Four of the five largest values indicate a possible non-exponential trend with several more events occurring in the tail of the distribution than would be predicted by the exponential model.

The assumption of exponential inter-arrival times has several benefits. The exponential model is well known, mathematically simple, and supported by all statistical and simulation software. Furthermore, the exponential distribution has a single parameter, \( \lambda = \text{rate} \), to define the shape of the distribution function. This parameter is easily estimated, as in (3.2). The probability density function and cumulative distribution function are shown below by (3.3) and (3.4), respectively.

\[
\hat{\lambda} = \frac{1}{\bar{x}_{\text{obs}}} \quad \text{where} \quad x_{\text{obs}} \quad \text{are observed inter-arrival times}
\]  \hspace{1cm} (3.2)

\[
f(x) = \lambda e^{-\lambda x}, \quad \forall \ x \geq 0
\]  \hspace{1cm} (3.3)
Exponential inter-arrival times also imply that arrivals are governed by a Poisson process, that the number of events in a specified time interval is governed by the Poisson distribution [65]. The Poisson distribution is also specified by the same single parameter, \( \lambda = \text{rate} \). This can be estimated as shown in (3.2), or by (3.5). The probability mass function and cumulative distribution function are shown below by (3.6) and (3.7), respectively.

\[
F(x) = 1 - e^{-\lambda x}, \quad \forall \ x \geq 0
\]  

(3.4)

\[\hat{\lambda} = \bar{y}_{\text{obs}} \text{ where } y_{\text{obs}} \text{ are observed events within a time interval} \]  

(3.5)

\[p(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad \forall \ x \in \mathbb{Z}^+ \text{ (non-negative integers)} \]  

(3.6)

\[F(x) = \sum_{i=0}^{x} \frac{e^{-\lambda} \lambda^i}{i!} \]  

(3.7)

Poisson processes have useful properties for simulation referred to as random splitting and pooling [65]. Random splitting means that if a Poisson process, with rate = \( \lambda \), generates events of multiple types \( (i) \), each of those event types is also a Poisson process with \( \lambda_i = p_i \lambda \), where \( p_i \) is the proportion of events of type \( (i) \).
Similarly, events from multiple Poisson streams may be combined by adding their rates together. Therefore one Poisson process may determine the number of events in a period, while a separate stochastic process specifies what types of events they are.

**Data Scrubbing**

The historic event data required some cleaning before proceeding. The events indicating no Army involvement were screened out. Assuming the contingencies are from a Poisson process, eliminating classes of events should change the overall rate of events, but not the nature of the process. All approximate, average, or ranges of values for maximum participation were arbitrarily assigned a single value. Remaining records with blank entries for Army participation were checked for open source reports on troop levels. After consideration, the remaining events with blank values for maximum participation were excluded from the data used. Although still useful for calculating the rate or frequency of events, it is likely that records with missing data are considerably smaller, on average, than the records not missing data. Since there is no convenient way to quantify the size of these contingencies or reliably group them with the contingencies of known size, the most reasonable course of action was to exclude them from the data set.

Next, events with maximum troop levels of less than 100 were screened out. For an Army of one million soldiers, this represents 0.01% of the force. This assumes that micro-contingencies have little impact on the overall force utilization and readiness, which is consistent with the other force structure studies. An additional threshold was
established for events where the Army supported domestic civil authorities with less than 50 troop-years of involvement. These small missions are often short duration local tasks involving the National Guard, and frequently employ voluntarily mobilized troops with little adverse effect on unit readiness or availability for subsequent operations.

Counter-drug operations were excluded due to policy changes limiting their use since their prevalence in the 1990s. Maritime interdiction operations were excluded because Army participation in maritime operations appears to be ancillary and only included as a matter of convenience when available. Finally, one additional record was excluded because it began well before the specified data collection period. This leaves 174 records in the data set to model the occurrence of contingency events for the Army, as shown in Table 3.1 below.

<table>
<thead>
<tr>
<th>Table 3.1 Summary of data screening for historic contingency operations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Records</td>
</tr>
<tr>
<td>No Army Involvement</td>
</tr>
<tr>
<td>Blank Size</td>
</tr>
<tr>
<td>Below Size Threshold (100)</td>
</tr>
<tr>
<td>Below Domestic Size Threshold (50 my)</td>
</tr>
<tr>
<td>Counter Drug</td>
</tr>
<tr>
<td>Maritime Interdictions</td>
</tr>
<tr>
<td>Outside Collection Period</td>
</tr>
<tr>
<td>Total Records Remaining</td>
</tr>
</tbody>
</table>

We chose to use event data from the entire post-Cold War period rather than partition it as growth, steady state, and contraction periods as in the RANGER IPOD
example. This was done to preserve a larger sample size, and also because we have no reason to assume any of these conditions is more likely for the coming decade. The number of contingency operations over time is displayed below in Figure 3.3. It indicates the proliferation of small scale contingency operations in the post-Cold war period. There is a peak in activity from 1999 to 2004, followed by less frequent operations during the 2005 to 2009 period that corresponds with Army surge operations in Operation Iraqi Freedom [67]. This implies that the U.S. may have avoided engagement in additional contingency operations while the Army was struggling with ongoing commitments, but additional data to support that theory is not available.

![Figure 3.3 Historic number of contingency events per year.](image)

The screened data set of Army contingencies plotted across time shows a growing trend of operations in the early 1990s, frequent operations from 1999 to 2004, decreased levels during peak operations in Iraq from 2005 to 2009, and potentially increasing level thereafter.
**Testing Assumptions**

We then re-tested our hypothesis (3.1): that the exponential distribution provides a reasonable model for event inter-arrival times of the remaining contingency data set. The same three tests were performed as before on the full data set. The histogram of the inter-arrival data in Figure 3.4 again shows the general shape of an exponential distribution. The quantile-quantile plot in Figure 3.5 also indicates a good fit for the small intervals, with some irregularity in the larger intervals, or tail section of the distribution. The chi-squared ($\chi^2$) goodness of fit test with 15 degrees of freedom yields a p-value of 0.168.

![Histogram of screened contingency event inter-arrival periods.](image)
We fail to reject the null hypothesis: that inter-arrival data is from the exponential distribution. Considering the inherently uncertain process of using messy historic data to make predictions of future events and the prevailing practice of using an exponential distribution to model event inter-arrivals, we accept this assumption moving forward for this analysis.

The average value of the inter-arrival data is $\bar{X}_{\text{obs}} = 45.22$ days, which corresponds to $\lambda = 0.02212/\text{day}$ or $8.078/\text{year}$. The standard deviation of data is $\sigma = 49.01$, which means that the single $\lambda$ parameter for the exponential distribution will provide a good estimate for both the mean and variance. Otherwise, we could consider an alternate distribution, such as the negative binomial, so that both mean and
variance could be estimated separately. A Poisson process with the rate specified provides a model of the number of contingency events per year.

3.1.2 Event Categorization and Distribution

*Category Factors*

The categorization of events is a subjective process. Real events may require multiple mission types, and individual operations as part of a larger operation may be classified individually for greater fidelity, or in aggregate to represent dependency. Also, a contingency may change substantially over time in terms of type and size. Ultimately, we chose to adopt the historic contingency classifications developed by CAA because it was from an authoritative source, and any construct we might create would be equally arbitrary.

The historic contingency data contains information about the geographic region. This is not important to our study, so that factor is not used in the model. Duration is not considered as a separate factor either as it did not show additional benefit in the RANGER IPOD study. We also suspect that duration is related to contingency type. Event durations were useful for aligning planning scenarios with the contingency categories, which follows in 3.1.3.

The maximum level of Army involvement is used as a measure of conflict size. Each of the category of events were ordered by size to consider the range of values involved and look for natural groupings or break points. We then divided event categories by size in order to better represent the different levels of troop demand,
attempting to avoid any size category from spanning more than an order of magnitude difference. The size designations are relative to the category type, not an independent factor. Large conventional contingencies were renamed major contingency operations, examples being Operation Desert Storm and Operation Iraqi Freedom.

The categories of contingency operations are shown in Figure 3.6 with the number of times each was observed during the data collection period. Because we have essentially applied a single factor classification system, event frequencies can be determined directly from the proportion of observations without the need to employ more complex models, as in [16].

![Figure 3.6 Observed contingency events by category.](image-url)
**Event Frequencies by Category**

Observations of each of the contingency category types over time are shown in Figure 3.7. Together with Figure 3.6, we see that conventional contingencies occur infrequently. Domestic civil support operations are by far the most common type of contingency, with many times more small operations than large, distributed throughout the analysis period. Enforcing sanctions does not often involve the Army as a primary mission, and has not occurred in some time. Humanitarian assistance missions are common throughout the period shown. Both homeland defense and irregular warfare operations appear to be clustered after the September 11\textsuperscript{th}, 2001 terrorist attacks. Non-combatant evacuations occasionally involve Army units, and are evenly distributed. Both peacekeeping and show of force operations are fairly common until 2005, corresponding with Army surge operations in Iraq. Strike operations rarely involve Army units, but in this case, Army helicopters, along with support and security personnel played a supporting role.

![Figure 3.7 Observations of each contingency type over time.](image)
Each of the event types are assumed to be uniformly distributed for the purpose of calculating frequencies for stochastic demand generation using a cumulative empirical distribution function. It would be acceptable to modify the rates or select a particular portion of the historic data in order to study a specific type of future, but we choose to use a general future representation, in keeping with our previous decisions.

The relative probability of each type of event can then be calculated, as shown below in (3.8). Calculated probabilities will be presented as Table 3.3 in the following section, after further development.

Let $\omega_i$ be a random contingency event of type $i$. Let $x_i$ be the count of $\omega_i$ observed. Let $n = \sum x_i \forall i$, the number of observations.

$$P(\omega_i \mid \omega) = \frac{x_i}{n}$$  \hspace{1cm} (3.8)

An empirical CDF can also be developed using the count of observations in each event category, $x_i$, as shown below in (3.9). The empirical CDF will also be presented later, in Table 3.3.
Let $I$ be the set of contingency types.

\[ F(\omega_j) = \frac{\sum_{i=1}^{j} x_i}{n} \text{ where } 0 \leq j \leq |I| \]  

(3.9)

The empirical CDF is used for the stochastic determination of event types, with the proportion of event types based on the historic contingency observations.

**Major Contingency Operations**

Typical Army force structure analysis assumes that two MCO events will occur contained within the analysis period, usually 15 to 20 years long [17] [20]. That is, the MCO events are selected from a uniform distribution to occur at uniformly random start times, such that both will end within time considered. This convention may have once been a simplification for computational expedience, but it is now a political issue tied to funding requirements for the Department of Defense. We consider MCOs in greater detail here because they provide the greatest demands on Army forces, or the worst case scenarios. Furthermore, they occur infrequently, with only two observations in the period considered here.

National defense policy has long been focused on winning a two-theater war, although the only time the U.S. faced such a threat was during World War II (WWII). This guidance was recently changed to winning one major theater conflict while deterring a second conflict, effectively reducing defense requirements with the intent of reducing defense spending. While conflict avoidance certainly reduces military requirements, it is
not without cost itself. It is reasonable to consider reducing costs by accepting risk in the ability to confront two simultaneous MCOs (a rare event), but it is a risk policy makers should fully understand.

The U.S. has participated in 11 MCO operations throughout its existence, if WWII is counted as 2 events, one for each theater. A timeline of those conflicts is shown below in Figure 3.8. The nature of warfare as well as the role of the U.S. in global affairs has changed immensely in the last 239 years, therefore considering data over that period as homogeneous is suspect. With that said, the conflicts appear to be spaced rather uniformly from the Spanish American War in 1898 to present.

![Figure 3.8 Timeline of U.S. Major Conflicts.](image)

Events appear more tightly spaced after the Spanish American War in 1898. The highlighted color for WWII in 1941 signifies two simultaneous major conflicts in separate theaters.

Using the period from 1898 to present, those 8 conflicts occurring over 116 years provide a rate of \( \lambda = 0.069 \text{ MCOs/year} \). The average rates for alternate intervals are shown below in Table 3.2. Depending on the interval selected, the rate of observed MCOs can vary from a low of 0.060 to a high of 0.083 per year. The value from 1898 to present provides a reasonable mid-range value, and will be used in subsequent analysis.
Table 3.2 Rate of MCOs over various time intervals in U.S. history.

<table>
<thead>
<tr>
<th>Period</th>
<th># MCOs</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990 - 2014</td>
<td>2</td>
<td>0.083</td>
</tr>
<tr>
<td>1964 - 2014</td>
<td>3</td>
<td>0.060</td>
</tr>
<tr>
<td>1950 - 2014</td>
<td>4</td>
<td>0.063</td>
</tr>
<tr>
<td>1941 - 2014</td>
<td>6</td>
<td>0.082</td>
</tr>
<tr>
<td>1917 - 2014</td>
<td>7</td>
<td>0.072</td>
</tr>
<tr>
<td>1898 - 2014</td>
<td>8</td>
<td>0.069</td>
</tr>
</tbody>
</table>

The new MCO frequency will replace the value derived from the primary data set. The expectation is that the new value from the extended MCO data is a better indicator of historic MCO occurrences. With only eight data points, a histogram would not be useful. The quantile-quantile plot shown below in Figure 3.9 indicates that the MCO inter-arrival times are not approximated as well by the exponential distribution as was the primary data set.
Figure 3.9 The Q-Q exponential plot of MCO inter-arrival times from 1898 to 2014. The plot shows a systemic deviation from the exponential distribution. There are too many observations clustered near the median and not enough in the tail section to be well approximated by the exponential distribution.

In considering other possible distributions, we were wary of highly parameterized distributions that could over-fit this small data set. We also wanted a model that had a rational basis to explain the phenomenon, not just a model that fit well. The inter-arrival times of MCOs, Figure 3.10 below, suggest that a normal or uniform distribution may fit the data.

Figure 3.10 The inter-arrival times of major U.S. conflicts from 1898 to 2014. This sample of data suggests a normal or uniform distribution.

Of those two, the normal has more appealing theoretical properties, namely that events could be generated outside the observed range. Any negative values, however, would need to be rejected, resulting in a truncated-normal distribution. A quantile-quantile plot of the MCO inter-arrival times using the normal distribution, shown in Figure 3.11, indicates a good fit.
Assuming a normal distribution, the parameters are estimated as $\mu = 14.5$, and $\sigma = 8.98$. Using the truncated normal distribution, random sampling would produce inter-arrival times less than 1 year about 1.4% of the time, indicating two nearly-simultaneous MCO events. Inter-arrival times less than 5 years occur about 9.7% of the time, indicating overlapping MCO events.

There are a number of possible explanations for a normal distribution of MCO observations. There may be an underlying aversion to engaging in a major conflict once already committed to another MCO. A decade and a half is also enough time to change political administrations, replace the vast majority of service members that had participated in the last war, change global economic outlooks, and allow new regional
powers and threats to emerge. Any or all of these may play a part in influencing the
distribution of major conflicts.

Without a better understanding of the underlying mechanisms that may be
influencing MCO occurrences, we are hesitant to deviate from our prior assumption that
all events can be approximated by a Poisson process. Furthermore, it is possible that we
are observing some type of conflict avoidance, though somewhat different from missed
demand, as we have defined it. We therefore maintain the assumption that modeling
MCO events with exponential inter-arrival times, as part of the overall Poisson event
generating process, is an adequate approximation to generate realistic deployment
demand profiles for the purpose of this research. It is also an improvement over current
MCO representation practices. We acknowledge that a separate process to generate
MCOs from a normal or other justifiable distribution is an interesting prospect, left for
future study or possible later extension.

The new MCO frequency of 0.069 events per year was used to replace the
previous value of 0.094, and from interpolation, $x_{MCO} = 2 \rightarrow 1.47$ observations in 21.4
years. With the new $n = 173.47$ the probabilities for each contingency type were
updated, as shown in Table 3.3.
The Poisson process defined in 3.1.1 provides the number of events per period. The empirical CDF in Table 3.3 can then be used to stochastically determine event types, using the splitting property described in 3.1.1. Sampling the empirical CDF for each period is done with replacement, as multiple occurrences of the same type of contingency may occur within a period. This provides a model of contingency occurrences by category type.

3.1.3 Planning Scenarios
The Defense Planning Scenarios (DPS) provide a common set of approved potential future conflicts, for planning, justifying requirements, and funding. The
services develop these scenarios further through detailed planning, combat simulation, and war gaming in order to determine the forces necessary for a successful outcome. Following the RANGER IPOD methodology [19], we would select which scenarios to incorporate and match them to the categories defined form the analysis of the historic data. The DPS and much of the derivative work are classified, however. We therefore apply the same general approach, but develop unclassified scenario data.

An excellent source of unclassified scenario data, specifically Army force demands, was developed and provided by the U.S. Army Training and Doctrine Command Analysis Center (TRAC). These scenarios were developed for the purpose of providing a realistic venue for analysis to an unclassified audience, such as allied nations. The unclassified scenarios are developed in the same fashion as the classified scenarios, except that the unclassified scenarios involve notional belligerent countries overlaid on U.S. terrain with notional military forces. TRAC provided data that included 2 MCOs and 21 SSCs. Each scenario provided the forces required each month as well as major unit types such as Infantry, Armored, and Stryker Brigade Combat Teams (BCTs); Combat Aviation Brigades (CABs); Field Artillery Brigades (Fires); and Support Brigades. More detailed unit information was available, but not necessary for this aggregated level model.

Another useful source of unclassified scenario information was the notional data developed by IDA for the SARA project [24]. They make no claim on the validity of the scenarios, but they were developed to be realistic for the purpose of illustrating the
value of their analytic approach. As described in 2.2.6, they aggregated the Army force structure into 20 different basic brigade sized unit types. This notional force structure was used to estimate total troop requirements for both the IDA and TRAC scenarios. The SARA notional data considered 12 different contingency types, providing corresponding force packages for three additional MCOs and eight SSCs.

These two sources of unclassified scenarios provided example force demands for a wide variety of contingency types.

**Stochastic Force Demands**
Each of the TRAC scenarios was assigned a category type and size from the classifications developed in the historic conflict data analysis above. Mission types and other identifying information were not included with the TRAC data. Scenarios were matched to categories by overall troop demand size, duration, and types of units required. MCOs and smaller conventional contingencies were readily identifiable by their size and combat units. Humanitarian assistance missions should not require many combat forces. Enforcing sanctions and peacekeeping operations should be long term operations. Shows of Force are typically short duration. In this way, the TRAC scenarios were categorized arbitrarily, but with a fair degree of confidence.

The three MCO demands developed by IDA were different sizes and lengths than the scenarios already included from TRAC. Of particular note, one was an air-sea scenario with ground forces in a supporting role, which is consistent with the current national strategy shifting focus toward Asia. These three MCO scenarios were added to
the collection for our demand generation. The SSCs were not added, as they did not appear to add significantly to the collection already assembled, either being redundant, or not fitting the size/type of contingency well where gaps in coverage were present.

For the categories that remained unmatched with any scenarios, notional force demands were developed using historic contingency data as well as the unclassified scenarios. The domestic missions supporting civil authorities were derived in this manner. Using infantry for security and general labor and enablers to provide support, services, and transportation, the force demand was scaled for each size using the average force size and duration from the historic contingency data analysis. A small humanitarian assistance mission was created by scaling down one of the TRAC scenarios. The irregular warfare demand was adapted from the IDA counter-insurgency scenario, scaled to match the historic size and duration. Additional non-combatant evacuation demands were added to augment the existing scenario with different unit types and varying sizes. A small peacekeeping operation was added by scaling down one of the large TRAC peacekeeping scenarios.

The probabilities for each contingency type, from Table 3.3, were then applied to the set of force demand scenarios, as shown below in Table 3.4. Some categories had more than one scenario assigned, so the likelihood was divided among the scenarios in each category, shown in (3.10). The MCOs were arbitrarily weighted to make a moderate conventional ground conflict and an air-
sea conflict with ground support more likely, while making a very large ground conflict less likely.

Let $w$ be the scenario weighting factor.
Let $i$ be the index of contingency.
Let $j$ be the index of planning scenarios.

\[
P(\omega_i) = \sum_j w_{i,j} P(\omega_j) \quad \forall i
\]

(3.10)

Where $\omega_j$ is the random occurrence of scenario $j$. 
Table 3.4 Contingency force demands and likelihood probabilities.
Abbreviated names correspond to the full names in Table 3.3. Durations are rounded up to next integer year and averaged over each one-year period to accommodate the model time interval. A 3 month demand for an IBCT would be reflected as $\frac{1}{4}$ IBCT for 1 year.

<table>
<thead>
<tr>
<th>Scenario Abbrev.</th>
<th>Source</th>
<th>Weight</th>
<th>Probability</th>
<th>Cumulative Probability</th>
<th>Max No. BCTs</th>
<th>Max Total Troops (k)</th>
<th>Duration (Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC(1)</td>
<td>TRAC</td>
<td>1/2</td>
<td>0.0086</td>
<td>0.0086</td>
<td>3.00</td>
<td>37.84</td>
<td>2</td>
</tr>
<tr>
<td>CC(2)</td>
<td>TRAC</td>
<td>1/2</td>
<td>0.0086</td>
<td>0.0173</td>
<td>1.00</td>
<td>13.43</td>
<td>1</td>
</tr>
<tr>
<td>DSCA-C</td>
<td>Historic</td>
<td>1</td>
<td>0.0058</td>
<td>0.0231</td>
<td>2.00</td>
<td>23.86</td>
<td>1</td>
</tr>
<tr>
<td>DSCA-L</td>
<td>Historic</td>
<td>1</td>
<td>0.0231</td>
<td>0.0461</td>
<td>0.50</td>
<td>4.22</td>
<td>1</td>
</tr>
<tr>
<td>DSCA-M</td>
<td>Historic</td>
<td>1</td>
<td>0.1557</td>
<td>0.2018</td>
<td>0.08</td>
<td>0.68</td>
<td>1</td>
</tr>
<tr>
<td>DSCA-S</td>
<td>Historic</td>
<td>1</td>
<td>0.2652</td>
<td>0.4671</td>
<td>0.02</td>
<td>0.14</td>
<td>1</td>
</tr>
<tr>
<td>ES(1)</td>
<td>TRAC</td>
<td>1/2</td>
<td>0.0086</td>
<td>0.4757</td>
<td>0.00</td>
<td>2.69</td>
<td>4</td>
</tr>
<tr>
<td>ES(2)</td>
<td>TRAC</td>
<td>1/2</td>
<td>0.0086</td>
<td>0.4844</td>
<td>0.50</td>
<td>3.22</td>
<td>4</td>
</tr>
<tr>
<td>FHA-L(1)</td>
<td>TRAC</td>
<td>1/5</td>
<td>0.0081</td>
<td>0.4924</td>
<td>0.00</td>
<td>5.83</td>
<td>1</td>
</tr>
<tr>
<td>FHA-L(2)</td>
<td>TRAC</td>
<td>1/5</td>
<td>0.0081</td>
<td>0.5005</td>
<td>0.00</td>
<td>5.00</td>
<td>4</td>
</tr>
<tr>
<td>FHA-L(3)</td>
<td>TRAC</td>
<td>1/5</td>
<td>0.0081</td>
<td>0.5086</td>
<td>0.00</td>
<td>1.50</td>
<td>1</td>
</tr>
<tr>
<td>FHA-L(4)</td>
<td>TRAC</td>
<td>1/5</td>
<td>0.0081</td>
<td>0.5167</td>
<td>1.00</td>
<td>8.93</td>
<td>1</td>
</tr>
<tr>
<td>FHA-L(5)</td>
<td>TRAC</td>
<td>1/5</td>
<td>0.0081</td>
<td>0.5247</td>
<td>0.00</td>
<td>2.00</td>
<td>1</td>
</tr>
<tr>
<td>HD-L</td>
<td>TRAC</td>
<td>1</td>
<td>0.0058</td>
<td>0.6228</td>
<td>1.58</td>
<td>17.68</td>
<td>2</td>
</tr>
<tr>
<td>HD-S</td>
<td>Historic</td>
<td>1</td>
<td>0.0634</td>
<td>0.6862</td>
<td>0.11</td>
<td>0.69</td>
<td>1</td>
</tr>
<tr>
<td>IW-L</td>
<td>Historic</td>
<td>1</td>
<td>0.0173</td>
<td>0.7035</td>
<td>5.00</td>
<td>52.46</td>
<td>6</td>
</tr>
<tr>
<td>IW-S</td>
<td>Historic</td>
<td>1</td>
<td>0.0231</td>
<td>0.7265</td>
<td>0.02</td>
<td>0.49</td>
<td>6</td>
</tr>
<tr>
<td>MCO(1)</td>
<td>TRAC</td>
<td>1/4</td>
<td>0.0021</td>
<td>0.7286</td>
<td>17.01</td>
<td>178.63</td>
<td>5</td>
</tr>
<tr>
<td>MCO(2)</td>
<td>TRAC</td>
<td>1/5</td>
<td>0.0016</td>
<td>0.7302</td>
<td>7.75</td>
<td>89.05</td>
<td>4</td>
</tr>
<tr>
<td>MCO(3)</td>
<td>IDA</td>
<td>1/5</td>
<td>0.0016</td>
<td>0.7319</td>
<td>13.50</td>
<td>165.38</td>
<td>6</td>
</tr>
<tr>
<td>MCO(4)</td>
<td>IDA</td>
<td>1/10</td>
<td>0.0008</td>
<td>0.7327</td>
<td>33.00</td>
<td>386.57</td>
<td>8</td>
</tr>
<tr>
<td>MCO(5)AS</td>
<td>IDA</td>
<td>1/4</td>
<td>0.0021</td>
<td>0.7348</td>
<td>2.50</td>
<td>42.82</td>
<td>3</td>
</tr>
<tr>
<td>NEO-L(1)</td>
<td>TRAC</td>
<td>1/3</td>
<td>0.0038</td>
<td>0.7386</td>
<td>0.00</td>
<td>2.14</td>
<td>1</td>
</tr>
<tr>
<td>NEO-L(2)</td>
<td>Historic</td>
<td>2/3</td>
<td>0.0077</td>
<td>0.7463</td>
<td>0.04</td>
<td>0.50</td>
<td>1</td>
</tr>
<tr>
<td>NEO-S</td>
<td>Historic</td>
<td>1</td>
<td>0.0288</td>
<td>0.7751</td>
<td>0.00</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td>PO-L(1)</td>
<td>TRAC</td>
<td>1/7</td>
<td>0.0058</td>
<td>0.7809</td>
<td>1.00</td>
<td>11.93</td>
<td>4</td>
</tr>
<tr>
<td>PO-L(2)</td>
<td>TRAC</td>
<td>1/7</td>
<td>0.0058</td>
<td>0.7866</td>
<td>1.00</td>
<td>9.10</td>
<td>2</td>
</tr>
<tr>
<td>PO-L(3)</td>
<td>TRAC</td>
<td>1/7</td>
<td>0.0058</td>
<td>0.7924</td>
<td>1.00</td>
<td>8.93</td>
<td>2</td>
</tr>
<tr>
<td>PO-L(4)</td>
<td>TRAC</td>
<td>1/7</td>
<td>0.0058</td>
<td>0.7982</td>
<td>1.00</td>
<td>12.12</td>
<td>4</td>
</tr>
<tr>
<td>PO-L(5)</td>
<td>TRAC</td>
<td>1/7</td>
<td>0.0058</td>
<td>0.8039</td>
<td>1.00</td>
<td>13.12</td>
<td>2</td>
</tr>
<tr>
<td>PO-L(6)</td>
<td>TRAC</td>
<td>1/7</td>
<td>0.0058</td>
<td>0.8097</td>
<td>2.00</td>
<td>22.84</td>
<td>3</td>
</tr>
<tr>
<td>PO-L(7)</td>
<td>TRAC</td>
<td>1/7</td>
<td>0.0058</td>
<td>0.8155</td>
<td>1.00</td>
<td>11.42</td>
<td>1</td>
</tr>
<tr>
<td>PO-S</td>
<td>Historic</td>
<td>1</td>
<td>0.0980</td>
<td>0.9135</td>
<td>0.05</td>
<td>0.47</td>
<td>3</td>
</tr>
<tr>
<td>ShOF-L(1)</td>
<td>TRAC</td>
<td>1/2</td>
<td>0.0202</td>
<td>0.9337</td>
<td>0.75</td>
<td>5.57</td>
<td>1</td>
</tr>
<tr>
<td>ShOF-L(2)</td>
<td>TRAC</td>
<td>1/2</td>
<td>0.0202</td>
<td>0.9539</td>
<td>0.58</td>
<td>4.31</td>
<td>1</td>
</tr>
<tr>
<td>ShOF-S</td>
<td>Historic</td>
<td>1</td>
<td>0.0404</td>
<td>0.9942</td>
<td>0.03</td>
<td>0.23</td>
<td>1</td>
</tr>
<tr>
<td>STR</td>
<td>TRAC</td>
<td>1</td>
<td>0.0058</td>
<td>1.0000</td>
<td>0.25</td>
<td>2.58</td>
<td>1</td>
</tr>
</tbody>
</table>
It would be logical to use the historic probabilities as a starting point to adjust the probability of scenarios based on actual intelligence estimates, expert opinion, and national defense strategy priorities. To our knowledge, no explicit likelihood estimates are assigned to the various planning scenarios. It may be possible for planners to provide relative likelihood ratings through a procedure such as pairwise comparison. For our analysis, however, the contingency and scenario likelihoods are unmodified, except as previously discussed for MCOs.

Using the empirical CDF shown in Table 3.3, we can stochastically select planning scenarios based on the frequency of historic contingencies of the same type. Each planning scenario includes a force list, or the number and types of units required over time in order to ensure a favorable outcome.

The notation introduced here for demands will be useful for describing concepts in the following sections.

Let \( i \in I \) be an index for demand (unit) type, where \( I = \{IBCT, SBCT, ABCT, CAB, Fires, Enablers, Total\} \).

Let \( \tau \) be the scenario specific time index.

Let \( d^\tau_i \) be a force demand indexed by time and type.

Let \( D(\omega) \) be a set of demands \( \{d^\tau_i\} \), such as for scenario \( \omega \).

The approach taken here was to analyze the historic contingency data, categorize it, match the available scenario demands to those categories, and fill any gaps with modified scenarios based on historic data. That was because we had an incomplete and unofficial set of scenarios to work with and the historic data was
considered more authoritative. If given a set of scenarios to be included, it is also reasonable to reverse the process by categorizing based on the scenarios and matching the historic observations to those scenarios to determine event frequencies. The analyst should beware of observed category types that have no corresponding approved scenarios. This may be a valid exclusion due to a future outlook that is different from the past or an omission from an oversight, lack of planning resources, or political/policy influence. Conversely, the analyst must also be cautious of scenarios that have no clear historic precedent. An arbitrary likelihood may have to be assigned based on surrogate data or expert opinion.

By mapping the planning scenarios with force demands to the contingencies, we can assign surrogate probabilities to each planning scenario. Each planning scenario contains a predetermined force requirement. This allows us to now to stochastically select scenarios, and generate deployment demands.

**Fixed Demands**

Although our approach focuses on producing a realistic stochastic deployment demand, there are situations where it should be augmented with a fixed or pre-specified demand. This could be to incorporate the requirements of an ongoing contingency, or account for planned missions affecting the usage and readiness of rotational forces.

Current (2014) conditions are used as the starting point for demand generation, but any desired existing demand can be specified. The Army has several small ongoing
commitments, listed in Table 3.5. These force demands are represented by adding a fixed requirement to every stochastic scenario generated. Our representation is simply uniform over time, but any desired profile may be specified. The corresponding demand summaries are provided below in Table 3.6.

### Table 3.5 Fixed ongoing demands for U.S. Army forces.
Troop levels for Afghanistan reflect the stated Presidential objective, not current levels.

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Troops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Guardian</td>
<td>Bosnia</td>
<td>13</td>
</tr>
<tr>
<td>Joint Guardian</td>
<td>Kosovo</td>
<td>721</td>
</tr>
<tr>
<td>JTF-B</td>
<td>Honduras</td>
<td>601</td>
</tr>
<tr>
<td>OEF-CCA</td>
<td>Caribbean/Central America</td>
<td>94</td>
</tr>
<tr>
<td>OEF-TS</td>
<td>Africa</td>
<td>79</td>
</tr>
<tr>
<td>Noble Eagle</td>
<td>US</td>
<td>320</td>
</tr>
<tr>
<td>Bright Star</td>
<td>Sinai</td>
<td>692</td>
</tr>
<tr>
<td>OEF</td>
<td>Afghanistan</td>
<td>9,800</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>12,320</strong></td>
</tr>
</tbody>
</table>

### Table 3.6 Summary of force demand data for fixed and ongoing operations.

<table>
<thead>
<tr>
<th>Scenario Abbrev.</th>
<th>Source</th>
<th>Weight</th>
<th>Probability</th>
<th>Cumulative Probability</th>
<th>Max No. BCTs</th>
<th>Max Total Troops (k)</th>
<th>Duration (Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix - OEF</td>
<td>Historic</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0.67</td>
<td>9.80</td>
<td>indef</td>
</tr>
<tr>
<td>Fix - Other</td>
<td>Historic</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0.4</td>
<td>2.52</td>
<td>indef</td>
</tr>
</tbody>
</table>

Fixed demands are referred to as $D_{\text{fix}}$ to distinguish them from the stochastic demands in (3.11). Fixed, or predetermined demands allow the stochastic demand generation process to account for different starting conditions and other expected requirements.
3.2 Stochastic Demand Generation
We have developed a model and the data necessary to generate stochastic deployment demands through discrete event simulation. In this format, the representation of complex systems is limited only by our understanding of its behavior and our ability to represent it mathematically.

Stochastic Demand Generator (SDG) Procedure
A Poisson arrival process determines the number of stochastic contingency events that occur in any year. The fixed rate parameter is determined from historic contingency data, though a non-stationary process could be implemented if needed. For each contingency event, an empirical distribution of scenario likelihood (developed from the historic contingency data) is used to designate a specific planning scenario. The deployment demand from each designated planning scenario (pre-determined from combat simulation and war gaming) is added to the running total along with the specified fixed demands (from current conditions). This process continues for a user specified number of periods, or years, to create a stochastic deployment demand. A depiction of this process is provided below in Figure 3.12. Each replication of the simulation will create one sample path or branch through the scenario tree, called a demand profile here to distinguish from the underlying planning scenarios.
The SDG procedure, as implemented, is outlined below. Demand type indices are omitted for clarity, as all demand types are treated the same. We use \( t \) to denote simulation periods, distinguishing it from \( \tau \) for the scenario specific time. An intermediary demand accumulator \( (\Delta) \) is used below to sum demands from scenarios beginning on the same simulation period \( t \).

- Time horizon \( H := 10 \)
- Event rate \( \lambda := 8.078 \)
- Define stochastic scenario probabilities and demands (Table 3.4)
- Define fixed scenario demands for all periods (Table 3.6)
- Set fixed demand as aggregated demand, \( D := D(\text{fix}) \)
- For each period, \( t = 1, \ldots, H \)
- Generate Poisson random number of events, \( n := \text{Poisson}(\lambda) \)
- Select \( n \) random planning scenarios \( (\omega) \) from empirical CDF, with replacement
- Initialize accumulator, \( \Delta := 0 \)
For each scenario selected \( (\omega_j) \), \( j = 1, \ldots, n \)

Add scenario demands for all scenario years, \( \Delta^\tau := \Delta^\tau + d^\tau (\omega_j) \forall \tau \in \omega_j \)

Add period specific demand to aggregated total, \( D^{t+\tau-1} := D^{t+\tau-1} + \Delta^\tau \forall \tau \)

Truncate aggregated demand, \( D := D^{1 \ldots H} \)

### 3.2.1 Simple SDG Example

To illustrate, we construct a three year example stochastic demand. Consider

Table 3.7 below with the set of example scenarios \( \Omega = \{A, B, C\} \). Scenario A demands 1 BCT and 5k enablers for 1 year. Scenario B needs 2 BCTs and 8k enablers for 2 years. Scenario C requires 5 BCTs and 20k enablers for 4 years. The probabilities of each scenario are shown in. Table 3.8

<table>
<thead>
<tr>
<th>Scenario A</th>
<th>Year 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCTs</td>
<td>1</td>
</tr>
<tr>
<td>Enablers</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario B</th>
<th>Year 1</th>
<th>Year 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCTs</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Enablers</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario C</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCTs</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Enablers</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
Table 3.8 Example scenario probabilities.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.6</td>
</tr>
<tr>
<td>B</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
</tr>
</tbody>
</table>

There is no fixed demand, so $D^t := 0 \forall t$.

Beginning with the first period, $t = 1$.

The Poisson number generator determines there are 2 events in the first year, $t = 1$.

From the empirical CDF, scenario A and scenario B are selected, $\omega_1 = A$, $\omega_2 = B$.

Summing period 1 demands, $\Delta^1 := d^1_A + d^1_B = \left[ \frac{1}{5} \right] + \left[ \frac{2}{8} \right] = \left[ \frac{3}{13} \right]$, and $\Delta^2 := d^2_B = \left[ \frac{2}{8} \right]$.

There is no pre-existing demand, so $D^1 := \Delta^1$ and $D^2 := \Delta^2$ (Table 3.9).

Table 3.9 Example period 1 demand from scenarios A & B.

<table>
<thead>
<tr>
<th>Year 1</th>
<th>A + B</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCTs</td>
<td>3 2</td>
</tr>
<tr>
<td>Enablers</td>
<td>13 8</td>
</tr>
</tbody>
</table>

The process is repeated for year 2, $t = 2$.

Now, $n = 1$, and $\omega_1 = C$.

$\Delta^1 := d^1_C = \left[ \frac{5}{20} \right] = \Delta^{2,3,4}$.

Therefore, $D^2 := D^2 + \Delta^1 = \left[ \frac{2}{8} \right] + \left[ \frac{5}{20} \right] = \left[ \frac{7}{28} \right]$, and $D^3 := 0 + \Delta^2 = \left[ \frac{5}{20} \right] = D^{4,5}$.

The resulting demand from the first two periods is shown in Table 3.10.
Finishing with $t = 3$.

$n = 2$, and $\omega_1 = \omega_2 = A$.

$\Delta^1 := d^1_A = \begin{bmatrix} 1 \\ 5 \end{bmatrix} + \begin{bmatrix} 1 \\ 5 \end{bmatrix} = \begin{bmatrix} 2 \\ 10 \end{bmatrix}$.

$D^3 := D^3 + \Delta^1 = \begin{bmatrix} 5 \\ 20 \end{bmatrix} + \begin{bmatrix} 2 \\ 10 \end{bmatrix} = \begin{bmatrix} 7 \\ 30 \end{bmatrix}$.

Truncating $D$ for $H = 3$, the resulting demand is shown in Table 3.11.

This represents the three years of stochastic demand produced by simulation, for this simplified example.
An example stochastic demand produced by the SDG simulation for an arbitrary 10 year period is shown in Table 3.12. This output from a single run represents one possible future demand profile, or “sample path.” The corresponding event list that determined the demand is provided below in Table 3.13. Stochastic demands may be quickly generated in this manner for as many years as desired using any scenarios and unit types contained in the data set.

Table 3.12 Simulated stochastic demand for a 10-year period.

<table>
<thead>
<tr>
<th>Year</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBCT</td>
<td>25</td>
<td>6.0</td>
<td>5.8</td>
<td>1.4</td>
<td>2.2</td>
<td>4.5</td>
<td>3.9</td>
<td>7.8</td>
<td>8.1</td>
<td>10.8</td>
</tr>
<tr>
<td>SBCT</td>
<td>0.3</td>
<td>4.7</td>
<td>4.3</td>
<td>1.9</td>
<td>1.0</td>
<td>1.8</td>
<td>1.0</td>
<td>4.0</td>
<td>4.0</td>
<td>3.5</td>
</tr>
<tr>
<td>ABCT</td>
<td>0.3</td>
<td>7.7</td>
<td>5.3</td>
<td>1.3</td>
<td>2.0</td>
<td>3.8</td>
<td>1.0</td>
<td>5.6</td>
<td>5.0</td>
<td>4.0</td>
</tr>
<tr>
<td>CAB</td>
<td>2.5</td>
<td>6.0</td>
<td>5.8</td>
<td>1.2</td>
<td>0.5</td>
<td>3.0</td>
<td>2.5</td>
<td>6.6</td>
<td>6.5</td>
<td>5.8</td>
</tr>
<tr>
<td>Fires</td>
<td>0.1</td>
<td>5.0</td>
<td>4.7</td>
<td>2.3</td>
<td>0.4</td>
<td>1.4</td>
<td>1.4</td>
<td>4.4</td>
<td>4.4</td>
<td>3.7</td>
</tr>
<tr>
<td>Enablers</td>
<td>113</td>
<td>100.5</td>
<td>95.4</td>
<td>410</td>
<td>39.9</td>
<td>55.4</td>
<td>48.3</td>
<td>86.4</td>
<td>97.9</td>
<td>114.1</td>
</tr>
<tr>
<td>Total</td>
<td>303</td>
<td>207.7</td>
<td>188.5</td>
<td>86.5</td>
<td>65.1</td>
<td>110.7</td>
<td>91.4</td>
<td>189.2</td>
<td>199.3</td>
<td>217.9</td>
</tr>
</tbody>
</table>

*IBCT = Infantry Brigade Combat Team*
*SBCT = Stryker Brigade Combat Team*
*ABCT = Armored Brigade Combat Team*
*CAB = Combat Aviation Brigade*
*Fires = Field Artillery Brigade*
*Enablers = Other Supporting Personnel (thousands)*
*Total = Sum of Personnel (thousands)*
Table 3.13 Simulated contingency event list.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSCA-M</td>
<td>FHA-L(1)</td>
<td>FHA-S</td>
<td>FHA-S</td>
<td>HD-S</td>
<td>DSCA-M</td>
<td>HD-S</td>
<td>IW-L</td>
<td>DSCA-M</td>
<td>FHA-S</td>
<td></td>
</tr>
<tr>
<td>DSCA-S</td>
<td>FHA-S</td>
<td>DSCA-M</td>
<td>ShOF-S</td>
<td>DSCA-S</td>
<td>FHM-L(4)</td>
<td>FHA-L(1)</td>
<td>IW-S</td>
<td>DSCA-M</td>
<td>ES(2)</td>
<td></td>
</tr>
<tr>
<td>HD-S</td>
<td>FHA-S</td>
<td>NEO-S</td>
<td>HD-S</td>
<td>NEO-S</td>
<td>IW-S</td>
<td>ShOF-L(1)</td>
<td>MCO(2)</td>
<td>PO-S</td>
<td>DSCA-C</td>
<td></td>
</tr>
<tr>
<td>PO-S</td>
<td>DSCA-M</td>
<td>ShOF-L(1)</td>
<td>PO-S</td>
<td>PO-L(2)</td>
<td>PO-S</td>
<td>FHA-L(4)</td>
<td>DSCA-S</td>
<td>DSCA-S</td>
<td>DSCA-S</td>
<td></td>
</tr>
<tr>
<td>ShOF-S</td>
<td>MCO(1)</td>
<td>DSCA-M</td>
<td>DSCA-S</td>
<td>ShOF-L(2)</td>
<td>FHA-S</td>
<td>DSCA-S</td>
<td>FHM-L(1)</td>
<td>FHA-S</td>
<td>FHM-L(1)</td>
<td></td>
</tr>
<tr>
<td>IW-S</td>
<td>PO-S</td>
<td>FHA-L(1)</td>
<td>IW-S</td>
<td>DSCA-M</td>
<td>HD-S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-S</td>
<td>FHA-L(1)</td>
<td>PO-S</td>
<td>ShOF-L(3)</td>
<td>PO-S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IW-S</td>
<td>DSCA-S</td>
<td>DSCA-S</td>
<td>DSCA-S</td>
<td>PO-L(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES(1)</td>
<td>PO-S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSCA-S</td>
<td>PO-S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSCA-S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CC = Conventional Contingencies  
FHA = Foreign Humanitarian Assistance  
MCO = Major Contingency Operations  
ShOF = Show of Force  
L = Large  
DSCA = Support to Civil Authorities  
HD = Homeland Defense  
NEO = Non-combatant Evacuation  
STK = Strike  
M = Medium  
ES = Enforcing Sanctions  
IW = Irregular Warfare  
PO = Peacekeeping Operation  
C = Catastrophic  
S = Small

3.2.2 Stochastic Demand Verification and Validation

Our goal is to produce realistic stochastic demands using the available historic data and unclassified planning scenarios, such that updated and classified data could be easily implemented in the future. We have no expectation that the Army will validate our model. Validating the stochastic demand generator is problematic because there is no good basis of comparison available. In this section we verify each component of the stochastic demand generation model, and investigate the model output including a comparison to the historic conflict data.

Process Verification

The historic conflict data that underlies this model we take as valid from an authoritative source, CAA. The assumption of a Poisson process for number of events per year was explained and tested above in Section 3.1.1. The process for developing an
empirical distribution of contingency types and matching them with contingency planning scenarios is based on validated and accepted procedures [19]. The scenarios used are unclassified, and not approved for official resource and requirements determination. They were, however, developed to be realistic by trustworthy sources, TRAC and IDA. Any scenarios that had no historic precedent would have been identified in the category to scenario matching process, as would any historic contingencies with no corresponding planning scenarios. The planning scenarios used here do not come with the same vetting and pedigree as the classified planning scenarios, however.

The Poisson event generator and contingency selection process produces eight events per year, on average, following the empirical distribution shown in Table 3.4. The resulting total demand, Table 3.12, was checked manually to ensure the program was implemented correctly and aggregating deployment demand as anticipated.

**Output Generation and Exploration**

We have addressed the stochastic demand model inputs as well as the procedure and implementation for generating stochastic demands. Next we run the simulation and investigate the SDG simulation output. We chose to replicate the stochastic demand generation simulation for \( N = 10^6 \) because the simulated observations were relatively inexpensive, and we wanted a reasonable chance of generating rare events. The simulation length was arbitrarily set for \( H = 10 \) years to cover the 7 year forecasting period of interest plus several sacrificial periods. The result
is one million stochastic demand profiles \((D_j)\) or sample paths through the scenario tree, where \(D_j = \left[ d_{ji}^t \right] \forall t, i \forall j = 1,\ldots, N \).

We wanted to order the demand profiles from greatest to least deployment demand, but we needed to choose a metric on which to sort the demand profiles. The sample output in Table 3.12 illustrates that each simulation provides a number of values. We chose to use the “total demand” (estimated total number of soldiers from number and types of units plus enablers), but any of the unit type demands or even a weighted scoring function could be used if supported by area of analytic interest. Still, we have 10 periods of total demand values for each demand profile. Two natural metric candidates to order the multi-period demand profiles are the average (3.12) and peak (3.13) total demands taken for each demand profile.

\[
\text{Average Total Demand} = \frac{\sum_{t=1}^{H} d_{\text{Tot}}^t}{H} = \overline{d}_{\text{Tot}} \tag{3.12}
\]

\[
\text{Peak Total Demand} = \max \left( d_{\text{Tot}}^t \right) \text{ for } t = 1 \ldots H \tag{3.13}
\]

Ordering the simulation observations by average total demand \((\overline{d}_{\text{Tot}})\) is shown in Figure 3.13. The average demand increases steadily through the first 97% of observations, then increases exponentially for the most demanding 3% of cases. This is
consistent with findings from SARA [24] and FMCA [20]. Figure 3.13 also indicates a strong correlation between peak and average demand for the observations displayed. Using the full set of simulated demand data, a 0.863 correlation factor was found between average total demand and peak total demand.

![Graph showing distribution of average total deployment demand](image)

Figure 3.13 Distribution of average total deployment demand. The data indicate a sharp rise in average total demand above the 97th percentile. There is also a clear correlation between the average total demand and peak total demand for a demand profile. 200 equally spaced observations are displayed to illustrate the relationships described.

Changing the sorting metric to peak total demand \(\left(\text{max}(d_{tot}^i)\right)\) yields Figure 3.14, below. The rate of increase for peak demand begins accelerating above the 90th percentile. The bump in the peak total demand curve near the 93rd percentile suggests a bimodal distribution. This is attributed to the most demanding planning scenario,
MCO(4), with a peak demand of 387 thousand troops. Using the probability 0.0008 from Table 3.4, a Poisson parameter of 8.1, and a 10 year simulation period, MCO(4) should occur in 6.5 % of cases, which is consistent with our observations.

Figure 3.14 Distribution of peak total deployment demand. Average total demand for the same 200 observations as used in Figure 3.13. The data indicate increasing peak total demand above the 90th percentile. There is an interesting bump in peak total demand near the 93rd percentile which is attributed to the most demanding MCO planning scenario. The correlation between the average total demand and peak total demand is also visible in this graph.

To provide a perspective of the complete demand profiles, several sample paths are plotted below in Figure 3.15, chosen by equally spaced percentiles of total average demand \( (\bar{d}_{tot}) \). Most of the demand observations are below 200 thousand, with only the 95th percentile and higher significantly surpassing that threshold. This is further indication that there is quite a bit of variation in the most demanding 5% of cases. This
A group of demand profiles appears to represent the lower 90% of the demand distribution well, but does a poor job of representing the most demanding cases.

![Figure 3.15 Stochastic demand profiles by percentile.]

These are a subset of the demand profiles plotted in Figure 3.13 and Figure 3.14. This reinforces the observation that the first 95% of cases increase incrementally, while the last 5% increase dramatically. Percentiles used are evenly spaced based on average total demand.

The histogram of average total demands in Figure 3.16 shows a highly skewed distribution. The upper range of the demand distribution is a critical range in determining the force size, however. Throughout recent history, the Army has been sized adequately for the majority of possible contingencies, leaving only a small proportion of cases where available forces may be insufficient. This represents accepted risk. It is reasonable to assume that the U.S. will continue to support an Army large enough to handle the majority of contingency missions encountered. We are therefore
particularly interested in the high deployment demand cases as they represent the reasonable decision space between cost and risk of insufficient forces.

![Histogram of average total demand observations.](image)

Figure 3.16 Histogram of average total demand observations. The frequency bins are arranged in 20 equally spaced intervals covering the observed range of average total demand values. This distribution is clearly skewed.

Figure 3.15 displays the demand profiles selected by evenly spaced percentiles. By changing the spacing, better coverage can be achieved in the upper demand region, as shown in Figure 3.17. This was accomplished by calculating 10 equally spaced values of average total demand \( \{0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95\} \), and selecting the demand profile \( (D_j) \) with the closest value for average total demand \( (\bar{d}_{\text{tot}}) \). Although used here to simply illustrate the shape and characteristics of the demand distribution, the skew of the demand distribution will be an important
factor in the following section where we consider approximating the demand
distribution. The set of demand profiles in Figure 3.17 provides a much better coverage
of the upper range of the demand distribution, while sacrificing fidelity in the low and
middle ranges, which are less critical for our problem of interest.

Figure 3.17 Stochastic demand profiles by percentile of total average demand.
These percentiles are spaced unevenly based on equal intervals of average total demand. This achieves better
coverage of the upper end of the stochastic demand distribution.

Finally, we look for trends in the demand over time. The simulated demand
values for a specified set of unequally spaced percentiles are depicted by Figure 3.18.
Each period of simulation data was considered independently \( \left( d_{tot}^{t,j} \right) \) to find the
desired percentiles. The period specific demand levels are each members from different
demand profiles, shown joined by lines for each common percentile level. The graph
illustrates the rising demand trend for the first five periods, leveling off after the sixth period. This effect is caused by the simulation starting conditions, modeled after 2014 current conditions, which are lower than the simulation steady state conditions. Therefore, demands tend to increase for the first several periods until the average steady state demand level is reached. This is consistent with our expectations of the stationary Poisson process with fixed planning scenario probabilities.

![Figure 3.18 Total demand levels using skewed distribution. Each period values are independent, members of different individual demand profiles. The points in each percentile are connected to show the increasing demand trend through period 5, leveling off after period 6. This trend is present because the simulation starts with current (2014) conditions with demand below the steady state average.](image)

Furthermore, the demand is correlated between periods, shown in Table 3.14 below. The rising correlation over the initial periods is also attributed to the starting conditions’ distance from the simulation steady state average.
Table 3.14 Correlation of total demand between periods.

<table>
<thead>
<tr>
<th>Periods</th>
<th>1-2</th>
<th>2-3</th>
<th>3-4</th>
<th>4-5</th>
<th>5-6</th>
<th>6-7</th>
<th>7-8</th>
<th>8-9</th>
<th>9-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.569</td>
<td>0.623</td>
<td>0.654</td>
<td>0.671</td>
<td>0.682</td>
<td>0.672</td>
<td>0.671</td>
<td>0.570</td>
<td>0.670</td>
</tr>
</tbody>
</table>

**Comparison to Historic Demands**

Having explored the simulation output to determine the general shape and magnitude of the stochastic deployment demand, we seek to validate the quality of the output. This is difficult, however, because we lack a good basis of comparison. The FMCA stochastic demand is based on the same historic data, and follows a similar procedure, but the planning scenarios and output are classified. Comparison with a historic period is not ideal because the planning scenarios are intended to represent the future, not replicate the past. Regardless, comparison to a historic period is our best option.

The historic conflict data spans 21 years and contains data for the maximum level of soldiers deployed. It does not have information on unit types, varying levels over time, or missed demand. We use the total deployment demand from the stochastic demand generator \(D_{tot}\) to compare with the historic deployment data.

The historic demand was calculated using the contingency start and end dates to determine the portion of each year a contingency was active, and multiplying that time by the maximum contingency demand level, summing for all recorded contingencies.
The stochastic demand generator was used to create 1,000 stochastic demand profiles, each 21 years long to match the historic period length. The results, shown in Figure 3.19, support the assumption that the historic deployment levels are reasonably represented by the stochastic demand generator. The historic deployment level generally stays within the region expected for stochastic demand, and displays similar variation through time. There does not appear to be sufficient evidence to conclude the historic deployments are not from the same distribution as the stochastic demands.

![Figure 3.19 Comparison of historic to simulated troop demands.](image)

*The historic demand shows high demands for Desert Shield / Desert Storm, as well as OEF/ OIF, with low demand in the inter-war period. Five additional random profiles are displayed in the background to provide a visual indication of the general shape and variability of the stochastic demand profiles.*
3.3 Application for Stochastic Programming

Stochastic programs represent uncertainty with random variables. Typical stochastic programming applications assume that a probabilistic description of the random variables is available [27], but that is not the case for our problem. This section seeks to define the random variable, deployment demand, for implementation in a stochastic program, with a suitable discrete probability distribution.

3.3.1 Stochastic Demand Sample Space

In the SDG simulation, there are a variable number of discrete random events per period, modeled by a Poisson distribution, to determine the deployment demand \( d^i_t \) for each unit type in each period. Each event may create demand that carries over to subsequent periods.

There is an unfortunate terminology overlap between Army planning and stochastic programming for the use of scenario that requires clarification. The Army uses planning scenario to describe a single contingency or military operation. When these contingencies are arranged over time, an integrated security construct (ISC) is defined. We have been using the term demand profile or sample path to describe the deployment demand vector \( D \) derived from an ISC. In the stochastic programming domain, however, the term scenario describes a realization of the random vector \( \xi \). For our problem, \( \xi \) is the deployment demand vector or demand profile generated from a stochastic ISC, defined below by (3.14). Throughout this section we discuss
scenarios in the stochastic programming sense, though we try to maintain the distinction.

Let \( i \in I \) be an index for demand (unit) type,
where \( I = \{IBCT, SBCT, ABCT, CAB, Fires, Enablers, Total\} \).
Let \( t \) be the period (or stage) time index,
where \( t = 1, \ldots, H \).
Let \( d^i_t(\omega^t) \) be force demand dependent on a random outcome.

\[ \xi = \left( d^i_t(\omega^t) \right) \forall t, i \]

Each demand realization \( (\xi) \) is of size \( |I| \times H \) or \( 7 \times 10 \) for our example.

It is convenient here to redefine \( \omega \) as a random event representing a combination of military contingencies within a year, in order to simplify our discussion.

Now, \( \Omega \) is the set of all possible combinations of contingencies. For stochastic programs, the focus is on the random variables \( (\xi) \), and a precise definition of random outcomes \( (\omega) \) is not as important [27]. A stochastic programming scenario \( (s) \) is defined in (3.15), using the random outcomes that lead to the realization \( \xi(s) \).

\[ s := (\omega^t, \ldots, \omega^H); \omega^t \in \Omega^t \] (3.15)

In our model, \( \Omega \) is the same for each period \( (t) \), but the resulting deployment demand \( \xi^t \) is not because some military contingencies have multi-period demand that
carry over to subsequent periods. From the SDG simulation, the demand that may be realized in any period depends on the random outcomes of previous periods, or
\[
\xi^t(\omega^{\tau} \forall \tau \leq t) \text{ for } t = 1, \ldots, H.
\]

There are 38 military contingencies defined for the SDG, and for this illustration, we assume that 0 to 14 events may begin in a period. Each contingency type may occur multiple times within a period, so we use sampling with replacement. Therefore, the combination of military contingencies (\(\omega\)) possible in any period is
\[
|\Omega| = C^n_r(\max(n, r)) = 10^{12}.
\]

Each scenario, from (3.15), provides one realization of random demand vector \(\xi(s) \triangleq d_i^t(s) \forall i, t\). If we define \(\Xi\) as the support of \(\xi\), and \(K\) as the total number of possible realizations of \(\xi\), then it follows that \(|\Xi| = K \geq 10^{12}\).

### 3.3.2 Approximation Methods

The SDG simulation yields an unimaginably large sample space, even for 2 periods. A straightforward application with such a large number of states would be impossible. We consider two approaches to deal with this problem. The first is to employ a stochastic programming method that employs sampling approximation, thereby limiting the number of alternative demand profiles to evaluate. The second is to approximate the stochastic demand function with a simpler empirical distribution of deployment demand.
Stochastic Programming Sampling Approximation Methods
Several sampling approximation methods for stochastic programming were introduced in Section 2.3.3. In order to implement the stochastic demand for these methods, the stochastic demand generation function would provide a sample of scenarios \((S)\) for the stochastic program to use. The SDG simulation provides one possible demand realization \((\xi(s))\) for each replication, or \(K = N\). A sampling approach would allow the use of the SDG simulation without further defining or manipulating the underlying stochastic process. Experimentation would be necessary to determine the amount of sampling needed to achieve reasonable accuracy for our problem.

Approximation of the Demand Function
The second option is to approximate and simplify the demand function.

Approximating the demand function is complicated, however. The goal of approximation is to find a smaller discrete demand distribution that still represents the properties of the full set. Unfortunately, there is no guarantee that the approximated distribution will maintain the same characteristics as the original. We will develop an empirical distribution, in stages, from the SDG simulation output already collected, and then compare the results.

The typical approach to develop a discrete empirical distribution is to make the desired number of observations or samples, order the observations by their values, and assume they are equally likely by assigning equal probability to each. A discrete
distribution is desirable for the stochastic programming application, rather than interpolating between observations. We must select a metric to order the observations. We chose to continue with the total demand, but as previously stated, any of the other values or a weighted scoring function may be used if appropriate for the study questions of interest.

Rather than sampling the simulated demand distribution to collect observations only for the number of levels desired in the empirical distribution, we will developed our empirical distribution from the simulation output data set generated for the demand validation in 0. This data set has $10^6$ observations covering 10 periods each. By using a large set of observations and specifying which percentiles to use in the empirical distribution, we can be reasonably assured that our observations are nearly evenly spaced or specify an unevenly spaced interval. In this case we use unevenly spaced intervals in order to increase fidelity in the region above the 90th percentile of total demand, as discussed in section 0.

If we were to simply pool the demand observations $(\xi_t)$ from each period we would have a sample of 10 million observed demands to construct our empirical distribution. This assumes that the observations from each period are homogeneous, however, ignoring the increasing demand trend over the initial 6 years, shown in Figure 3.18. We therefore begin by considering each year separately.
Method A – Independent Period Demands with 10 Levels per Period

Method A treats each period independently, assuming that the demand realized in each period is only dependent on events occurring in that period:

\[ \xi^t(\omega^t) := d^t_j(\omega^t) \quad \forall i. \]

We create an empirical distribution for each period with \( L = 10 \) discrete demand levels in order to provide a good approximation of the demand distribution. A random event \( \omega_A^t \) for Method A represents the selection of one of the discrete demand levels, and is no longer tied to specific military contingencies, as before. This means that \( K^t = L = 10 \quad \forall t \). The demand vector \( \xi^t(\omega_A^t) \) is still the deployment demand realization for period \( t \).

Based on our understanding of the demand distribution from section 0 and the nature of the AFSP, we will use observation intervals are unevenly spaced, with smaller intervals in the upper range of the demand distribution. First, the observations from each period are put in rank order. Let \( j = 1, \ldots, N \) be the index of demands within a period \( \xi^t_{\text{obs}} \) ordered by \( d^t_{\text{tot}} \) such that \( d^t_{\text{tot},j} \leq d^t_{\text{tot},j+1} \quad \forall j, t \).

Next we partition the observations into \( L \) intervals with lower \( j_{a(i)} \) and upper \( j_{b(i)} \) bounds for each interval, as shown below in Table 3.15. The set of observations in each interval is \( J_i = \{ j_{a(i)}, \ldots, j_{b(i)} \} \) such that there is exactly one \( J_i \) of which each \( j \) is a member.
Let $\omega_{A,t}^j$ be a random event in period $t$ that leads to the discrete demand level $\xi_i$ in Method A. The probability of each demand level (Table 3.15) is proportional to the size of the observation interval, described by (3.16). The discrete demand level is the average of the observations in the interval, described by (3.17). The values for total demand by level ($\xi_{tot,t}^j$) are shown in Table 3.15.

$$P\left(\xi_i^j \left(\omega_{A,t}^j\right)\right) = \frac{\left|J_{l_i,t}\right|}{N} \forall l_i,t$$

(3.16)

$$\xi_i^j \left(\omega_{A,t}^j\right) := \left(\overline{d}_{l_i,j,t}^j \forall i,j\right) \forall l_i,t$$

(3.17)

In order to formalize the approach for Method A, the procedure is outlined below.
Simulate and collect demand observations \((\xi_{\text{obs}})\) (here \(N = 10^6\))

Specify intervals \((j_{a(l)})\) and \((j_{b(l)})\) \(\forall l\) to define the empirical distribution (here \(L = 10\))

Determine discrete level probabilities using (3.16)

For each period \(t = 1…H\) (here \(H = 10\))

Order observations \((\xi^t_{\text{obs}})\) from least to greatest (here \(d^t_{\text{tot}}\) was used)

For each interval \(l = 1\) to \(L\)

Calculate the empirical demand \((\xi^t_l)\) using (3.17)

With our selected design specifications, Method A provides 10 discrete demand levels for each period of the model. This means that \(K^t_A = 10\), or for all 10 periods,

\[K_A = 10^{10}\].

To create a sample demand profile or scenario \((\xi)\) using the Method A discrete empirical distribution, each period requires a unique uniform random number to choose a discrete demand level using the empirical CDF for that period.

The procedure to generate a demand profile using a Method A randomly generated scenario \((\xi(s_A))\) is provided below.

For each period \(t = 1…H\) (here \(H = 10\))

Generate \(U\) (uniform random number from 0 to 1)
Look up discrete demand value using CDF(A); \( \xi^t \left( U^t \right) \)

Stochastic demand, \( \xi \left( s_A \right) = \left( \xi^t \left( U^t \right) \right) \forall t \)

The possible demand values for each period of the Method A discrete empirical distribution \( \left( \xi^t \left( s_A \right) \right) \) are shown below in Figure 3.20, along with the observed minimum and maximum from the SDG data for reference. We observe that the deployment demands tend to rise until period \( t = 6 \), where the system appears to reach steady state. This reflects the data produced by the stochastic demand generator simulation in Figure 3.18, due to the chosen simulation starting conditions.

![Figure 3.20](image)

**Figure 3.20** Discrete total demand levels by year for Method A.
Uneven spacing achieves better coverage of the upper range of the demand distribution. The demand values are connected by lines for readability, but are not from the same demand profiles as in previous figures. The stochastic demand simulation starts with no major ongoing conflicts, and demands tend to increase until a steady state is reached in period 6.
An example of demand profiles generated in this manner is depicted in Figure 3.21, below. These demands are not dissimilar to those in Figure 3.15 (excluding the maximum and minimum reference values), which showed the demand profiles evenly spaced by percentile of total average demand. From observation, this appears to be a reasonable approximation of the original stochastic process. The demand levels and variation appear consistent with the stochastic demand generator simulation.

![Figure 3.21 Deployment demand profiles generated using Method A. Only seven demand profiles are shown for readability.](image)

Method A, however, assumes each period is independent, contrary to our findings in Table 3.14. The inherent correlation of time series data is ignored in this approach, which may have a detrimental effect on the force size analysis. Furthermore, $K_4 = 10^{10}$ possible scenarios are still too many to enumerate and evaluate. An
empirical distribution with fewer levels would be required in order to implement this method without a sampling procedure.

**Method B - Independent Period Demands with Varying Levels per Period**

For the AFSP that we are concerned with, the initial several periods are more important than the final periods. We now exploit this property to further develop an implementable approximation method.

Method B builds upon the previous procedure. Like Method A, it treats each period independently, but uses a smaller, variable number of discrete distribution levels that decrease in later periods, $L_t \geq L_{t+1}$. Assuming that the optimization model will have about 100 decision variables per period/scenario and that commercial solvers running on a PC should be able to find optimal solutions to problems with up to $10^6$ decision variables, we will consider an approximation with $K_B \leq 10^4$ scenarios. This should allow the extensive form of the stochastic program to be solved directly.

For this method, we must choose the number of discrete levels ($L_t$) to use in the empirical distribution for each period. Again considering a 10 period problem, there are many possible designs within our budget. We arbitrarily chose to develop the one shown below in Table 3.16. This design provides greater fidelity in the early periods, which are more important, and less fidelity at the final periods, which will be discarded. The product of the levels of design for each period provides the cardinality of the set of possible random vectors, shown in (3.18).
Table 3.16 Method B discrete empirical distribution levels. 
(\| = |\Omega|) for each period

\[
\prod_{t=1}^{H} L' = 4^3 \times 3^3 \times 2^2 \times 1^2 = 6,912 = K_B
\] (3.18)

The percentiles of demand covered by each discrete level are unevenly spaced, as before, to provide better approximation of the upper demand region. Table 3.17, below, shows the interval of observed demand percentiles covered by each segment of the empirical distribution \( \frac{j_{a(l)}}{N} \) to \( \frac{j_{k(l)}}{N} \) for \( l = 1,\ldots,L \) where \( L = \{1,2,3,4\} \).

Otherwise the procedure for determining the empirical demand levels \( \xi_t^l (\omega_{l,l}) \forall t, l \) is the same as for Method A, above.

Table 3.17 Method B empirical distribution partitioning for each demand level.
The discrete total demand values \( \left( \xi^t_{\text{Tot}} (s_B) \right) \) derived for Method B are shown below in Figure 3.22, for the percentile ranges listed in Table 3.17.

![Figure 3.22 Empirical distribution values for method B.](image)

This is consistent with Table 3.16 for number of levels per period and Table 3.17 for the percentile range covered by each interval. Total demand values are averaged over each interval. This is evident in the slight jump in average demand for the highest interval each time the number of levels is decreased.

Each of the possible demand scenarios was enumerated \( \left( K_B = 6.912 \right) \), so random sampling is not necessary. The joint probability for each scenario was calculated using the event probabilities in Table 3.17, shown in (3.19).

\[
P\left( \xi (s_B) \right) = P\left( \xi (\omega_B^1, \ldots, \omega_B^H) \right) = \prod_{i=1}^{H} P\left( \xi^t_B (\omega^i_B) \right) \tag{3.19}
\]
In a stochastic programming application, each of the scenarios could be evaluated, and weighted according to scenario probability. An example of scenarios from Method B is provided in Figure 3.23. The demand generated by Method B is noticeably coarser than Method A (Figure 3.21), and different scenarios tend to have many data points in common.

![Figure 3.23 Example demands generated by method B. Plotted values are perturbed to uncover masked values.](image)

The coarseness of the method B distribution may limit the effectiveness of our stochastic program. Also, Method B, like method A, does not account for correlation of demand between periods. It does, however, provide a reasonable number of scenarios which may be enumerated and evaluated.
Method C – Correlated Period Demands with Varying Levels per Period

Continuing to develop the approximation procedure, Method C addresses the correlation of demand between periods. We begin with the same variable discrete demand level design \( L \) shown in Table 3.17 and demand values developed for method B \( \xi(s_B) = \xi(s_C) \), shown in Figure 3.22. Therefore, \( K_B = K_C = 6,912 \). The adaptation for Method C is the development of conditional probabilities for state (or demand level) transitions between periods in a Markovian fashion, as in

\[
P\left(\xi_C^t (\omega_C^t) \mid \xi_C^{t-1} (\omega_C^{t-1})\right)
\]

To replace the independent probabilities used previously. In this approximation, it is assumed that only the state of demand in the current period is necessary for determining the state of demand in the next period.

In determining the transition probabilities, we began with the same SDG simulation data set of \( 10^6 \) observations. Taking two adjacent periods \( (t - 1, t) \), the pairs of observations \( (\xi_{obs}^{t-1}, \xi_{obs}^t) \) are partitioned separately by \( d_{\text{Tot}}^{t-1} \) and \( d_{\text{Tot}}^t \), respectively, into the levels specified by Table 3.17 from Method B, while maintaining the observation pairing. Within each partition, the transition probability is then the proportion of observations from level \( l^{t-1} \) that are paired with observations from level \( l^t \), as defined in (3.20). This is repeated for each pair of adjacent periods. The resulting transition matrices are shown below in Table 3.18.
Let \( l \) be the discrete demand level index for time \( (t) \).
Let \( l' \) be the discrete demand level index for time \( (t - 1) \).

\[
P\left( \xi^t_C \left( \omega^t_{C,l} \right) \mid \xi^{t-1}_C \left( \omega^{t-1}_{C,l} \right) \right) = \frac{\left| J^t_{l'} \cap J^t_i \right|}{\left| J^{t-1}_l \right|} \quad \forall l', l, (t - 1, t) \tag{3.20}
\]

Table 3.18 Method C Markovian transition probabilities.

<table>
<thead>
<tr>
<th>Period</th>
<th>( A )</th>
<th>( B )</th>
<th>( C )</th>
<th>( D )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Period 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.9</td>
<td>0.09</td>
<td>0.009</td>
<td>0.401</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Period 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.946</td>
<td>0.047</td>
<td>0.007</td>
<td>0.401</td>
</tr>
<tr>
<td>B</td>
<td>0.541</td>
<td>0.442</td>
<td>0.013</td>
<td>0.404</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>0.857</td>
<td>0.135</td>
</tr>
<tr>
<td>D</td>
<td>0.298</td>
<td>0.628</td>
<td></td>
<td>0.974</td>
</tr>
<tr>
<td><strong>Period 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.948</td>
<td>0.045</td>
<td>0.007</td>
<td>0.401</td>
</tr>
<tr>
<td>B</td>
<td>0.520</td>
<td>0.462</td>
<td>0.014</td>
<td>0.404</td>
</tr>
<tr>
<td>C</td>
<td>0.010</td>
<td>0.865</td>
<td>0.117</td>
<td>0.408</td>
</tr>
<tr>
<td>D</td>
<td>0.216</td>
<td>0.708</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td><strong>Period 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.948</td>
<td>0.044</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.515</td>
<td>0.467</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.005</td>
<td>0.872</td>
<td>0.123</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.244</td>
<td>0.756</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Period 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.948</td>
<td>0.044</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.516</td>
<td>0.467</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.006</td>
<td>0.807</td>
<td>0.187</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Period 6: A | 0.549 | 0.044 | 0.007 |
B | 0.513 | 0.469 | 0.018 |
C | 0.006 | 0.810 | 0.185 |
D |

Period 7: A | 0.548 | 0.052 |
B | 0.516 | 0.484 |
C | 0.008 | 0.993 |
D |

Period 8: A | 0.548 | 0.052 |
B | 0.507 | 0.333 |
C |
D |

Period 9: A | 1 |
B | 1 |
C |
D |

Period 10: A | 1 |
B |
C |
D |
The probability of each scenario \(\xi(s_c)\) was calculated as the joint probability of its transitions, shown by (3.21).

\[
P(\xi(s_c)) = \prod_{t=1}^{H} P\left(\xi'_t (\omega'_{C,t}) \mid \xi_{C}^{t-1} (\omega'_{C,t-1})\right)
\]  

(3.21)

Although \(\xi(s_b) = \xi(s_c)\), with discrete demand levels shown by Figure 3.22, \(P(\xi(s_b)) \neq P(\xi(s_c))\). An illustration of this difference is provided by Figure 3.24, which contains a sample of demands from Method C \(\xi_{\text{Tot}}(s_c)\). Compared with the method B example in Figure 3.23, it is apparent the scenarios with greater correlation between periods are now more likely.
Recall that for Method C, \( K_C = 6,912 \) possible scenarios, which is a reasonable number to evaluate. The example transition matrix in Table 3.18 indicates that some of the possible transitions had no observations in the \( N = 10^6 \) simulation replications. Therefore, our model approximation in method C assigns any possible scenario \( (\xi(s_c)) \) containing those transitions a probability of zero as well. All told, there are 1,800 instances where \( P(\xi(s_c)) = 0 \). Furthermore, there are many additional scenarios that are extremely unlikely; there are 1,281 instances where \( 0 < P(\xi(s_c)) \leq 10^{-9} \). The prevalence of low probability scenarios is shown by Figure 3.25 as histogram of scenario probability frequencies along a \( \log_{10} \) scale with the cumulative density function superimposed. The majority of scenarios contribute little to the CDF, as displayed in Figure 3.26.
Figure 3.25 Histogram of Methods C scenario probabilities. The majority of scenarios are extremely unlikely, and contribute very little to the overall distribution.

Figure 3.26 Cumulative distribution function for Method C. The first 63 scenarios (<1%) account for 95% of the probable outcomes.
**Method D – Consolidated Scenarios from Correlated Demands with Varying Levels**

We continue by investigating the effects of further simplifying our approximating distribution by consolidating low probability scenarios. Beginning with the distribution developed in Method C, for Method D we combine the 5,635 scenarios with \( P(\xi(s_{c,i})) \leq 10^{-6} \) into one consolidated scenario. The relative probabilities are used to calculate weighted average demands from across those scenarios, then made discrete to conform to the scenario tree based on defined demand levels for each period, as shown in (3.22). The probability of the consolidated scenario is given by (3.23).

Let \( s_{D,consol} \) be the consolidated scenario

Let \( S_{C,rare} \) be the set of Method C scenarios where \( P(\xi(s_{c,i})) \leq 10^{-6} \)

\[
\bar{\xi}^t(S_{C,rare}) := \left\{ \sum_{s_c \in S_{C,rare}} P(\xi(s_{c,i})) \times \xi^t(s_{c,j}) \right\} \quad \forall t
\]

\[
\xi^t(s_{D,consol}) := \min \left\| \bar{\xi}^t(S_{C,rare}) - (\xi^t(\omega^i_{C,j})) \right\| \quad \forall t
\]

\[
P(\xi(s_{D,consol})) = \sum_j P(\xi(s_{c,j})) \quad \forall s_{c,j} \in S_{C,rare}
\]
This results in $K'_D = 1,278$, or 18% of $K_C$. The probability of the consolidated scenario, from (3.23), is 0.00036. The consolidated scenario accounts for many high demand, low probability scenarios. Therefore the averaged demand values for this scenario are well above the overall average. The resulting CDF for method D is depicted in Figure 3.27.

![Figure 3.27 Cumulative distribution function from method D. This distribution begins identically to method C, but over 80% of the lowest probability scenarios were aggregated into one single scenario for Method D.](image)

The empirical demand distribution levels used for method D are identical to methods B and C, $\xi^i \left( \omega^i_{B,i} \right) = \xi^i \left( \omega^i_{C,i} \right) = \xi^i \left( \omega^i_{D,i} \right)$, excluding the consolidated scenario, as shown in Figure 3.22. Also, the random sample of demand profiles for method D is
the same as method C, as only the extremely rare scenarios were consolidated. This is shown in Figure 3.28, below, with the consolidated scenario colored black.

![Figure 3.28 Example demands generated by Method D. Many rare scenarios are consolidated into one, artificially selected and shown here in black. The other scenario demands are identical to method C. Plotted values are perturbed to uncover masked values.](image)

Method D is meant to provide performance and characteristics similar to Method C while requiring only ½ the number of scenarios. Any noticeable difference between the methods is expected to be well within the threshold of accuracy for our model. This may then be taken as a computational savings, or used to redesign and increase the number of discrete levels in the empirical demand distribution to increase model fidelity.
3.3.3 Approximation Method Comparison

In order to evaluate the performance of each approximation method, we compared each of the methods with the SDG simulation output used throughout this section. We used three criteria to evaluate them. Average total demand is a measure of the magnitude of the demand. Standard deviation is a measure of the variability of the demand level. Correlation between periods provides a measure of the relationship of demand levels between periods.

To review, a summary table of the methods discussed is provided in Table 3.19.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>K</th>
<th>L</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDG</td>
<td>Simulation from military planning scenarios and historic data</td>
<td>&gt; $10^{12}$</td>
<td>Nearly continuous</td>
<td>SDG*</td>
</tr>
<tr>
<td>A</td>
<td>Independent period demands; 10 levels/pd</td>
<td>$10^{10}$</td>
<td>10</td>
<td>Independent</td>
</tr>
<tr>
<td>B</td>
<td>Independent period demands; varying levels/pd</td>
<td>6,912</td>
<td>4 to 1</td>
<td>Independent</td>
</tr>
<tr>
<td>C</td>
<td>Correlated period demands; varying levels/pd</td>
<td>5,112</td>
<td>4 to 1</td>
<td>1 pd correlation</td>
</tr>
<tr>
<td>D</td>
<td>Consolidated; correlated period demands; varying levels/pd</td>
<td>1,278</td>
<td>4 to 1</td>
<td>1 pd correlation</td>
</tr>
</tbody>
</table>

SDG*: Correlated demands generated by simulation of overlapping deterministic military planning contingencies.

Monte Carlo simulation was used to generate $N = 10^5$ sample demands $\xi(s)$ with $H = 10$ periods each by using the approximation distributions developed by methods A through D. For methods B through D, the sample average corresponded to the weighted average calculated from the respective empirical distribution within 1% of
the demand value, justifying the sample size. Method A was not enumerable, and therefore a weighted average of the distribution could not be calculated.

**Demand Magnitude**

The total average demand is shown below in Figure 3.29. Each of the approximation methods represents the average total demand well. They are nearly indistinguishable from the SDG average.

![Figure 3.29 Average total demand comparison of demand approximation methods.](image)

**Demand Variation**

Next we consider the standard deviation from each method. Figure 3.30 illustrates the comparison. Method A replicates the baseline variation very well. Methods B, C, and D all under-represent the variance of the baseline distribution due to the small number of discrete demand levels \((L)\) used. The standard deviation drops to
zero in periods 9 and 10 where only one demand level was modeled. Considering the performance of method A, the under-representation of standard deviation is not a flaw of the approach, but rather a limitation based on the small number of demand levels used to reduce the overall number of possible scenarios for $K_B, K_C, K_D$.

![Figure 3.30 Standard deviation comparison of demand approximation methods.](image)

**Demand Correlation between Periods**

Last, we evaluate the correlation between the demands of adjacent periods each method provides. Methods A and B provide no correlation, as shown in Figure 3.31. Methods C and D both used the conditional probabilities based on Markovian transitions. Here they are shown to under-represent the baseline correlation, attributed to the small number of discrete demand levels used in the empirical distribution and the single period correlation approximation. The zero correlation between years 8 and 9, as
well as the perfect correlation between years 9 and 10 are both due to the single demand level used in the last two periods of methods C and D.

![Figure 3.31 Correlation comparison of demand approximation methods.](image)

**Comparison Summary**

In summary, the original stochastic demand generator provides a realistic deployment demand that can be used for our stochastic programming model, but will require a solution approximation method that involves sampling. Several distribution approximation methods were developed here that should allow a direct solution of our stochastic programming model. Method A is not implementable in its current form because it still generates too many possible scenarios. Methods B, C, and D are all related, using a variable level distribution to generate a much smaller, and manageable number of scenarios. Method C is an improvement over method B, adding correlation
between adjacent periods. This requires additional work to develop the conditional probabilities, but should have no adverse effects on computation or runtime of the stochastic programming model. Method D consolidated many low probability scenarios to simplify the empirical distribution developed in method C. It used less than \( \frac{1}{5} \) the number of scenarios with no discernable performance degradation. Method D still under-represents standard deviation and correlation, but performance should be improved if the number of scenarios represented can be increased.

The most important considerations, such as the overall stochastic program performance and solution running time will be presented in Chapter 5, after the stochastic program model is implemented.
This chapter provides an exposition of the multi-stage stochastic programming model used in this study. We begin by outlining the behavior and relationships that are central to the AFSP. We then define the Army as a system to be modeled. Next, we develop a mathematical model of the Army and consider the special structure it provides. Finally, we implement that model for optimization as a multi-stage stochastic program.

4.1 Key Relationships
This section identifies critical aspects of the system under investigation for the purpose of studying the AFSP. These shape both how we define the system of interest as well as the type and composition of the model that results.

Cost and Risk
The economic force size we endeavor to define is inherently a trade-off between the cost of the force and the risk of not meeting requirements. We are interested in base (or routine peacetime) costs, the supplemental (or overseas contingency operations) costs, and non-monetary costs and penalties. Penalties are assessed for unsatisfied deployment demand. We will seek to develop the relationship between force costs and the likelihood of meeting deployment demands.
There is also an important relationship between base and non-base costs. Force structure decisions can reduce base costs at the risk of higher supplemental and penalty costs. There is always pressure to reduce base costs (which equates to the planned budget), but doing so may result in higher overall expected costs and penalties. Helping to identify that balance point is another goal of this research.

Furthermore, in considering force structure changes we need to take a long term approach. Force structure changes take years to implement, while force demands fluctuate much more quickly. Myopic decisions may reduce costs in the near term, but have significant detrimental effects on costs and capabilities for a long time.

*Dynamic Decisions Under Uncertainty*

Future contingencies and the resulting demand for the deployment of forces are the primary sources of uncertainty in this system. Although there are other uncertain factors that could be considered such as budget changes or employment costs, the deployment demand is the only stochastic element incorporated in the model.

Force structure decisions are made on a yearly basis. These decisions involve the size and composition of the workforce as well as the number and type of units. This results in the activation and deactivation of units. Because of the long implementation time of these changes, the force structure decisions are made well in advance of knowing what the deployment demand will be when the designed force is available.

Once a deployment demand is realized, additional recourse decisions may be made. These include the number and type of units to deploy, policies affecting the
readiness and utilization of units, and policies affecting the retention of current service members. Force structure updates may then be made based on the newly available deployment demand information, which will then affect future capabilities and deployment decisions.

Throughout this process, the work force is managed by recruiting or incorporating new personnel, while existing personnel are permitted to separate in order to maintain and shape the workforce over time. Workforce targets are also set on a yearly basis, in coordination with other budgeting and force structure decisions.

As a result, the decisions made each year span from long term strategic decisions for future years to more tactical decisions for the current year.

**Operational vs. Institutional Forces**

The operational force provides deployable combat units, while the institutional force provides administration, services, and supporting functions. They compete for resources in terms of funding and manpower. The institutional force exists to create and support operational forces, and the operational force cannot function or maintain itself without institutional forces. The institutional capacity limits not only the size of the force that can be supported, but also the rate at which the force can be expanded. Furthermore, military manpower may be used in the institutional force at a lower cost than operational units, and transferred back to the operational force when force expansion is necessary.
Total Workforce Composition

The Army uses military (active and reserve) personnel, DA civilians, and contractors to meet manpower requirements. Most analysis is only focused on military manpower, however. Within the institutional force, organizations are likely to have a mixture of personnel types. Although authorization documents specify the number of each, by type, the reality is that many duty positions could be filled by multiple personnel types, and that the proportions of authorizations do change over time. Our model allows the relative proportions of the workforce to change over time to meet changing conditions.

Each component of the workforce has different costs, capabilities, and characteristics. Military manpower is only generated through a closed personnel system. Therefore, shortages (especially officers) may take a decade or more to correct. Military manpower is necessary to constitute deployable combat units. Although relatively expensive in peacetime, military manpower is the most cost effective option during deployments.

Civilians provide a less expensive labor option with an open hiring system. Civilians are rarely used for deployments, but when they are used to augment deployments, they are subject to substantial surcharges. Additionally, they leave vacancies in their parent organizations which imply further indirect costs.

Contractors provide the most flexibility to increase or decrease manpower quickly. Costs are similar to, or slightly higher than, civilians. Contractors also provide
niche services that are not economical for the Army to maintain internally, as well as large scale services that leverage economy of scale discounts from the private sector. Deploying U.S. contractors is very expensive, however. Personnel indigenous to the region of conflict (local nationals) can also be contracted to support deployed forces. They are usually inexpensive, but may only perform limited tasks and require supervision.

**Composition of Forces**

We use composition in a general sense, to discuss the component of a unit (active or reserve) as well as unit type or specialization. Units may be designated as active or reserve components, and formed from personnel of that component. Active and reserve units have different costs, readiness, and utilization rates. Reserve units are less expensive while in a part-time status, but take longer to mobilize, and are designed to deploy less frequently than their Active Duty counterparts.

Unit types are considered as well, sometimes referred to as force mix. This is the number of Infantry BCTs, Aviation Brigades, Support Brigades, etc., in each component. Unit types are mission specific, and one type of unit cannot generally be substituted for another without significant training and mission modification.

Our model allows the number of each type of unit to vary, but holds the personnel authorizations of each unit type as constant. The composition of the force is considered for total force design, as well as for selecting units for deployment.
**OPTEMPO (Operations Tempo)**

OPTEMPO, for our purposes, is a utilization rate of units being deployed.

Increasing the proportion of the force that is deployed results in increased OPTEMPO.

This means more frequent or longer deployments and greater stress on the force. That stress manifests in various ways including increased repair and maintenance costs for equipment, increased difficulty in planning for and managing Army systems, and increased rates of service members separating from the Army.

### 4.2 System Definition

The Army may be viewed as a complex amalgamation of systems and subsystems. This section provides the scope and definition of the system that will be modeled in order to address the AFSP.

#### 4.2.1 System Diagram

In an effort to find a model that is both manageable and realistic, we developed the system diagram shown in Figure 4.1. In this system, we focus on two major functions within the Army: to produce, develop, and manage the workforce; and to organize that workforce (along with other resources) into units capable of providing the services and accomplishing the missions required.
Looking at the personnel system first, the Army has a large, multi-faceted workforce with personnel constantly entering and leaving. The Army recruits and trains new personnel from the general population with the support of Institutional Forces (both recruiting and training, and general support). New personnel are added to the workforce while those seeking to leave are separated from the Army workforce.

The size and composition of the workforce present constraints on the organization of units and their allocation. The force structure (unit organization) is adjusted based on evolving needs and resources available. Operational Forces are then identified to deploy, as necessary, while Institutional Forces provide support services to the entire force. Notably, the recruiting and training function presents a capacity limit for the induction of new personnel into the workforce. The number of personnel seeking to separate is modeled as a function of the size of the force, and whether or not
they were deployed. Each of these functions consumes substantial monetary resources. Additionally, both are subject to policy constraints, laws, and regulations.

Considering the personnel system further, Figure 4.2 provides greater detail for the *Provide Workforce* function. Here we show the sub-functions *Produce New Personnel*, and *Shape Workforce*. Shaping the Workforce creates a need for new personnel, which controls the production of new personnel. There are also several other mechanisms available to shape the workforce: retention incentives, stop loss policy, separation incentives, and involuntary separation mechanisms. Each of these influences the number of personnel leaving the Army workforce, or separations.
Additional detail of the *Provide Force Structure* function is provided in Figure 4.3.

The force structure is managed by activating and deactivating units, as required by policy, demand, and budget. The Army mans, trains, and equips the force internally, with the support of Institutional Units. The Army is organized as an Operational Force (combat units capable of deployment), and Institutional Force (organizations that typically remain at home station and provide services and supplies for the entire force).

As required, the Army selects and provides units to deploy. The organization of units
and deployment status (labeled force structure here), provide the basis for costs, personnel separation demand, and the impetus for further organizational changes.

Figure 4.3 Detail of the Provide Force Structure function. Forces are used to organize operational and institutional units. The operational units can fulfill deployment requirements, while the institutional forces are necessary to support the entire force, as well as to perform recruiting and training functions.

These diagrams provide an organizational outline and graphical depiction of the key aspects of the Army that need to be incorporated in the model, as identified in Section 4.1. Each function, input, output, resource, and requirement shown in Figure 4.1 through Figure 4.3 are incorporated in the AFSP model.
4.2.2 Temporal Model of Periods and Decisions

A distinct advantage of stochastic programming is the ability to create recourse models. This type of model allows us to explicitly consider the uncertainty of future states or events, and to identify whether decisions must be made before the uncertainty is resolved, or if decisions may be delayed in anticipation of better information.

The Army must be large enough to meet an uncertain future deployment demand and sustain itself. The cost for upkeep and potential deployment of the Army should be balanced against the penalty for not meeting the full deployment demand. The suite of decisions necessary to define the force structure \( x^t \) take considerable time to plan, approve, and implement. These decisions are made in advance of knowing the deployment demand \( \xi^{t+1} = d^{t+1}(\omega^{t+1}) \) (3.14). Once \( \xi^{t+1} \) is realized, the Army makes deployment decisions, as well as force structure updates. These recourse decisions \( x^{t+1} \) are made in the subsequent period.

The decision and information cycle repeats to create a multi-period problem, as depicted in Figure 4.4. For this instantiation, we use a fixed period of 1 year. Deployment decisions affect the current period demand, while force structure decisions, due to the implementation delay, affect subsequent periods.
A multi-period formulation provides the ability to account for the uncertainty of future conditions and the timing of decisions while also considering the conditions and opportunities that may occur several periods in the future, thereby preventing myopic decisions.

4.3 Mathematical Model
A multi-period stochastic program formulation provides the modeling flexibility necessary to represent the Army functions and behaviors we require as a mathematical model for the AFSP. This section presents that mathematical model, based on the system outlined in Section 4.2.

4.3.1 Variable Definition
Basic stochastic programming notation was introduced in Section 2.3.2. Here, the specific variables for our AFSP model are defined.

**Variable Indices**
The indices shown in
Table 4.1 provide categorical factors for the decision variables that will follow. The indices used are for unit type \((i)\), personnel type \((j)\), personnel status \((k)\), unit component \((l)\), unit status \((m)\), equipment status \((n)\), cost type \((o)\), model period \((t)\), and relative time differentials \((\tau)\).

Table 4.1 Summary of variable index definitions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Index Set Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>unit type</td>
<td>{ib, sb, ab, av, fb, en, rt, gs, tt} (IBCT, SBCT, ABCT, CAB, Fires, Enablers, Recruit &amp; Train, General Spt, Transients/Students, Individual Ready Reserve)</td>
</tr>
<tr>
<td>(j)</td>
<td>personnel type</td>
<td>{ao, ae, ro, re, civ, ctr, ln} (Active Officer, Active Enlisted, Reserve Officer, Reserve Enlisted, DA Civilian, Contractor, Local National)</td>
</tr>
<tr>
<td>(k)</td>
<td>personnel status</td>
<td>{tot, av, dep, dep1, dep2, dep3, hs, mob, ir, tng, sep, vsep, rinc, sl, sinc, siv} (Total, Available, Deployed (1,2,3), Home Station, Mobilized, Individual Ready Reserve, Training, Separating, Voluntary Separation, Retention Incentive, Stop Loss, Separation Incentive, Involuntary Separation)</td>
</tr>
<tr>
<td>(l)</td>
<td>unit component</td>
<td>{ac, rc, civ, ctr, ln} (Active, Reserve, DA Civilian, Contractor, Local National)</td>
</tr>
<tr>
<td>(m)</td>
<td>unit status</td>
<td>{tot, av, dep, dep1, dep2, dep3, hs, mob, act, dact} (Total, Available, Deployed (1,2,3), Home Station, Mobilized, Activating, De-activating)</td>
</tr>
<tr>
<td>(n)</td>
<td>equipment status</td>
<td>{ex, new} (existing, new)</td>
</tr>
<tr>
<td>(o)</td>
<td>cost type</td>
<td>{bs, sup, pen} (base, supplemental, penalty)</td>
</tr>
<tr>
<td>$t$</td>
<td>time period</td>
<td>${1, \ldots, H}$</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>$\tau$</td>
<td>relative time differential</td>
<td>${1, \ldots, H}$</td>
</tr>
<tr>
<td>$\omega$</td>
<td>random outcome</td>
<td>discrete levels based on distribution</td>
</tr>
</tbody>
</table>

The model represents several different types of notional units ($i$). Each category is mutually exclusive. Infantry, Stryker, and Armored Brigade Combat teams (IBCT, SBCT, ABCT), as well as Combat Aviation Brigades (CAB), and Fires Brigades (Fires) are based on the notional units defined for the SARA model [68] using the Army’s table of organization and equipment (TOE). Enablers are not a distinct unit type, but rather an amalgamation of support units to reduce the complexity of the model. Enablers are modeled as a proportional mix of supporting units as listed by [68] using the Army TOEs and measured in thousands of personnel rather than number of units like BCTs. The remaining unit categories are also counted in thousands of personnel. Recruiting and training organizations are responsible for maintaining a stream of new personnel into the workforce and transitioning them from civilians to soldiers. General support represents the remainder of the institutional force that is not involved in recruiting or training, such as acquisitions, medical support, and installation services. Trainees, transients, holdees, and students (TTHS) is an Army accounting category for soldiers who are not otherwise assigned to a unit. For this model, TTHS does not include soldiers in initial training, who are accounted for separately.
Workforce personnel are distinguished by personnel type \((j)\), and each category is mutually exclusive. Each is accounted for in thousands of personnel. Military personnel are identified as Active or Reserve Component, and further designated as Officer or Enlisted. Department of the Army Civilians (DAC) are civilian government employees. Contractors provide additional labor to the Army workforce through contracts with private industry. Local Nationals (LN) are people indigenous to the region of operations, contracted to perform limited tasks. They are distinguished from U.S. contractors by limited capabilities and reduced costs.

Disposition of the workforce is further determined by personnel status \((k)\). All statuses are in thousands of personnel. Total personnel is the sum of mutually exclusive partitions. Available personnel may be assigned to units, which excludes personnel in training or separating. Available personnel can be further divided into deployed or home station status. Deployed personnel is the sum of deployment levels, each representing incrementally higher utilization rates. The deployment levels are an arbitrary design decisions, roughly equivalent to 1:3, 1:2 and 1:1.4 (15 out of 36 months deployed) for Active Duty (deployed : home station) ratios, and 1:5, 1:4, 1:3 Reserve Component ratios. The mobilized Reserve personnel category is used to account for soldiers activated to perform institutional duties, whereas the deployed Reservists category implicitly assumes mobilization for deployment. The Reserve component uses a number of soldiers who are on full time or active duty status, known as Active Guard Reserve (AGR). AGR soldiers are represented by the model with a minimum number of mobilized
reservists. The separations category is an adjusted count of personnel leaving the Army, after accounting for retention and separation policies. Voluntary separations are the unadjusted number of personnel desiring to leave the Army (quit or retire), based on a proportion of the force and deployment status. Retention incentives reflect bonuses paid to reduce voluntary separations. Stop loss is a temporary policy to cease voluntary separations, though some are still required for medical, disciplinary, and regulatory reasons. A separation incentive is a bonus paid to increase voluntary separations. Involuntary separations are akin to laying off a portion of the workforce.

Unit component \((l)\) is related to personnel type \((j)\). Operational units are purely Active or Reserve Component consisting of the corresponding personnel type of both officers and enlisted. Institutional organizations are also accounted for as though they are pure components, though in reality they often have a mixture of personnel types. The model treats the Institutional units in aggregate, however, thereby allowing the proportion of each component to vary to allow for the dynamic shaping of the workforce over time.

Unit status \((m)\) corresponds to personnel status \((k)\). Available units include deployed, home station, and mobilized units while excluding units activating or deactivating. The deployment, home station, and mobilized categories are analogous to the personnel status defined above. Units activating and deactivating adjust the number of available units, but are not directly related to personnel training or separating.
Equipment status \( n \) is used to track new equipment procurement requirements. Each unit must have a matching equipment set, and when existing stocks are exceeded, new equipment sets must be purchased. This model does not account for the sale or disposal of equipment.

Cost type \( o \) distinguishes between base, supplemental, and penalty costs. Base costs are predictable for a specified force structure. Supplemental and penalty costs are dependent on deployments, with the penalty costs representing non-monetary considerations.

Model time periods \( t \) are in unit of years, and are displayed as superscripts to distinguish from other indices. A secondary time index \( \tau \) is used when a variable time differential from \( t \) must be specified.

The stochastic deployment demand values are dependent upon the specific discrete probability distribution used. The level of demand within that distribution is indicated by a random event \( \omega \).

**Parameters**

There are several classes of parameters used in the model. General constraint parameters in Table 4.2 are multiplied by one of the decision variables, and are typically proportions or ratios. Other parameters, including resource constraints, cost factors, deployment demands, and discount rates are listed in Table 4.3.
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{1j}^{\tau}$</td>
<td>average proportion of new personnel completing initial training in period $t-\tau$</td>
</tr>
<tr>
<td>$\alpha_{2j}^{\tau}$</td>
<td>average proportion of separating personnel unavailable period $t-\tau$</td>
</tr>
<tr>
<td>$\zeta_{j,k}$</td>
<td>proportional contractor growth rate limit</td>
</tr>
<tr>
<td>$\rho_{j,j',k}$</td>
<td>proportional limits for total and deployed populations between components</td>
</tr>
<tr>
<td>$\nu_{j,k}$</td>
<td>proportion of population intending to voluntarily separate from the Army</td>
</tr>
<tr>
<td>$\xi_{1j}$</td>
<td>proportion of voluntarily separations willing to accept a retention incentive</td>
</tr>
<tr>
<td>$\xi_{2j}$</td>
<td>proportion of available personnel willing to accept an early separation incentive</td>
</tr>
<tr>
<td>$\xi_{3j}$</td>
<td>proportion of available personnel subject to involuntary separation</td>
</tr>
<tr>
<td>$\xi_{4j}$</td>
<td>proportion of voluntary separations that must be permitted to separate</td>
</tr>
<tr>
<td>$\beta_{j,l}^{\tau}$</td>
<td>average proportion of activating unit available in period $t-\tau$</td>
</tr>
<tr>
<td>$\eta_{1j,l}$</td>
<td>proportional activation rate limit for units</td>
</tr>
<tr>
<td>$\eta_{2j,l}$</td>
<td>proportional deactivation rate limit for units</td>
</tr>
<tr>
<td>$\psi_{i,l,j',m}$</td>
<td>proportional limits for available and deployed units between components</td>
</tr>
<tr>
<td>$\alpha_{3j}^{\tau}$</td>
<td>average proportion of new personnel in training status for period $t-\tau$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>student to training unit ratio</td>
</tr>
<tr>
<td>$\theta_{j,k}$</td>
<td>proportion of man-years in TTHS for personnel, status dependent</td>
</tr>
<tr>
<td>$a_{i,j}$</td>
<td>personnel authorizations for each unit type</td>
</tr>
<tr>
<td>$\varepsilon_{i,m}$</td>
<td>status dependent manning level requirement for units</td>
</tr>
<tr>
<td>$\gamma_{1i,j,m}$</td>
<td>general support requirement factor for units</td>
</tr>
<tr>
<td>$\gamma_{2j,k}$</td>
<td>general support requirement factor for personnel</td>
</tr>
<tr>
<td>$\delta_{i,l}$</td>
<td>proportion of the available force that is deployable for each category</td>
</tr>
<tr>
<td>$\nu_{i}$</td>
<td>proportion of non-deployed forces that may be mobilized (excluding deployment)</td>
</tr>
<tr>
<td>$\phi_{i,l,j,m}$</td>
<td>proportion of usable deployment time to satisfy requirements</td>
</tr>
</tbody>
</table>
The general constraint parameters are used to define the linear constraint relationships throughout the model. They are indexed as necessary based on the decision variable they are paired with. None of the general constraint parameters in this model are period dependent, though several take different values relative to the current time period, indicated by \( \alpha^{t-\tau} \). For example, when a particular constraint requires the value of \( \alpha \) for the current and two previous relative periods to \( t \), it would use \( \alpha^{t-\tau} \) for \( \tau = 0 \) to \( 2 \). These are useful when multiple time periods of the same decision variable are part of the same constraint.

Each of the general constraint parameters will be introduced individually in the following section, in the context of the constraint to which they apply.
Resource constraints are indexed as necessary based on the variables of the constraints where they apply. The budget and personnel end strength limits may be specified by period to match actual policy or any desired conditions. The resource constraint parameters will be introduced in the following section, in the context of the constraint to which they apply.

Cost factors are objective function coefficients to determine the cost of decisions. Cost factors are fixed for this model, and do not vary over time or by stochastic event. Each cost factor is indexed by cost type \( o \) as well as the indices to which it is matched.
Stochastic deployment demand \( \left( d_i^t \right) \) provides the uncertainty in this model. It is indexed by time period \( (t) \), as well as by unit type \( (i) \).

Finally, the discount rate parameter \( (r^t) \) is used to modify the relative weight of costs over time, as part of the objective function. Force structure decisions made for the near term are more critical than more distant future decisions. The conditions of the long term decisions are more uncertain, and there is more flexibility to make adjustment in future decision cycles. The discount rate parameter is used to model that behavior. Any arbitrary weighting scheme may be specified, though we use the model similar to fixed compounding interest where \( 0 < r \leq 1 \) and \( t \) represents the exponential power congruent with the time period.

**Decision Variables**

The decision variables defined for this model indicate the number of personnel \( (x) \), number of units \( (u) \), amount of unmet deployment demand \( (y) \), equipment sets \( (e) \), and total costs \( (z) \). Total costs \( (z) \) are implemented as a linear expression rather than an independent decision variable. Table 4.4 provides a list of the decision variables.
### Table 4.4 Summary of variable definitions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{j,k}^t$</td>
<td>personnel</td>
<td>personnel type, status</td>
</tr>
<tr>
<td>$u_{i,j,m}^t$</td>
<td>units</td>
<td>unit type, component, status</td>
</tr>
<tr>
<td>$y_i^t$</td>
<td>unmet deployment demand</td>
<td>unit type</td>
</tr>
<tr>
<td>$e_{i,n}^t$</td>
<td>equipment sets</td>
<td>unit type, status</td>
</tr>
<tr>
<td>$z_o^t$</td>
<td>total cost (objective variable)</td>
<td>cost type</td>
</tr>
</tbody>
</table>

Personnel variables ($x_{j,k}^t$) are indexed by personnel type (such as active duty enlisted or civilian), and personnel status (such as available, deployed, or initial training). All personnel variables are in units of thousands of personnel.

Unit variables ($u_{i,j,m}^t$) are indexed by unit type, component, and status. Unit type indicates the primary function of the unit (such as infantry, support, or training). Unit component is analogous to personnel type. Operational units are usually purely active or reserve, but institutional organizations may be a mixture of personnel types. Although the model accounts for units as if they are purely one component, it treats them in aggregate to allow for mixtures of personnel. Unit status is analogous to personnel status, except that units are activated and deactivated rather than recruited or separated.

Unmet deployment demand is represented by ($y_i^t$) for each unit type. The number of equipment sets ($e_{i,n}^t$) for each unit type, indicate new and pre-existing
equipment. The total cost variables \( z^t \) are used to aggregate costs for the objective function, by period and type.

### 4.3.2 Objective Function

With decision variables defined, we present the general form of the AFSP beginning with the objective function. The stochastic program’s objective function (4.1) minimizes the cost of the recourse function which is dependent on random future event outcomes indexed by \( \omega^t \). The discount factor \( r^t \) reduces the importance of future decisions, consistent with less certainty about future conditions and more flexibility to make budgetary changes. Although the cost types in this model are simply added, a different weighting scheme could easily be incorporated depending on study needs.

Equation (4.2) defines the recourse function as the expected value of future decisions based on the realization of the random vector \( \xi^t(\omega^t) \), containing the deployment demand, and the current period’s decisions.

Equation (4.3) defines a generic period \( t \) value function based on the previous period decisions \( (x^{t-1}, u^{t-1}, y^{t-1}, e^{t-1}) \) and the realization of random variables \( (\xi^t) \). The future costs are captured in the recourse function \( Q^{t+1}(x^t, u^t, y^t, e^t, \xi^{t+1}(\omega^{t+1})) \).

Finally, the terminating recourse function for period \( H \) is defined by (4.4). In this fashion, the discounted expected value of each future period of the model is nested within the objective function.

The multi-period stochastic program objective function can be written as
\[
\min z = \sum_o \left( \sum_j \sum_k \left( c_{o,j,k} x_{j,k}^1 \right) + \sum_i \sum_l \sum_m \left( v_{o,l,m} u_{i,l,m}^l \right) + \sum_i \left( q_{o,j} y_i^l \right) + \sum_i \sum_n \left( f_{o,i,n} e_{i,n}^l \right) \right) + \\
Q^2 \left( x^1, u^1, y^1, e^1, \xi^2 (\omega^2) \right),
\]
(4.1)

where the period \( t \) recourse function is

\[
Q^t \left( x^{t-1}, u^{t-1}, y^{t-1}, e^{t-1}, \xi^t (\omega^t) \right) = E_{\xi} \left[ Q^t \left( x^{t-1}_{j,k}, u^{t-1}_{i,l,m}, y^{t-1}_i, e^{t-1}_{i,n}, \xi^t (\omega^t) \right) \right] \forall t,
\]
(4.2)

and the period \( t \) value function is

\[
Q^t \left( x^{t-1}, u^{t-1}_{i,l,m}, y^{t-1}_i, e^{t-1}_{i,n}, \xi^t (\omega^t) \right) = \\
r^t \sum_o \left( \sum_j \sum_k \left( c_{o,j,k} x_{j,k}^t \right) + \sum_i \sum_l \sum_m \left( v_{o,l,m} u_{i,l,m}^t \right) + \sum_i \left( q_{o,j} y_i^t \right) + \sum_i \sum_n \left( f_{o,i,n} e_{i,n}^t \right) \right) + \\
Q^{t+1} \left( x^t, u^t, y^t, e^t, \xi^{t+1} (\omega^{t+1}) \right) \forall t < H.
\]
(4.3)

The terminating period \((H)\) value function, however, does not contain an
additional recourse function:
4.3.3 Personnel Constraints

Now we present the various classes of constraints pertaining to personnel management and behavior. Each of these constraints involves personnel decision variables \((x_{j,k})\) exclusively.

**End Strength Continuity**

The total personnel, of each type, depends on the previous population as well as gains and losses in the current period. Contractors do not have a training period, and may be separated at will, therefore they are not included here.

\[
x_{j,tot}^t = x_{j,tot}^{t-1} + x_{j,lng}^t - x_{j,sep}^t \quad \forall j = \{ctr, ln\}
\]  

(4.5)

**End Strength Cap**

The end strength cap \((h_j^t)\) represents an arbitrary limit set on the total number of personnel, by component. For contractors, this is applied to institutional (home...
station) functions only. This constraint may be enforced or removed depending on the nature of the analysis.

\[
\begin{align*}
    x_{ao,\text{tot}}^t + x_{ac,\text{tot}}^t & \leq h_{ac}^t \\
    x_{re,\text{tot}}^t + x_{re,\text{tot}}^t & \leq h_{rc}^t \\
    x_{\text{civ,\text{tot}}}^t & \leq h_{\text{civ}}^t \\
    x_{\text{civ,hs}}^t & \leq h_{\text{civ}}^t
\end{align*}
\] (4.6)

**Personnel Availability**

Personnel availability defines the average work force available for assignment to a unit within a period, excluding those in initial training or in the process of separating.

The parameter \( \alpha_1^T \) indicates the average proportion of period \( (t - \tau) \) new personnel completing initial training, becoming available for assignment. The parameter \( \alpha_2^{\tau} j \) indicates the average proportion of period \( (t - \tau) \) where separating personnel are unavailable. In practice, the addition of new personnel to the available workforce and loss of separating personnel are apportioned over one or two years, depending on the length of training or transition, which varies by personnel type. Additionally, U.S. Military Academy and R.O.T.C. cadet populations are represented by a fixed growth component \( (g_{j,k}) \).
Personnel Availability Categories

Available personnel limits the total of status sub-categories for deployment, home station, and mobilized (for Reserves).

\[ x_{j,av}^t = x_{j,av}^{t-1} + \sum_{\tau=0}^{3} \left( \alpha_1^\tau x_{j,ing}^{t-\tau} - \alpha_2^\tau x_{j,sep}^{t-\tau} \right) + g_{j,lng} \quad \forall j \neq \{ctr, ln\} \]  \hfill (4.7)

Soldier Deployment Total

The sum of the soldier deployment status sub-categories defines the deployment total for each soldier type.

\[ x_{j,av}^t = x_{j,dep}^t + x_{j,hs}^t + x_{j,mob}^t \quad \forall j \neq \{ctr, ln\} \]  \hfill (4.8)

Civilian and Contractor and Availability

The contract work force is flexible, with no training requirement or lengthy separation process. This model distinguishes between U.S. or allied contractors used on a routine basis, and local national contractors used at whatever locale to which the
Army deploys units. The contractor growth rate is limited by a proportion of the previous force, plus a fixed amount. All contractors in the model are available for assignment, and the contractor workforce may be reduced without explicit constraints. The proportional component limiting contractor growth is $\zeta_{j,k}$, while the fixed component limiting contractor growth is $g_{j,k}$.

$$x_{j,k}^i \leq \zeta_{j,k} x_{j,k}^{i-1} + g_{j,k} \quad \forall (j,k) = \{(\text{civ, av}), (\text{ctr, hs}), (\text{ctr, dep})(\text{ln, dep})\} \quad (4.10)$$

**Population Ratios**

The population ratios reflect the bounds of acceptable population sizes within relation to one another to manage the composition of the force. This constraint applies to both the total personnel strength, as well as the deployed force. It represents both political requirements, as well as practical necessities. For example, many specific combat support unit types exist largely in the reserve component. Also, the Army may desire to provide the Reserve Component with deployment experience, even if it is more expensive than Active Component alternatives. The proportional population ratios between components are represented by $\rho_{j,j',k}$. This constraint also applies to combinations of populations, for example, Active Officers and Active Enlisted represent all Active Duty personnel when combined.
Recruiting Floors
A steady stream of soldiers must be recruited each year even when force size is declining in order to maintain the proper rank and experience structure necessary for the Army’s hierarchal system. The recruiting floors represent the minimum number of recruits necessary per year, as a proportion of the respective soldier category \(\zeta(0_j)\), to maintain a healthy workforce.

\[
\begin{align*}
\rho_{j,j',k} x^t_{j,k} &\leq x^t_{j',k} & \forall j, j' \neq j, k = \{\text{tot, dep}\} \\
\rho_{1,t',k} \sum_{j \in I} x^t_{j,k} &\leq \sum_{j' \in I'} x^t_{j',k} & \forall k = \{\text{tot, dep}\}; \text{ where } I \subseteq J
\end{align*}
\] (4.11)

Voluntary Separations
Soldiers and Civilians routinely retire or choose to seek different employment.
We assume a fixed separation rate for each personnel type, allowing for deployed soldiers to have higher separation rates. Civilians deploy on a voluntary basis, and
should not impact separation rates. The proportion of a population intending to voluntarily separate is indicated by \( \rho_{jk} \) for each personnel type and deployment status.

\[
x^{t}_{j,sep} = \sum_{k} \rho_{jk} x^{t-1}_{j,k} \quad \forall j \neq \{ctr, ln\}; \quad k = \{dep1, dep2, dep3, hs, mob\}
\]  

(4.13)

**Retention Incentives**

There are several population control mechanisms available to the Army. The first one we present is the retention incentive. The Army may offer cash bonuses in order to reduce the number of enlisted personnel who would otherwise voluntarily separate.

The parameter \( \zeta_{1,j} \) is the proportional limit of personnel voluntarily separating that are willing to accept a retention incentive.

\[
x^{t}_{j,inc} \leq \zeta_{1,j} x^{t}_{j,sep} \quad \forall j = \{ae, re\}
\]  

(4.14)

**Stop Loss**

When necessary, the Army may suspend the voluntary separation of soldiers.

This policy is known as stop loss. This option typically has a high penalty cost. Stop loss and retention incentive personnel cannot exceed voluntary separations.
Separation Incentives

Like the retention incentives, personnel who might otherwise continue service can be offered incentives to separate early to reduce the force size quickly. This option applies to the available population of soldiers and civilians, represented as a proportion of the available force by $\zeta^2_j$. It is unnecessary for contractors.

\[
x_j^{i,sl} \leq x_j^{i,sep} - x_j^{i,rinc} \quad \forall j = \{ao, ae, ro, re\}
\]  

(4.15)

Involuntary Separations

Both soldiers and civilians can be involuntarily separated, or fired. This option requires centralized planning and approval along with a selection process, which requires a lead time. The proportion of the available force that can be involuntarily separated is represented by $\zeta^3_j$. This is unnecessary for contractors, whose contracts can be allowed to expire.

\[
x_j^{i,rinc} \leq \zeta^2_j x_j^{i,rv} \quad \forall j = \{ao, ae, ro, re, civ\}
\]  

(4.16)
Separation Minimum

Some military personnel will always be permitted to leave the workforce for a variety of administrative and regulatory reasons. For civilians, employee turnover is necessary to keep the workforce in balance and adjust to internal changes in job specialties and locations. This is established by a minimum proportion of voluntary separations \( \zeta_{j} \) that must be permitted to leave service.

\[
x_{j,sep}^t \leq \zeta_{j} x_{j,sev}^t \quad \forall j = \{ao, ae, ro, re, civ\} \tag{4.17}
\]

Population Controls

The four control mechanisms above influence the overall separation rate. The resulting separation total is defined here. The lag time of involuntary separations is reflected in (4.19), and may be adjusted as necessary. Additionally, officer candidates are selected from the enlisted ranks, reflected in the second form of the population control constraint. Recall that USMA and ROTC cadets are modeled with a constant parameter \( g_{j,mg} \).

\[
x_{j,sep}^t \geq \zeta_{4_j} x_{j,sep}^t \quad \forall j = \{ao, ae, ro, re\} \tag{4.18}
\]
4.3.4 Unit Constraints

This section outlines the classes of constraints necessary for modeling unit behavior. Each of these constraints involves unit decision variables \((u_{i,j,m}^t)\) exclusively.

**Unit Continuity**

Available units are determined from the previous number of units, plus units activated, and minus units deactivated. Unit activations and deactivations are decided in advance. Activation may take one or more periods, indicated by \(\beta_{i,j}^\tau\) as the proportion of an activating unit available in a specified period. Deactivations are modeled as one period.

\[
x^t_{j,sep} = x^t_{j,sep} - x^t_{j,sl} + x^t_{j,sinc} + x^{t-1}_{j,siv} \quad \forall j = \{ao, ro, civ\}
\]

\[
x^t_{j,sep} = x^t_{j,sep} - x^t_{j,rinc} - x^t_{j,sl} + x^t_{j,sinc} + x^{t-1}_{j,siv} + x^{t-1}_{j,ing} \quad \forall (j, j') = \{(ae, ao), (re, ro)\}
\]

\[
u_{i,j,av}^t = u_{i,j,av}^{t-1} + \sum_{\tau=1}^{3} (\beta_{i,j}^\tau u_{i,j,act}^{t-\tau}) - u_{i,j,act}^{t-1} \quad \forall i, l = \{ctr, ln\}
\]

**Units Available**

Any units activating or deactivating are unavailable. Available units are either deployed, at home station, or possibly mobilized in the case of reserve units.
Unit Growth Rate Limit

These constraints provide a practical limit on the rate at which new units can be activated. The activation rate limit \( \eta_{1,i,j} \) is proportional to the current force size plus a constant \( g_{1,i,j} \). Contractors are excluded because they are limited by the number of personnel available. TTHS units are groups of individuals, not activated and deactivated.

\[
\begin{align*}
    u^i_{t,i,av} &= u^i_{t,i,depl} + u^i_{t,i,hrs} + u^i_{t,i,mob} \quad \forall i, l = \{ctr, ln\} \\
    \eta_{1,i,j} &\leq u^i_{t,i,act} + g_{1,i,j} \quad \forall i = \{tt\}, l = \{ctr, ln\} \\
    u^i_{t,i,ht} &\leq (1 + \eta_{1,i,j}) u^{i-1}_{t,i,ht} + g_{1,i,j} \quad \forall i = \{tt\}, l = \{ctr, ln\}
\end{align*}
\] (4.22)

Unit Reduction Rate Limit

The unit reduction rate constraints provide a practical limit on the rate at which units can be deactivated \( \eta_{2,i,j} \) based on the size of the current force plus a constant \( g_{2,i,j} \). This constraint parallels the unit growth rate limit, above.

\[
\begin{align*}
    u^i_{t,i,act} &\leq \eta_{2,i,j} u^i_{t,i,av} + g_{2,i,j} \quad \forall i = \{tt, tte\}, l = \{ctr, ln\} \\
    u^i_{t,i,hrs} &\geq (1 - \eta_{2,i,j}) u^{i-1}_{t,i,hrs} - g_{2,i,j} \quad \forall i = \{tt, tte\}, l = \{ctr, ln\}
\end{align*}
\] (4.23)
Unit Deployment Total

The model represents three discrete levels of unit deployment utilization, or OPTEMPO. The deployment total is the sum of each of the deployment categories

\[ u_{i,l,dep}^t = \sum_m u_{i,l,m}^t \quad \forall i, l; m = \{dep1, dep2, dep3\} \tag{4.24} \]

Unit Component Ratios

Component ratios \((\psi_{i,l,l',m})\) limit the number of units of one component with respect to another. This constraint applies to available, deployed, and mobilized units. Similar to the preceding personnel component ratios, these constraints may apply to groups of units, such as all military (Active and Reserves), as well as all non-military (Civilians and Contractors).

\[ \psi_{i,l,l',m} u_{i,l,av}^t \leq u_{i,l',av}^t \quad \forall i; l \neq l', m = \{av, dep, mob\} \]

\[ \psi_{i,l,l',m} \sum_{l \in I} u_{i,l,av}^t \leq \sum_{l' \in I} u_{i,l',av}^t \quad \forall i, m = \{av, dep, mob\}, I \subseteq L \tag{4.25} \]
**Unit Deployment Limit**

The number of units that may be deployed, by deployment category, is limited by the proportion of the available force \( \delta_{i,m} \) for each component. Contractors are excluded here because their deployments are limited by personnel available and the proportion of the total deployment force.

\[
\begin{align*}
    u^t_{i,m} & \leq \delta_{i,m} u^t_{i,av} & \forall i = \{ac, rc\}; l; m = \{dep1, dep2, dep3\} \\
    u^t_{i,m} & \leq \delta_{i,m} u^t_{i,av} & \forall i = \{civ\}; l; m = \{dep\}
\end{align*}
\]  

(4.26)

**RC Unit Mobilization Limit**

The model implicitly considers mobilization requirements for deploying units. This constraint governs the use of reserve units in support of domestic authorities or for institutional support roles, based on a proportion of forces available \( \nu \).

\[
\begin{align*}
    u^t_{i,rc,dep} + u^t_{i,rc,mob} & \leq \nu \left( u^t_{i,rc,av} - g3_{i} \right) + g3_{i} & \forall i
\end{align*}
\]

(4.27)

**RC Full Time Support Minimum (AGR)**

The Reserve Component has a number of support personnel who are on full time active duty. This is represented in the model by a minimum number of mobilized
reserve institutional support units. The level of support is fixed because it is a rather political decision, though it could be written as proportional to the size of the Reserve Component or the number of Reserve Component units, if desired.

\[ u_{i,rc,mo}^t \geq g_3^i \quad \forall i \]  

(4.28)

4.3.5 Constraints Linking Personnel and Units

This collection of constraints defines the model relationships between personnel \( (x_{i,j,k}) \) and units \( (u_{i,j,m}) \).

**Recruiting and Training Capacity**

The recruiting and training capacity constraint ensures that there is adequate institutional support for the recruiting and training load. The training requirement for all personnel types must be less than or equal to the training support available from all recruiting and training units. Multi-year training programs are represented by including training populations from previous years in the workload summation and adding the fixed cadet (officer candidate) populations. The proportion of period \( t-\tau \) that trainees are in training status \( \left( \alpha_{3,t}^j \right) \) accounts for training time requirements. The number of training and support unit members needed for each student is \( \sigma \). Training requirements for the continuing development of soldiers is included as well.
where \( j \in \{ao, ae, ro, re\}; i \in \{tto, tte\}; j' \in \{ro, re\}\)

\[
\sum_j \sigma_{j,ng} \left( 4g_{j,ng} + \sum_{\tau=0}^3 \alpha_j^\tau x_{j,ng}^{i-\tau} \right) + \sum_i \sigma_{i,ac} u_{i,ac,av} + \sum_j \sigma_{j,av} x_{j,av}^{i} \leq \sum_l u_{r,l,hs} \quad \forall l
\]

(4.29)

**TTHS Account**

TTHS stands for trainees, transients, holdees, and students. It is the Army’s accounting category for soldiers who are not assigned to normal unit, for a variety of reasons. We have accounted for initial entry training separately, but use this category to represent active duty personnel undergoing routine reassignments \( (\theta_{1,j,k}) \) as well as lifecycle professional development training requirements \( (\theta_{2,j,k}) \). Here we have allowed for the normal demand as a proportion of personnel at home station, as well as deferred demand from deployments.

\[
u_{i,hs}^j \geq \sum_k (\theta_{1,j,k} + \theta_{2,j,k}) x_{j,k}^{i-1} \quad \forall (i, j) = \{(tto, ao), (tte, ae)\}, k = \{hs, dep\}
\]

(4.30)


**Personnel Assignment**

Each unit type has an authorized number of personnel, by type and component \( (a_{i,j}) \). The number of units in the force structure is limited by the amount of personnel available to man them. Manning levels \( (e_{i,l,m}) \) vary, depending on the unit status. Institutional units also have prescribed authorizations, often incorporating a mixture of components. They are modeled here in aggregate, however, by separately accounting for each component contribution. Deployed contractors are not necessarily from the same pool as non-deployed contractors, therefore they are treated differently, as shown in (4.31). There is an inherent relationship between the personnel type \( (j) \) and the unit component \( (l) \). Active Component officers and enlisted \( (j) \) both comprise Active Component units \( (l) \). The same is true for the Reserve Component. Therefore specifying the personnel type \( (j) \) as in (4.31) implies knowledge of the unit component \( (l) \).

\[
\sum_{i} a_{i,j} \sum_{m} e_{i,l,m} u_{i,l,m}^l \leq x_{j,av}^l \quad \forall j = \{\text{ctr, ln}\}, j \rightarrow l \quad (j \text{ implies } l)
\]

\[
\sum_{i} a_{i,j} \epsilon_{i,l,m} u_{i,l,m}^l \leq x_{j,av}^l \quad \forall j = \{\text{ctr, ln}\}, j \rightarrow l, m = \{\text{hs, dep}\} \quad (4.31)
\]
**Personnel Deployment**

The number of deployed personnel is determined from the units that are designated to deploy. The model used tiered deployment categories to account for increased stress and cost of soldiers at higher OPTEMPO rates, but civilian deployments are assumed voluntary. Personnel authorizations per unit \( (a_{i,j}) \) are as above. The Reserve Component personnel required for mobilization uses the same constraint structure. Like personnel type \( (j) \) and the unit component \( (l) \) from (4.31), personnel status \( (k) \) and the unit status \( (m) \) have a special relationship. In the case of the specific index values used in (4.32) both indices must use the same value, such as deployed or mobilized.

\[
x^i_{j,k} = \sum_i a_{i,j} u^l_{i,j,m} \quad \forall j = \{ao, ae, ro, re\}; j \rightarrow l; k = \{dep1, dep2, dep3\}; k \rightarrow m
\]
\[
x^i_{j,k} = \sum_i a_{i,j} u^l_{i,j,m} \quad \forall j = \{civ\}; j \rightarrow l; k = \{dep\}; k \rightarrow m
\]
\[
x^i_{j,k} = \sum_i a_{i,j} u^l_{i,j,m} \quad \forall j = \{ro, re\}; j \rightarrow l; k = \{mob\}; k \rightarrow m
\]

(4.32)

**General Institutional Support Requirements**

Every person in the Army and every unit (including institutional units) require institutional support. This is allocated on a per capita \( (\gamma_{2,j,k}) \) and per unit \( (\gamma_{1,i,j,m}) \) basis to determine the minimum institutional force size.
4.3.6 Other Constraints

This section of constraints includes unmet demand, equipment requirements, budget restrictions, and non-negativity constraints.

**Equipment Set Limit**

Each unit must have an equipment set. Equipment may be ordered at the same time a decision to activate a unit is made.

\[ \sum_{i} \sum_{l} \sum_{m} \gamma_{i,l,m} u_{i,l,m}^t + \sum_{j} \sum_{k} \gamma_{j,k} x_{j,k}^t \leq \sum_{l} u_{g,n,l,s}^t \]  

(4.33)

**Equipment Purchase**

Once the existing equipment has been allocated, new equipment must be purchased in order to create more units. It is assumed that the delivery of equipment is built into the unit activation delay.

\[ \sum_{l} \sum_{m} u_{i,l,m}^t \leq e_{i,ex}^t + e_{i,new}^t \quad \forall i; m = \{av, act, dact\} \]  

(4.34)

\[ e_{i,ex}^t = e_{i,ex}^{t-1} + e_{i,new}^{t-1} \quad \forall i \]  

(4.35)
**Deployment Demand**

The deployment demand is either met with the required unit types, or unmet demand is accumulated and a penalty is assessed. The proportion of usable time \( (\phi_{i,m}) \), for each deployment type, is accounted for by the deployment parameters.

\[
d^t_i = \sum_l \sum_m \left( \phi_{i,m} u_{i,m}^t \right) + y^t_i \quad \forall i,l,m = \{dep1, dep2, dep3\}
\]  

\[(4.36)\]

**Budget**

Budget constraints \( (b^t) \) pertain to the base portion of the budget. They may be implemented or omitted depending on the nature of the analysis.

\[
z_{hs}^t \leq b^t
\]

\[(4.37)\]

**Non-negativity**

All decision variables are restricted to non-negative values.
4.3.7 Multi-stage Considerations

Multi-stage problems present some unique challenges. The variables and constraints outlined above must be generated for each period of the problem, and for each scenario represented. The scenario information about random events must be carefully organized in order to ensure that each periods' decisions only use the information available at that period, i.e., decisions cannot anticipate future scenario information. To that end, each decision defined in section 4.3.1 was given an additional scenario index, so that a distinct decision variable is created for each variable type, period, and scenario considered.

There are various methods for controlling the scenario information. We chose to represent scenarios in their entirety, and use explicit non-anticipative constraints to manage the scenario information.

This section first addresses non-anticipativity, then the need to form a single, consistent multi-period plan based on a scenario-dependent multi-period stochastic solution.

$$
\begin{align*}
  x_{j,k}^t & \geq 0 \\
  u_{i,m}^t & \geq 0 \\
  y_i^t & \geq 0 \\
  e_{i,n}^t & \geq 0
\end{align*}
$$
Non-anticipativity

Implementability was introduced in Section 2.3.1, where the decisions at any node in the scenario tree must be consistent for all scenarios that incorporate that node. Solutions may not differentiate outcomes based on events that have not yet occurred; the model may not anticipate future random outcomes based on information not yet available.

Non-anticipativity constraints can take a number of forms. The notation and constraints developed here are consistent with [38], the seminal work in scenario aggregation. We defined a scenario as a series of random events (3.15). Then \( S \), where \( s \in S \) for all \( s \), is the finite set of scenarios. If two scenarios \( (s, s') \) are indistinguishable based on the information available at time \( t \), let them belong to the same scenario bundle \( (A_{\omega_1, \omega_t}) \). The index on \( A \) specifies a scenario path up to period \( t \), which corresponds to a node in the scenario tree at period \( t \). This index was not included in the original notation, but added for clarity.

Furthermore, let \( A^t \) be the set of all scenario bundles at time \( t \), such that \( A_{\omega_t} \in A^t \) and \( A^{t+1} \subseteq A^t \). A simple form of the non-anticipativity constraints is expressed by (4.39) and applies to each decision variable.

\[
x^t(s) = x^t(s') \quad \forall (s, s') \in A_{\omega_1, \omega_t}, \forall A_{\omega_1, \omega_t} \in A^t, \forall t \quad (4.39)
\]
This leads to a sequential implementation of (4.39), where each constraint uses the variables of two consecutive scenarios within the same bundle. A simple example is shown in (4.40) where \( H = 3, \omega' \in (0,1) \forall t \), and \( x^l \) represents the vector of all decision variables for a period. The decision variables are indexed by scenario which is designated by the random events that define that scenario \( (s_{\omega' \ldots \omega''}) \).

\[
\begin{align*}
&x^0(s_{000}) = x^0(s_{001}) \\ &x^0(s_{001}) = x^0(s_{010}) \\ &x^0(s_{010}) = x^0(s_{011}) \\ &x^0(s_{100}) = x^0(s_{101}) \\ &x^0(s_{101}) = x^0(s_{110}) \\ &x^0(s_{110}) = x^0(s_{111}) \\ &x^1(s_{000}) = x^1(s_{001}), A_{00} \\ &x^1(s_{001}) = x^1(s_{011}), A_{01} \\ &x^1(s_{010}) = x^1(s_{011}), A_{00} \\ &x^1(s_{100}) = x^1(s_{111}), A_{11} \\ &x^1(s_{110}) = x^1(s_{111}), A_{11} \\ \end{align*}
\]

(4.40)

There is no need for non-anticipative constraints for the final period, as each scenario bundle will contain only a single scenario and there is no additional information that the model may anticipate.
**Multi-period Policy (simple)**

We desire a multi-year plan as a result of solving the AFSP. Not all decision variables are necessary, or appropriate for this plan. For the AFSP, the overall number of personnel and available units must be specified in a plan up front. Recourse variables, those decided in response to the unfolding of events, should be set after the demand is known. The number and types of units to deploy are examples of recourse variables. The decision variables relevant to a multi-year plan are referred to as policy variables here.

The non-anticipative constraints ensure that a single variable solution is produced for each scenario in the scenario bundle \( (\forall s \in A) \). The policy variables, however, must be decided several periods in advance, and therefore have a different, more restrictive form of non-anticipative constraints than the recourse variables.

A simple solution would be to require all decision variables that are in the policy variable set to take on the same values period wide, regardless of scenario, for the policy time horizon \( (P) \). Such an approach does not permit the model to take advantage of any information regarding the scenario or demand events, and is equivalent to making all policy decisions at time \( t = 0 \). This can be accomplished efficiently using the previously defined scenario bundles.

\[
x' (A) = x' (A') \quad \forall A_{w, \ldots, w} \in A', \ t \in P
\]  

\[\text{(4.41)}\]
In our model, the policy variables are the personnel totals by type, and units available by type, as shown below.

\[
\begin{align*}
  x_{j,tot}^t & \quad \forall j, t \leq P \\
  u_{i,l,av}^t & \quad \forall i, l, t \leq P
\end{align*}
\]  

(4.42)

While (4.41) ensures that a consistent policy is produced, we find that it is also more restrictive than the actual system we are modeling. In reality, although a consistent policy for \( P \) periods must be decided at \( t = 0 \), those decisions can be revisited and incrementally adjusted each year, after new information is available. This modeling deficiency is addressed by the following alternative set of enhanced policy constraints.

**Multi-period Policy (enhanced)**

Here the previous policy constraints (4.41) are adapted to better reflect the true system. Decisions must still be made at \( t = 0 \) regarding all policy decision variables up to \( t = P \). Now the decisions at each period may modify the previous policy variables, within specified limits set by previous decisions along the scenario path using the scenario bundles.
This approach requires some additional notation. The vector of policy variables is represented as \( x \). We define \( (T) \) as the period about which a decision is being made. The scenario bundle identifies the information available at the decision point, and therefore implies the period in which the decision was made. Finally, \( x^{(T)}(A_{\omega_1...\omega_T}) \) specifies the policy variables relevant to period \( T \) determined at node \( A_{\omega_1...\omega_T} \). The scenario information available at a decision node is synonymous with a scenario bundle designation.

There are two constraint types required. The first inequalities, (4.43), provide bounds on the updated value of a policy variable, while the second inequalities, (4.44), require all decision variables within a scenario bundle to equal the policy variable for that bundle.

\[
\begin{align*}
    a^{(T-i)}x^{(T)}(A_{\omega_1...\omega_{T-i-1}}) &\leq x^{(T)}(A_{\omega_1...\omega_T}) \leq b^{(T-i)}x^{(T)}(A_{\omega_1...\omega_{T-i-1}}) & \forall T \leq P \\
\text{where } a &\leq 1 \leq b
\end{align*}
\]

(4.43)

\[
\begin{align*}
    x^{(T)}(A_{\omega_1...\omega_T}) &= x^t(s) & \forall s \in A_{\omega_1...\omega_T}; t = T \leq P
\end{align*}
\]

(4.44)

The policy bounding constraints (4.43) sequentially link the final policy variable for each scenario bundle, back to the original policy decisions made before any
information was available. Each successive period away from a policy decision allows the policy value to be modified incrementally, within the bounds provided \((a, b)\), from the earlier policy decision. In our implementation, \((T - t)\) is an exponent, so that no deviation from the plan is permitted one period in the future, but increasing flexibility is provided in each subsequent year.

The policy to decision variable constraints (4.44) ties the final policy variable at each scenario bundle (or decision node) to the related decision variable at that node. These were implemented by leveraging the standard non-anticipative constraints that were previously emplaced and equating the first decision variable in each scenario bundle to the policy variable.

This process is illustrated by a three-period example with two possible outcomes per random event in Figure 4.5.
Figure 4.5 Three period example of enhanced policy constraints. Highlighted equations trace the constraints for \( s_{100} \) decisions back to the original policy decision.

### 4.3.8 Model Characteristics

The resulting extensive form of the MSSP described above is a linear program (LP). Because we are concerned with tractability of the model solution, integer variables and non-linear expressions were not considered for inclusion at this time.

Although integer variables could be useful for further model development, it is not necessary to form a useful strategic level model, as we intend to demonstrate.
The only stochastic element in this model is the deployment demand \( d_t' \) which manifests as a right-hand-side (rhs) constraint value as shown in (4.36). Therefore, this model also has fixed recourse, that is, the constraint coefficients of the recourse variables which comprise the recourse matrix \( W \) are fixed. The objective function coefficients are also fixed. Under these conditions the stochastic program objective function is convex [27]. Furthermore, Jensen’s inequality is valid, which states that for any convex function \( f(\xi) \) of \( \xi \), \( E f(\xi) \geq f(E \xi) \). This provides a lower bound for the optimal solution.

The AFSP model does not possess complete recourse, where feasible decisions at any period will lead to a feasible solution in the subsequent period. In a simpler form, the model can be reformulated to have complete recourse because any deployment demand that cannot be met is penalized through the objective function. The constraints necessary to manage the workforce, unit assignments, deployments and budgetary restrictions, however, can lead to infeasible problem solutions, even when feasible decision variable values can be found at earlier periods. The lack of complete recourse would complicate the implementation of a nested decomposition algorithm.

4.4 Modeling Summary
The AFSP model provides a multi-period representation of the Army personnel and force structure systems. It is designed to find the minimum cost policy to manage
the total Army workforce and unit organizational structure over time, while considering the uncertain nature of future deployment demands.

A variety of constraints control the behavior of the model and decisions it may consider. Proportional constraints govern the relative size of workforce components as well as unit types. The recruiting, training, and support capacity of the institutional force limits the growth and size of the operational force. Training and reset requirements as well as logistical realities limit the availability of deployable forces. Budgetary and manpower constraints can be enforced or ignored depending on the nature of the study questions.

Special non-anticipative constraints and policy constraints carefully control how scenario information is used by the model to ensure accurate representation of the planning and decision process as well as an implementable solution.

In order to meet our stated objectives, the model accounts for both planned (base) and contingency (supplemental) costs, as well as assessing likelihood of deployment demand going unsatisfied. Non-monetary costs, such as missed demand are accounted for by a penalty cost. The model solution recommends how the workforce and force structure should change incrementally over a number of periods rather than simply offering an objective force without consideration of how to achieve it. The long lead time of force structure changes is explicitly modeled. It accounts for the support and generation capacities of the institutional force to create and maintain the operational force, though these relationships are immature at this time. Ongoing
research sponsored by the Army Staff should yield better understanding of these systems, and enable better modeling. The model considers the need for and interaction of military, civilian, and contract personnel, though we recognize the need for better data concerning contractors. Finally, the model also allows the flexibility to adjust readiness policies (ARFORGEN cycles) in order to increase the supply of deployable units in times of high demand.

The model focuses on the development and management of the workforce, the force structure by number and type of units, and the deployment of those units to meet contingency demands. Only a very cursory treatment of equipment procurement cost is provided, and readiness requirements are assumed constant. These are areas that warrant further study in order to better represent the total cost of the Army and the internal competition for funding and personnel.

This model, like any other, is a simple representation of the true system. It is expected that stakeholders of any future analysis based on this model would review and adjust the model to meet their own study requirements, as well as update the parameters, cost factors, and starting conditions with new information. Despite that caveat, the AFSP model provides a unique strategic level perspective on Army workforce and force shaping decisions that should prove insightful.
This chapter contains the numerical results from experimentation with the AFSP model. It outlines model implementation and solution techniques before proceeding to compare demand distributions from Section 3.3.2, select model settings, and draw insights for Army force size requirements.

The AFSP model is designed to recommend a multi-year force size plan. Therefore we are primarily concerned with the overall size of the workforce and the ability to meet deployment demands. The AFSP is not intended to manage or recommend troop deployments or make detailed force structure decisions, though it does make simple decisions in order to assess the force size and unmet demand metrics we are interested in.

5.1 Model Implementation

5.1.1 Environment

The AFSP model described in section 4.3 was implemented using Python 2.7.5 with Gurobi 6.0.2 optimizer. A scripted language, such as Python, is preferable over standard mathematical programming packages in order to provide the control and flexibility to generate the stochastic program as well as implement specialized solution algorithms when necessary.
The program reads scenarios from a data file that contains the name, probability, and demand vector for each scenario in the scenario tree. The scenario name, as described in 4.3.7, contains all the information necessary for the model to reconstruct the scenario tree.

All experimentation was performed on a 64bit PC with a quad-core 2.2GHz processor and 8GB of RAM.

5.1.2 Parameter Estimation
Determination of the cost factors and other parameters necessary to implement the AFSP model was not meant to be a central part of this research effort, and is therefore only given brief consideration here.

Cost Factors
Costs were determined primarily from three sources. AM COS (Army Military – Civilian Cost System) is a validated Army personnel costing model. The Army Forces Cost Model is a validated Army costing model for units with modules for both base costs, and contingency operations (supplemental costs). Finally, the Army Planning, Programming, Budgeting, and Execution Data Warehouse (PPBE-DW) provides planned and actual costs covering the entire Army budget.

Base personnel costs are from AMCOS tables, including federal (non-Army) costs, aggregated for Officers and Enlisted Soldiers weighted by personnel authorizations of each rank. Civilian costs were likewise aggregated to arrive at a single average cost. Contractor costs are not easy to determine, as there is no direct link between contract
costs and number of personnel required, but only the service provided. Efforts are ongoing within the Army and DoD to better understand contractor costs. For our purposes, we use civilian cost as a surrogate, with an added surcharge for the flexibility and risk assumption that a contractor provides along with a profit margin.

Supplemental personnel costs are provided by the Army Contingency Costing Model (ACM) within Forces. Items from the cost estimate most attributable to personnel were used to estimate the deployment supplemental costs. Deployed civilian and contractor rates are determined from their base cost (above) multiplied by an additional deployment factor from anecdotal deployment salaries. Penalty costs for stop-loss and involuntary separations are arbitrarily assigned, proportional to base salary.

Operational unit base and supplemental costs are provided by the Army Forces Cost Model. All model units are based on the brigade notional force structure developed for SARA [68]. The *Enabler* unit type is a weighted average of the support unit types based on Army inventory, and grouped into units of 1,000 soldiers each, since the size of each brigade type varies. All institutional, installation, and personnel costs are apportioned to the units by the model, therefore personnel costs must be removed for our application. Unit activation and deactivation costs were also provided by Forces. There are no penalty costs for units.

Institutional units are not included in the Forces model, and the Army budget does not distinguish organizational costs from the services they provide. Therefore, a
forces unit cost was generated with all operational equipment removed to estimate the sustainment costs of an institutional unit. Like the enablers, the units are generic groups of 1,000 personnel each.

New equipment set costs are also provided by Forces, using the full authorized equipment sets for each unit. The enabler equipment cost is a weighted average of support units in the current inventory.

Unmet deployment demand is assessed an arbitrary penalty cost. For our purposes, it is more convenient to consider a single unmet demand index score, rather than each of the individual unit types. The unmet demand index used here is based on the standard deployment cost of each active component unit type with personnel costs included, normalized to the cost of an IBCT. A scaling factor for this penalty score will be considered in numerical analysis.

Resource Constraints
Both budget constraints and manpower constraints were taken from the PPBE-DW programmed values for years 2014 through 2020, and held constant thereafter. The budget values were modified with an 8% increase to overcome model feasibility issues. In the Army’s current downsizing, rather severe resource reductions have required careful management and policy modification. These are the types of decisions we would not want the model to assume without human oversight, therefore it is desirable that the AFSP should find these conditions infeasible. Additionally, the non-Army costs included in the personnel estimates had to be accommodated.
General Parameters
Parameters for personnel proportions were taken from PPBE-DW positions where possible. For example, the proportion of military personnel in the institutional force during the operational surge in Iraq is taken as a minimum value. Others are taken from the FY14 position, and afforded reasonable range to set upper and lower bounds for the parameters.

Voluntary separation rates are based on a simple representation of the current Army rank structure and approximate experience level. This is also the basis for determining continuing professional schooling requirements for TTHS as well as training unit resources. The cost and effectiveness of retention incentives was set arbitrarily, and should be updated with data from actual retention programs.

Training times and availability of soldiers are based on average training requirements for officers (warrant and commissioned) and enlisted personnel. Civilian hiring delays are based on the typical proportion of positions authorized, but vacant. Available manpower is averaged over the year, but personnel constraints are represented as end strength or the target strength at the end of the year. These constraints are parameterized assuming a linear transition from one target strength to the next.

The unit activation rate represents a variety of real growth limitations in terms of how many units may be activated per year. Although this model parameter was arbitrarily set at a 0.1 proportion, real force planning data should be used.
The combat unit types represent actual brigade size units, except for the enablers, which are an aggregate of many different types of support units. Because these support units are represented in aggregate, it was much easier to satisfy the enabler deployment demands than the other specific unit types. Therefore, an artificial parameter was added to reduce the availability of enablers, and adjusted so that enabler availability and utilization was in line with the other unit types.

Personnel requirements for each unit type were extracted from the Forces Cost Model because they were readily available once the notional force structure was built. The personnel data in the model is provided by the Army Force Management Support Agency, the authoritative office responsible for documenting force structure requirements.

The student to trainer (and recruiter) ratio is based on the FY14 number of authorized recruiting and training related positions and the estimated number of student-years of training required. Likewise, the total institutional support requirement is based on the FY14 force structure and institutional support authorizations. Support requirements were apportioned to each unit type based on their yearly estimated operating cost. The relationship of institutional and operational forces is a current topic of study within the Army. RAND developed a good categorization and mapping of institutional functions [23], and CAA has promising work underway using linear regression to identify the predictive factors for support requirements. This could
provide the basis for a much more detailed and accurate institutional representation in the future.

**Deployment Representation**

The AFSP models units in aggregate rather than as distinct entities. Therefore, ARFORGEN (unit training and deployment cycles) is considered implicitly, unlike most simulation models where the units and rotations are explicitly represented. To accomplish this, three deployment levels were considered for the ASFP model to represent a steady state rotational requirement, as well as moderate and high surge levels. Recall that full surge ARFORGEN is not modeled here because it requires the case by case modification of policy.

The deployment levels used in the AFSP model are based on deployed : home station ratios of 1:3 (9:27 months), 1:2 (12:24 months) and 1:1.4 (15:21 months) for Active Duty units, and 1:5 (12:60 months), 1:4 (12:48 months), 1:3 (12:36 months) for Reserve Component units.

Assuming that a total of one deployment month is consumed in travel (both directions) and transition with the outgoing unit if necessary, deployment mission availability is directly calculated. Operational costs are determined using the contingency operations module of the Forces Cost Model for each of the cases above. All Reserve Component deployments assume three months of mobilization time to complete pre-deployment training requirements and de-mobilization. Combined with
the travel and transition requirements, a one year Reserve Component deployment
covers 8 months of mission requirements.

All costs and deployment utilization times were normalized to a one year time
period to be consistent with the AFSP model requirements, using one year time periods.
For example, a steady state Active Component deployment (1:3) has a \(\frac{8}{9}\) deployment
utilization ratio, and \(\frac{9}{8}\) unit-years of deployment would be necessary to meet each
year of deployment demand. The model accounts for unit-years of deployment time in
an aggregated fashion, rather than the actual number of distinct units necessary to
deploy.

Each deployment level represents a segment of a piecewise-linear function of
deployment cost. Deployment demands are filled using the steady-state (dep1) method,
until no more units are available. Then the model may use the surge policies to deploy
additional forces if required. Because longer deployments are more efficient (smaller
proportion of training, travel, and transition times) penalty costs are used to make the
higher deployment levels less desirable. This is consistent with reality, although long
deployments are less expensive, the Army chooses to use more frequent, shorter
deployments for the sake of morale and distributing experience, so long as demand can
be met under these less efficient policies.

In the future, integer constraints may be used to restrict each model period to a
single readiness policy which would be more accurate and intuitive for parameterization
and interpretation. Integer constraints were avoided here, however, where model tractability is a primary concern.

**Starting Conditions**

The AFSP model requires a set of values for each of the decision variables in order to represent the model starting point. Several long lead time items, such as soldier recruiting / training and unit activations require multiple previous period decisions.

The starting conditions for personnel and unit amounts in our implementation of the AFSP model are based on FY14. Total personnel authorization values for soldiers and civilians were obtained from the PPBE-DW. Unit activations and deactivations, as well as recruiting and separations were set to represent a force at steady-state.

An equivalent number of contractor work-years is difficult to determine given current accounting practices and data available. This is another area of ongoing research within the Army, and not one that we intend to solve here. For the AFSP starting conditions, we assume that the number of contractor man-years was equal to the number of civilian authorizations for general peacetime operations. Although this is a reasonable assumption, it is not precise. The model can indicate the general need to increase or decrease contractor levels, but better data concerning contractor use is necessary before the figures are reliable.

Although any starting conditions of interest may be used and analyzed using the AFSP model, care must be taken to ensure the starting values are consistent with the constraint parameters. Modification of the starting conditions without consideration of
the constraint parameters will likely lead to infeasibility in the initial period of the problem. Once this issue is resolved, the model maintains feasibility unless unreasonable budget or manpower restrictions are imposed.

5.2 Initial Trials and Solution Methods
This section briefly outlines the initial models and methods used to develop solutions used throughout the subsequent experimentation. Additional detail regarding software settings for implementation is provided in Appendix A for the interested reader. All experimentation was performed using the same set of starting conditions and constraint parameters.

5.2.1 Initial Models
A variety of simple model instances were used during the early testing and validation period. Models were tested for up to 10 periods in order to ensure we could produce a reliable plan for up to 6 periods (a 5 year plan beginning 1 year in the future).

The expected value (EV) problem is a simple, single scenario (deterministic) formulation using the expected value (or mean value) of each of the stochastic components (2.7). The 10 period EV formulation solved in less than a second.

Additional single scenarios were developed and solved one at a time, as in a wait and see (WS) approach (2.6). Constant demand WS scenarios were useful for model verification, as discussed in Section 5.3.1.
Method Z

A simple demand distribution was created for the initial stochastic trial, and named Method Z here to provide a convenient name for reference. Method Z is similar to Method A (Section 3.3.2) except that the observations are partitioned into equal sections so the average demand derived from each partition has the same probability of occurring. This method is discussed further in Section A.2.2.

5.2.2 Solution Methods

Solution of extensive form models up to 8 periods long was possible using default optimizer parameters. By using the barrier method only, without solving for an extreme point solution, models with 10 periods and 1,000 scenarios were solvable. This approximation method produced very accurate solutions that were both feasible and implementable. Objective function values were accurate to 7 significant digits, and the decision variables varied less than 0.0003%. Informal testing showed that models of up to 1,500 scenarios were generally solvable using the barrier method only approximation.

Increasing the system memory would expand the range of solvable problems. In addition, several promising solution techniques, such as progressive hedging, were identified in Section 2.3.3 which could aid in solving larger problem instances.

5.3 Model Output Verification

In this section we verify the model by examining the output from single scenarios as well as stochastic runs.
5.3.1 Verification Scenarios

In order to verify the model was operating as intended, we considered several scenarios, each run independently. This is often referred to as a wait and see (WS) problem, as described earlier.

These scenarios were based on the demand data from the SDG used in Section 3.2.2 to develop the probability distribution approximations. Four scenarios are used, each at a different constant percentile of demand, for each period. Level 0 (L0) is the average of all scenario demands, this is also the mean value (EV) demand level.

Level 1 (L1) is the average of the top 10% of demand. Similarly, Level 2 (L2) is the average of the top 1% of demand, and Level 3 (L3) is the average of the top 0.1% of demand. The demand levels for the total demand index are shown by Figure 5.1.
Figure 5.1 The 4 levels of deployment demand for single scenario model verification.

5.3.2 Single Scenario Verification
Model parameters such as the penalty factor for missed demand are arbitrarily set at 5, which will be addressed later. We do not try to draw insights for the Army from these results, but rather, we look for decision values that are reasonable, and trends that behave as we would expect. This is a form of face validation.

A sample of recommended workforce levels is provided in Figure 5.2. They show that workforce requirements increase as the deployment demand increases, as is expected.
Enlisted soldier strength is shown, but Officer strength follows the same pattern, only on a smaller scale. The Level 2 values are identical to the Level 3 values, and therefore obscured in Figure 5.2. This suggests that the Level 2 demand forces the model to its limits, and cannot adapt further or quicker to accommodate the higher demand in Level 3.

We see a reduction of reserve forces, even at the highest demand levels. This is caused by the divestiture of Reserve Component institutional personnel. The model does not currently recognize a need or advantage of Reserves in the institutional force,
therefore it opts for the less costly forms of labor, and reduces the institutional Reserve forces to minimum levels. This behavior could be controlled with additional knowledge and data concerning Reserve institutional forces.

It also appears that contract labor is more flexible than the civilian workforce from the faster rate of change, as well as more expensive than the civilian workforce from the greater overall reduction in the more peaceful Level 0 scenario.

Finally, we observe a drop-off of personnel levels, especially noticeable with the Reserve Component. This is attributed to the model termination anomaly that we intend to mitigate through sacrificial ending periods.

Expected costs and unmet deployment demand are shown in Figure 5.3. Base costs track closely with total personnel levels. Supplemental costs increase at each level as deployment demand (Figure 5.1) increases for each level. Penalty costs are highly correlated with missed demand. At this setting, penalty costs dominate base and supplemental costs.
The sharp increase in supplemental costs for Level 2 & 3 demands are due in large part to the increasing number of contractors being deployed. The contractor effect is also seen in the penalty cost and unmet demand. For high demand levels, Military forces are immediately deployed at peak sustainable rates and it is several years before additional units can be activated. Deployed contractors continue to ramp up over the first 3 years, reducing the missed demand and resulting penalty costs.
A sample of unit force size recommendations are displayed in Figure 5.4. At low demand levels (L0) Active units are deactivated to a greater extent than Reserve units, and at high demand levels, Active units are increased to a greater extent than Reserve units. This conforms to our expectations, that Active forces are more adaptable. The time delay between seeing an increased demand and units becoming available is also apparent.

Figure 5.4 Comparison of selected unit recommendations from single scenario models.
It seems unusual that the number of AC IBCTs is greater for moderate demand (L1) than for high demand (L2&3). The model gains more benefit from concentrating on increasing enabler units in these circumstances. This is due in part to a proportional constraint between the number of contractors that can be deployed and the number of military enabler units.

The AFSP model appears to be providing reasonable recommendation that are intuitive and justifiable for these single scenario examples. An additional scenario was run for the highest 0.01% of observed deployment demands (L4). It was not included in the discussion above because the decision variable values were identical to L2 & L3 while only the penalty cost and unmet demand values differed.

5.3.3 Stochastic Scenario Verification
Model verification was extended to the stochastic demand context. Although we reserve formal testing of the penalty weighting parameter until later, we do want to ensure that the model responds appropriately to changes in unmet demand penalty costs. To look at both of these issues, we evaluated the Method Z demand distribution (Section 5.2.1) under two different penalty factors.

The single scenario results were simple to report, as there was only one set of decision variables per period. For the Method Z model used here, there are a set of decision variables for each scenario, or $\Omega = 2^H = 1,024$ scenarios. The results reported below are the averages of the decision variables for each scenario weighted by the scenario probability.
A sample of the results of the Method Z model solutions are shown in Figure 5.5, which confirmed our expectations. We observed a declining force size with Method Z, similar to the (L0) single scenario because Method Z does not include high demand levels or have much variation in demand. There is an increased expected cost from the stochastic model solution over the deterministic. For each personnel and unit type, the high penalty version was at or above the standard penalty level. Operating costs were higher with the increased penalty, and increasing the penalty factor eliminated unmet deployment demand for the Method Z high penalty case. There were, however, some residual penalties for personnel reduction decisions.
5.4 **Comparison of Probability Distributions**

This section compares AFSP model results generated by implementing the empirical distribution approximation methods developed in Section 3.3.2. The goal of these experiments is to draw inferences regarding the impact of each on the quality of AFSP solutions and select a distribution for further experimentation, not to draw inferences about the Army force size or structure.
5.4.1 Even vs. Skewed Probability Distributions (Method Z vs. Method A)

Method Z is an example of an even distribution, where the observations are partitioned into equally sized groups, and probability of the demand levels developed are all equal. Based on the shape of the demand distribution from Section 3.2.2, we hypothesized that a skewed distribution like Method A [Section 3.3.2] would be more effective on our limited budget.

Our modeling standard is for 10 periods, consistent with our stated goals. A 3 level model would require $3^{10} = 59,049$ scenarios, which is well beyond our capability to generate or solve at this time. Therefore, this experiment is limited to 2 possible demand levels at each period for $2^{10} = 1,024$ scenarios.

We know that the demand based on the upper half of observation for Method Z will not represent the high demand levels well. We expect that Method A, based on a 90%/10% partition of demand observations will provide a distinctly different solution which more accurately represents higher demand levels.

Across all personnel categories, the Method A workforce levels were at or below the Method Z levels, as seen in Figure 5.6 Total Personnel. Expected costs diverge after period 3, with the Method Z model predicting higher costs. We did not expect that the force size and cost would be lower for Method A, but we attribute that to the selection of the penalty parameter. With only a 10% chance of experiencing the high demand state at any period, it appears less expensive to pay the penalty than to maintain the Army necessary to meet demand.
The most striking difference is in unmet demand. As we predicted, the Method Z distribution leads to a suppressed expectation of unmet demand, both in terms of average and the maximum unmet demand. We are interested in studying model behavior under the highest deployment requirements, therefore, it is important to use a probability distribution that will produce high demand scenarios.

Figure 5.6 Comparison of Method Z vs. Method A model solutions. The average and maximum unmet demand are markedly different between the two probability distribution methods.
Although anecdotal, we observed an increase in solution time for the model formed using the Method A distribution, which we consider significant. We attribute this to the disparity of scenario probabilities which affects the objective function coefficients.

![Figure 5.7 Anecdotal computational times for Method Z and Method A model solutions.](image)

### 5.4.2 Fixed vs. Declining Demand Levels (Method A vs. Method B)

Method A does not provide much fidelity in the way of demand levels. For a 10 period problem, on our computational budget, we can only afford 2 demand levels per decision node. Since the initial periods of the problem are of greatest interest, and we intend for the final periods of the problem to be sacrificial, it would be nice to allocate model fidelity accordingly. Using Method A, doubling scenarios each period, the sacrificial periods are very expensive. Method B uses additional demand levels in the early periods of the model, reducing to a single demand level at the end.
The Method B design developed in Section 3.3.2 called for 6,912 scenarios. This design is too large to generate the AFSP model. Therefore, alternative designs were developed for testing with the AFSP model (Table 5.1).

<table>
<thead>
<tr>
<th>Design</th>
<th>Pd 0</th>
<th>Pd 1</th>
<th>Pd 2</th>
<th>Pd 3</th>
<th>Pd 4</th>
<th>Pd 5</th>
<th>Pd 6</th>
<th>Pd 7</th>
<th>Pd 8</th>
<th>Pd 9</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1152</td>
</tr>
<tr>
<td>B2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>972</td>
</tr>
</tbody>
</table>

The Method B1 and B2 designs should provide better solutions for the initial model periods than the previous Method A. The quality of decision recommendations should decline in later periods. A selection of metrics comparing the solutions using Method A, B1, and B2 are provided in Figure 5.8.
Figure 5.8 Comparison of Method A (fixed levels) with the Method B (diminishing level) models.

The Method B2 model’s force size and expected cost are very similar to Method A. The Method B1 model recommends a smaller force size with a correspondingly lower cost, much the same as we observed the Method A force size was smaller than that from the Method Z model. This is consistent across all personnel types, though only the total force is shown here. The smaller force requirement observed using Method B1 is attributed to the greater model fidelity of 4 demand levels, though the reason 3 levels are not significantly different from 2 is unclear. Unit recommendations (not depicted) followed the same general pattern as personnel, with Method B2 very similar to
Method A until the single level demand, and Method B1 showing the same characteristics, but at slightly lower levels.

The average unmet demand from all three methods begins similarly. It is obvious, however when the Method B1 and B2 distributions reduce to a single average demand level. In both Method B1 and B2 cases, there is a noticeable increase in the average unmet demand just before dropping to zero. The models reduces force size early in anticipation of reduced demand levels across all scenarios, resulting in increased missed demand for that period.

The stepwise reduction in demand levels is clearly visible in the maximum unmet demand graph. There are not corresponding discernable jumps in the force size and force structure recommendations until there is only 1 demand level at the end of the design.

Having seen the noticeable effect of using the single level demand in the final periods, we questioned the benefit of including the final periods. In order to investigate, another model was generated and solved based on the Method B1 probability distribution, truncated at 6 periods just before the single scenario periods.

Many of the recommended decision levels between the 6 and 10 period Method B distribution models were very similar, including total personnel, total cost, and unmet demand. The selection of decision variables in Figure 5.9 highlights the difference, however. The initial periods are similar, but the second half is suspect. Producing only 2 or 3 periods of reliable decisions does not meet our problem criteria. Therefore, the
single level demand periods at the end are necessary in order to extend the number of useful decision periods. Furthermore, we favor Method B1 over B2 because it appears to provide a better estimation of the extreme demand levels with little if any detrimental effect over the first 5 periods.

![Figure 5.9 Difference between solutions using 10 period and 6 period Method B1 distribution.](image)

The solution times for this experiment are shown in Figure 5.10. Although Method B1 only has 20% more scenarios than Method B2, it required double the
computational time. The truncated Method B1 model solves considerably quicker than the full version, due to the reduction in model size.

Figure 5.10 Anecdotal solution times for the Method B experiments.

5.4.3 Independent vs. Correlated Time Period Demands (Method B vs. Method C)

The correlation of demand between periods was evident from Section 3.2.2. Accounting for the period demand correlation in the probability distribution of the demand variable should improve the accuracy of the AFSP model. This correlation should encourage the multi-year adjustment of force structure, either up or down, rather than just riding out variations expecting that they will be short lived. Given the number of years it takes to adjust the force structure, and the limited time horizon of our model, we were uncertain whether or not the inclusion of correlated demands would have any noticeable effect on the AFSP solution. The additional work to develop
the correlated joint probability distributions is performed prior to generating and solving the model, so it should have little effect on solution times.

The same demand levels and scenarios from Method B1 were used to generate the Method C1 distribution. The only difference is that the probabilities have been adjusted to reflect a single period correlation as observed from the simulation data (Section 3.3.2). The Method C experimental results provided in Figure 5.11 demonstrate that correlation has a clear effect on model results.

![Figure 5.11 Comparison results from Method B1 and Method C1 experiments.](image_url)
The total personnel graph shows higher personnel levels resulting from the model where Method C1 was used. This is the trend across all personnel and unit types. The increased force size generates higher expected costs. Supplemental costs are higher using the Method C1 distribution because it is more likely to have to sustain consecutive high levels of deploying units, which is more expensive. Similarly, the average unmet demand shows that it is more difficult to meet the deployment demand under the correlated demand conditions, even though the demand levels are identical and force levels are higher in the Method C model. The seemingly anomalous change in unmet demand for Method C1 at period 4 is due to the transition to the single demand level where the model correlation is perfectly independent (0), then perfectly correlated (1). The dip in supplemental cost at period 5 is also due to the transition to the single average demand level.

We conclude that the correlated demands are important for accurate AFSP model representation.

5.4.4 Complete vs. Consolidated Demand Probability Distributions (Method C vs. Method D)
In developing the transition matrices for the correlated probability distribution in Section 3.3.2 there were several cases where no observations within our data set resulted in scenario probabilities of zero. There were many more instances where the scenario probabilities were very small. We supposed that by eliminating all scenarios with probability zero, and consolidating scenarios with small likelihoods, that we could create a new probability distribution that was less expensive to analyze, and acceptably
similar to the original. Our goal was to find a computational savings without sacrificing much in the way of the richness of the low probability, high demand scenarios.

The simulated data set included $1 \times 10^6$ observations, so the cutoff for consolidation was arbitrarily set at $1 \times 10^{-6}$ for a scenario’s probability. The demands for all consolidated scenarios were averaged, weighted by probability, and the consolidated probability was added to the scenario corresponding to that averaged demand level (Section 3.3.2). This procedure was performed on the Method C1 distribution, resulting in a Method D1 probability distribution that included 420 scenarios of the original 1,152 Method C1 scenarios.
The comparison of AFSP model solutions using the Method C1 & D1 probability distributions (Figure 5.12) shows that the recommended force levels and expected costs are indistinguishable. The personnel and unit level recommendations are nearly identical for all design variables. Only the unmet demand metrics show a difference between the two. Using Method D1, the average unmet demand is slightly over-estimated, while the maximum unmet demand is under-estimated. This effect is intuitive, as some of the most extreme cases were among the most unlikely which were consolidated, but the average demand was maintained, causing the average unmet
demand to increase slightly. We deemed this an acceptable trade-off and will use the computational savings (Figure 5.13) to create higher fidelity designs.

![Anecdotal times from Method C1 and Method D1 experiments.](image)

**Figure 5.13** Anecdotal times from Method C1 and Method D1 experiments.

**5.4.5 Increasing the Number of Demand Levels in Probability Distribution Designs (Method D)**

Using Method D probability distributions to construct the scenario trees, we are now able to increase the number of demand levels in the designs. Based on the earlier experimentation, our target was for 10 periods with 1,000 scenarios. The single scenario validation runs showed that the demands based on the highest 0.1% of the observations were beyond the capability of the modeled current force structure to react further (without making policy changes). Therefore, rather than adding a higher demand level, we looked to increase the accuracy in the middle to late periods, as the desired policy window ends. Two additional designs were developed for this experiment (Table 5.2).
Method D3 has a longer, more gradual transition from 4 to 1 levels of demand. Method D4 provides 4 demand levels through the first 5 periods, then drops off more quickly to a single demand level.

### Table 5.2 Alternative Method D demand design levels.
The number of scenarios from Method D ($K_0$) are significantly reduced from the scenarios with Method C ($K_C$).

<table>
<thead>
<tr>
<th>Design</th>
<th>Pd 0</th>
<th>Pd 1</th>
<th>Pd 2</th>
<th>Pd 3</th>
<th>Pd 4</th>
<th>Pd 5</th>
<th>Pd 6</th>
<th>Pd 7</th>
<th>Pd 8</th>
<th>Pd 9</th>
<th>$K_0$</th>
<th>$K_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>420</td>
<td>1152</td>
</tr>
<tr>
<td>D3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1277</td>
<td>6912</td>
</tr>
<tr>
<td>D4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>945</td>
<td>6144</td>
</tr>
</tbody>
</table>

Methods D3 and D4 produce models with very similar results. The summary charts in Figure 5.14 are representative of most of the personnel and unit force size types.

Figure 5.14 Similarity of model solutions using Method D designs.
There are several however, where we do observe differences. Some of the most pronounced cases are displayed in Figure 5.15. The average unmet demand using Method D3 is lower than when using Method D4. They begin to diverge noticeably after period 3, when Method D3 transitions to using 3 demand levels, therefore Method D4 is seen as more accurate. Method D3 clearly gives a better estimate of unmet demand for period 7, though this period would not be reported or used in the decision policy. Method D4 also appears to provide a better maximum unmet demand estimate that Method D3 for periods 3 & 4, but again, not 7. The Contractor and IBCT levels show some variation between the two methods, though primarily after period 5, making them inconsequential to our 5 period policy requirement.
Based on these results, the Method D4 design will be used for further experimentation. Method D4 has a longer initial window of higher fidelity, and the degraded ending periods do not seem to adversely affect periods 0 through 5. Additionally, the model using Method D4 provides reasonably quick solutions (Figure 5.15). It is also encouraging that the Method D1, D3, D4 all yield similar decision values for the first several periods, and are not overly sensitive to an arbitrary design choice.
5.5 Setting Model Parameters

Having selected a probability distribution (D4) to represent the stochastic deployment demand for the AFSP Model, this section considers the appropriate model parameter settings for subsequent analysis.

5.5.1 Number of Periods

The initial goal of the AFSP model was to provide recommendations for up to six periods in the future (a five year plan starting one year in the future). We have modeled 10 periods up to this point in order to mitigate any model termination anomalies, with the final periods being sacrificial. Here we investigate the importance of the sacrificial periods and quality of models with shorter time horizons.

For this experiment, we varied the number of periods modeled, beginning with 10, and truncating the model by one period each subsequent run, with all other settings being held constant. We expected that each of the models will begin identically, with
the shorter models deviating by a greater amount in later periods. The longer models should generally provide better decisions than the shorter models.

Most of the results were very similar for all model lengths. The total personnel, base and supplemental costs, and unmet demand metrics are nearly indistinguishable through period 5, as seen in Figure 5.17.

Figure 5.17 Similarity of results between models of differing lengths.
For some decision variables, however, the difference was quite noticeable, as in Figure 5.18. The deviation of the shorter models from the 10 period baseline was particularly pronounced for the personnel and unit metrics for Reserve Officers and Active Component Institutional Support, respectively. Where differences were appreciable, the longer models recommended lower force levels.

![Figure 5.18 Differences in results between models of varying lengths.](image)

The 7 period model is not reliable throughout our 6 period requirement window. The extent of deviation cannot be known without running longer models. The 8 and 9 period models provide good estimates in most cases for the model we tested. We found that though solution times tend to increase with the number of periods modeled, it is worth maintaining the 10 period model.
5.5.2 Policy Variable Settings

The policy variables are used to describe the decision variables that must be decided up front in order to produce a multi-year plan (Section 4.3.7). Our formulation of the AFSP model considers the personnel end strength, by type, and the number of available units, by type, to be policy variables. This section comprises two experiments. The first considers different implementation methods, while the second deals with varying policy flexibility.

Policy Methods

This experiment compares the results using no policy constraints, simple policy constraints, and enhanced policy constraints. Until this point, all decision variables reported have been expected values, or weighted averages based on the probability of the scenario to which they pertain. Here we will compare both the expected values of the decision variables, as well as the policy variables.
We begin with the decision variables. With no policy constraints, or non-anticipative constraints only, the model is less restrictive than the true system, and the simple policy constraints are more restrictive than the true system. The enhanced constraints should be between those two bounds. Additionally, the enhanced constraints with 0% flexibility is identical to the simple policy implementation. The flexibility level of 3% for the enhanced policy constraints was chosen as a reasonable value based on typical fluctuation between initial and final budget levels.

The expected values of the personnel and unit variables are very similar, as shown in the Total Personnel graph (Figure 5.20). The expected cost did not vary much either, with all the difference coming from supplemental costs. The unmet demand graphs are very telling, however. Under the unrestricted Policy 0, the model is better able to meet demands, resulting in higher supplemental costs. Policy 1, the simple policy constraints, is identical to Policy 2, enhanced constraints, with 0% flexibility. This is intuitive, but served as further verification that the constraints were implemented correctly. The model with enhanced policy constraints (Policy 2) and 3% flexibility is between the bounds set by the other two policies.
We experienced increased solution times while using policy variables, but not much variation between the different types of policy variables (Figure 5.21).
The policy variables from these cases (Figure 5.22) show that the model with simple policy constraints creates a plan (or policy) very similar to the model with no policy constraints, using the weighted averages. Again the simple policy constraint case is identical to the enhanced with 0% flexibility. The model with enhanced policy constraints and 3% flexibility, however, behaves differently. It creates a more conservative (higher force level) plan, and is able to adjust the plan year by year to achieve the same expected force levels as the other methods, as we observed in Figure 5.20. The trend is the same across personnel and unit types, with total personnel, total civilians, Active Infantry Brigade Combat Teams, and Reserve Armor Brigade Combat Teams used here as examples. The cost and unmet deployment demands are not included here because they are inherently weighted averages, not policy variables.
It is interesting that each of the methods shown here produced solutions with very similar expected decision variable values, but quite different policy variables. Recall that it is the policy variables that we ultimately seek to use the AFSP model to produce a five year plan.

**Policy Flexibility**

This experiment looks at the effects of varying the flexibility parameter of the enhanced policy constraints. Using Method D4 for the demand distribution and the enhanced policy constraints, the policy flexibility parameter is varied. The flexibility
parameter determines how much the decisions made in the actual period may vary from the original policy, compounding each year. The flexibility rates used are 0% through 4% in 1% increments.

The previous experiment showed that although the expected values of the decision variables did not change much, the enhanced policy constraints produced a more conservative plan with higher force levels, and was better able to meet deployment demand. The same should be true here, with each increment of increased policy flexibility increasing force levels in an orderly fashion.

![Figure 5.23 Comparison of unmet demand and cost for various policy flexibility levels.](image)

The average unmet demands (Figure 5.23) confirm that the models with increased flexibility are better able to meet deployment demands. Also, the expected costs, both base and supplemental, vary little with policy flexibility.
Comparing the policy variables to the expected value of the decision variables (Figure 5.24), we see results similar to the previous experiment. Under these conditions, increasing flexibility produces a plan with higher force levels (left column), but the averaged decision variables are similar (right column) across all cases. This trend holds across personnel and unit types.

Figure 5.24 Comparison of policy variables with expected decision variables.
5.5.3 Discount Factor
The discount factor is used as part of the objective function to weight the relative importance of near term decisions higher than the more distant future decisions. This models the reality of the Army’s decisions, as there is less flexibility to adjust the near term decisions and expectations about more distant future conditions are less certain.

The discount factor is applied as in a financial model, with the power of the discount increasing in each subsequent period \( (r^t) \). Table 5.3 displays the discount factors, where 1 represents no discount, 0.9 represents a moderate discount rate, and 0.8 represents a high discount rate.

<table>
<thead>
<tr>
<th>Discount</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.9</td>
<td>1.00</td>
<td>0.90</td>
<td>0.81</td>
<td>0.73</td>
<td>0.66</td>
<td>0.59</td>
<td>0.53</td>
<td>0.48</td>
<td>0.43</td>
<td>0.39</td>
</tr>
<tr>
<td>0.8</td>
<td>1.00</td>
<td>0.80</td>
<td>0.64</td>
<td>0.51</td>
<td>0.41</td>
<td>0.33</td>
<td>0.26</td>
<td>0.21</td>
<td>0.17</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Since future decisions are discounted, larger discount factors produce solutions with larger future force structures, as shown in Figure 5.25. Likewise, expected costs increase slightly (in terms of non-discounted costs), and unmet demand decreases with the larger force design.
Figure 5.25 Comparison of discount factors.

The discount factor ultimately reflects a perception concerning the uncertainty of future conditions, and is subjective. However, the model solution is sensitive to the discount factor selected. Although we do not offer an objective measure for setting the discount factor, a reasonable value such as 0.9 models reality better than ignoring this concept by setting the value to 1. This weights decisions 5 periods in the future close to 50%. The experimentation that follows is performed using a discount factor of 0.9.
5.6 **Analysis of Army Force Size**

This section describes experiments with penalty factors and resource constraints to develop insights for U.S. Army force size.

5.6.1 **Penalty Factor for Unmet Demand**

*Varying Penalty Factors to Compare Policies*

The penalty factor for unmet demand is an arbitrary parameter that relates that penalty to costs in terms of dollars. It is a proportional scaling factor based on the cost of deploying an Active Component Unit for a year (Section 5.1.2). Penalty costs are unit type specific, although we report missed demands in a single metric of IBCT equivalents.

This experiment varies the penalty factor to show the effect it has on recommended policy variables as well as unmet demand. We also wish to find a value that reasonably reflects current conditions. The model used the Method D4 probability distribution, and policy flexibility of 3%. Trials up to this point used a penalty factor of 5. The model was solved for zero penalty, then we began doubling the original factor up to a level of 160. We expect that increasing penalty factors will increase force levels and costs while decreasing unmet demand.

The results in Figure 5.26 clearly show that increasing the penalty factor leads to solutions with less unmet demand and higher costs.
At the highest penalty levels, the model recommends that the Army keep the current force levels, and even grow slightly. With a penalty factor of 20 and below, there is a clear trend to reduce force size. The examples in Figure 5.27 are representative of the other personnel and unit types.
**Detailed Exploration of Unmet Demand**

The average and maximum unmet demand provide a basis for relative comparisons, but they do not provide a complete picture. We created a cumulative distribution function of unmet demand for the various penalty factors using the probabilities of the scenarios and the unmet demand observed for period 4. Period 4 was used because it uses the maximum number of demand levels (4) and is far enough in the future to show good separation between the model settings.

The cumulative distribution function of unmet demand (Figure 5.28) clearly reflects the underlying discrete demand distribution, especially for the 0 penalty model. The interesting section of this plot is enlarged as an inset to the right. It shows that the various force designs, derived from the different penalty levels, should meet all demands more than 80% of the time. Between the 90th and 99th percentile cases there is a good differentiation between the amounts of unmet demand from the various force
designs. Above the 99\textsuperscript{th} percentile, there is always significant unmet demand regardless of the penalty parameter setting, though the larger designs do fare better. Response for these high demand cases requires policy modification and application of elements which are not modeled here (such as strategic weapons as well as diplomatic and economic means).

Figure 5.28 Cumulative distribution function of unmet demand by scenario probability. Left graph shows full function with inset area detail provided on right.

As an artifact of the discrete demand distribution, unmet demand is overestimated near the 90\textsuperscript{th} percentile, and underestimated near the 99\textsuperscript{th} percentile. In the middle, the 95\textsuperscript{th} percentile should provide a good approximation. The same is true for the 99.5\textsuperscript{th} percentile.

To interpret Figure 5.28, if we desired the most economic force plan to provide 95\% confidence that the demand for no more than 10 IBCT equivalents will go unmet,
we would choose the policy resulting from the model with a penalty factor of 40. If a maximum unmet demand of 5 IBCTs was desired, again with 95% confidence, we could choose the 80 penalty factor design, or solve a model for a penalty factor of 60 to determine if it was adequate.

**Cost to Risk Relationship**

Since the expected costs of these designs is also provided by the model, we can use that to compare the risk of unmet demand as well. These costs are useful for comparisons of the policies, but should be interpreted carefully. The costs and parameters reflect the year from which they were derived. The model here was parameterized based on FY14, where the Army was under considerable pressure to reduce costs. Those reductions were largely in the readiness and modernization areas, as the manpower, while declining, is slower to adjust. The current form of the AFSP model does not account for changing rates of readiness and modernization, and therefore, may under-represent costs based on personnel and units once a better balance is reached. By updating the parameters each period, the recommended policies should not deviate so much as to cause erratic changes from year to year, though multi-period cost totals should deviate somewhat.

The relative expected costs of the policies created in this penalty factor experiment are shown in Figure 5.29. Each group of expected costs are indexed by the average unmet demand score at the 95% confidence level. This approach provides a way to relate the expected cost of a force design to the risk of missing deployment.
demand at a specified confidence level. The underlying deployment demand probability
distribution may need to be adjusted in order to consider other confidence levels
accurately.

![Figure 5.29 Five-year expected cost of policies organized by 95% confidence index of unmet demand.](image-url)

**Selection of Appropriate Penalty Factors**

For further analysis, we want to select a penalty factor that reasonably reflects
real-world conditions. To do this, we will match planned manpower levels to the
modeled policies. It would be possible to seek expert opinion in this regard, if we had
access, but the concept of a penalty factor is abstract, and further complicated by
different individuals’ view of risk. Using cost is also unadvisable because in addition to the cost issues identified above, our model seeks to minimize cost without consideration of any political ramifications. Unfortunately, all government bureaucracies seek to justify and spend their entire budget each year, lest they be reduced in the future and not have resources available when actually necessary. Therefore, we will use active component military manpower levels as the benchmark to match.

The Army’s 2015 budget plan called for Active Component force levels, which began at 490,000, to drop to 450,000 in three years. This is similar to the force level recommended by the model using a penalty factor of 35 (Figure 5.30). Since this constraint was imposed upon the Army, we consider this the risk level analogous to National leadership, external to the Army. Let us assume that the Army would prefer to maintain the current Active Component force level of 490,000, which is similar to the plan derived using a penalty factor of 80.

Relating that in the terms developed above, our National leadership should have a tolerance for a 5% likelihood of not being able to meet the deployment demand of 6 or more IBCT equivalents (nearly an Infantry Division). Otherwise their policies may not be in concert with their risk preference. For the Army leadership, we expect that number to be near 3 IBCT equivalents, or half the amount of the National leadership.

Assume that decision makers believe that the suite of military planning scenarios represents conflicts in the likely future and that their decisions about the Army size do
not conflict with their beliefs. Then, force requirements generated from the AFSP model with a penalty factor of 35 should reflect the National leaderships’ preferences, while the requirements from the penalty factor of 80 model should reflect Army leaderships’ preferences.

Next, we will investigate the effects that externally imposed resource constraints, both budgetary and manpower, should have on policy decisions.

![Figure 5.30 Active Component force levels from models with varying penalty factors.](image)

5.6.2 Resource Constraints

The National leadership uses resource constraints to limit the force size and cost of the Army. This is essentially forcing compliance with their risk preference. The
previous experiments have been performed without budget or manpower (end strength) constraints. This section explores the effect those additional constraints have on the model solution. Both the end strength and budgetary constraints are based on Army reported values in the 2015 budget submission. The end strength constraint applies to the Active Component, Reserve Component, and DA Civilians. The budgetary constraint applies to base funds only.

Our expectation is that both types of constraints will decrease force structure and increase the likelihood of missing deployment demand. We are unsure which of these constraints will be binding, however.

The results shown in Figure 5.31 were generated using the standard Method D4 probability distribution with a penalty factor of 80. This is meant to represent the force decisions that the Army should make while conforming to resource constraints. Each graph depicts the unconstrained case, end strength constraints only, budget constraints only, and both types of resource constraints. The graphs in the right column represent the same data as the graphs in the left, but rescaled to better show differences.

The total personnel plot shows that the budget constraints are forcing rapid personnel reductions for the first two years. Thereafter, the end strength constraints become more restrictive. Interestingly, the externally imposed end strength reductions occur over the first 3 periods only, but the conditions created lead the model to recommend further force reductions, even though the Army prefers higher force levels when unconstrained. These additional force reductions are mainly contractors.
Figure 5.31 Selected policy metrics comparing resource constrained solutions. Graphs on left show full scale for perspective, while graphs on right are re-scaled to provide greater detail.
The Active Component, shown in the second set of graphs, is the primary focus of current force reductions, as they are the most expensive type of manpower during peacetime operations. Again, we observe that the budget reductions are driving personnel reductions faster than the end strength constraint in the first two years. Thereafter, Active Component forces level out, and budgetary and end strength constraints are generally in agreement.

Not depicted above, the model recommends Reserve Component force levels lower than the end strength requirement in all cases. Even when forced to make manpower cuts in other areas, and unconstrained by budget, the model still reaches solutions with a smaller Reserve Component. The DA Civilian levels hold steady near their strength cap for all cases.

The lowest set of graphs from Figure 5.31 shows the expected base budget levels. This provides further evidence that the budget is more constraining for the first two periods, but the end strength restricts force size further in the subsequent periods. The expected budget with end strength constraints is below the budget constraint. This suggests a recognition that the Army needs to adjust its posture after reductions and re-invest in readiness and modernization programs.

The average unmet demand in Figure 5.32 shows that the unconstrained solutions are better able to meet deployment demand than the resource constrained cases. The sharp budget cuts in period 0 and period 1 which cause the rapid decrease in
force size also noticeably impact the ability of the Army to respond to contingencies in the near term.

The last diagram shows the total cost of each policy above with the base, supplemental and penalty costs (at the 80 factor level) aggregated. It indicates that the unconstrained solution is the cheapest overall, while the budget constrained solutions have high expected penalty costs in the near term, resulting from the increased likelihood of missing deployment demand. The relative size of the penalty cost indicates a significant near term risk in force projection capability.

![Graph showing unmet demand and total cost for resource constrained solutions.](image)

**Figure 5.32** Unmet demand and total cost for resource constrained solutions.

One final consideration is the effect that different risk preferences may have on the force size policy. In this experiment we compare the resource constrained solutions from models with the 35 and 80 penalty factor levels. We expect the higher penalty factor to recommend a larger force level, at higher cost.
Figure 5.33 Comparison of resource constrained solutions with differing penalty factors. Graphs on left show full scale for perspective, while graphs on right are re-scaled to provide greater detail.

The results (Figure 5.33) support our expectations. Personnel levels were slightly higher across each category when a larger penalty factor is used, even with the same end strength constraint for both cases. The larger force also costs more, as seen in the base budget graphs. This could suggest that Army internal decisions may tend to preserve more manpower, at the expense of other programs, than the National leadership might otherwise prefer.
If the National leadership that imposes both budgetary and end strength constraints on the Army truly has a risk preference as we described, as well as a keen understanding of Army force structure issues, then these dual constraints may be a way to force Army decisions to better align with the National leaderships’ risk preferences.

5.7 Value of the Stochastic Solution

Developing and solving the stochastic form of the AFSP model, also referred to as the recourse problem (RP), requires considerable effort. In order to determine if the stochastic element of the AFSP is necessary or cost effective to consider, as we have done in this thesis, we evaluate the value of the stochastic solution (VSS). The VSS (5.1) is the difference between the RP solution value, and a deterministic reference point. The expectation of the expected value solution (EEV) often serves this purpose. For the EEV, the deterministic expected value (EV) problem is solved, and the EV decision variable values, $\bar{x}^i$, are then used in stochastic model to determine what the expected cost would be of implementing the EV solution under realistic conditions, as defined by (2.7), (2.8), and (2.10)

\[ VSS = EEV - RP \]  

(5.1)

For this comparison, we will use the specified policy variables (Section 4.3.7) only. We assume that the recourse variables can still be determined in the typical
fashion as events unfold. Furthermore, we will consider the policy variables both as fixed decisions (with a 0% flexibility parameter) and as updatable decisions (with a 3% flexibility parameter). For the RP we use the model instance from the previous section with the Method D4 demand probability distribution, penalty factor of 80, policy flexibility rate of 3%, and both budgetary as well as end strength constraints.

First, we will illustrate the difference numerically using the objective function values, then we will display selected metrics graphically with discussion as we have done previously.

The objective function value (4.1) is a measure of total cost (base, supplemental, and penalty) discounted by period, over the 10 period model time horizon. The units are in dollars, but that does not translate directly into budgetary costs, as noted above. It does provide our relative measure of goodness for a solution, however.

The objective values and VSS are displayed in Table 5.4. There is a large potential value of that stochastic solution in fixed EEV case, though part of that can be attributed to policy differences. Using the policy constraints with a 3% flexibility factor per year, the RP still provided almost 5% savings over implementation of the corresponding EV solution.
The graphs in Figure 5.34 show a significant divergence in policies over a 5 year period. Total personnel levels are policy variables, therefore the EEV with fixed policy variables, and EEV with 3% flexibility all contain the same values, by design. Recall that the policy variables set at period 0 are displayed below, but not any adjustments from subsequent periods. The average unmet demand of the EV solution expects no unmet demand, for its single scenario, which we know to be unrealistic. Evaluated under conditions of uncertainty, as shown by the EEV values, there is a significant difference. The RP solution provides the best policy here for reducing unmet demand.

The RP policy results in the highest expected base cost. The fixed EEV policy costs are both derived from the same reducing force structure policy. The EEV with 3% policy flexibility adjusts force structure upward each year to be closer to the RP policy values. Even still, the realized force from the EEV with 3% flexibility cannot return to the RP policy values.

Finally, the total cost graph shows that the RP solution provides a more economical solution policy than the EEV solutions.
The EV policy values clearly do not provide a good force plan. Even with policy flexibility, the force structure cannot recover fully. If implemented for consecutive years, the results could be disastrous.

Figure 5.34 Comparison of the expected value solution to the recourse problem solution.

5.8 Results Summary
In this section the extensive form of the AFSP model was implemented, where all scenarios are solved simultaneously, weighted by their likelihood, and the solution must
be both feasible and implementable. Data for the model was obtained from the Army Forces Cost Model, the Army Military-Civilian Cost System, and the Army Planning Programming Budgeting and Execution Data Warehouse, focusing on FY14 as a baseline.

We showed that the multi-period stochastic program form of the AFSP model is solvable for 10 periods and over 1,000 scenarios. This is adequate to generate reliable solutions for 6 periods (0 through 5). Additional memory would make solution of larger problem instances possible.

The probability distribution used to represent the stochastic deployment demand can have a large effect on the problem solution. The distributions developed in Section 3.3.2 that are skewed toward the higher demand range provide a better representation for the AFSP, especially when we can only afford to use a small number of demand levels. Increasing demand level fidelity in the earlier periods at the expense of sacrificial periods at the end was also shown to be beneficial. The correlation of deployment demand between periods is also important to include for an accurate representation. Furthermore, very low probability scenarios may be consolidated without negative effect on the AFSP solution, facilitating the use of higher fidelity demand level designs and larger scenario trees.

Experimentation showed that the use of all 10 model periods (4 sacrificial periods) was beneficial for generating reliable solutions. Furthermore, the incorporation of the policy variable flexibility rate did not have much effect on the policy level values, but was important for the accurate representation of unmet demand.
We estimated the penalty factor for unmet demand from the current manpower policies to show a potential difference in risk preference internal and external to the Army. We demonstrated how the level of unmet demand in each scenario can be used to develop the cumulative distribution function of unmet demand and a 95% risk tolerance level for missed demand. The model solutions at varying penalty levels was also used to develop the relationship between expected force costs and the likelihood of unmet demand.

The AFSP model was used to recommend force size requirements under unconstrained conditions, as well as determine policies conforming to external budgetary and manpower constraints. For the conditions analyzed, the budgetary constraints were binding for the first two periods, but personnel constraints were more restrictive thereafter. We also showed that using different penalty factors (our surrogate for risk preference) will lead to different force size policies, even under the same budget and end strength constraints.

Finally, we compared the stochastic AFSP solution (RP) to a deterministic solution evaluated over the same distribution of demand scenarios (EEV). There is a substantial potential benefit to solving the RP and considering the stochastic nature of deployment demand in general.
CHAPTER 6 CONCLUSIONS

In this thesis, we outlined the complex, dynamic nature of determining an appropriate force size for the U.S. Army under the uncertainty of future contingency events. In consideration of this problem we included the military, civilian, and contract segments of the work force, both the operational (deploying) forces and capacity of institutional forces to generate and support the total force, and an appreciation for how the workforce and force structure evolve over time.

Through study of the Army force size problem (AFSP), we developed a model of deployment demand and used that model to develop probability distributions that could be implemented in stochastic programming. The AFSP model developed in this thesis focuses on determining an economical multi-year force size policy for each work force segment under uncertain future deployment demands while accounting for aggregate brigade level unit types. Experimentation showed the importance of risk preference and policy flexibility as well as appropriate parameter settings to represent each.

The remainder of this chapter is organized as follows. In Section 6.1 we identify and summarize the most important conclusions from this thesis. In Section 6.2 we
conclude with areas in which this work may be extended as well as potentially interesting directions for related research.

6.1 Contributions

The research described in this thesis required both the novel application of existing methods as well as the development of new ones.

Military force structure analysis is a well-studied problem. Most methods, however, seek to develop an optimal future operational (combat) force using stochastic simulation, then design the remainder of the force to support that operational force within resource constraints, through iteration if necessary. Finally, plans to transition from the current force to the objective force are developed either concurrently or iteratively.

Through the approach developed in this thesis, the operational and institutional forces, along with all components of the workforce are considered simultaneously for optimization. Furthermore, rather than developing a single objective force, we solve for the force size for each year of a multi-year plan, allowing for dynamic force shaping and policy adjustment to occur over time as new information becomes available. Although there have been previous limited applications of stochastic programming in force structure analysis, this is first to use a multi-stage stochastic program to develop dynamic force sizes in pursuit of a multi-year plan.

We drew upon published techniques to develop a stochastic model of small scale contingency events using current military planning contingencies in conjunction with
historic conflict frequencies, and extended these methods to include major contingency operations. This provides a modeling alternative to arbitrarily specifying the number or frequency of major contingencies. The implementation of this model was referred to in Chapter 3 as the stochastic demand generator (SDG).

We developed a practical methodology in Chapter 3 for generating approximate discrete empirical deployment demand probability distributions appropriate for use in stochastic programming applications. This is an area currently lacking in the stochastic programming literature, where uncertainty inputs are assumed available outside of the model development. Although the distributions used in this thesis were based on simulation output from the aforementioned SDG, any validated demand data may be used, making these techniques readily adaptable for use with classified data and deployment demand models, as well as completely different problem domains.

Study of the shape and characteristics of the deployment demand distribution led to the development of several discrete probability distribution designs, which were tested using the AFSP model in Chapter 5. The use of a skewed distribution was critical for an accurate representation of deployment demand with a small number of scenarios. Adjusting the model fidelity to provide more demand levels for the initial periods of analysis while sacrificing detail in the ending sacrificial stages improved the deployment demand representation for this problem. Incorporation of demand correlation between periods further improved model accuracy. Finally, the consolidation of extremely rare events reduced the number of scenarios to evaluate.
without adverse effect on the solution, thereby facilitating better demand designs with more levels of demand.

The AFSP model developed for this thesis is unique in that it established the behavior and relationships between the operational and institutional forces while considering the total Army workforce, not just soldiers. The model deals in aggregate personnel and units based on a notional brigade level force structure. This allowed the consideration of different personnel and unit types at a strategic level without being overwhelmed by the myriad of unique personnel and unit characteristics necessary for detailed force planning. This approach also facilitates the implementation of implicit adaptive force readiness policies through a piecewise linear function rather than an explicit fixed policy. The AFSP model also incorporates an adaptation on standard non-anticipative constraints in order to produce a multi-year policy (plan). These policy constraints account for the system’s flexibility to make adjustments in the future while maintaining admissibility and implementability of the current plan.

We determined in Chapter 5 that although the small joint probabilities used to generate skewed demand distributions can cause numerical issues for optimization software, we can achieve adequate results efficiently from barrier method solutions without requiring extreme point solutions.

We showed an effective approach for approximating an infinite horizon problem with a finite period model. Simply adding additional sacrificial periods can be very expensive as the scenario tree grows exponentially. By using a diminishing number of
demand levels per period we could effectively mitigate adverse termination effects while limiting the growth of the scenario tree.

Chapter 5 also developed insights for Army force planning. A cost to risk relationship based on unmet demand for each scenario was evaluated. We observed that demands below the 80th percentile should be universally met by all policies considered, while demands above the 99.5th percentile could not be adequately met by any of the design policies, indicating an important range for further analysis. We showed that tolerance for unmet demand at a 95% confidence level could be used to compare and select a force design, while indicating the size and relative expected cost of that design.

Additionally we provided an analysis of Army external budget and manpower constraints, showing where each was more restricting as well as the effects of differing risk preferences. This could lead to a better understanding of how external constraints are impacting Army force decisions, or perhaps an additional argument to relax one or both constraints.

Finally, in Chapter 5 we showed the importance of incorporating the stochastic nature of deployment demand for the AFSP. The value of the stochastic solution (VSS) shows a significant difference between the policies generated from solving the stochastic AFSP model over a naïve deterministic solution.

The true value of this approach for the Army is not that it produces new force designs that are cheaper to maintain or better able to meet deployment demand. It
provides an inexpensive way to holistically consider force size changes and the effects of new external constraints without committing the hundreds of man-years required to generate an alternative force plan. Although crude in comparison to typical force structure analysis, the AFSP model has the potential to yield useful insights for policy makers which could then inform subsequent, more detailed planning and analysis.

6.2 Future Research
There are several areas related to this research that could lead to additional valuable contributions.

The discrete empirical probability distributions developed in Chapter 3 are all based upon a single stochastic variable, or index. We were fortunate in that each demand type was highly correlated with this index, so much so that we assumed they were proportional. It would be a much more complicated case to consider multiple independent stochastic variables and still develop discrete joint probability distributions that yield manageable scenario trees. Although not required for the AFSP application, it may be necessary in other problem domains.

Additionally, further refinement of the demand levels and partitioning used for the AFSP problem may produce further improvements for this problem. Determining the useful limits of scenario consolidation or scenario bundling would also be important. The study of stochastic components with differently shaped distributions could prove beneficial for other problem domains.
The AFSP model from Chapter 4 is a work in progress. Greater detail is needed to define the relationships between the institutional and operational force in particular, better data regarding contractors is required, and separating the different enabler types into distinct units would also be helpful. Additional force readiness policies under consideration may also be incorporated to conduct additional analysis in the future.

The AFSP model is designed to minimize cost while assessing a cost penalty for unmet demand. It could be valuable to re-formulate the model to minimize unmet demand on a fixed budget. Furthermore, probabilistic or chance-constraints could be incorporated to represent unmet demand as a percentage or likelihood rather than with a penalty parameter, thereby fundamentally changing the modeling philosophy of the problem.

Finally, the AFSP model was originally envisioned as a tool to analyze the total Army’s resource allocations. Substantial effort would be needed to extend this model to accurately include readiness (training and infrastructure) and modernization (equipment procurement and upgrades). Such a model should prove useful for Army policy makers.
APPENDIX A  NOTES ON MODEL TRACTABILITY

This appendix discusses the solvability of the AFSP model, the initial trials, tuning parameters used, and solution conditions established for experimentation.

A.1 Experimental Conditions
All experimentation was performed on a 64bit PC with a quad-core 2.2GHz processor and 8GB of RAM. The memory proved to be a limiting factor in size of the model that could be generated as well as solved.

The models were generated and solved using Gurobi Optimizer 6.0.2 in a Python 2.7.5 shell.

A.1.1 Design Limitations
A significant amount of effort is required to estimate the model parameters and establish feasible model conditions for a given set of model starting conditions. Therefore, all experimentation here was limited to the same set of starting conditions and model parameters, except as explicitly discussed in experimental design.

Our initial goal was to solve up to 10 model periods in order to create a 5 year plan, beginning a year in the future, with several sacrificial ending period to mitigate model terminating anomalies. Model runs of more than 10 periods were not performed.
A.2 Default Solution Times

The solution times presented here and throughout the remainder of the chapter are anecdotal rather than rigorous. They still provide a useful relative comparison of computational effort required to solve the models concerned, but the times reported reflect one observation of a single instantiation of the model.

A.2.1 Expected Value Problem

The AFSP model was first run on the expected value (EV) problem, where the mean value of each stochastic variable is used in place of the probability distribution in order to create a simple deterministic problem using a single mean value scenario. The EV problem for 10 periods was solved in less than 0.1 seconds.

This is typical for any single scenario run. The 10 period EV problem had 2,260 variables, 2,910 constraints, and was readily solvable. This suggests that a scenario decomposition method may be useful if necessary.

A.2.2 Method Z

Initial stochastic runs of the AFSP model were performed using a simple demand distribution with two levels or possibilities at each period. One level was the demand average of the lower half of observations generated for Section 3.2.2, and the other for the upper half. This produces a $2^H$ scenario tree structure with all scenarios having equal probability. The equal probability demand distribution is referred to here as Method Z, as it was not included in the earlier development of demand distributions.

The Method Z model was solved for an increasing number of periods using default optimization parameters. Although small versions of the model solved quickly
(less than 1 second for 4 periods) the solution times grow exponentially (Figure A.1). The 9-period model timed out after 1,800s (½ hour) of running. It is clear that default parameters would not produce quick solutions for a 10 period problem using Method Z.

![Figure A.1 Anecdotal solution times for increasing numbers of periods. Solution times increase exponentially as the number of periods increase.](image)

**A.3 Optimizer Tuning**

Experimentation with Gurobi optimizer parameters was performed on an 8 period model using the 2 level Method Z probability distribution, as above. This model has 465k variables and 795k constraints.

**A.3.1 Optimization Algorithm (Method)**

Gurobi documentation indicates that solution method is the most influential factor in LP models [69]. The available solution methods are: automatic, primal simplex, dual simplex, barrier, concurrent, and deterministic concurrent. In automatic (default)
mode the concurrent method is used with multiple solution methods are starting simultaneously on different threads. For the AFSP problem, it chooses barrier and dual simplex, and with more than 4 periods, we find that the barrier method consistently solves first. The deterministic concurrent method ensures the same solution each time, but is slower than the standard concurrent method.

The model was run using each of the solution methods and a ½ hour time limit to produce the results in Figure A.2. Both the primal and simplex methods did not reach optimality within the time limit, coming within 1% and 5% of the optimal objective value, respectively. This indicates that the barrier method is the most efficient algorithm for solving this problem.

![Figure A.2 Anecdotal solution times for different optimization algorithm. The interior point, barrier, method provides the fastest solutions for the AFSP.](image-url)
All further testing was performed using the barrier method parameter setting. Not only was it significantly faster than the default (concurrent) method, but selecting a single algorithm will save memory which becomes important for larger problems.

A.3.2 Other Optimization Parameters
A built in automated tuning algorithm was used to identify any other potentially useful parameters for consideration. The automated tuner was run for 10800s (3 hr) and selected three additional parameters: PreDual, PrePasses, and ScaleFlag. Follow up experiments were performed to investigate each.

The first, PreDual, forces the optimizer to form and pre-solve the dual problem. In automatic mode, the optimizer chooses whether or not to pre-solve the dual problem. It can be turned off, turned on, or pre-solve both the primal and dual on separate threads. Figure A.3 shows that forcing the optimizer to pre-solve the dual problem is advantageous.
Figure A.3 Anecdotal solution times comparing dual problem pre-solve settings. Pre-solving the dual problem improves solution times for the AFSP model.

The second, PrePasses, manually limits the number of pre-solve passes thereby limiting pre-solve time. No appreciable benefit was achieved over the default automatic setting (Figure A.4). Furthermore, we were suspect of how a fixed number of pre-solve passes may affect larger, more complex models, so PrePasses was left at the default automatic setting.
Finally, ScaleFlag controls model re-scaling, which is meant to improve the numerical properties of the constraint matrix, but may increase constraint violations [69]. Improvement was only seen in conjunction with pre-solving the dual problem (Figure A.5). Although restricting the scaling was only seen to have a small impact here, it was left on.
Figure A.5 Anecdotal solution times for the scaling parameter. No scaling was slightly faster than normal scaling when used in conjunction with pre-solving the dual problem.

A.3.3 Barrier Method Crossover to Simplex

During the course of these experiments, we observed that the optimizer was spending half of the solution time crossing over from the barrier solution to a more exact primal/dual solution. This additional experiment investigates crossover options.

Once the barrier objective tolerance is reached, the crossover procedure is used. The crossover procedure may be performed by pushing the dual variables to their bounds then the primal variables, or vice versa. A final clean-up phase may then use primal or dual simplex. The crossover can also be turned off or left in automatic mode. For our problem, beginning with the dual variables is advantageous, but the clean-up phase is inconsequential (Figure A.6). Both procedures beginning with primal variables timed out after 5 minutes.
Figure A.6 Anecdotal solution times for the barrier crossover method parameter. Beginning with dual variables works better for the AFSP model than primal first. Eliminating the crossover phase offers significant computational savings while yielding very good solutions. Both crossovers beginning with primal variables timed out.

Omitting the crossover completely saves considerable time, but is not as precise as the solutions after crossover. A comparison of each of the decision variables, with and without the crossover procedure was performed. This showed that variable values from the barrier only method deviated by at most 0.0003% from the barrier with crossover variable values. The default barrier tolerance of $1 \times 10^{-8}$ was adequate. We deemed this an acceptable solution approximation moving forward which with testing, and will revisit this assumption once the final version of the model is established. Although removing the crossover procedure is not necessary for the 8-period problem here, it becomes increasingly important as the model get larger.

Model runs up to 10 periods are now possible, as shown by Figure A.7. The 10-period model here with 2.3 million variables and 4 million constraints is the first
model discussed here where the system RAM was insufficient and additional virtual memory was required.

![Figure A.7 Anecdotal solution times with increasing numbers of periods. Using the selected optimization tuning parameters reduces run time significantly over default settings. The default 9-period run did not complete in 1800s. The 10-period default run was not attempted.](image)

**A.3.4 Solution Verification**

After selecting the Method D4 distribution for continued experimentation, we revisited our earlier assumption that we were obtaining high quality solutions by solving with the barrier method. We attempted to solve the Method D4 model beginning with the barrier method and then crossing over to an extreme point solution using simplex (through Gurobi’s built-in algorithms), but the optimizer produced a numerical error after 8 hrs. Repeating with the smaller Method D1 model, we obtained similar results after an hour. An attempt using dual simplex was failing to converge after 8 hours, and was terminated. An additional run was performed using the Method D1 model with the
numerical focus parameter increased, and all other settings at default. After 40 hours of running, the objective function had 2 significant digits of accuracy and was progressing very slowly so the trial was terminated.

This issue was taken up by the Gurobi technical staff, and although they were able to replicate the numerical issues, they could not offer additional tuning parameters to improve the solver performance. The numerical issues stem from the objective function, presumably due to the wide range of scenario probabilities that scale the cost parameters in the model. The model we provided has been added to Gurobi’s bank of test problems for future development and improvement.

Solving the smaller Method D1 model using Cplex took about 4 hours when beginning with the barrier method, and 10 hours when restricted dual simplex. The objective function from each was the same to 10 significant digits. The Cplex runs were performed on a different system, and run times are not directly comparable.

Comparing with the original Method D1 approximate solution, which we achieved in under 10 minutes, with the Cplex objective function, we found that they matched to 7 significant digits. This accuracy is adequate for our purposes, and is beyond our capability to accurately estimate model parameters.

A.4 Practical Model Size Limits
The initial experimentation showed that the extensive form of the AFSP model can be solved (approximated) efficiently, but the number of scenarios must be limited. Further informal testing indicated that although models with 1,000 scenarios are quite
tractable, models with 1,500 scenarios take considerably longer to solve and make extensive use of virtual memory. Models with 1,800 or more scenarios were not solvable on our testing platform, as the optimizer runs out of memory.

The research of this thesis is limited to solving and improving the extensive form model through manipulation of the empirical probability distribution approximation. The literature review indicated that progressive hedging methods [38] as well as sampling based decomposition methods [37][43] have the potential to address larger model formulations with a greater number of scenarios.
APPENDIX B  PROBABILITY DISTRIBUTION FOR 2-STAGE APPROXIMATION

One additional method was developed to represent the stochastic demand, but not implemented for experimentation. Different from the methods presented in 3.3.2, this method selects complete demand profiles from the full simulated data set for evaluation by the stochastic program. This is analogous to a quasi-Monte Carlo approach for simulation. Because the scenarios selected are not necessarily connected as in a typical stochastic programming scenario tree except at the root node, or starting period, this would amount to a 2-stage approximation approach of the problem.

B.1 Method E – Subset of SDG simulation scenarios
Method E carefully selects a subset of demand profiles from the SDG output $\left( \xi(s_{SDG}) \right)$, assuming the selection of scenarios $\left( S_E \subset S_{SDG} \right)$ is an adequate representation of the full set. The scenarios selected $\left( S_E \right)$ would then be evaluated whole by the stochastic program model. A benefit of this approach is that each of the demand profiles naturally occurred from the original demand simulation and with a sufficient sample, should exhibit the same qualities and characteristics of the full demand distribution.
This approach amounts to a 2-stage approximation of the original problem, however. After the first stage decisions are made, the entire scenario is effectively revealed, making the remaining stages deterministic.

To create $S_E$, rather than random sampling from the SDG or our set of validation scenarios used throughout this section, we chose specific values to achieve the coverage of the upper range of the distribution that we desire. The scenarios ($s_{SDG}$) are ordered by $d_{Tot}$ such that $d_{Tot,j} < d_{Tot,j+1} \forall j$. We then found 1,000 evenly spaced intervals, in terms of total average demand ($\overline{d}_{Tot}$), between the minimum and maximum observations. In this regard, the selection bears some resemblance to a simple quasi-Monte Carlo process.

These desired average total demand values were then cross-referenced with the set of SDG scenario demands in order to find the closest match. Many demand value intervals at the upper end of the distribution selected the same representative scenarios because the observations are sparse in that region. The intervals in the upper region were therefore increased to limit each scenario to be matched with no more than one demand interval because duplicating scenarios in $S_E$ would not provide any additional model fidelity. The Method E procedure for selection of scenarios is outlined below.
Let \( j = 1, \ldots, N \) be the ordered scenario index of \( S_{SDG} \) such that
\[
\overline{d}_{Tot} (s_j) \leq \overline{d}_{Tot} (s_{j+1}) \quad \forall j.
\]
Let \( K_E = |S_E| \leq N \), the number of scenarios to select for \( S_E \).
Let \( k = 1, \ldots, K_E \) be the scenario index of \( S_E \).
Let \( L \) be the theoretical ideal set of \( \overline{d}_{Tot} \) for \( S_E \).
Let \( l = 1, \ldots, K_E \) be the index of \( L \).
\[
L := \left\{ \overline{d}_{Tot} (s_N) - \overline{d}_{Tot} (s_1) \times (l - 0.5) \right\} \quad \forall l \quad \text{(equal intervals over range of } \overline{d}_{Tot} \text{)}
\]
\( S_E = \{\} \)
For \( l = 1, \ldots, K_E \): (select scenarios for \( S_E \))
\[
S_E := S_E + s_j \text{ where } \arg\min_{j \in S_{SDG}} \left| L_l - \overline{d}_{Tot} (s_j) \right|
\]
For \( k = K_E, \ldots, 2 \): (adjust scenario selection so that all \( s_E \) are unique)
If \( \overline{d}_{Tot} (s_k) \leq \overline{d}_{Tot} (s_{k-1}) \), then \( s_{k-1} := s_{j-1} \) (where \( s_k \in S_E \mapsto s_j \in S_{SDG} \))
\( \xi(s_k) := \xi(s_j) \) where \( s_k \in S_E \mapsto s_j \in S_{SDG} \) (assign scenario demand values)
\[
P\left( \xi(s_k) \right) = \frac{j(s_{k+1} \mapsto s_j \in S_{SDG}) - j'(s_{k-1} \mapsto s_j \in S_{SDG})}{2N} \quad \forall k \quad \text{(probability)}
\]

The resulting cumulative density function for Method E is shown in Figure B.1.

The average total demand of Method E scenarios are displayed in Figure B.2.
Scenarios with lower average demand have smaller index numbers. The original equally spaced intervals are shown in red. The adjusted intervals are shown in blue, where every scenario observed from the upper range of the distribution is included in the sample.

The original linear portion was designed to cover the demand range evenly. The non-linear portion was adjusted to compensate for the availability of data in the higher demand region.
Method E does not use discrete demand levels as we have seen in the previous approximations. Each individual scenario retains its inherent nearly-continuous demands calculated by the stochastic demand generator simulation. A random sample of scenarios from method E is shown in Figure B.3. This sample displays the average demand, variation, and correlation we expect.

![Figure B.3 Example demands generated by method E.](image)

Each demand profile was selected and incorporated, in whole, from the stochastic demand generator simulation.

**B.2 Approximation Method Comparison**

This section repeats the analysis of Section 3.3.3 to include Method E. The total average demand is shown below in Figure B.4. Method E deviated from the baseline more than the others, attributed to the manner in which the demand profiles were
selected. Method E scenarios were selected based on total demand, averaged across all periods, while methods A through D considered each period separately.

![Graph showing total demand comparison of demand approximation methods.](image)

Figure B.4 Average total demand comparison of demand approximation methods.

Figure B.5 illustrates that Method E replicates the baseline variation very well.
Last, Figure B.6 demonstrates that method E replicates the baseline correlation very well.
Method E takes a different approach than the previous Methods A through D by selecting entire scenarios from the underlying simulated data set. This approach would approximate our multi-stage problem as a two-stage problem.

Although Method E was not implemented in this analysis, it would be interesting to evaluate the quality of solutions derived from this approach with respect to the Method D4 experiments performed in Chapter 5.
REFERENCES


BIOGRAPHY

John C. Checco earned a Bachelor of Science in Chemical Engineering from Lehigh University, Pennsylvania, in 1996 and was commissioned as an Army Officer. During the course of his duties, he obtained a Master of Science in Engineering Management from the University of Missouri for Science and Technology in 2001 and a Master of Science in Operations Research from Kansas State University in 2007. He has been working as an Army Operations Research and Systems Analyst since 2005.