9-2022

Retention Prediction and Policy Optimization for United States Air Force Personnel Management

Joseph C. Hoecherl

Retention Prediction and Policy Optimization for United States Air Force Personnel Management

DISSERTATION

Joseph C. Hoecherl, Maj, USAF

AFIT-ENS-DS-22-S-062

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-DS-22-S-062

RETENTION PREDICTION AND POLICY OPTIMIZATION FOR UNITED STATES AIR FORCE PERSONNEL MANAGEMENT

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

Joseph C. Hoecherl, BS, MS

Maj, USAF

August 19, 2022

AFIT-ENS-DS-22-S-062

RETENTION PREDICTION AND POLICY OPTIMIZATION FOR UNITED STATES AIR FORCE PERSONNEL MANAGEMENT

DISSERTATION

Joseph C. Hoecherl, BS, MS
Maj, USAF

Committee Membership:

Dr. Matthew J. D. Robbins
Chair

Dr. Raymond R. Hill
Member

Dr. Brett J. Borghetti
Member

Dr. Adedeji B. Badiru
Dean, Graduate School of Engineering and Management

Abstract

Effective personnel management policies in the United States Air Force (USAF) require methods to predict the number of personnel who will remain in the USAF as well as to replenish personnel with different skillsets over time as they depart. To improve retention predictions, we develop and test traditional random forest models and feedforward neural networks as well as partially autoregressive forms of both, outperforming the benchmark on a test dataset by 62.8% and 34.8% for the neural network and the partially autoregressive neural network, respectively. We formulate the workforce replenishment problem as a Markov decision process for active duty enlisted personnel, then extend this formulation to include the Air Force Reserve and Air National Guard. We develop and test an adaptation of the Concave Adaptive Value Estimation (CAVE) algorithm and a parameterized Deep Q-Network on the active duty problem instance with 7050 dimensions, finding that CAVE reduces costs from the benchmark policy by 29.76% and 17.38% for the two cost functions tested. We test CAVE across a range of hyperparameters for the larger intercomponent problem instance with 21,240 dimensions, reducing costs by 23.06% from the benchmark, then develop the Stochastic Use of Perturbations to Enhance Robustness of CAVE (SUPERCAVE) algorithm, reducing costs by another 0.67%. Resulting algorithms and methods are directly applicable to contemporary USAF personnel business practices and enable more accurate, less time-intensive, cogent, and data-informed policy targets for current processes.

Dedicated to my wife and children. Our little farm has been my refuge during this writing process; thank you for your sacrifices before and during this program that enabled this research.

I also dedicate this work to my father. While he did not live to see me enter or complete this program, I spoke to him often during my master's thesis research and time on Air Staff about the ideas that would lead to this research. The things he taught me as a father and the experiences he shared as an enlisted soldier and officer in the US Army helped develop what would become the principles described in this document.

Additionally, I dedicate this work to my mother, who taught me at an early age that I should never fear questioning things that are true, that the truth holds up to scrutiny. This has set the stage for a lifetime of attempted scientific inquiry.

Finally, I dedicate this work to the enlisted men and women of the US Air Force. While the officer corps has a greater level of focus on its human capital management and smaller numbers of personnel to manage, the enlisted force is far more likely to live by the model outputs in this work, for better and worse. Getting these systems right has become very near to my heart these past few years, and I hope that the results of this work can improve the time that you have volunteered to spend in our Air Force by some small amount.

Acknowledgements

Thanks first to my advisor, who has so patiently guided my research into sequential decision-making for nearly a decade, and my committee members, who were invaluable in guiding and refining the ideas and writing that culminated in this dissertation.

Thanks to Dr. Jerry Diaz for seeing the potential for something better back in 2014, then creating the opportunity to build it. Thanks to Col Jim Barger as my "Yoda," providing unmatched knowledge of the interrelationships in the manpower and personnel systems, and John Sanzone for his guidance exploring MilPDS data.

Thanks to Mr. Doug Boerman and the past and current AF/A1PF and A1XD members for the years spent on these analyses to make the AF better. Special thanks to Greg Renner for automating enough that I could build something new, Sean Ritter for working with me to unravel our history, Stefan Zavislan for helping me craft my white papers, and Zack Hornberger for helping me to think even bigger.

Thanks to the enlisted CFMs and CMSgt Cristina Gutierrez for years of discussion and feedback. The math can't matter until we solve the right problems and take care of our people; you provided me with perspective and I can't thank you enough.

Thanks to the many analysts, action officers, and leaders whose feedback and constructive criticism enabled this work, from AF/A1, AF/RE, AFPC, AF/A3, AF/A4, SAF/SA, AF/A10, HQ AETC, AFRS, Navy OPNAV N81, Army G1, and SAF/CO.

Thanks to Lt Gen Kelly and Lt Gen Grosso, who set the stage for this research.

This research was funded in part by the Omar N. Bradley Research Fellowships.

Joseph C. Hoecherl

Table of Contents

Page

Page

List of Figures

Figure Page

List of Tables

RETENTION PREDICTION AND POLICY OPTIMIZATION FOR UNITED STATES AIR FORCE PERSONNEL MANAGEMENT

I. Introduction: US Air Force Human Capital Management

> "Credibility of personnel policies and management practices suffers when the reasons for their existence are not clearly defined or understood by all members of the force. Increasing the visibility of the logic behind personnel policies promotes acceptance and understanding."
>
> - The USAF Personnel Plan (Dixon Plan), 1979

1.1 Fundamentals of Manpower and Personnel (Why Does This System Exist?)

The United States Air Force's (USAF) manpower and personnel systems frequently act in complex and counterintuitive ways that are difficult to understand and measure without detailed knowledge of the subject. Unlike many functional areas wherein the USAF can look to industry to find solutions to common problems, the USAF has a fundamentally different personnel problem than the majority of businesses in the private sector, given the necessity to comply with Congressional end-strength expectations in combination with its unique structure. With the exception of medical and legal personnel, this current structure of the USAF requires it to develop its people from the beginning, developing the knowledge, skills, and abilities to be a fighter pilot, F-22 crew chief, remotely piloted aircraft (RPA) pilot, or cyber operator, for example, from the ground up. Unlike human resource planning in many areas, the USAF's primary personnel problem is not how quickly it can hire

a person with the prerequisite skillset. The USAF problem is how to address current and emerging manpower requirements with new accessions (i.e., newly hired, entry-level personnel) and existing, experienced personnel (with various skillsets and levels of experience). In large part, training and force management decisions made 5, 10, or 20 years ago determine the number of people with specific skillsets and levels of experience.

One key difference between the modern USAF and the historical approach to fielding military personnel is the role of people in fielding technological capabilities. Although not universally true, the combat capability of much of the military has historically depended on its ability to field substantial numbers of recently trained junior personnel. The USAF, in contrast, fields much of its combat capability via technologically complex systems. An F-22 maintenance crew chief is a most effective subject matter expert once qualified as a 7-level craftsman, after gaining multiple years of experience. Without the correct number of 7-level maintainers, aircraft cannot fly, even when an abundance of recently trained 3-level apprentices are available. Recently recruited and trained infantry soldiers can have a direct and consequential impact on Army readiness, but the same is not true for the vast majority of Air Force specialties. Therefore, the USAF must continually grapple with complex and long-term consequences of manning challenges—even when these challenges stem from internal or external factors many years ago.

In order to discuss this topic effectively, a clear set of definitions must be proffered. First, when discussing manning, we are referring specifically to the number of permanent party inventory divided by the number of authorizations on the Manpower Programming and Execution System Unit Manpower Document (MPES-UMD, also referred to as the UMD). As such, manning itself does not address whether the number or type of authorizations are correct or what the relationship to the original

unfunded requirements might be. Moreover, all references to manpower relate to the process for planning and funding requirements (i.e., spaces), whereas all references to personnel relate to the process of fielding human capital (i.e., faces). One set of human capital challenges arises when disconnects occur between the numbers of faces and spaces, observed as shortages.

1.2 Guiding Principles for an Idealized Human Capital System

Before delving into how the AF system works now, we need to be oriented to how the system should work in theory. This enables an examination of where current manpower and personnel policies and practices fall short and what solutions to those disconnects could look like.

- Principle 1: The system should utilize the human capital it has presently to the greatest effect by:
    1. maximizing commanders' flexibility to make resourcing decisions to execute current operations while
    2. making specific decisions on where to not apply resources (colloquially: to take risk) and
    3. communicating that risk to the commanders who are not receiving the required resources so they can adapt to this decision.
- Principle 2: The system should enable senior leaders and the U.S. Congress to make decisions about future force composition and understand the cost and consequences thereof. These decisions manifest as manpower authorizations but are only relevant in how they affect future human capital (personnel). Colloquially: without a service member to fill it, no authorization has ever contributed anything to mission effectiveness... ever.

- Principle 3: The system should plan and execute precise, intentional policy decisions to shape future human capital resources to meet future force composition. These data-informed policy decisions should consider the likely range of outcomes associated with those policies. We should measure these outcomes against alternatives to assess whether required human capital will be available to enable operations over time (readiness/lethality), and whether this availability over time can be sustained. We should measure outcomes and compare these to predictions to continuously refine models, assess model confidence, and identify lessons learned.

- Principle 4: The system should enable smart talent management opportunities, balanced with executing current operations and meeting future needs. Talent management is a dynamic rather than fixed process. Defined requirements should largely guide policies for the development of individuals, but enabling people to meet their full potential may require developmental experiences that are tailored, difficult to quantify fully, or beyond the awareness of those who establish requirements.

- Principle 5: The system should function well enough to minimize strain on the service members within this system. Stresses from unnecessary bureaucracy, inadequate support, or clearly inequitable or inefficient business practices create negative consequences for performance, retention, satisfaction, and engagement.

1.3 The Human Capital Analytic Pyramid

Given the complexities of managing the USAF manpower and personnel system, the broader problem must be identified and clearly scoped. To that end, we propose the human capital analytic pyramid, which helps depict the range of granularity of the

human capital problem with the relationships between these decomposed problems and the principles identified above. This pyramid shows a way of framing the problem where difficulty increases as one ascends upward through its layers. Each layer grows progressively more complex while interacting with the layers above and below.

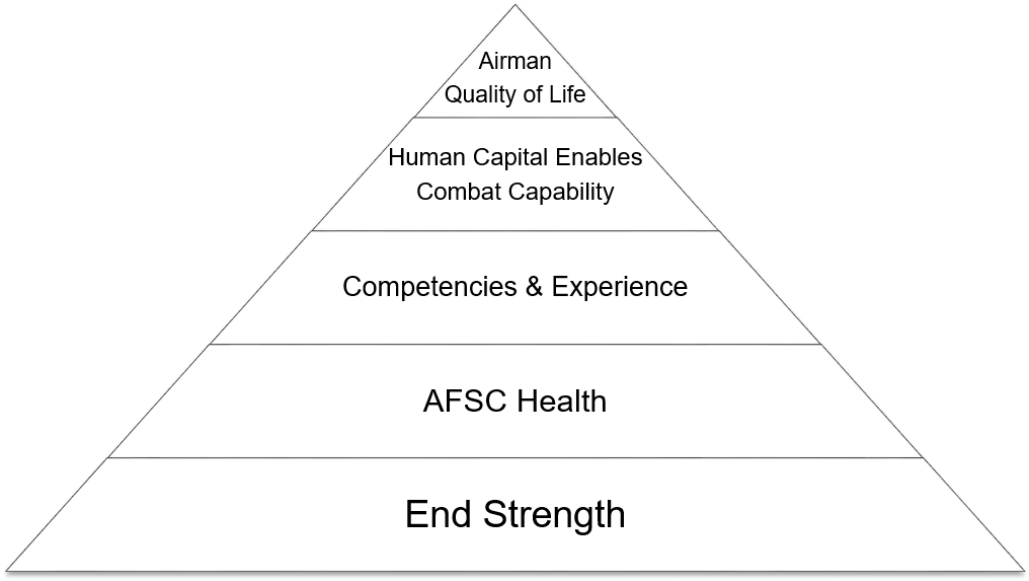

Figure 1. Five Levels of Human Capital Analytic Pyramid

The first layer is end-strength management, wherein the AF can achieve success simply by ensuring it has the Congressionally authorized total number of people irrespective of training or skillset composition. Even this problem is not trivial, but it is an order of magnitude easier than even one level deeper, Air Force Specialty Code (AFSC) Health, which is measured primarily by overall AFSC permanent party manning. This AFSC Health layer is easier to manage than the next, which requires not only the right mix of AFSCs, but also whether the personnel in those AFSCs have the correct experience and competencies, historically measured by grade or skill level for the enlisted force. Another level further includes whether the personnel available are adequate to enable some level of combat capability. Finally, we include a nebulous "Airman Quality of Life," which captures many different, independently important

features such as morale and culture. Given the scope and complexity of this level, no single, good, representative measure relates. However, it is important to remember this level exists, as the effects of the higher levels play a large role in influencing airman experience, and it in turn plays a large role in influencing every other level.

As an illustrative example, when end-strength is managed aggressively (Level 1), it makes it difficult to manage AFSC and grade manning (Levels 2 and 3), which affects people's lives, feelings of security, and workload (Level 5). This impact can then be felt in retention, which in turn influences every level below, cascading through combat effectiveness, experience, AFSC manning, and end-strength. Although we cannot entirely separate any of these layers, there are ways to measure each that provide different insights to the human capital problem.

1.3.1 Level One: Total Personnel (End-strength)

The first problem in managing human capital is to have the chosen total number of people in the system. In the USAF, analysts describe this problem in terms of managing end-strength. The USAF traditionally simplifies this problem by managing how to finish a given fiscal year with the desired number of personnel.

The primary lever for solving this problem is determining an appropriate, aggregate number of accessions. Efforts to shape retention should not be dismissed entirely, although these efforts typically have a far smaller impact than outside factors such as changes in economic factors or Airmen's impressions of AF culture and standard of living. Additionally, retention incentives change by relatively small amounts each year in comparison to aggregate compensation. For these reasons, the USAF manages end-strength primarily through accessions, especially when total end-strength is flat or growing. When slight cuts are required, it can also reduce the force size through reductions in the number of accessions. For significant cuts in end-strength, it can

use force management actions such as Reduction in Force (RIF) and force shaping boards; however, these tools can have significant negative impacts both to the Airmen selected as well as those not selected, who experience the career uncertainty when meeting these boards. Although the AF must be able to manage its resources, it should not cavalierly embrace policy levers that create pain and frustration for its personnel and compromise the trust and security of its Airmen.

When building active duty, or Regular Air Force (RegAF), accession plans, USAF analysts determine the aggregate number of accessions each year not by examining the number of accessions needed for individual AFSCs in the RegAF, but by considering the aggregate funded end-strength target. This change in end-strength is the combination of expected attrition from the force and the desired level of growth or decline in total personnel. In an environment with unstable budgets and desires for different end-strength targets, accession levels change across the AF on a regular basis. As an added complication, trainees attending Basic Military Training (BMT) and Initial Skills Training (IST) have some level of washout rate resulting in a departure from the USAF. This washout rate describing departures is not to be confused with a washback rate, which describes trainees who simply move back a class, or washout rates describing transfers that result in an Airman moving to another IST pipeline, which does not result in a loss to the AF. This washout rate describing departures means that for every additional accession, the estimate of Airmen losses also increases slightly due to the potential for this new recruit to depart the AF before the end of BMT or IST. Figure 2 depicts this relationship.

The implementation of policies to manage end-strength creates *bow waves* (i.e., overages compared to steady state) and *bathtubs* (i.e., shortages compared to steady state) whenever the Air Force changes its end-strength by a significant margin. Each one of these decisions continues to impact experience, readiness, and aggregate reten-

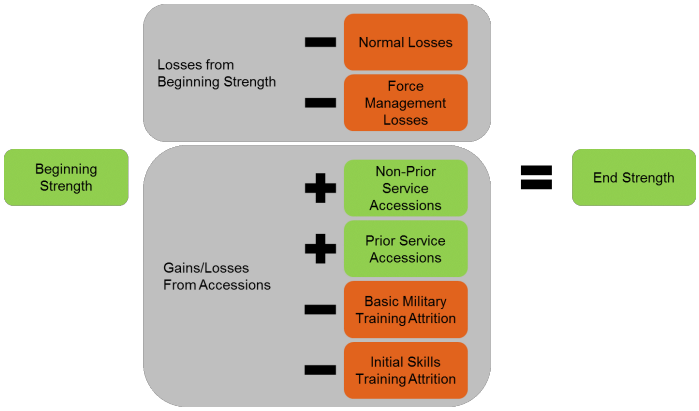

Figure 2. End-strength is determined by beginning strength, gains, and losses.

tion for over 20 years as Airmen age through the system.

1.3.2 Level Two: AFSC Health

It should be apparent that having the right total number of people in the RegAF is necessary but insufficient to field the human capital needed to deliver effective combat capabilities. As we progress to the next level, we now consider whether the total number of permanent party personnel with a given AFSC matches the total authorizations for that AFSC summed across the UMD for each Major Command (MAJCOM). Permanent party personnel are fully qualified personnel who are not in a designated Student, Transient, and Personnel Holdee status.

Further background is required to understand the nuances of the career field manning aspect of USAF Human Capital Management. Unlike with end-strength management, the USAF regularly struggles to achieve its primary objectives at this level, averaging approximately 12,000 enlisted AFSC shortages per year over the last two decades. Fundamentally, this HCAP level is about having the right number of personnel within each AFSC. However, two complications arise. First, only permanent party personnel can fill positions on the UMD, which do not include those still in IST prior to arrival to their first duty station. Second, positions on the UMD change over

time.

Career Field Health and the USAF Sustainment Model

The Career Field Health (CFH) approach to force management leverages the sustainment model, which is a steady state representation of each AFSC given they achieved 100% manning by accessing and retraining a consistent number of personnel every year. Officer and enlisted AFSCs both utilize the Career Field Health approach, but there are some differences. The officer methodology accounts differently for those serving outside their core AFSC. For this reason, this discussion describes sustainment in terms of the enlisted force, as that is the baseline for both methodologies. Any differences between the two approaches are noted by exception. There are several desirable features of this approach. First, it uses the AFSC's own retention behavior by years of service (YOS) over the last five years as a predictor for future behavior. Years of service show substantial predictive power for retention behaviors; Airmen make many transitions at key points in time that are generally stable in relation to years of service. Retention here describes the observed probability of an Airman remaining in the force for an additional year given their number of completed years of service. For example, the likelihood of departing the service after two years remains consistently low, as the enlisted Airman entered service under a four or six year enlistment contract and has at least two years of obligated service remaining. At four to six years of service, as the service member completes the first enlistment or active duty service commitment (ADSC), we observe substantially lower retention. At 18 years of service, retention rises dramatically due to the incentive of a defined benefit retirement plan only available when the service member reaches 20 YOS. We observe a sharp drop in retention at retirement eligibility. Finally, AFSC sustainment maps retention within the AFSC, not within the USAF. Thus, the sustainment model con-

siders an individual retraining out of an AFSC as a loss for that AFSC because that Airman will no longer meet one of those AFSC's authorizations.

The second set of behaviors that feed the sustainment line derives from the gains distribution describing when individuals complete training and become permanent party members in their career field. This distinction bears mentioning because the delay into an AFSC depends not only on their own training pipeline, but any other pipelines that the trainee did not successfully complete prior to the final AFSC.

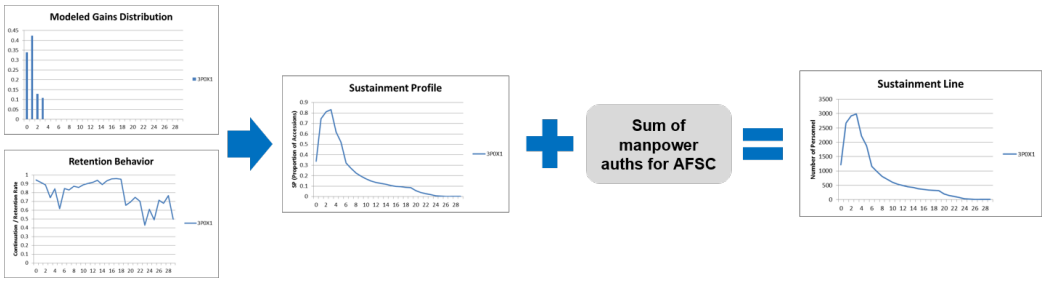

Figure 3. Process to Build AFSC Sustainment Line

With both of these behaviors mapped, the USAF creates a sustainment profile showing the probability of a single Airman making it to any given year of service in that AFSC. This line goes up when personnel arrive into the AFSC and down as they depart the AFSC or leave the AF. Once the shape of the line has been determined, it is scaled upwards until the area under the curve is equal to total authorizations.

Another key insight from the sustainment model is the sustainment requirement for accessions. This quantity captures the number of individuals who would need to graduate IST each year to sustain the career field. We represent this target with the black dashed line and the number in the black rectangle on the left side of the chart below.

Accessing individuals at the level of the sustainment requirement still requires all existing bathtubs (shortages) and bow waves (overages) to age through the system prior to returning the AFSC to full health over a 20-30 year lifecycle. Conversely,

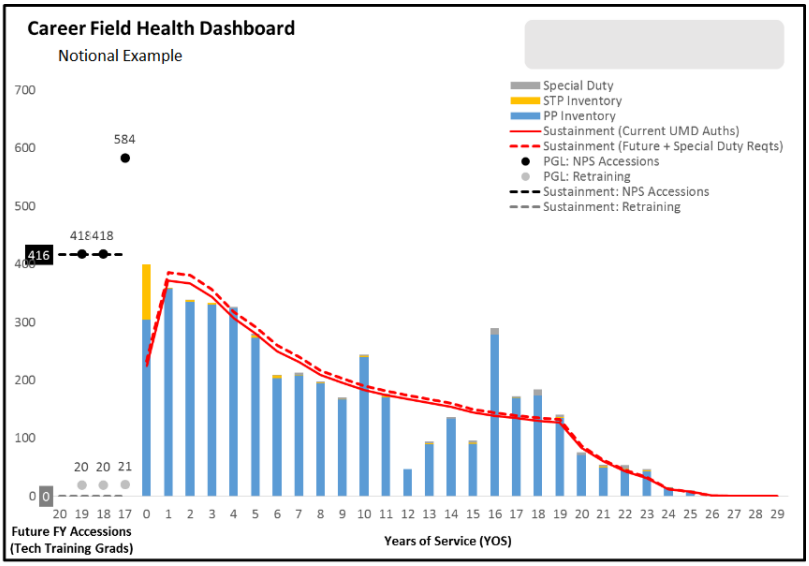

Figure 4. Career Field Health Chart with Sustainment Line and Inventory

deviating from the sustainment accessions target builds the next set of bathtubs and bow waves, although doing so may address aggregate manning problems in the short term. Accessions policies and other force management policies such as retraining policies, retention bonuses, and high year of tenure (HYT) waivers are used to move AFSCs as close to 100% manning as possible.

There are several benefits of keeping an AFSC close to its steady state as defined by sustainment. Absent dramatic changes in retention or requirements, this distribution of inventory results in the same aggregate retention, the same required number of accessions, the same experience ratios and associated distribution of labor, and the same upgrade training burden each year. The USAF invests substantial resources (e.g., personnel and infrastructure) to execute a steady state level of recruiting and training for accessions each year and it is extremely expensive to dramatically adjust the number of accessions. However, deviating significantly from this sustainment distribution results in varying numbers of personnel hitting the same retention decisions each year, simultaneously driving large swings in the required number of accessions to offset losses and maintain 100% AFSC manning.

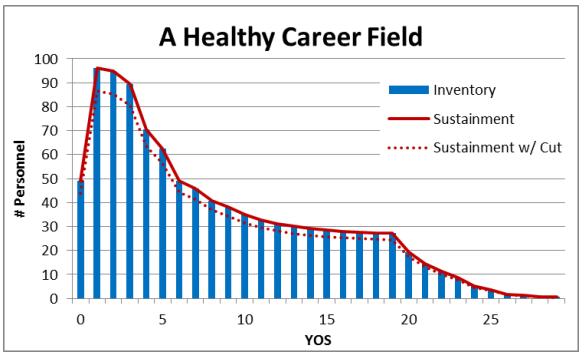

Figure 5. Vignette AFSC Starting State

The need to avoid varying too far from sustainment must be balanced with the desire to correct manning in a reasonable amount of time when changes do occur. One consistent feature of the USAF manpower and personnel system is the need to continually change the mix of AFSCs within the service based on changing programmatic requirements or decisions made within the MAJCOMs. As these needs change, the demand signal for AFSCs often changes rapidly, and frequently with little advance warning on the UMD. To illustrate the dynamics of this cycle, a notional AFSC is examined. As shown in Figure 5, the AFSC starts out perfectly healthy with 100% manning. A 10% reduction is applied to the UMD requirements and force management programs remove 10% of personnel through a combination of separations, retirements, transfers to the reserve components, or retraining, as shown in Figure 6.

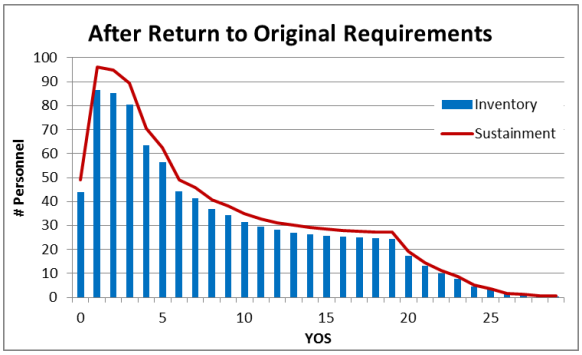

Figure 6. Vignette AFSC After Authorization Growth

After some period of time, additional authorizations are added to return the AFSC

to its original size, causing it to now be 90% manned. For some AFSCs, retraining Airmen into the specialty is a viable option under these circumstances. However, other AFSCs require a high degree of technical knowledge and are not helped by retraining in personnel with AF experience but not the specific technical knowledge required. Accordingly, the first option to grow the AFSC is to recruit and train new Airmen to the sustainment target accessions level. This avoids overloading the pipeline and the need to arbitrarily reduce accessions to other career fields (a second-order effect of end strength management), only grows the training requirement by 12%, and avoids creating a bow wave that will continue for 25-30 years. As seen in Figure 7, after two years of this policy, manning has only improved from 90% to 91.4% and is implicitly on a 20 year get well plan. This outcome is undesirable for commanders facing new and accelerating mission requirements today.

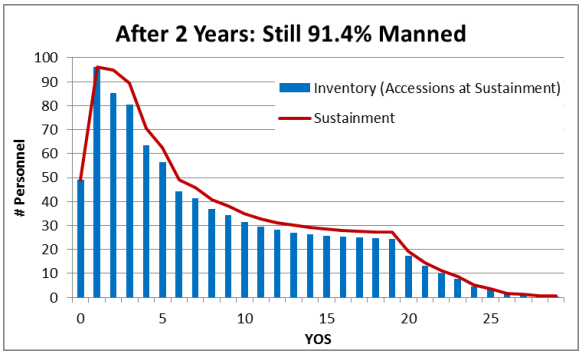

Figure 7. Vignette AFSC Steady State Get-Well Plan

Alternatively, consider the two year get-well plan shown in Figure 8. To achieve 100% manning in only two years, the training pipeline must expand capacity by 78% immediately, exceeding programmed instructors and training resources for this schoolhouse, while reducing accessions for other career fields below their sustainment target. Such a surge compromises the grade structure, causing the mid-level supervisors who remain in the inventory to be burdened with a significantly higher on-the-job-training (OJT) workload to train the new, inexperienced Airmen in ad-

dition to ensuring mission accomplishment. Moreover, the newly trained personnel cannot complete the same duty requirements of the mid-level supervisors previously cut, so the same level of manning (100%) would actually reduce mission effectiveness compared to the personnel prior to the cut.

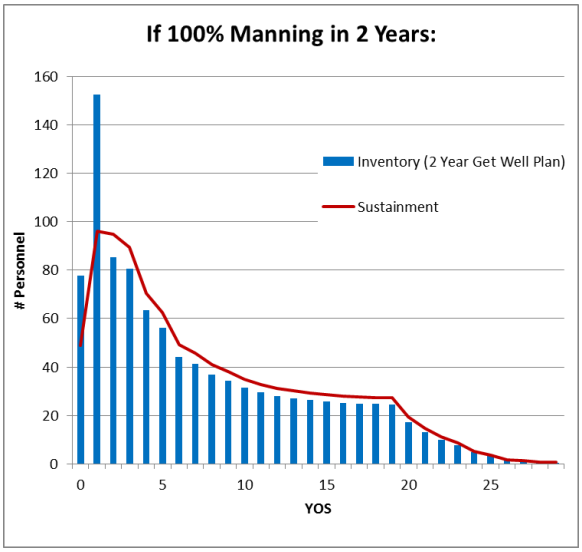

Figure 8. Vignette AFSC Two-Year Get-Well Plan

AFSC Shortage Root Causes

The aggregate effect of all of these factors is a substantial number of shortages in the RegAF that endure to a varying degree from year to year; the enlisted force has averaged about 12,000 since 2000.

Shortages can be roughly quantified according to root cause within two broad categories. The first category represents unfunded manning disconnects that result in the aggregate enlisted permanent party personnel being fewer than the aggregate enlisted UMD authorizations. This category also includes temporary disconnects from surges in the number of students when end-strength is growing. This category of shortages can be observed as the difference between the total number of shortages and the total number of overages in Figure 9. When excess overages exist, shortages can

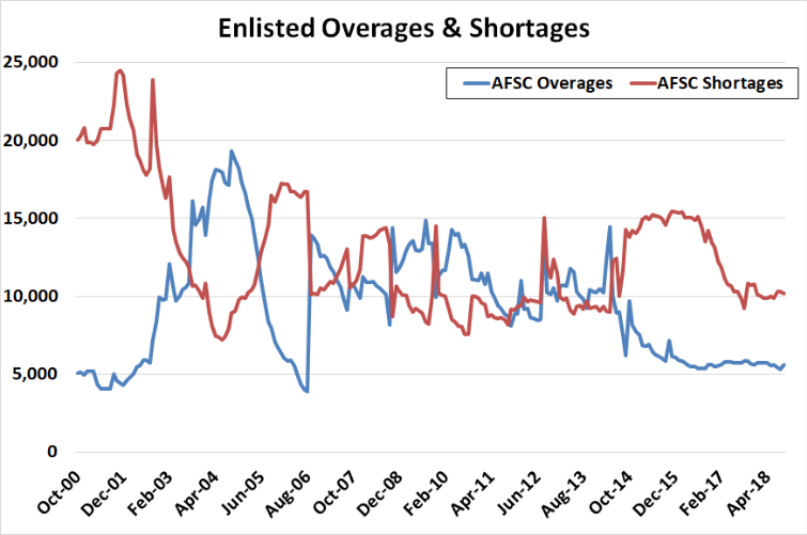

Figure 9. Enlisted Overages and Shortages

be filled by force management actions (e.g., retraining and accessions). When excess overages do not exist, the only way to solve shortages is to make funded end-strength levels match authorizations, either by reducing the total number of authorizations or adding end-strength.

The second category represents funded manning disconnects (i.e., disconnects due to overages), which include pipeline constraints and training execution problems, retention changes and modeling limitations, or manpower authorizations not projected sufficiently in advance for the AF to access and train personnel to fill them. Considering these overages and shortages comprehensively, we can assess the broad performance of the personnel system with regards to enlisted AFSC manning. We observe a slight positive trend from 2000-2015, with the annual total disconnect remaining largely constant over time. In 2016, two changes occurred simultaneously. The AF began to grow its end-strength, and the enlisted sustainment model was rebuilt to improve enlisted force management policies. At this point, we observe a decrease in the unfunded manning disconnect, as the aggregate end-strength moved closer to the required number of personnel to meet the aggregate authorizations. Meanwhile,

the AF's improved force management policies preserved a level of efficiency only previously possible when there were large funding disconnects. When AFSCs are all manned well below 100%, the odds of having excess personnel in any AFSC is low. However, when AFSCs are manned at 100% on average, every person must be in the correct AFSC to avoid overages.

1.3.3 Level Three: Competencies and Experience

While not well understood, it should be apparent to most Airmen that shortages are a problem that the USAF should make every effort to solve. The next level of the pyramid demonstrates no such clarity. Many assume that measuring requirements by grade and then using force management policies to shape the force to meet these requirements would be the next step. However, in the next section we discuss why this is not a viable path.

More broadly, the USAF is attempting to define the competencies and experience that Airmen require to effectively complete their jobs, which is not necessarily captured by grade. If the USAF desires additional experienced Airmen in an AFSC, simply promoting more junior personnel to a higher grade does not solve a problem with missing experience – the same Airmen are still completing the mission. Alternatively, YOS provides some measure of how long someone has had the opportunity to learn their craft, but does not capture aptitude, attitude, or capability.

Primary skill level is a better proxy for the enlisted force, although not all AFSCs utilize primary skill levels in the same way. Additionally, there is no "requirement" to measure primary skill level against, as UMD authorizations only specify control skill level, which progresses more slowly. A common misunderstanding is the difference between control skill level and primary skill level. Control skill level is driven by grade and is the metric being measured for skill level manning. However, primary skill level

reflects the level of qualification of the Airman. For example, a SSgt maintainer (control skill level is 5) who is certified for 7 level duties (primary skill level is 7) can meet the unit commander's 7 level requirements to sign off on aircraft to generate sorties, even though the SSgt's control AFSC will continue to show as a 5 level. This maintainer will remain a 5 level until the Airman is promoted to TSgt at which time both the primary and control AFSCs will carry the 7 level.

Because primary skill level is achieved prior to control skill level, commanders frequently execute the required mission with disconnects in control skill manning if there are adequate personnel who have achieved higher primary skill levels. Figure 10 shows an example of an AFSC's skill manning using control and primary skill levels, with manpower authorizations in red and personnel in blue. In the pictured example, we see that when considering only control AFSC, there appears to be a shortage of 7 levels. However, when considering primary AFSC, we see that there are plenty of 7 levels, and the excess of 7 levels can help to meet the apparent shortages in 5 & 3 skill levels. Thus, when discussing skill manning, the conversation is truly about either grade manning (i.e., control skill manning), or a meaningless comparison of personnel and their primary skill level to manpower requirements and the required control skill level. This mismatch prevents the current construct (at least in this format) from providing meaningful feedback on whether current upgrade timelines can meet USAF requirements.

Furthermore, "shortages," such as we may quantify them, are frequently the result of policy choices made years or decades ago, and significant limitations exist regarding what the USAF can do to solve such manning problems. If the USAF faces a shortage of competence and experience in an area, it frequently does not have adequate policy levers to address this problem with agility. Wherever the USAF has the opportunity to maximize learning and development of competencies, it attempts to construct

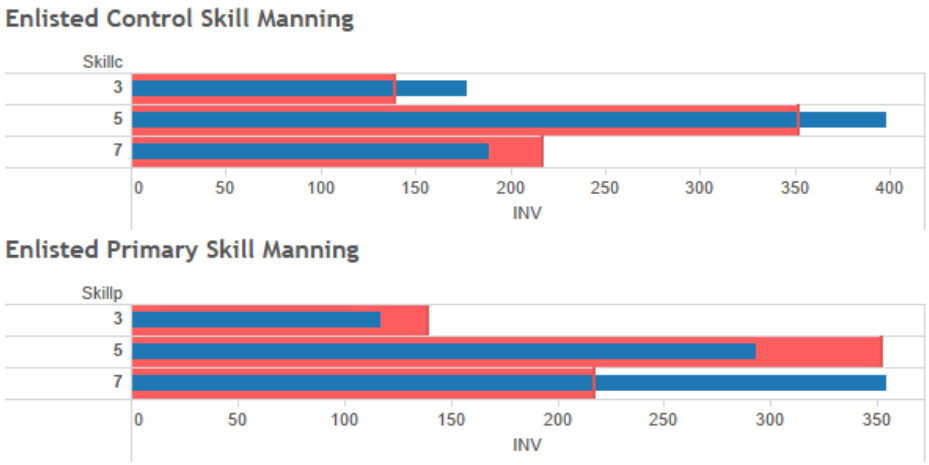

Figure 10. AFSC Skill Manning Example

policies to do just that. In general, as the USAF is always looking to maximize the learning and competencies it is developing in its Airmen. As such, the policy question that the Air Education and Training Command (AETC) commander, in their force development role, is challenged with is how to increase learning more generally, not just temporarily boost learning to solve a crisis. The USAF can also influence retention behavior to some degree, but Airmen's aggregate compensation, outside opportunities, and satisfaction with the USAF typically dwarf the retention incentive provided by comparatively small Skills Retention Bonuses and similar programs (Joffrion and Wozny, 2015). Finally, retraining individuals into AFSCs only helps if the experience shortfall requires generalized AF experience instead of technical capacity only gained working in the AFSC.

The goal of this section is to provide structures that truly increase competencies and experience where possible, avoid optimizing policies to "solve" flawed or inadequate metrics, avoid policy decisions today that will drive additional dilemmas 10 years from now, and save the substantial amount of wasted effort and resources the USAF dedicates institutionally to solving non-existent problems in our metrics. Prior to delving into modifications to the system, we must examine some additional

background on system behaviors that affect experience levels in the USAF.

AFSC Grade Management: The Case for Sustainable Grade Structures

In addition to the overarching grade structure issue, a second business process disconnect results in commanders chronically not receiving personnel with the AFSC and grade mix that has been authorized on the UMD. This disconnect must be considered separately from temporary fluctuations in grade manning that will be solved over time. The AFSC may not have sufficient personnel to meet the sustainment requirement in a specific year of service; this is temporary and will be solved over time as bathtubs and bow waves age through the system and the USAF utilizes force management policies to solve these disconnects. The second and more serious disconnect is systemic and arises when there is a substantial difference between the sustainable grade distribution for an AFSC and the distribution of grades MAJCOMs place on their UMDs.

The historical grade review process ensures that MAJCOMs are keeping aggregate grades distributed correctly on the UMD; this process allocates to each MAJCOM a share of the overarching grade structure, while the collective requests by the MAJCOMs determine the distribution of this share by AFSC. However, MAJCOMs frequently request distributions of grades that by AFSC are not feasible when combined with other MAJCOM requests. As an extreme example in a non-prior service accessions AFSC, requesting all E-5s or all E-6s in an AFSC is obviously a request that cannot be satisfied; E-3s and E-4s must exist to grow into the more senior grades. The sustainment model defines a historical normal for an AFSC by YOS; we can also observe the historical probability of an Airman being in a specific grade given their level of experience as measured by YOS, shown as grade sustainment to the right.

The existing sustainment line for the AFSC can be combined with the corresponding historical grade distribution to determine what the approximate number of personnel in each grade will be in the long run if the AFSC is manned at 100% and distributed according to the sustainment line. This represents the distribution by grade that an AFSC can achieve assuming no substantial changes in retention.

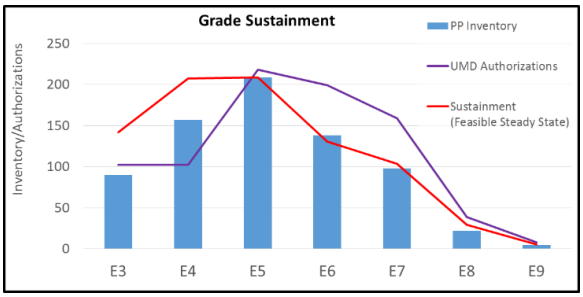

Figure 11. AFSC Grade Sustainment Example

In order to examine this disconnect, the year of service based sustainment line discussed above is converted to a grade based model. This allows for a comparison of current inventory, sustainment (a feasible steady state inventory), and UMD authorizations for each grade. Consider Figure 11. The relationship between the blue bars (current inventory) and the purple line (UMD authorizations) is what is reported as grade manning. The red line (sustainment) demonstrates the feasible steady state grade manning based on current retention behavior. Where the purple line (UMD authorizations) departs from the red line is a manifestation of an infeasible grade distribution on the UMD. In this particular example, there are substantial grade manning problems for this AFSC. However, these particular problems are largely systemic, a result of a desired career pyramid that is not feasible with normal retention patterns.

As described earlier, the mechanisms to shape retention are generally weak and highly constrained. In the absence of better options, current policies only offer two mechanisms to meet unsustainable grade structures. The first is a hybrid grade

structure; supplementing non-prior service accessions with retrainees increases the aggregate AF experience in the career field by adding personnel with higher YOS, although this does not address needs for technical experience gained in the AFSC. The second option is to promote more Airmen at junior levels of experience (also known as promoting to requirements). The problem with this second approach is that the commander still receives the exact same Airmen with the exact same experience and competencies; we have simply manipulated the metric by increasing their rank and cost, which does very little to help accomplish the mission under most circumstances. We recommend that grade reviews include guidance on the desired grade distribution within a specific AFSCs as well as the aggregate for each MAJCOM. While this appears to be a restriction on the MAJCOMs, in reality, this informs the MAJCOMs what levels of experience will be available. This empowers the MAJCOMs to make decisions about which positions may be more appropriate for more junior personnel, instead of creating an infeasible wishlist (which cannot be met), then effectively delegating to the AF Personnel Center (AFPC) the MAJCOM's decision on how to distribute their personnel.

1.3.4 Level Four: Human Capital Fielded as Combat Capability

Readiness and Lethality

The next level of assessment for the HCAP is the USAF's ability to field combat capabilities with our assigned personnel. Aside from classification issues, military readiness is a nuanced subject, difficult to effectively measure, and not necessarily suited to simplistic metrics (Betts, 1995; Harrison, 2014).

This level becomes more complex for several reasons. The first is that personnel must be assigned to units by AFPC based on several different prioritization schemes by the commanders or HAF staff, depending on whether the personnel are officer or

enlisted and whether the personnel are rated aircrew or non-rated. This assignment process is a complex process requiring careful balancing of changing AFSC and grade requirements, retention, move cycles, and individual personnel considerations.

Another complexity is that the RegAF, Air Force Reserve (AFR), and Air National Guard (ANG) all deploy together to field capabilities to the joint commander, requiring metrics that are not constrained to the RegAF. Ideally, any metrics in this domain would need to first quantify what capabilities the USAF had fielded the human capital to support. The USAF's current personnel readiness metrics, however, simply report whether it has fielded what has been funded, yielding efficiency metrics for the personnel system instead of true measures of readiness.

P-Ratings

The primary driver of poor readiness measures are unit P-ratings, an assessment of whether units have adequate personnel to accomplish their mission. Although defining an appropriate metric for determining whether the USAF has adequate personnel is outside the scope of this research, the inappropriate use of P-ratings to provide insight for resourcing decisions deserves attention. P-ratings are derived by comparing available personnel to authorized manpower requirements on the UMD for specific combinations of AFSCs and skill levels.

A key feature of this metric is its use of the current resourcing decisions as a baseline. An example may help illustrate the problem with this feature. A unit with poor readiness measures due to poor manning is being considered for additional resourcing to solve their readiness problem. As they add manpower authorizations, the unit's readiness measures (i.e., P-Ratings) initially get worse, not better, as the personnel system lags in filling the new positions. As time passes, unless the personnel system has grown more efficient in some way, the unit's readiness measures return to

the old baseline despite the increase in actual readiness provided by the additional personnel filling the new authorizations. What has been measured for P-Ratings cannot reflect whether the resourcing decision was appropriate or what capability has been procured by these additional personnel; it only reflects the efficiency of the system in filling those positions, regardless of whether those positions are adequate to provide the combat capability needed from the unit. Without increasing funded end-strength relative to total authorizations or relying on prioritization to cannibalize other units, the only way to boost P-Ratings with a resourcing decision is to reduce authorizations in a unit. P-Ratings would temporarily rise in such a situation because the loss of existing personnel from the unit will lag behind the resourcing decision. This is counter to most decision-makers' intuitive understanding of such a system.

1.3.5 Level Five: Airman Quality of Life

The fielding of human capital as combat capability is the primary success condition for the manpower and personnel enterprise. However, successfully fielding combat capability is not a sufficient success metric for the personnel themselves. There is an entire level of complexity to Airmen's experience that goes well beyond the functions they enable.

Like each previous level in the human capital pyramid, this level is both affected by the levels below it and in turn affects the levels below it. Airmen's experiences in their unit are greatly affected by force management actions to manage end-strength, manning levels of their own AFSC and support AFSCs, and the competency and experience of their fellow Airmen at every level. Positive or negative experiences like this drive retention, performance, attitudes towards risk and innovation, culture, and much more. A comprehensive view of the management of human capital cannot neglect these aspects that have a substantial impact on any relevant measure of

success.

This level is incredibly broad, including leadership, culture, messaging, and mission as well as business practices. Additionally, many of the descriptors at this level also have significant impacts directly on readiness and lethality of units, making it difficult to parse where one level ends and the other begins.

1. **Clarity of Purpose**

   Airmen are naturally driven by meeting both AF and personal needs, which vary by Airman. Impacts between these two aspects can interact with each other; Airmen are far more willing to put up with poor conditions when the mission is clear and has the support of the personnel involved (Siebold, 2006). However, when both of these suffer at the same time, the negative effect is compounded.

2. **High Public and Intra-Service Esteem**

   One boon to USAF recruiting and retention is that the US military is perceived by the US public as winners and public servants, and remains the most trusted institution in the US (Kennedy, 2018). This aids the development of a culture of reinforcement and value; high public esteem acts as a form of non-monetary compensation. This also affects who joins the military and then how those individuals respond to military bureaucracy, compensation, and cultures within the military (*Recruiting and Retention of Military Personnel*, 2007).

   High public esteem can be offset within the service by degrading or mistrustful behavior. Those who join to become a part of something larger may then chafe at a risk-averse leadership "treating them like children," an oft-heard complaint among members. Leadership must continually strike a fine balance between managing risk for the personnel and treating their personnel with a greater

level of trust and respect, knowing that a non-zero number of individuals in any large organization will abuse this trust.

3. **Organizational Culture**

   The nebulous concept of culture includes leadership styles, command climate, levels of personnel and leadership homogeneity or diversity, work-life balance, and much more. This research does not provide a comprehensive overview of all aspects of culture but makes note of the importance of this difficult-to-measure aspect of USAF life. Once again, like the other levels of the HCAP, culture can have major impacts on readiness and lethality. This culture can also be self-reinforcing once established. Some cultural aspects may be easily identifiable as clearly positive or negative. Other aspects may create tradeoffs for the organization's mission effectiveness or simply be a matter of preference for the individuals involved (Siebold, 2006). Notably, cultural aspects of AF organizations can greatly influence an organization's ability to be diverse and inclusive either positively or negatively. This can create a cascade of effects through the other layers of the HCAP and directly impact mission effectiveness (Lim, 2015).

4. **Compensation**

   Airmen have their own financial goals and considerations. Some Airmen are profit maximizing, with marginal income directly impacting retention likelihood. Others are satisficing, requiring some base amount of compensation to meet their and their family's needs, with limited impact to retention beyond that personal pay requirement. While historical pay rates were much lower than current levels, years of pay raises have increased military compensation to be competitive with compensation for private sector employees with similar edu-

cational attainment (Smith et al., 2020). A recent RAND study found military personnel's pay to outstrip their peers in the private sector (Asch, 2019), but the military's equitable pay system also doesn't allow for high and low performers to be compensated at different levels as the private sector does (Hoecherl, Schulker, Hornberger and Walsh, 2022). Thus, higher pay may be a critical driver of a military's effectiveness, if it enables the retention of a higher performing talent pool. Additionally, direct comparisons do not account for the substantial negative impact of a military career on a spouse's earnings (Hosek and Wadsworth, 2013).

Interestingly, providing credentials and experience that can result in a high level of compensation in the private sector may also be perceived positively as an additional form of compensation, acting as a pull both to reduce and increase retention. However, this benefit is reduced if the benefit is transparently transactional. Decision-makers must balance compensation policies carefully, ensuring that the taxpayer receives a benefit for additional expense, while also avoiding becoming too risk averse in developing its most valuable resource: its people.

5. **High Performing Organizations vs Poor Bureaucratic Processes**

The USAF bureaucratic processes impact all Airmen. The effectiveness and efficiency of these processes create work (negative compensation) for Airmen. When this work becomes too onerous compared to the compensation, Airmen exit the system.

Bureaucracy, counter to its use in the pejorative sense, is absolutely necessary to make any large organization function. However, bureaucracy's bad reputation results from a correct assessment that many bureaucracies fail to remain responsive to the objectives of the organization. When the system demands work

or sacrifices of its members without reason, this quickly results in cynicism and skepticism, even of valid requirements.

In the age of social media, every policy is scrutinized and analyzed by those with a vast access to part of the relevant information. In this environment, clear communication of the "why" for different policy changes is critical. Historically opaque, increases in transparency (intentional or otherwise) have resulted in an awareness and amplification of any missteps by the organization. This new transparency cannot and should not be reversed, but it does amplify the importance of additional organizational transparency and a continuous effort to improve systems. Many criticisms are the result of confusion; consistent messaging, with regard to both specific policies and the rationale behind them, is vital to preventing this frustration from festering.

As many systems move to the cloud and AF/A1 makes many of the AF's core processes more streamlined and user-friendly, the underlying data must be captured and used to continue improving the experience of Airmen at every step of their time in the AF. Every minute spent struggling with the Defense Travel System or being unable to solve pay discrepancies echoes through the AF's ecosystem, impacting retention, satisfaction, core competencies, and eventually the ability to execute its combat mission.

1.4 Research Questions

Many facets of the current personnel system are products of historical development and may no longer be relevant; much of the system is ready for redesign. However, the current system is so complicated and the language to describe what is happening so imprecise that intelligent, knowledgeable people talk right past each other. Additionally, causality is frequently difficult to attribute; is poor performance in a unit

due to a local leadership problem, inexperienced Airmen, a low manning level, inadequate authorizations compared to manpower requirements, or inadequate manpower requirements to start with? The underlying causality is impossible to fully determine in some cases, but end-strength management and AFSC management are not wicked problems; they are problems of mathematics and system design. The goal of this work is to simplify and solve these problems, so that knowledgeable experts and leaders can examine the remaining simplified but still wicked problems.

To this end, we propose the following research questions:

1. How can the USAF use MilPDS and publicly available data to accurately and precisely predict monthly retention behavior over a 12 month period? The answer to this question directly impacts Level 1 of the HCAP, end-strength management.

2. How can the USAF improve the quality of accessions policies implemented by AFSC to reduce AFSC shortages and improve AFSC manning. The answers to this question directly impact Level 2 (career field manning) and Level 3 (competencies and experience) of the HCAP.

3. How can the USAF improve the quality of accessions policies across all components implemented by AFSC to reduce AFSC shortages and improve AFSC manning? What policies that significantly impact AFSC manning need to be managed differently or start being managed? How do we ensure good solutions to those policies? Within this research, we confine the scope of this question to Level 2 (career field manning) and Level 3 (competencies and experience).

1.5 Research Contributions

This research makes the following three contributions, collectively addressing the research questions in Section 1.4.

1. We develop, test, and compare multiple statistical machine learning methods to predict USAF retention accurately. Accurate predictions of retention are important because instability in retention modeling drives unnecessary changes to AETC and Air Force Recruiting Service (AFRS) accessions and recruiting decisions; or unnecessary overages and costs; or shortages and gaps in readiness. This work makes a novel contribution by developing a new, partially autoregressive feature and constructing a designed experiment for hyperparameter values for both multilayer perceptrons and random forests for a novel problem.

2. We design, develop, and test novel approximate dynamic programming (ADP) and reinforcement learning (RL) algorithms that determine high-quality accessions personnel policies. Manning is a function of personnel gains, personnel losses, and authorizations change. Voluntary retention rates are difficult to increase and decreases can require force management actions. Impacting the rate of authorizations change requires business process changes, and some courses of action require Congressional approval. This leaves accessions and retraining policies to control personnel gains and losses, though the effect of changing retraining policy is much more difficult to model. We formulate this problem as a Markov decision process, develop a direct lookahead policy modification of Concave Adaptive Value Estimation (CAVE), and develop an alternative parameterized deep reinforcement learning approach to generate high-quality policies for accession decisions with high dimensionality while maintaining a low computational demand. We also test the effects of potential cost functions

on the policies generated to inform further model development.

3. We design, develop, test, and compare multiple sequential decision-making approaches for determining high-quality personnel policies. This contribution extends the work proffered in Contribution 2 by considering a new, larger problem set, including RegAF, AF Reserve, and Air National Guard personnel. The USAF fields its personnel from all three components when presenting forces to the joint commander to execute operations, meaning that the ability of each component to meet its human capital needs is critical to collective mission accomplishment. Moreover, each component shares training resources and competes for many of the same recruits to meet their manning needs, but current coordination of policies is largely ad hoc. Improving these policies directly improves USAF personnel readiness instead of the more limited problem of RegAF manning. First, we extend the RegAF's benchmark equilibrium sustainment model to the AFR and ANG, then formulate this larger problem as a Markov decision process. We extend the CAVE approach to this larger problem and test performance across a range of hyperparameters. This extension creates an expanded state and action space and an opportunity to design algorithms that can scale efficiently to larger problems. Finally, we consider a novel algorithm modification to the CAVE approach which leverages a perturbation and retraining process to improve solution quality at the expense of additional computation and test the performance of this modification across multiple hyperparameters.

1.6 Organization of the Dissertation

This dissertation is organized as follows. In Chapter II, we answer Research Question 1 with Contribution 1, a set of statistical machine learning algorithms to predict USAF retention and enable better end strength management. In Chapter

III, we answer Research Question 2 with Contribution 2, a set of deep reinforcement learning algorithms to improve RegAF accessions and improve RegAF career field manning. In Chapter IV, we answer Research Question 3 with Contribution 3, a set of further developed deep reinforcement learning algorithms to improve a broader set of RegAF and AFR personnel policies and improve USAF readiness. Finally, in Chapter V, we summarize the dissertation and discuss our assumptions, limitations, and drawbacks of our proposed models. We also identify extensions for future work.

This dissertation provides a suite of models to provide improved personnel policies, enabling more effective, efficient recruiting and training pipeline, improved RegAF career field manning and fewer shortages, and improved Total Force career field manning and fewer shortages.

II. Partially Autoregressive Machine Learning: Development and Testing of Methods to Predict United States Air Force Retention

This chapter has been published in Computers and Industrial Engineering (Hoecherl, Robbins, Borghetti and Hill, 2022).

2.1 Introduction

The quality of the personnel in the United States (US) military, especially for its enlisted personnel, provides a substantial strategic advantage compared to most other nations. Although many factors play a causal role in this improved quality, two of the most important are its relatively high compensation and the all-volunteer force structure (Rostker and Yeh, 2006). To maximize this strategic advantage, political leaders must carefully balance the high costs of quality personnel with the opportunity cost to organize, train, equip, and field these forces. The aggregation of these balancing decisions determine the total number of personnel in the force at the end of the year in each military service - known as the authorized end strength.

To meet Congressionally-mandated end strength targets, military planners must plan to recruit and train new personnel to achieve any desired change in end strength as well as replace personnel who choose to depart. Because the US does not use compulsory service and personnel can choose to leave at specified windows of time within their service, planners cannot ascertain the exact number of retained personnel in advance. With average annual personnel costs exceeding $100,000 per person per year, even slight deviations from the planned personnel totals can result in dramatic cost overruns, complicating attempts to responsibly manage the larger budget. When retention estimates are significantly off, the Air Force Recruiting Service and Air Education and Training Command must also adjust their recruiting and training

plans, sometimes with very short notice. Poor estimates incur increased expenses as recruiters and trainers must either let purchased capacity go unused or increase capacity for which no one planned or budgeted, frequently at a higher cost than if the required capacity had been correctly planned.

The United States Air Force (USAF) personnel retention problem (PRP) is to predict how many aggregate personnel in the USAF at a specified point in time will remain in the USAF until another, future specified point in time. To this end, our research answers the following specific questions:

1. Of the personnel currently in the USAF, what is the total number of personnel that will retain for another 12 months?

2. How many personnel will depart each month?

These questions are examined via survival analysis, comprised of regression problems and attendant solution procedures to predict how long a process continues before ceasing. Survival analysis problems exhibit similar features to regression problems, but they can be solved by either estimating the retention or survival rates, or alternatively, by estimating the number of persons who survive based on some set of features, including starting inventory. Each approach offers different ways to leverage the underlying problem structure to improve solutions. Regardless of the approach selected, models applied to the USAF PRP must produce 12 numerical regression outputs predicting the total proportion of personnel in the force at a given time period who remain in the USAF over the next 12 months (i.e., the aggregate retention rate). This requirement differs from a classification problem, wherein the models must determine which individual airmen would depart.

The Military Personnel Data System (MilPDS) includes information on personnel and their characteristics across each component of the USAF, including active duty (Regular Air Force), Air National Guard, Air Force Reserve, or USAF civilians

(MilPDS Dataset, 2021). We track personnel longitudinally in each dataset and measure the rates at which these personnel choose to stay or depart. For this research, we generate 51.5 million individual monthly observations of personnel retention behavior from 2010-2021 excluding some non-representative retention data from involuntary force management programs in 2014. Unless otherwise noted, all years referenced in this chapter refer to the US government's fiscal year, which ends on 30 September.

For the USAF PRP, we select models for the purpose of minimizing mean aggregate absolute prediction error, which measures how well aggregate predictions match aggregate numbers of retained personnel across all prediction lengths. Although several models enable the identification and analysis of influential features, this chapter prioritizes methods most suited for predictive purposes. Future work may prioritize a different set of models for the purpose of inference.

Much of the previous work on USAF personnel retention predicted annual retention behavior over a period of years. Monthly models with shorter prediction lengths can use a much broader range of variables to predict retention because these models do not have to simultaneously predict how such variables will change over time. Simpler statistical methods have performed well on variants of the USAF PRP with longer prediction lengths (Schofield et al., 2018; Pujats, 2020), but models with greater capacity will likely prove to be more effective for the shorter term predictions addressed in this research.

This chapter provides two novel methodologies to predict monthly retention. Both methodologies leverage both greater model capacity as well as autoregressive structure traditionally limited to smaller, highly structured models. We compare the performance of these models to the USAF's current best known model for predicting monthly enlisted retention rates.

2.1.1 Proposed Contribution

This chapter provides three primary contributions. First, we seek to construct a machine learning approach that generates higher quality predictions for the USAF PRP compared to the benchmark Kaplan Meier (KM) model. We test several approaches on novel real-world, USAF training and validation datasets across a range of hyperparameters. The final superlative models leverage a separate USAF test dataset to develop an unbiased estimate of improvement in absolute prediction error.

Second, we propose the use of a multilayer perceptron (MLP) with a partially autoregressive feature for survival analysis problems predicting future behavior of population subgroups. This feature uses the previous time step's retention observation for a larger cohort to predict the next time step's retention observation for smaller specific cohorts. By using the larger cohort, many of the problems with sparse combinations of features can be avoided, allowing machine learning approaches with far more capacity and flexibility to be fielded and much more detailed subpopulations to be used, while still maintaining some of the advantages of an autoregressive approach.

We call this modification a partially autoregressive neural network (PARNet). PARNets provide a modified neural network structure that blends the advantage of an autoregressive structure for time series data with the flexibility of a traditional feedforward MLP without the traditional weaknesses of time series-specific machine learning approaches. Instead of using a large number of lags or previous sparse observations of specific subgroups, we include a set of features with the retention observations for 1 year prior to the observation (effectively a single lag of 12 time steps). To avoid the sparseness problem, we include the observed retention for the parsimonious KM approach already in use instead of a similar approach to build an observation for each specific cohort. This parsimonious model only uses years of service (YOS) and months until the expiration of term of service (ETS), so all subgroups

with feature vectors that match this observation will use the same observation as an additional feature. This approach creates an autoregressive pattern. The neural network can then reliably use a past observation for a much larger cohort to predict trends and changes over time that affect the smaller cohort. The granular variables explain differences from this larger cohort. This approach should prove particularly useful for modeling problems wherein large shocks may affect the system, but different subgroups respond in different ways. Importantly, this differs from previous work on hybrid approaches using MLPs and autoregressive approaches because the use of subgroups changes both the value and the sparsity of previous retention observations based on the number of features included.

Third, we examine whether this partially autoregressive approach can improve random forest regression (RFR) predictions as well. This inclusion allows the RFR to effectively weight observations more heavily for periods that have similar retention levels. Additionally, for problems of appropriate structure, the autoregressive approach provides a measure of closeness that may allow groups with similar features but different retention observations to help inform each other when retention trends drive retention behavior from one feature set to be similar to a different feature set's past retention observations. This approach generalizes to fewer problem structures than the PARNet because it will only improve modeling estimates if the retention pattern is otherwise structured similarly. However, for appropriately structured problems, the ensemble nature of the RFR approach may still provide superlative performance. We call this second approach a partially autoregressive random forest (PARFor).

We compare the results of each of these approaches using a full factorial experimental design to sample across selected hyperparameters for the PARNet, MLP, PARFor, and RFR models.

The rest of this chapter is organized as follows. Section 2.2 describes the specifics

of the USAF PRP, historical approaches to survival analysis and statistical machine learning, the sources and methods used to clean the data, and the methodology of our examination of two novel machine learning methods and a benchmark. Section 3.5 describes algorithms' performance on a validation dataset across all variables and hyperparameters tested, directly compares the top-performing model of each type, and examines the superlative models' performance on a test dataset. This chapter finishes with a description of the remaining work and initial conclusions.

2.2 Materials and Methods

2.2.1 USAF Problem Description and Business Practices

A solution to the USAF PRP must meet two requirements. First, it must produce a single accurate estimate of the aggregate enlisted and officer retention for a 12 month interval as of the beginning of the fiscal year. If the model meets this requirement, then confident planning can minimize the cost of disruptions to USAF recruiting and training organizations attempting to bring in personnel to replace those leaving. Second, it must produce quality estimates for monthly prediction intervals from 1 to 11 months. This is important for two reasons:

- As the fiscal year progresses, it is important to be able to update the prediction of losses for the remainder of the year.
- While Congressional guidance is provided in the form of a target for end strength, personnel costs such as pay and benefits are incurred each month a person is in the military. For this reason, accurate estimates of how end strength will rise and fall throughout the year are important for stable, accurate budget planning.

For the USAF PRP, although creating high-quality retention predictions for individual subgroups within the population is correlated with creating high-quality

retention predictions for the force in aggregate, the primary success metric for this model is how the aggregate accuracy performs workforce-wide. To assess the relevant accuracy of predictions, all errors are calculated by comparing the aggregate retention for a specific month's estimate to the observed (actual) aggregate retention. We examine these error metrics both by prediction length and by calculating the mean absolute prediction error across all prediction lengths. This allows us to examine how the cumulative statistical bias of individual predictions translates into the general accuracy level of the total predictions. Future models used for inference to identify the impact of changing subgroups and features would find the subgroup errors much more relevant to building a quality model.

This chapter limits its scope to the enlisted portion of the USAF PRP because there are many more enlisted personnel than officers, and both are managed separately. Enlisted personnel have slightly different characteristics, and their decisions to remain or depart are based on different incentives and policy constraints than the officer population. Traditionally, a second model is used to estimate the number of additional retention losses for individuals who enter the force after the prediction date but depart before the end of the prediction interval. The retention estimates for this model only include personnel already in the system. For time periods including multiple months, this model does not address individuals who enter the USAF after some number of months, then either retain or depart.

2.2.2 Review of Statistical Machine Learning Approaches

Different statistical machine learning approaches leverage different underlying structures. Herein, we review a subset of relevant approaches, some of which we will test and evaluate later in this chapter. We limit our examination to approaches that predict a rate ranging from 0 to 1 for subpopulations with a given feature set.

This rate estimates the proportion of airmen with this feature set that remain in the personnel system for a specified additional number of months.

Kaplan Meier

KM survival estimates have the useful property of effectively handling noisy or sparse data wherein a small number of variables may have a highly nonlinear effect on the dependent survival variable (Meeker and Escobar, 2014). However, these survival rates cannot extrapolate or interpolate survival rates for any combinations that have not already been observed due to a lack of parameters or any measure of distance. Hence, a KM approach cannot use any variables that may change over time or trend because the model would then lack both a direct observation to inform its estimate or a method to interpolate between existing observations.

The USAF's current loss modeling approach uses 12 separate KM estimates of the retention rate for existing personnel with specified combinations of features over the 12 respective prediction intervals. This predicted retention rate estimates the proportion of personnel with the given feature set that remain after the specified prediction interval. This approach leverages concepts from an underlying loss model based on YOS as proposed by Hoecherl et al. (2016). However, the USAF approach extended the variables to include months until separation date, provided the personnel had filed separation paperwork. Due to time constraints, USAF personnel management analysts tested only the final outputs of this model and observed a satisfactory and much improved aggregate error of less than 1.5% for annual losses and 0.2% for annual retained personnel using a 2015 test dataset. A notional example of two combinations of the two predictor variables and the 12 predicted retention rates is shown in Table 1.

Months Until Sep Date	Years of Service	Months											
		1	2	3	4	5	6	7	8	9	10	11	12
0	8	.32	.11	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
5	8	.97	.95	.95	.94	.48	.12	.06	.01	.01	.01	.01	.01

Table 1. Notional examples of KM retention estimates for feature groupings

Random Forest

A random forest is an ensemble method based on decision trees originally introduced by Ho (1995), which Breiman (1996) refined with a bagging approach. Originally developed for classification, Breiman (2001) extended its use to regression, and RFR has proven to be an effective machine learning approach for a number of problems (Géron, 2019). Unlike a single decision tree, which partitions the feature space to develop an estimate for an observation, random forests are collections of decision trees formed by allowing each decision node in each tree to randomly select a subset of features and then search for the best partition among just those features. This approach decorrelates the partitioning process in the set of trees, increasing the diversity of the individual trees in the random forest. Diversity in the forest helps ensure the model generalizes well to observations not already included in the training data. By examining how the inclusion of each feature impacts tree leaf purity (the lack of diversity of the observations contained within each leaf) throughout the forest, a second benefit of RFR is the ability to identify which features are most important for making accurate predictions.

Autoregressive Approaches for Time Series

One approach to predicting future values of time series data is the use of autoregressive approaches. Instead of observing correlations between the variable of interest and potential explanatory variables over time as in traditional regression approaches, autoregressive approaches seek to use information from previous observations to pre-

dict changes in the variable of interest (Wooldridge, 2016). This is especially useful for time series datasets where observations are often not independent and identically distributed. Autoregressive approaches deal with this problem by using some number of previous observations of the variable of interest as explanatory variables.

Vector autoregression, commonly annotated as VAR(p), is one of several time series forecasting methods that use previous observations with p lags to predict future evolution of interconnected variables (Sims, 1980). Vector autoregression with exogenous variables (VARX) extends this basic approach to allow the modeling of systems in which some of the explanatory variables are not affected by the primary variable of interest. These approaches can struggle with large models due to the number of parametric terms required, which creates problems with the number of degrees of freedom (Bernanke et al., 2005). Another limitation of vector autoregression is the need to only train with data that includes a historical record of p lags. This limitation becomes problematic when a relatively small dataset includes censored data, which then effectively censors all later data for p time steps in the future. Certain time periods of USAF retention data are unrepresentative due to large force management actions that perturb natural retention behavior in ways that are not easily modeled. One example of this phenomenon is the approval of large numbers of early retirements as well as reduction in force and force management boards in 2014 to comply with the fiscal constraints of the US government's sequestration policy.

Given the likelihood of some seasonality over the course of the year for personnel retention, 12 or 24 lags are likely to be the minimum number required to create an effective model using monthly data. Especially when considering 24 or more lags, this additional censoring can result in a potentially dramatic reduction of available training data for datasets with a limited history like the USAF PRP. One additional problem for autoregressive datasets is the limitation on the quantity of variables used.

With any large number of explanatory variables, the low frequency of observations with high numbers of years of service means that observations can be quite sparse for specific combinations of features. Many combinations of features may only have a few observations over time, so any methods that rely on a certain number of lags must severely restrict the number of variables to ensure a number of observations greater than 0 for all combinations of features. Other time series-oriented competitors to the VAR approach such as the autoregressive integrated moving average (ARIMA) and vector autoregressive moving average (VARMA) models share these censoring and sparsity-related weaknesses.

Multilayer Perceptrons (Artificial Neural Networks)

MLPs capitalize on many of the strengths of the other statistical machine learning methods previously described (McCulloch and Pitts, 1943; Chollet, 2021). With appropriate network size and hyperparameter selection, MLPs are flexible enough to map highly nonlinear functions. This flexibility provides the model capacity to handle complex problems whereas approaches with lower capacity struggle. Because MLPs are a parametric approach that constructs weights based on observed features, they do not require exact observations of every combination of features like KM. Although methods exist to examine the effect of different features on the final prediction, the large number of trainable parameters in most MLPs makes using them for inference and model understanding difficult. Nevertheless, MLPs are often able to perform better than other machine learning approaches precisely because of that level of capacity. Indeed, Hornik et al. (1989) proved that a single layer perceptron of sufficient size can approximate any continuous function for any arbitrary level of accuracy, leading to the title of "the universal function approximator."

Multilayer perceptrons can be used for survival analysis problems with the inclu-

sion of a sigmoid activation function on the output layer, which forces outputs to range from 0 to 1. Any other activation function that forces outputs from 0 to 1 can also be used for this problem structure, although the sigmoid activation function provides the benefit of being continuously differentiable.

As implied by the name, nonlinear autoregressive neural networks (NARNets) use an autoregressive structure similar to that of the VAR(p) model, but they instead use an MLP structure to determine appropriate parameters (Chakraborty et al., 1992). This approach has been shown to function well for many time series problems and continues to evolve (Triebe et al., 2019). However, NARNets share some of the same problems as the VAR(p) model described previously due to their reliance on lagged observations. When certain years of data are censored, the retention observations for these approaches require additional censoring to avoid contaminating training data with the censored retention behavior in the lagged observations. These approaches also generally require each retention observation to have a past observation with the appropriate number of lags. When modeling subgroups within a population, this structure requires the use of a limited number of variables to avoid problems with sparsely populated subgroups with inconsistent observations and partially diminishes the benefit of high-capacity approaches like MLPs.

Taskaya-Temizel and Casey (2005) provide a detailed comparison of autoregression-neural network hybrids; many of these approaches seek to use the nonlinear strengths of the neural network structure to fit the residuals of a classical autoregression approach. These approaches still retain the censoring and sparsity problems discussed here for other classical autoregression techniques.

2.2.3 Partially Autoregressive Feature Selection

In addition to features directly recorded in MilPDS, the PARNet and PARFor approaches include an autoregressive feature. Simply including the previous observation for the subpopulation with the identical feature set is not possible if no cohort with that feature set existed at the previous time step. This is a significant problem at this level of detail but becomes more problematic with every additional variable as the subpopulation sizes steadily decrease. At the extreme, with enough variables, all cohorts have a size of 1. For this reason, we select a larger cohort for which the subpopulation shares some features, but is also defined by few enough variables to reliably generate an observation at each time step. Because the current KM model has historically worked well under most conditions, we constructed the autoregressive feature to use the same combination of YOS and months until separation date. Hence, all subpopulations with a given combination of YOS and months until separation date include the single month retention observation for this larger cohort at the previous time step.

A common alternative approach is to use econometric data to develop a predictor of changing retention behavior. Including a partially autoregressive feature has two notable advantages over this approach, although they may be used in tandem. First, an autoregressive approach may prove more robust when there are multiple trends occuring at the same time, because it can capture the net effect of different variables even when the data does not provide a way to measure which variable is causing the aggregate trend. This specifically helps in the case identified here, wherein many USAF policies have changed over time, frequently without a consistent documentation captured in a single quantitative dataset.

2.2.4 Data Partitioning: Validation and Test Approach

USAF personnel policy has changed dramatically over time; events such as the transition to an all-volunteer force, the fall of the Soviet Union, the attacks of September 11, 2001, and the US government's policy of sequestration starting in 2014 all drove both immediate and lasting changes in USAF policy and personnel retention. Additionally, changes in compensation, mission, and culture occur slowly over time and create very different retention choices at various stages of the USAF's history.

The MilPDS personnel retention data extracts extend from September 1992 through September 2021. In advance of any other data processing, we partition a test dataset that uses only the observations from September 2020 to generate an estimate of model performance. Notably, the test dataset occurs during a notably unstable period in labor economic conditions. The last portion of 2020 fiscal year yielded an upward shift in retention due to the rapid change in economic circumstances related to the COVID-19 pandemic. The conditions of the economy from a labor perspective began to accelerate through 2021 as worker shortages yielded an upward pressure on wages and unemployment fell. Although this 2021 data may pose a substantial challenge to any set of retention models, this test dataset should allow us to determine whether the new models are able to robustly use partially autoregressive features and personnel data to successfully identify retention changes compared to previous retention modeling methodologies.

Next, we seek to ensure models can generalize well enough to predict behavior of the current force, despite changes in enlistment contracts, pay, benefits, and many other policies over time. We must either include variables that capture these changes or appropriately censor the data to avoid confusing the varying relationships of explanatory variables from periods with different policies. Aggregate retention, measured as the proportion of personnel in the USAF at the beginning of the fiscal

year who remain on active duty in the USAF 12 months later, shows two different sets of retention behavior during this period as shown in Figure 12. Notably, years with nonrepresentative policies to involuntarily separate personnel (red squares) and involuntarily retain personnel (green square) significantly alter natural retention behavior and must be censored without some other method to account for this nonrepresentative pattern of behavior. As one senior personnel analyst quipped: "Nothing boosts retention like making it illegal to get out of the military" (Barger, 2017).

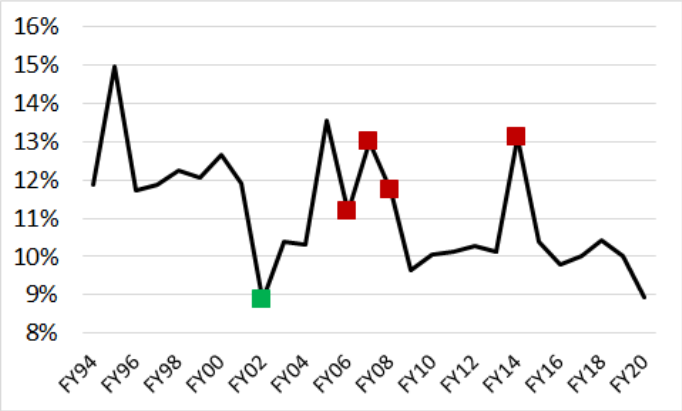

Figure 12. High loss rates in the 1990s followed by lower loss rates after financial crisis

It is unclear whether the change in loss rates from the 1990s to the recent years is due to economic changes between the roaring economy of the dot com boom and the tepid recovery after the 2007 financial crisis, changes in mission or compensation, cultural changes, or the changing messaging to airmen as the US government drew down USAF end strength dramatically throughout this period. Regardless, the periods are sufficiently different to merit treating post-2007 retention separately. We censor 2006-2008 and 2014 due to the forced losses in those years. We also censor 2009 and 2015 data to prevent contamination due to the inclusion of lagged retention observations in the partially autoregressive feature. After this censoring, we retain 2010-2013 and 2016-2020 retention observations to meet our training and validation needs.

One danger of using monthly data for predictions spanning multiple months is that longer prediction lengths sample the same personnel retaining in a single month for multiple observations. An 11-month retention observation from September 2016 and another 11-month retention observation from October 2016 are observing most of the same retention behavior. For example, a single airman in the force who retains from September 2016 to September 2017 would be represented as 2 separate observations of 11-month predictions for a single 12-month period. To avoid problems with repeated evaluation of the same portions of time, we use only the September observations from each fiscal year, providing non-overlapping observations for training and validation and also preventing any overlap with the test dataset.

We set aside two distinct, non-overlapping validation sets. The second validation set is a traditional validation dataset consisting of the monthly observations of starting inventory for September 2019 (i.e., 2020 data); this dataset is used to select the superlative model after all models across all hyperparameter combinations are fully trained. Although this data does contain a retention shift from the COVID-19 pandemic, this change in environment only manifested in the last 6 months of the fiscal year and retention effects experience some lag as reenlistment contracts are often signed in advance of the actual departure of an airman. This approach ensures that we avoid favoring large capacity models that simply overfit the training data yet fail to generalize to unseen data. Simultaneously, selecting models that perform well on both the dataset used during training and the second validation dataset should also help avoid underfitting. Residual analysis of predictions on the second validation dataset will help confirm if these models are appropriate.

Several of the models proposed in this research work best by using some validation data or process as part of their internal training process to build good models. For example, this implementation of the MLP uses a validation set and a version of early

stopping. During training, this approach selects the model weights that generalize best prior to further learning leading to overfitting on the training data. However, no data used in the process to train the models can help inform which models should perform the best because the models are already being fit to that data. For this reason, we partition both a test dataset and two separate validation datasets. After data cleaning and transformation, the training set and first validation set are split pseudo-randomly from the observations of starting inventory through September 2018. Of this data, 80% are used for training and 20% for initial validation. This initial validation set is either used to help ensure training generalizes for MLP and PARNet or folded back into the training data the RFR and PARFor models. Valid concerns exist with regard to using randomly split validation data for time series estimation because the observations in the first validation set will temporally overlap some of the observations in the training data, resulting in some level of cross-contamination. However, appropriate time series approaches would preserve the most recent data, which is also likely to be the most valuable data, entirely for validation. Because this first validation dataset is being used to help the models better generalize but is not being used to select the best model, the cross-contamination is an acceptable tradeoff to ensure we maintain enough data for effective training. The final size of each dataset is shown in Table 2.

Dataset	Individual Observations
Training and Validation 1	2,069,339
Validation 2	263,976
Test	265,369

Table 2. Number of final observations in each dataset given selected features

2.2.5 Military Personnel Data and Generation of Retention Observations

The MilPDS extracts used for this research do not contain retention observations themselves, but instead record longitudinally which personnel are on active duty each month and their respective personnel details. Because social security number is one of the recorded variables, we can examine when specific individuals enter, retain, and depart active duty and what variables are recorded at each of these stages. After censoring non-representative retention years, the training, validation, and test datasets contains 2,598,684 individual data points from 10 years of retention data, each containing 12 retention observations corresponding to the prediction lengths of interest. This research attempts to predict total numbers of personnel who will retain over different time intervals, so individual classification of retention behavior is unnecessary for high quality predictions. Instead, we group individuals with identical sets of features and attempt to predict how many of the group will retain at each time interval. Because retention observations are based on the rate of subgroups with a given feature set retaining, the number of observations is reduced as individual retention observations are translated to the retention observations of subgroups with the same combination of features. As an example, consider a cohort of 10 individuals with the same feature set at a given time. If 9 retain and 1 departs for the prediction interval of interest, these 10 individual observations become a single observation with a retention rate of 0.9.

Prior to grouping data points to create retention observations, this dataset requires several data preprocessing steps.

1. Blank entries are grouped by a common flag for each variable. In some cases, this represents a true similarity, such as a missing separation date suggesting that a person has not submitted separation paperwork. In other cases, this

grouping may represent a common error, such as pay date errors for reservists participating in a Voluntary Limited Period on Active Duty tour.

2. Less common observations for categorical variables are grouped together as well. Some variables have many ways to categorize a field, but the numbers may be sparse and have few personnel by which to judge a likely retention rate, especially when spread across the other variables. To reduce this sparsity, we used a minimum cutoff of 1% of the monthly observations.

3. Each categorical string variable was converted to multiple dummy variables representing each possible value (i.e., one-hot encoding was implemented).

4. All ordinal and interval data was normalized (scaled to range from 0 to 1) without standardization. While standardization is a common approach for data with significant outliers, each of these distributions only included integer values ranging from 0 to 14 or 0 to 30, meaning that a linear scaling would best allow a machine learning approach to differentiate the effects of observed values.

5. All dates in the future are translated into an integer measurement of the months until such an event happens. Since this set of models will only predict retention over a relatively short prediction interval (12 months), the maximum value for such observations is set to 14.

The features of interest (i.e., military personnel variables) are shown in Table 3.

These variable are selected to maximize known explanatory relationships. These include the following categorical and boolean variables.

- **AFSC (Air Force Specialty Code: Career Field)**

  Different specialties have different cultures, expectations, and economic opportunities outside the service, which drives different retention patterns.

Variable	Type	Processed Features
YOS (Years of Service)	Interval	1
Gender	Categorical	1
Race	Categorical	6
AFSC (Career Field)	Categorical	24
Grade/Rank	Categorical	10
Reenlistment Eligibility	Boolean	1
Separation Paperwork Filed	Categorical	2
Months Until Separation Date	Interval	1
Months Until ETS	Interval	1
Months Until HYT Cutoff	Interval	1

Table 3. Military personnel variables

- **Grade/Rank**

  Grade provides some measure of performance and compensation, which affects outside earning potential.

- **Gender**

  Female airmen leave at higher rates than their male peers early in their career, making gender an important consideration for the probability of retention.

- **Race**

  Although not as clear of a relationship as gender, race still has substantial predictive power for retention observations.

- **Reenlistment Eligibility**

  Some airmen may not be eligible to reenlist for a number of reasons. Even in absence of separation paperwork being filed, an impending expiration of term of service (ETS, i.e., the end of the enlistment contract) without reenlistment eligibility increases the probability of separation.

- **Separation Paperwork Filed**

  If they have filed their separation paperwork and have a separation status, they have signaled their intention to leave.

In addition, we include the following interval variables.

- **Years of Service**

  YOS correlates closely with significant career milestones and retention decisions; airmen reach the end of their first enlistment at four to six years and reach retirement eligibility at 20 years. Some airmen are not eligible to reenlist in their current career field; upcoming retention decisions may require volunteering to cross-train to another career field to remain in the USAF. This variable is recorded as the integer value of completed years of service, ranging from 0 to 30 for enlisted airmen.

- **Months Until Separation Date**

  Once paperwork has been filed indicating that a service member intends to depart, the date of the intended departure is also recorded. The difference between the current date and this variable is recorded in integer months, with a maximum value of 14.

- **Months Until ETS**

  An ETS occurs when the current enlistment runs out, driving a stay or go decision. If the airman reenlists, the ETS is extended into the future. If an ETS is very close but no separation paperwork has been filed, then the airman has probably not made a decision, although some airmen may wait to see if they become eligible for bonuses or similar policies. The difference between the current date and the ETS is recorded in integer months, with a maximum value of 14.

- **Months until HYT Cutoff**

  High Year of Tenure (HYT) cutoffs indicate the maximum YOS an airman of a specified grade can have before being forced to exit the service. This ceiling is a

mechanism to prevent personnel from remaining in the force if not continuing to progress in rank. The difference between the current date and the HYT cutoff is recorded in integer months, with a maximum value of 14.

After encoding each of the variables according to these rules, the vector representing a combination of these features is defined as $f \in \mathcal{F}$ where $\mathcal{F}$ is the set of all possible feature vectors.

As individual observations aggregate to form retention observations of groups, the number of observations declines proportional to the size of those groups. With this selection of features and the data cleaning methods employed, the original 2,598,694 individual data points transforms to 405,137 subgroup data points. The final size of each dataset is shown in Table 4.

Dataset	Subgroup Observations
Training	255,868
Validation 1	63,968
Validation 2	42,729
Test	42,572

Table 4. Number of final observations in each dataset given selected features

After transforming categorical variables with one-hot encoding, 48 total binary, ordinal, and interval features are produced for each observation, plus a single partially autoregressive feature. Feature vectors are annotated as f and the set of all possible feature vectors is $\mathcal{F}$. Specific time periods are annotated as t and the set of all time periods is $\mathcal{T}$. We define starting and surviving inventories of personnel as

$$S_{f,t} = \text{number of personnel with feature vector } f \text{ at time } t, \text{ and} \tag{1}$$

$$S_{f,t,\tau} = \text{number of remaining personnel starting with feature vector } f \text{ at time } t \text{ after } \tau \text{ time steps.} \tag{2}$$

Our retention estimation approaches capitalize on one aspect of the problem structure: with certainty, the rate of surviving personnel over τ time steps from time t with a given feature set will fall between 0 and 1. A retention rate is then defined as

$$\begin{aligned} r_{f,t,\tau} = & \frac{S_{f,t,\tau}}{S_{f,t}}, \text{ where } 0 \leq r_{f,t,\tau} \leq 1, \\ & \forall f \in \mathcal{F}, t \in \mathcal{T}, \tau \in \{1, 2, ..., 12\}. \end{aligned} \tag{3}$$

Predicted retention rates are denoted as $\hat{r}_{f,t,\tau}$.

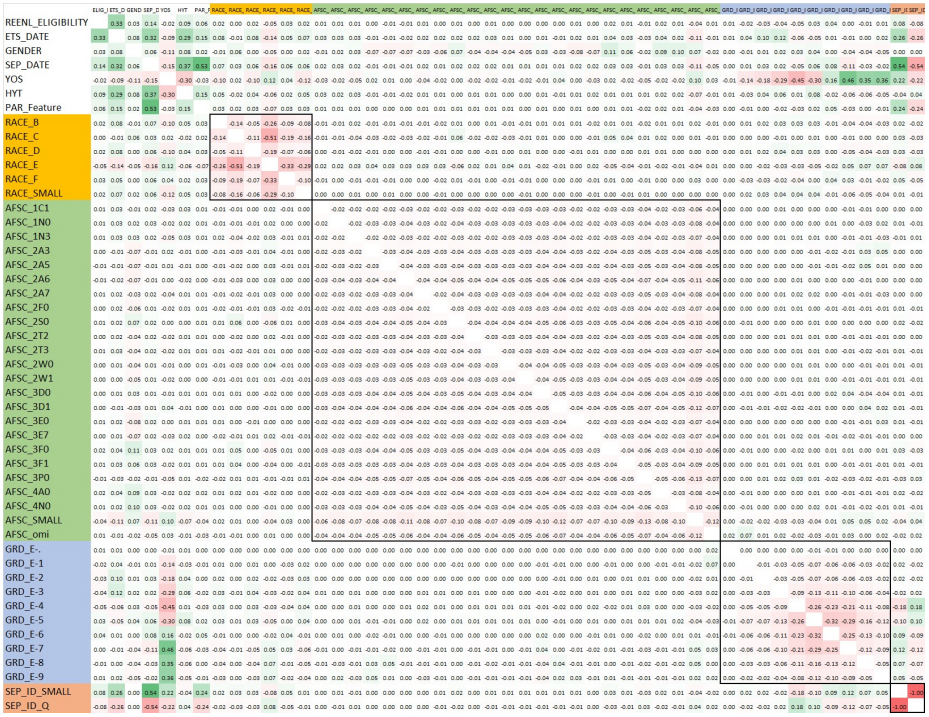

Figure 13. Correlation between transformed input variables for training dataset ranging from -1 to 0.54

We tested the correlation of the input variables in the training data, shown in Figure 13. One-hot encoded transformations of the same variable (i.e., the dummy variables) showed small negative correlations with each other, shown in the boxed areas of the figure. Additionally, YOS showed a correlation with several variables, including grade and separation ID. Grades generally progress over time, as does YOS,

so this matches our expectations. Enlistment contracts define a required term of service and HYT policies force personnel out at higher YOS if they have not progressed quickly enough to a higher grade, so some relationship is expected with these variables as well. Separation data and separation ID also predictably showed a strong relationship with each other and the partially autoregressive feature. The generalizability of models can suffer when the effects of different variables cannot be distinguished from each other due to multicollinearity. To observe whether this is a problem, we compute the variance inflation factors (VIFs) for each non-categorical variable, shown in Table 5. VIFs measure how each variable affects multicollinearity and are commonly used to diagnose problems for ordinary least squares regression problems. Most of the VIFs are in desirable ranges, although YOS is somewhat high at 6.7. Because it remains below the threshold of 10 recommended by Menard (2001), we proceed with these variables.

Variable	VIF
Reenlistment Eligibility	1.15
Months Until ETS	1.41
Gender	1.12
Months Until Separation Date	2.59
YOS (Years of Service)	6.71
Months Until HYT Cutoff	1.59
Autoregressive Variable	1.42

Table 5. Variance inflation factors for each non-categorical variable

Traditionally, classification problems need to utilize balanced datasets to create high-quality machine learning models. Regression problems have not faced the same issues, but our problem constrains predictions to a small range from 0 to 1 in the same way as historical classification problems. Even with the relatively parsimonious feature selection, we still observe many observations at the extreme values of 0 and 1. In such a case, we know that residuals will not be normally distributed, so highly imbalanced datasets may result in low-quality machine learning models. We tested

the transformed subgroup observations, and retention observations of 100% composed less than 64% of the totals, despite many of the aggregate retention rates being greater than 90%. For this reason, we did not explore resampling or selectively sampling our training data. The distribution of retention observations is shown in Figure 14.

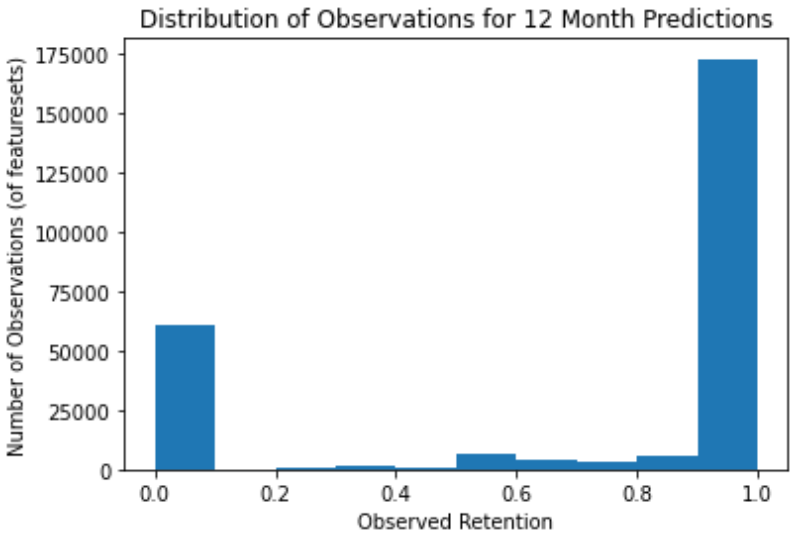

Figure 14. Histogram of 12-Month prediction interval retention observations in training dataset

Aggregate predictions of total retained personnel are constructed by summing the product of the predicted retention rates for a given combination of features with the number of starting personnel for that same combination of features, across all possible combinations of features:

$$\hat{\rho}_{t,\tau} = \sum_{f \in \mathcal{F}} \hat{r}_{f,t,\tau} S_{f,t} \quad \forall t \in \mathcal{T}, \tau \in \{1, 2, ..., 12\}, \tag{4}$$

while the observed aggregate retained personnel is denoted as $\rho_{t,\tau}$. Finally, we define the absolute aggregate prediction error for a given time t and prediction length τ as

$$E_{t,\tau} = |\hat{\rho}_{t,\tau} - \rho_{t,\tau}|. \tag{5}$$

This measure provides the means to assess the goodness of any model predictions because it directly measures a model's usefulness to the research sponsor. Because $\tau \in \{1, ..., 12\}$ the size of the output layer for our MLP approaches is 12. For random forest approaches, we generate 12 separate models to produce the 12 outputs.

The initial retention code was developed in SAS, the native language of the datasets and their Air Force Personnel Center caretakers. Datasets were then imported to Python for further cleaning and subsequent analyses.

2.2.6 Hyperparameter Selection for Computational Experiments

In RFR models and various forms of neural networks for problems of our size, a significant driver of model quality is the selection of hyperparameters that tune how the model is structured and optimized. For our approaches using a feedforward neural network, we consider the following hyperparameters.

Given the importance of both generalizability and computational demand, we select a large batch size of 8,192 and seek to use a high learning rate. We implement the *1cycle* approach to scheduling learning rate (Smith, 2018) with the maximum learning rate set according to the test recommended by Géron (2019) over five epochs. This test begins training over some small number of epochs, steadily increasing the learning rate at each iteration to observe how high the learning rate can rise before training diverges and the loss begins rising dramatically. In order to automate this test, we find the minimum loss value during this training and set the maximum loss rate to 90% of the value of the corresponding learning rate. The minimum loss rate is then set to 10% of this value. Smith (2018) recommends using stochastic gradient descent with weight decay as the optimizer and using a weight decay value that allows the highest learning rate. We test three recommended values, 0, 0.001, and 0.01, to explore if any consistent relationship exists for this problem structure. We also

test a version of each architecture with and without momentum. Architectures with momentum use the momentum scheduling approach described by Smith (2018). All models use binary cross entropy for the loss function because the predicted outcomes are probabilities ranging from 0 to 1.

Another important architecture design issue is determining the superlative combination of activation function and regularization approach. The regularization approach is particularly important for time series problems as we attempt to find models that generalize well to future observations and avoid overfitting noise in the training data. We examine two activation functions: exponential linear units (ELU) (Clevert et al., 2015) and scaled exponential linear units (SELU) (Klambauer et al., 2017). The use of *1cycle* is a form of regularization, so we test both of these approaches without additional regularization as well as with appropriate techniques for both. For ELU activation functions, we consider batch normalization (Ioffe and Szegedy, 2015) and Monte Carlo Dropout (Gal and Ghahramani, 2016). For models with Monte Carlo dropout, we test 2 configurations: 1 with dropout after each hidden layer and 1 with dropout after only the final hidden layer. For SELU activation functions, we test *AlphaDropout*, the modification to traditional dropout proposed by Klambauer et al. (2017). This approach maintains the mean and variance of the outputs for each hidden layer, preserving desirable properties of the SELU activation function.

Finally, we test architectures with hidden layers ranging from two to five and between 25 and 100 neurons per hidden layer, with discrete settings of 25, 50, and 100. We conduct a full factorial experiment with 10 replications of each of these sets of features, indicated in Table 6. While these additional replications could be used to sample a wider collection of hyperparameter settings, replications enable observation of whether differences in performance are due to the hyperparameter settings or random noise generated by the algorithm's stochastic starting conditions.

If the addition of a partially autoregressive feature is a valuable addition to the MLP modeling approach for this problem, we should observe the superlative models consistently using this feature.

For all MLP models, we use the Tensorflow API (Abadi et al., 2015) for development, training, and testing.

Hyperparameter	Settings
Activation Function and Regularization	ELU, no regularization ELU, batch normalization ELU, MC dropout on each hidden layer ELU, MC dropout on final hidden layer SELU, no regularization SELU, *AlphaDropout*
Weight Decay	0, 0.001, 0.01
Momentum	Scheduled, No momentum
Hidden Layers	2, 3, 4, 5
Neurons/Hidden Layer	25, 50, 100

Table 6. Hyperparameters for MLP and PARNet models

We next consider the hyperparameters for our random forest models. Although random forest models are less sensitive to hyperparameter selection due both to their structure and their nature as an ensemble learner, correct selection of hyperparameter settings can still have a significant effect on superlative performance. We consider three hyperparameters of interest: number of decision trees, maximum tree depth, and the use of bootstrapping, as shown in Table 7. Like the MLP approach, we replicate each group of hyperparameter settings 10 times using different seeds to ensure reproducability. Based on preliminary empirical testing, the minimum observations to split a node was set to five and the minimum observations per node was set to two. The models were trained using the Random Forest Regression module from *Scikit-Learn* (Pedregosa et al., 2011); all other hyperparameters used the module's default settings.

Hyperparameter	Settings
Number of Trees	100, 250, 500
Maximum Depth	10, 25, 50
Bootstrapping	Yes, No

Table 7. Hyperparameters for RFR and PARFor Models

2.3 Results and Discussion

We seek a modeling approach that improves prediction performance versus the benchmark. Approaches that produce robust results with different randomized starting conditions are preferable to those requiring multiple restarts to find a high performing model, but the primary success criterion is performance as measured by mean absolute aggregate error.

2.3.1 Validation Results for MLP and PARNet Models

Once all training completes, each model generates a set of predictions for the second validation dataset, predicting the probability of retention for the population with each set of features for the next 1-12 months. As seen in Figure 15, both the random forest approach and the MLP approach generated models that outperformed the benchmark, shown in the green shaded portion of the chart. Moreover, both approaches showed improved performance for the highest performing models when including the partially autoregressive feature.

As seen in Figure 16, the quality of the MLP predictions varied significantly across all architectures, though only some architectures generated high quality predictions. Three of the architectures failed to produce any models that could defeat the benchmark, including both approaches without an additional form of regularization. A fourth architecture using ELU activation functions and Monte Carlo Dropout on all layers generated only a single model that defeated the benchmark, which appeared to be an outlier.

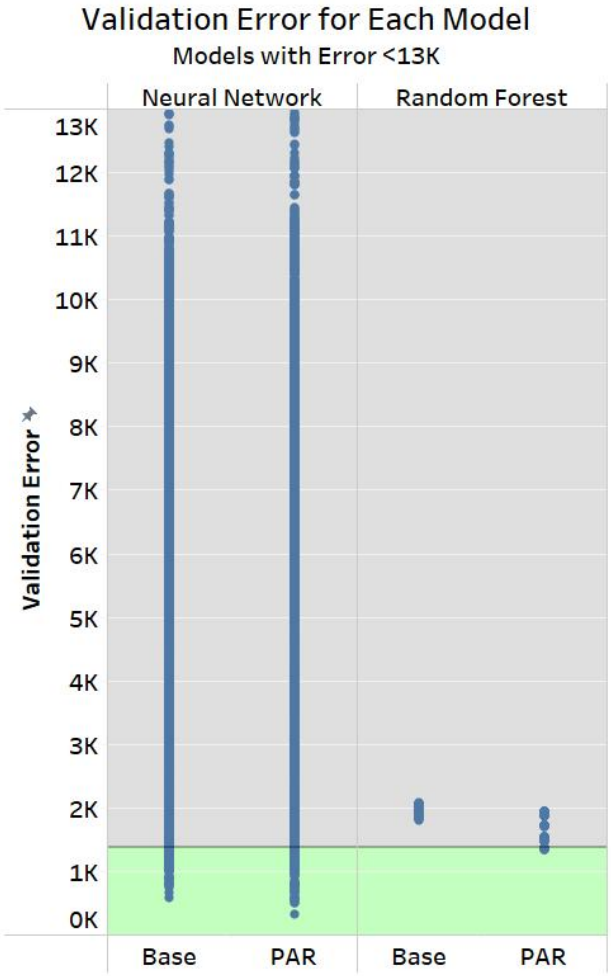

Figure 15. Overall performance varies, but multiple approaches produce models that outperform the benchmark of 1,383.3 (shown in green)

Examining the top performing models more closely in Figure 17, the architecture using ELU activation functions and batch normalization generates only a few models that defeat the benchmark, contrasting with the large number of better performing models generated using SELU activation functions and *AlphaDropout*. Moreover, for this architecture, the inclusion of the partially autoregressive feature appears to improve prediction performance for the best-performing models.

Figure 18 shows the performance results of the best models using the SELU activation function, *AlphaDropout* regularization method, and the partially autoregressive

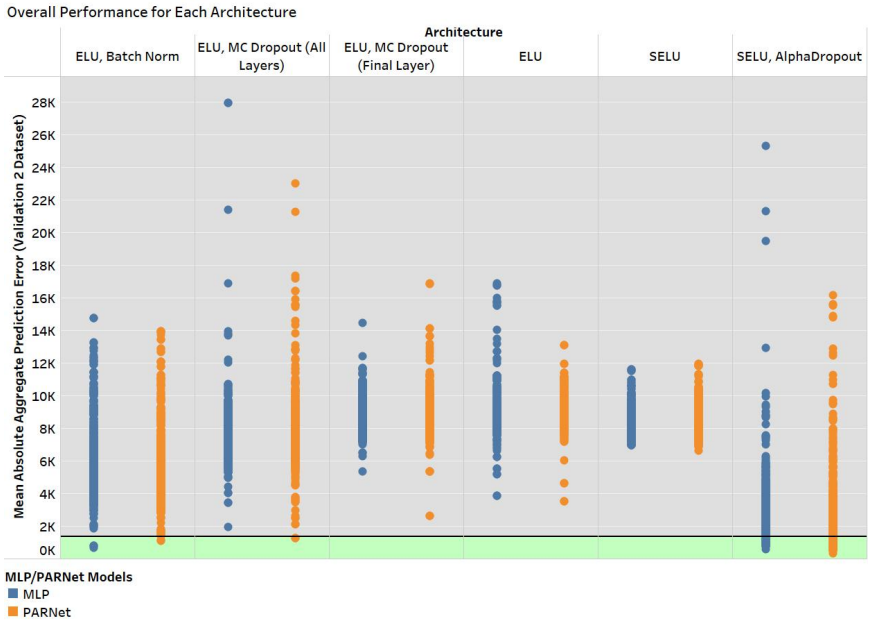

Figure 16. While each architecture had a wide range for quality of predictions, only 3 produced models that outperformed the benchmark

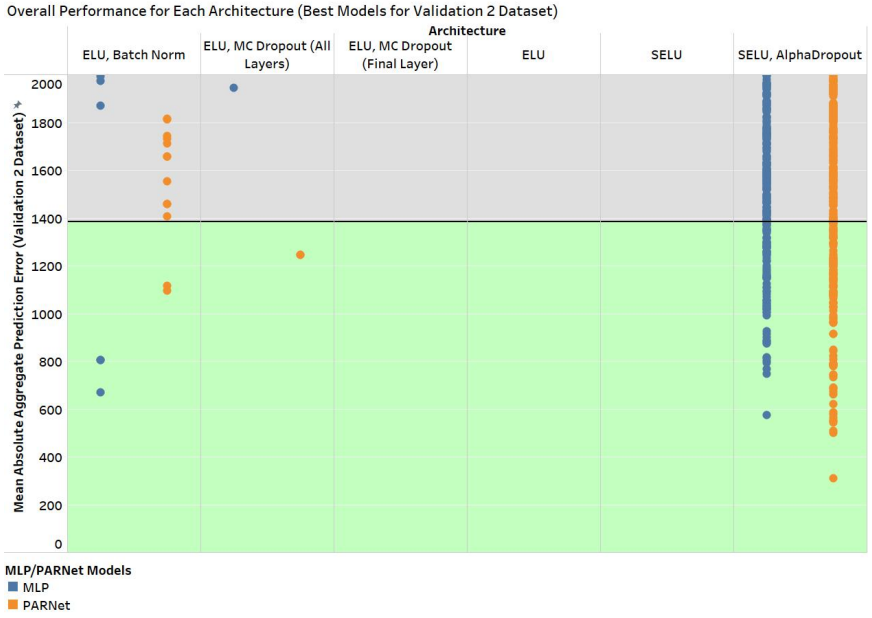

Figure 17. SELU with *AlphaDropout* and the partially autoregressive feature produces the best-performing models

feature. None of the other hyperparameters show a consistent relationship with solution quality, although many of the highest performing models used 25 neurons per hidden layer, suggesting that this smaller size may improve generalization for su-

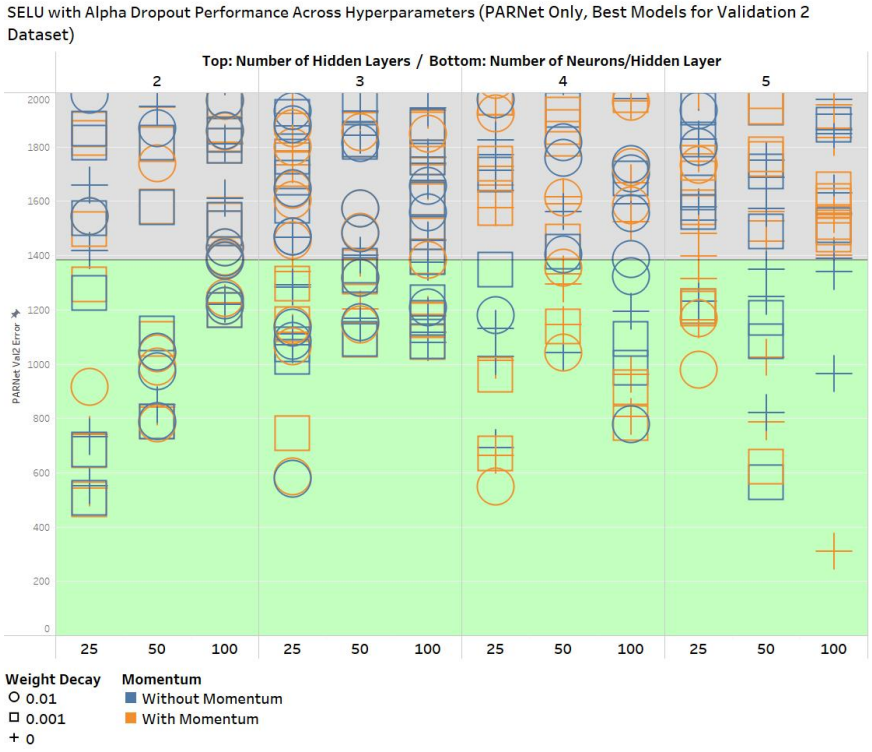

Figure 18. While the best model uses the largest architecture, many of the best models used the smallest number of neurons per hidden layer tested

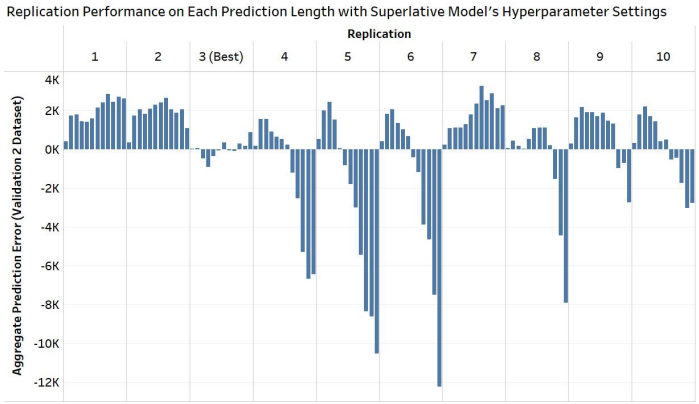

Figure 19. Best combination of hyperparameters showed inconsistent performance, suggesting that the difference in solution quality depends on pseudo-random initialization values

perlative models trained on this problem. Additionally, models with 3 hidden layers generated the largest number of models that outperformed the benchmark for each number of neurons per hidden layer.

Because each set of hyperparameters is used to generate 10 models, we seek to

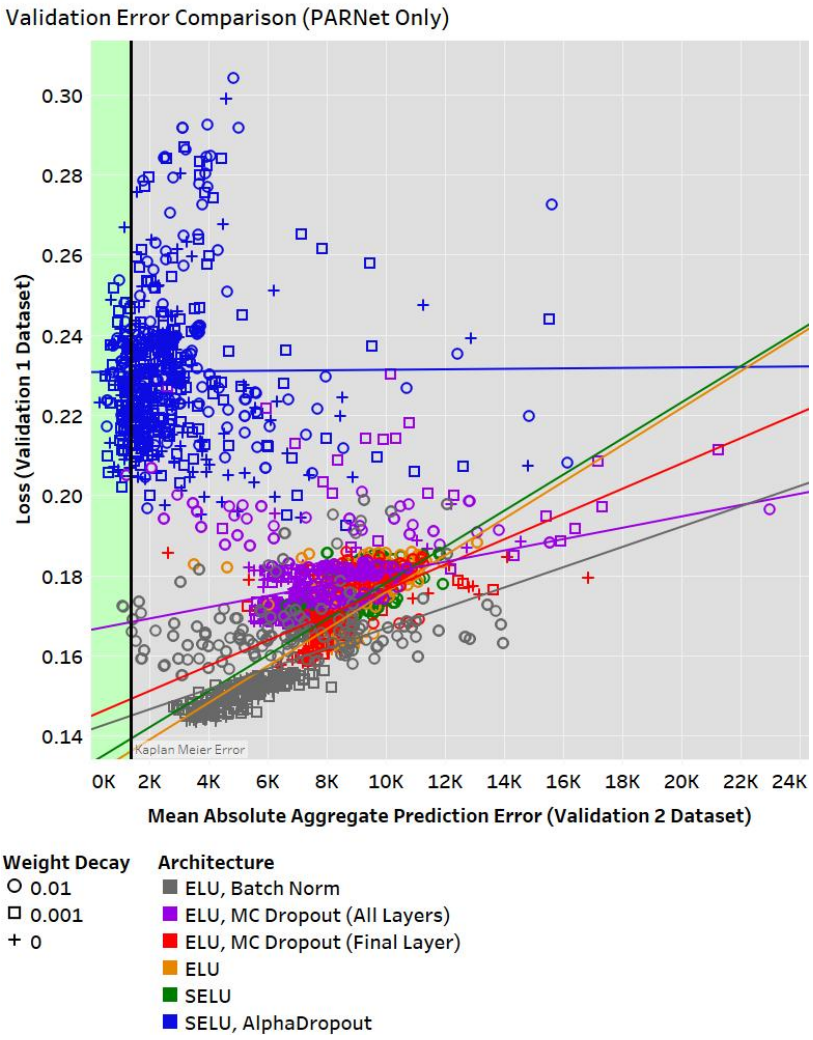

Figure 20. Top performing architecture for Validation 2 dataset shows minimal relationship between validation loss during training with Validation 2 performance

check whether this particular set of hyperparameters that generated the top model is consistently high-performing or whether such an approach requires a significant number of models trained with different starting weights to find a high-quality model. In Figure 19, we see that the superlative model's hyperparameter settings do not produce consistently high quality predictions. This suggests that the activation function and regularization method are very important, but that considerations such as computational demand can drive other hyperparameter settings without substantial worsening of prediction quality as long as a sufficient number of models are trained

to find a high-performing model. We focus on the best model generated, but future investigation of the USAF PRP should focus on the smaller model architectures for increased computational efficiency with minimal loss of solution quality.

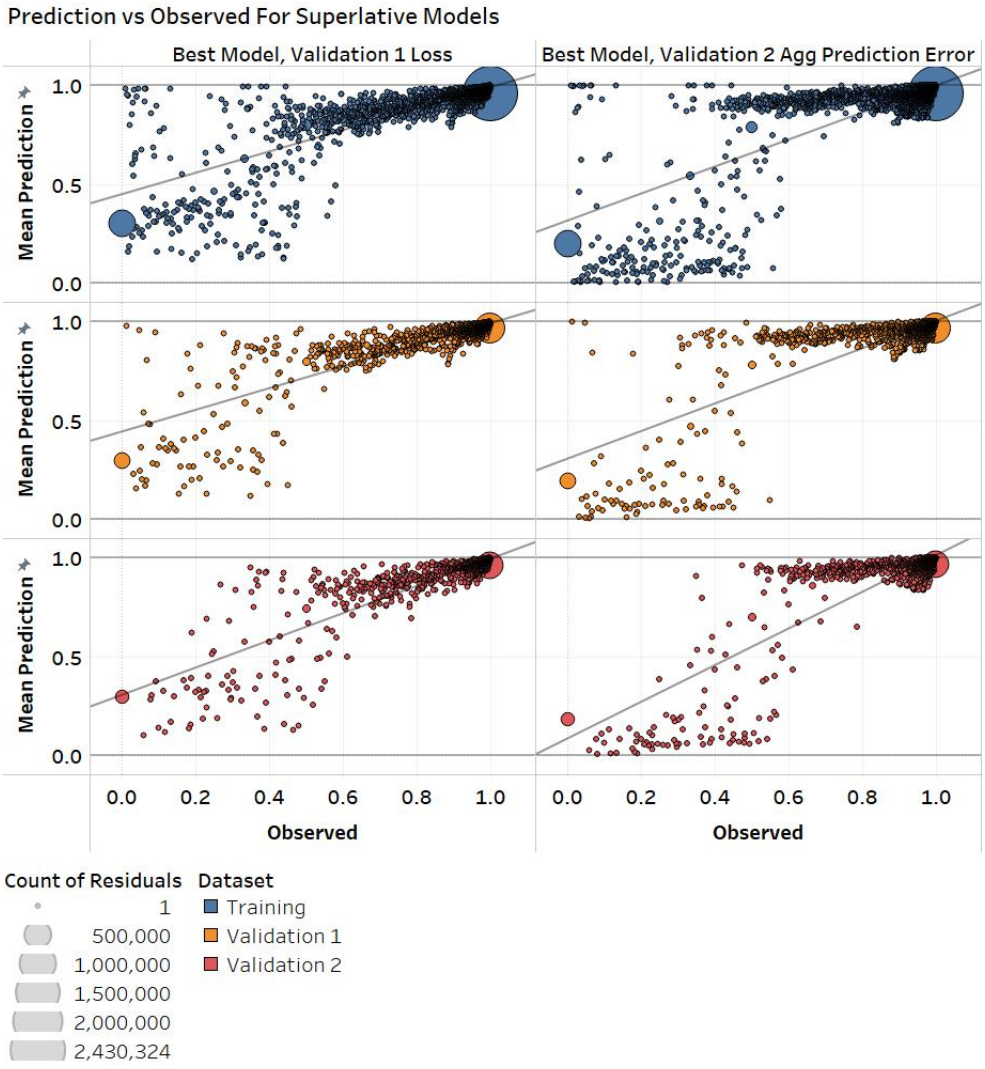

Figure 21. Best model for aggregate error in Validation 2 dataset demonstrates increased individual errors but reduced aggregate statistical bias

When generating these models, the use of separate validation datasets for training and for model selection helps to examine of how the prediction error on the first validation dataset used during training correlates with the aggregate prediction error for the second validation dataset. Generally, one would expect these errors to closely

correlate if the first set generalizes well to the second. In Figure 20, this proves true for all but one combination of activation functions and regularization method. In the case of the top performing architecture, we observe generally higher errors for the first validation set as well as a statistically insignificant relationship between the two errors.

We observe evidence for the first set generalizing well to the second in the results from the other architectures. However, consistent errors in a single direction, also known as statistical bias, could cause error in the first validation set to be low but aggregate error in the second validation set to be high if the small errors were consistently in the same direction. This is a significant concern for prediction of rates, because prediction errors cannot be symmetric for rates close to 0 or 1, making statistical bias particularly sensitive to the distribution of observations. Indeed, we observe this exact phenomenon in Figure 21, which compares the residuals for the highest performing model as measured by the loss for the first validation dataset to the residuals for the highest performing model as measured by the aggregate prediction error for the second validation dataset.

Next, we examine the performance of the approaches using a random forest structure on the second validation dataset. As we see in Figure 22, the approach is also able to outperform the benchmark and demonstrates more consistent performance across replications but fails to match the performance of the approaches using a neural network architecture. We observe that the inclusion of the partially autoregressive feature consistently demonstrates superior performance across all hyperparameter settings, as seen in Figure 23, wherein the validation error for each replication appears above the diagonal line representing parity between the two approaches.

Although the primary selection criterion is model quality, computational effort remains an important consideration. As seen in Figure 24, the MLP models require

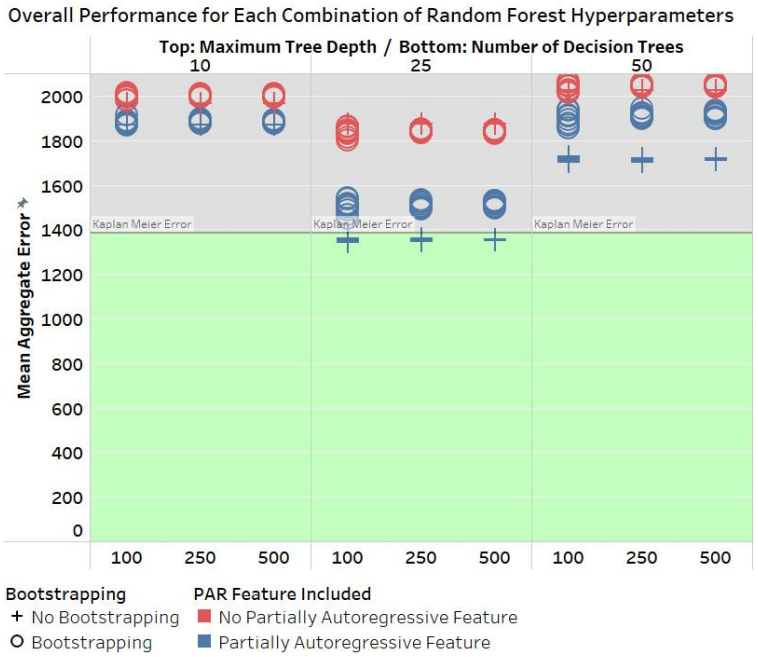

Figure 22. Superlative random forest architectures consistently outperform the benchmark but fail to match highest performing MLP models

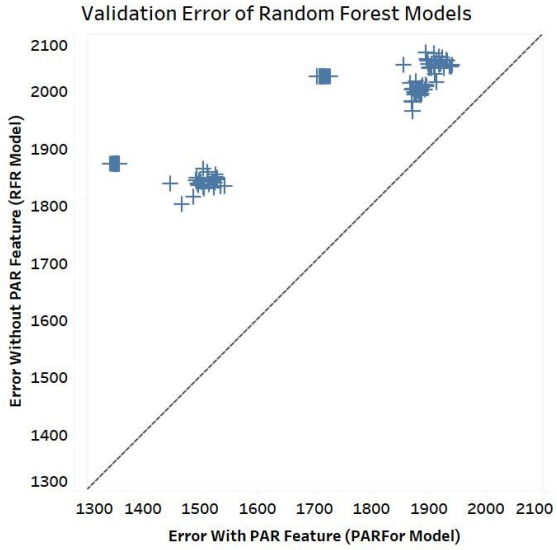

Figure 23. Random forest models with the partially autoregressive feature performed better (i.e., attained decreased validation error) than those without the feature across all replications.

relatively less training time, attaining times ranging from 9 to 31 seconds. However, a practitioner must train many models to generate a high quality prediction. Con-

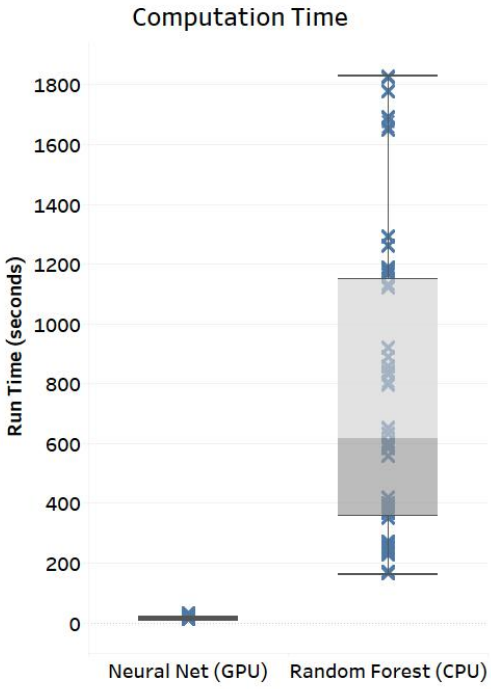

Figure 24. With the tested hyperparameters, training individual MLP models require less computation time (9-31 seconds) than the RFR models (162-1,829 seconds).

versely, the RFR models require a relatively greater amount of time to train, attaining times ranging from 162 to 1,829 seconds with the middle two quartiles ranging from 358 to 1,148 seconds, but consistently converge to models of similar quality given the same hyperparameters. The MLP models were trained on an NVIDIA Quadro RTX 8000 GPU while the RFR models were trained in parallel on an Intel Xeon CPU E5-2680 v3 at 2.50 GHz with 24 cores. Because RFR training ran on the CPU and MLP training ran on the GPU, a precise comparison regarding computational effort should be avoided.

With these results, we select the superlative model to be the highest-performing replication of the PARNet model with SELU activation functions and *AlphaDropout*. We also select the highest performing MLP model with SELU activation functions and *AlphaDropout* to measure the effect of including the partially autoregressive feature.

2.3.2 Test Results for Superlative Model

Given an inability to differentiate improvement in models based only on loss function, we directly apply the same superlative models with and without the partially autoregressive feature to the test dataset to estimate performance improvement of these models over the benchmark KM model without further training. Moreover, we leverage this dataset to assess the effect of the inclusion of the autoregressive feature for MLP structures. Like the validation results, the test observations are not independent, so confidence intervals do not provide an appropriate means to assess the significance of our findings due to the limitations of the data.

Model	Mean Error Reduction versus KM
PARNet	34.82%
MLP	62.78%

Table 8. Mean reduction in absolute aggregate prediction error on test dataset shows both models outperformed the benchmark, but the inclusion of the partially autoregressive feature resulted in a smaller improvement

Both the MLP and the PARNet substantially improve prediction quality compared to the current benchmark model, as indicated in Table 12. However, contrary to expectations based on the validation results, the MLP provides the highest quality predictions for the test dataset. A likely cause of this deviation is that the test dataset included economic data during the recovery from the COVID-19 pandemic. This data reflected a dramatic decline in economic opportunities and an associated increase in USAF personnel retention, followed by a reversal in labor market conditions as wages spiked and unemployment dropped. The earlier validation results suggest that the superlative algorithmic configurations with a partially autoregressive feature performed well in a retention year with a more stable trend, while the changes in retention during the test set will measure how much the algorithm can improve performance over the KM benchmark in a very different environment.

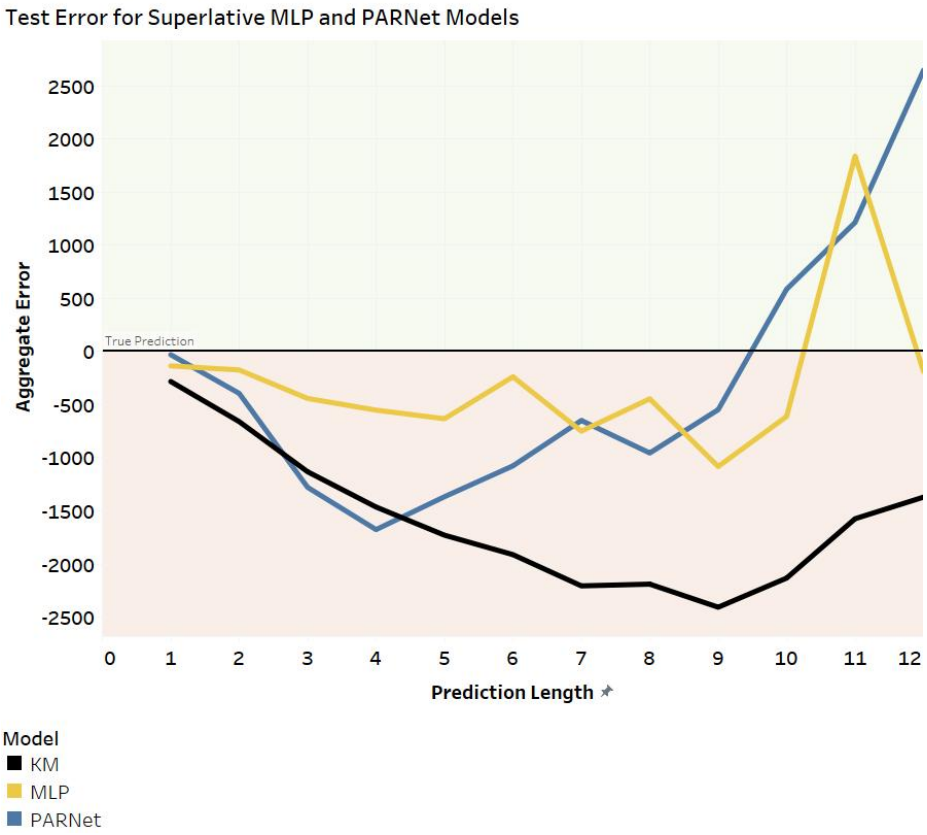

Figure 25. Mean test error of superlative models by prediction length

As expected, the PARNet's improvement in error rates in Table 12 is reduced from the validation results due to the rapid changes in underlying retention behavior caused by the pandemic. We see in Figure 25 that retention shifted direction in the ninth month, where the slope of the error in the KM model reverses direction. This change significantly worsened the prediction of the PARNet model, which beat the benchmark but did not beat the MLP without the partially autoregressive feature. The MLP appears to be generalizing quite well with the exception of the prediction from month 11. The PARNet error is higher than the MLP for early months, but a large spike in the observed error over the last three predictions appears to be caused by the shift in underlying retention behavior due to accelerating economic conditions. Because this shift occurred between the prediction and the observations

of the last three months, no indicators of the reversal were available in the partially autoregressive feature to help inform this estimate. The test dataset does appear to confirm that both the PARNet model and the basic MLP can provide generally superior performance to the current benchmark, even in difficult conditions.

2.4 Conclusions and Future Work

We have demonstrated multiple models that generate higher quality predictions for the USAF PRP compared to the current benchmark model. In addition, we showed that the inclusion of a partially autoregressive feature can reduce modeling error for multiple types of well-tuned machine learning algorithms during periods of consistent trend, although we were unable to confirm that this approach improved performance during periods of rapidly changing economic conditions or measure a positive impact using test data with these conditions. While the partially autoregressive MLP approach's results on the test dataset demonstrate sensitivity to changes in the trend direction compared to a model trained without the partially autoregressive feature, the chosen model still significantly outperformed the current USAF model serving as the benchmark. As additional training data is collected that includes changes in trend direction, both this approach and the MLP without the partially autoregressive feature are likely to improve in performance beyond the current measurement. Additionally, future test data without such a substantial shift in trend direction should show substantially improved results, although such theorized improvements in performance should be understood to be a measure of performance under those differing, economically steady conditions.

While the best PARNet model reduced mean absolute aggregate prediction error by 34.82% in the test dataset, most combinations of hyperparameters and replications failed to beat the KM benchmark during the validation process. This approach

currently requires a large number of neural networks to be trained to find a small number of high performing models. Further work should more finely examine the hyperparameter space near the winning combination and examine how robust those settings are for different time periods.

The primary problem with the approach provided in this chapter is the statistical bias of the estimates. The models that perform best as measured by binary cross-entropy have a consistent statistical bias that negatively impacts the quality of aggregate predictions. The architecture selected is not fitting the data better than the other architectures; it simply is fitting the data with less statistical bias. This must be addressed prior to operationalizing this model. Several methods are available to address this issue. First, reducing the number of features included until cohorts are significantly larger would ensure that fewer observations are at the extreme ends of the distribution at 0 and 1, decreasing the likelihood of consistent statistical bias in one direction. However, this statistical property comes at the cost of restricting the specific variables that have the most explanatory power. Second, the loss function can be modified to overweight penalties in one direction when consistent statistical bias is detected. Both of these approaches should be explored in further work.

Resampling data to address the imbalance between high and low retention observations may prove helpful to improve model training and prediction quality for low retention observations, but this will not address the statistically biased residuals for extreme values, which will remain imbalanced in the real world applications. Moreover, because this may increase error for the large number of high-retention observations, this may worsen problems with statistical bias.

In addition to the data used from MilPDS, AF policy variables and national-level econometric variables can provide a proxy measure for the individual opportunities and compensation available in the broader US labor market. To measure the value

of opportunities and compensation within the USAF, variables measuring personnel policies can be constructed, although most existing documentation of these policies is not stored in easily extractable formats or in a single location. Complicating the use of machine learning approaches, these policies are appropriately implemented to shape outcomes and not randomly designed to observe the exact effect of these policies, and thus create an endogeneity problem when attempting to model their effects. For these reasons, it is difficult to discern what caused retention behavior to change as well as to predict retention behavior when these underlying variables change. Further complicating the use of econometric variables, economic conditions often change slowly and only change direction every few years, making it difficult to model data spanning only short time periods. As policies change in the USAF, personnel from long ago may not retain similarly to airmen in the force today, making it difficult to use the entirety of data spanning long time periods. This shorter dataset used for our models was also limited by a single econometric trend during the training data; future machine learning work over the next few years will benefit from the natural experiment of a large economic shock from COVID-19. Early testing with econometric variables generalized poorly, but the inclusion of this natural experiment in a training dataset is likely to enable much better future performance.

Notably, this reversal in trend direction was marked by significant external factors that generated changes in macroeconomic variables. Some portion of this change would be captured simply by updating projections as the year progressed, still providing awareness of the retention impacts in advance. In addition, an operational deployment of such a model would not be blindly administered; analysts observing macroeconomic indicators can implement models using features that appear likely to improve performance. Further work could establish specific markers of trend instability based on macroeconomic indicators that can be used to select models that perform

best in the current environment. In addition, both of the superlative models showed the best performance for at least 25% of the prediction lengths. In combination with uncertainty about future trend stability, this suggests that the inclusion of an ensemble model using predictions from multiple types of machine learning models may provide better and more robust solutions than any individual model. Additionally, while the random forest approaches only beat the benchmark by a small margin in the second validation dataset, the inherent robustness of such approaches may be a valuable contribution to such an ensemble. Future work should examine both the inclusion of multiple models as well as strategies for creating diverse, high-quality models to contribute to this ensemble. Finally, although sequence-based approaches like long short term memory networks and other recurrent neural networks face many of the same problems associated with modeling heavily censored sequences, as described in Section 2.2.2, future work should verify that these approaches are unable to replicate or enhance the level of performance provided by the techniques proposed in this research.

III. Reinforcement Learning Approaches to Improve United States Air Force Accession Policies

3.1 Introduction

Each year approximately 9-13% of active duty personnel in the United States Air Force (USAF) depart the service. The USAF must develop policies to replace these personnel while meeting specialized skillset needs. The USAF manages a myriad of specific skill requirements for its personnel, primarily via career field designations, indicated by specific Air Force Specialty Codes (AFSCs). The Air Force corporate structure, United States Congress, and Air Force major commands each play a role in funding specific skillsets. This funding for current and future personnel is recorded as programmed manpower authorizations. Each year, the USAF must recruit and train many individuals for each required skillset; recruits entering each career field are called accessions.

Compared to historical military forces, much of the modern USAF's human capital is dedicated to fielding highly complex warfighting systems, which can take years to fully learn and operate effectively. Rapid changes in organizational experience levels due to large fluctuations in personnel in different year groups can substantially affect mission accomplishment. This effect can be partially measured by examining how the inventory in each AFSC matches the authorizations by grade for that AFSC. The number of accessions entering each AFSC impacts this distribution of experience because most skillsets in the USAF must be developed from the beginning for junior personnel. For this reason, policies that determine the accession level for each AFSC must be carefully constructed to meet both short and long term human capital needs. Not having adequate personnel with each skillset can have serious national security implications, so the long-term effects of these policies deserve careful consideration.

This chapter provides improved methods to accomplish this task.

This research formulates a closed workforce replenishment problem (WRP) to represent the USAF human capital decision context (i.e., without allowing for outside hiring to meet senior requirements), constructing a Markov decision process model to capture the effects of policies on outcomes. The USAF's Military Personnel Data System dataset is leveraged to design a realistic, high-quality state transition simulation. We present and test several reinforcement learning methods for developing high-quality enlisted accession policies for the Regular Air Force (i.e., active duty) that meet current and future manpower requirements as well as comply with and identify pipeline constraints. We measure the success of our proposed modeling and solution approaches by comparing simulated policy results to those obtained using the USAF's current benchmark equilibrium policy.

In pursuit of this goal, we propose the following methodological contributions. We propose and test a solution procedure that constructs a direct lookahead (DLA) policy using Monte Carlo simulation and a modification of Concave Adaptive Value Estimation (CAVE) (Godfrey and Powell, 2001). This extends previous work by using accessions constraints and relative values to solve a knapsack problem that determines the composition of total accessions for each time step. We also propose and test a second solution procedure that constructs a parameterized dynamic policy using approximate value iteration with Deep Q-Networks. This approach uses target networks and a replay buffer to stabilize learning. This parameterized approach overcomes computational limitations for searching enormous action spaces by using the same policy structures as the subject matter experts currently developing policies.

Defining the ideal state, wherein the number of personnel with each combination of AFSC, years of service (YOS), and grade match the corresponding number of manpower authorizations, allows a clear formulation of the WRP. However, ambiguity

exists when comparing the relative goodness of other possible states. Typically, the number of personnel within each AFSC and grade will differ from the ideal as defined by manpower authorizations. To examine this issue further for stakeholders, we test two candidate cost functions. This examination allows us to determine model robustness, gain insight about these function's effects on policies and outcomes, and present results that can be used to determine appropriate cost functions for future work.

The rest of this chapter is organized as follows. Section 3.2 describes the USAF WRP and the real world data and processing used to characterize system behavior. Section 3.3 formulates the Markov decision process model representing the USAF workforce system behavior. Section 3.4 describes how we developed and tested optimization approaches to find high-quality policies (relative to current practice) for this system. Section 3.5 describes the results from our computational experiments. This chapter finishes with implications for policy development and a description of the remaining work.

3.2 U.S. Air Force Workforce Replenishment Problem and Data

The USAF experiences constant programmatic change, which causes attendant changes in the mix of skills and competencies required from its workforce. Decisions regarding the appropriate personnel levels for these requirements are programmed and recorded as manpower authorizations by career field, designated by AFSC. As accession, retraining, and retention policies take time to plan, fund, and implement, the gap between the change in requirements and the change in numbers of personnel due to policies such as recruiting, training, and assigning personnel manifests as an AFSC shortage.

The quantity of personnel inventory with each skillset is an outcome of flows

through each AFSC, YOS, and grade, as shown in Figure 26. Policies seeking to affect this inventory must either change the rate of personnel departing the system, crosstraining between each AFSC within the system, or arriving into an AFSC from outside the system. Changing retention levels can require significant levels of investment, especially because most retention incentives must compensate both the additional individuals being retained as well as those who would have retained already. Moreover, retention incentives at the levels typically offered to enlisted AFSCs often have only small effects on personnel decisions (Joffrion and Wozny, 2015). Cross-flows (i.e., moving personnel from a donor AFSC to a receiving AFSC) can alleviate problems, but involuntary retraining policies come with a cost to retention because they create an opportunity for personnel to decline retraining and leave the service entirely. Because personnel depart the USAF each year and must be continually replaced, selecting an appropriate level of annual accessions for each AFSC is the policy that most directly impacts AFSC shortages.

The USAF has explored making the service more open to experienced outside talent because of this lack of flexibility in the current system, but this idea has yet to become a normal feature of the service's human capital lifecycle. Until such change occurs, methods to determine accession policies for each AFSC have effects that last for decades as the personnel in that cohort age through the system. Policies bringing in too many personnel can be offset by force management actions in later years. Policies bringing in too few are not so easily counterbalanced, especially in areas where training is either expensive, lengthy, or constrained.

Each year, the USAF first selects a level of total accessions to maintain aggregate end strength and comply with Congressionally-mandated end strength constraints. Even slight overages result in large military personnel expenditures and require offsets from other areas, so current practices prioritize end strength management, then

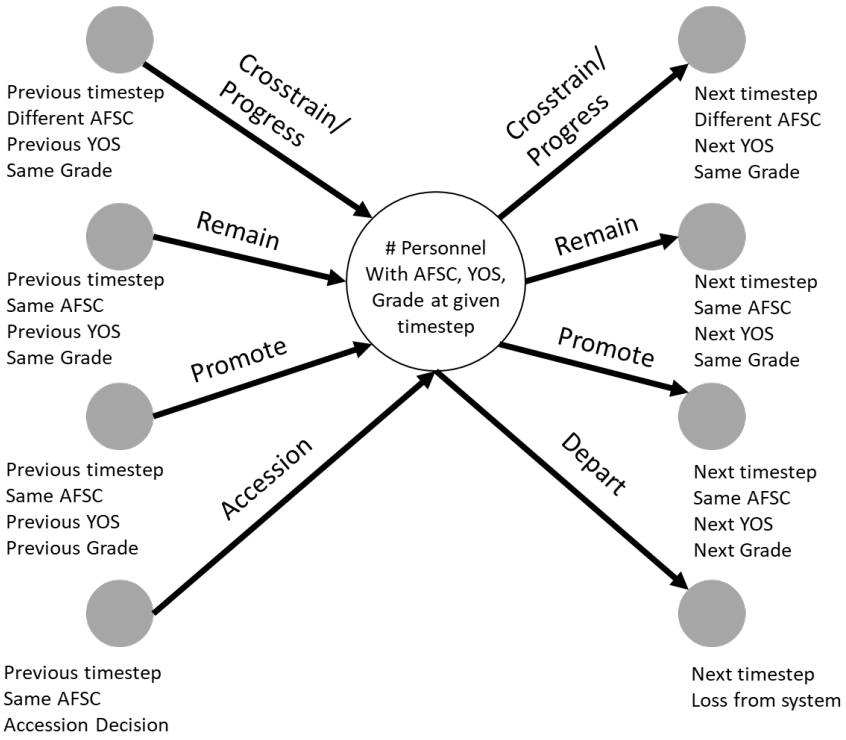

Figure 26. The number of personnel with each combination of attributes depends on the flows into and out of this state from adjacent states with combinations of AFSC, YOS, and grade at each time step.

address AFSC shortages within those budget constraints. After the aggregate accessions level is set, the staff of the Deputy Chief of Staff, Manpower, Personnel and Services, Headquarters USAF (AF/A1), works with the AFSCs' respective career field managers to divide this total among the various AFSCs. This research provides methods to rigorously develop a USAF-level accession policy for all AFSCs.

The USAF currently uses an equilibrium accession level for each AFSC as a starting point to develop current policy. This equilibrium level is developed using the USAF's Officer and Enlisted Sustainment Models. These models determine targets for the long term sustainment of an AFSC. These sustainment targets indicate the number of accessions desired for an AFSC to be 100% manned on average over an infinite time horizon if the USAF never adjusted its accession policy. USAF analysts develop these targets by measuring retention by AFSC and years of service over a

five-year period, excluding retention observed from years with non-representative force management actions, then projecting this behavior over a 30-year career by YOS. The sustainment model scales this projected profile based on the number of manpower authorizations at the furthest year recorded in the authorizations programming. This quantity is the 5th year personnel target programmed in the Manpower Programming and Execution System Unit Manpower Document (MPES-UMD, alternately referred to as the UMD).

Although the sustainment target provides a useful baseline for accessions policy, sustainment targets effectively produce a 20-30 year "get well plan," i.e., a plan for correcting AFSC imbalances due to personnel shortages and overages. Such a plan requires too long a time to make a meaningful impact and, in a dynamic environment with constantly changing requirements, the recovery time is even greater. This environment necessitates development of a set of accession policies that can more aggressively address AFSC shortages. Currently, this process requires a team of analysts to build these targeted accession policies in a time-intensive process. The process to validate these accession targets has historically relied on publishing these targets, then receiving feedback from the myriad supporting training schoolhouses to identify constraints. This process is iterated until a set of accession targets is both desirable and feasible within the current set of constraints. Fortunately, Air Education and Training Command has begun streamlining this process, and a new opportunity presents itself to improve the quality and timeliness of the planning processes. Because the sustainment target is the baseline for this manually developed target, we use it as the baseline policy for comparison when evaluating policies developed by our solution approach. We seek to determine high-quality accessions policies relative to the currently practiced baseline policy. Given the 7,050 dimensions of AFSC and YOS combinations for the inventory state space at each time step for the WRP

instance of interest, the use of conventional dynamic programming or approximate dynamic programming techniques that require an explicit listing of potential states prove intractable for this problem without some form of aggregation, decomposition, or parameterization.

To appreciate the need for a policy that can tailor accessions to emerging requirements without waiting for existing personnel cohorts to entirely depart of the system, it helps to observe the high rate of change for the manpower authorizations in many AFSCs. The authorization change for one example AFSC in a growing mission is shown in Figure 27. Each line represents a snapshot of the UMD in that year, with the first point on the line showing the actual number of authorizations in that year. The remainder of each line shows the projection of the future size of the AFSC based on the programming at that time. As seen in Figure 27, much of the growth in such AFSCs is not programmed in advance. This lack of anticipation requires agile policies that adapt to these changes quickly without violating training pipeline constraints or compromising long-term AFSC health due to large fluctuations in career field experience caused by disproportionately small or large year group cohorts.

The presence of non-stationary, stochastic demand (i.e., authorizations) suggests the need for deliberate modeling of this problem feature and an algorithmic approach to devise policies that display more resilience to this changing demand signal. However, anticipating future policy changes and developing policies to build a force that differs from that which is funded through Congressional and AF corporate structure inputs would effectively undercut senior decision-maker authority and the US Congress's legal oversight authority. Policies must address requirements as currently funded, but also induce a responsive structure resilient to change.

All transition behavior is measured using longitudinal measurement of 5 years of transitions in the USAF's Military Personnel Data System (MilPDS Dataset, 2021),

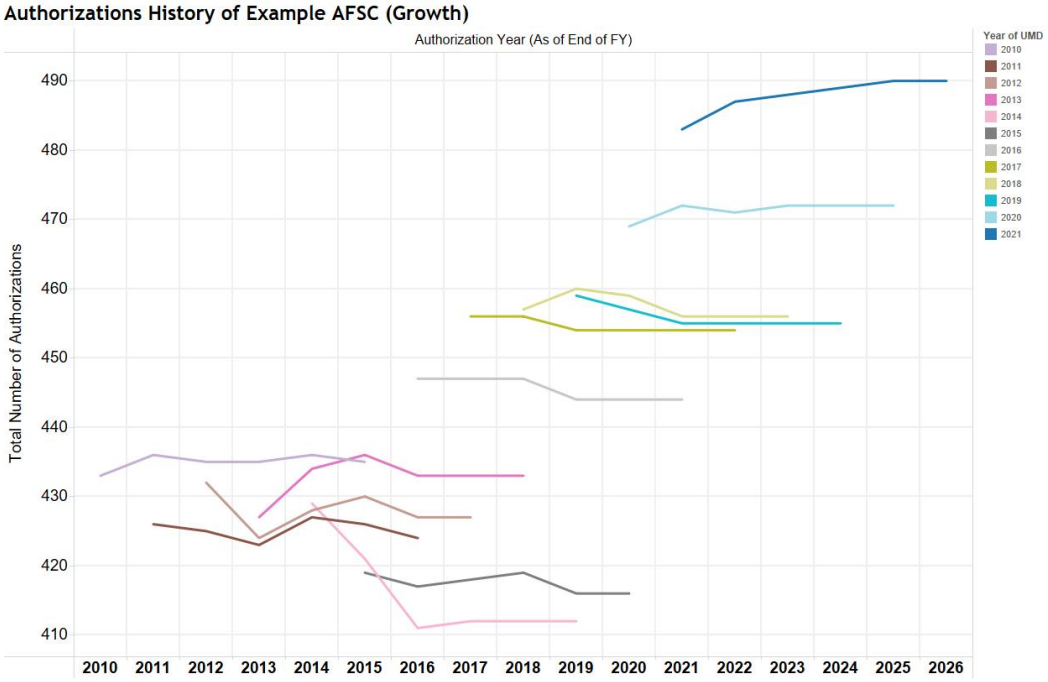

Figure 27. Many snapshots of the programmed authorizations level for the next five years change significantly when comparing the later years of programming with actual programming in that year

comprising 1,196,433 individual retention observations. Historical transition behavior is included for AFSCs that have changed their specified code based on the contents of the Enlisted Classification Directories over the last few years. Some AFSCs progress from a specific AFSC with a junior skill level to a related AFSC with a more senior skill level without requiring any kind of retraining action. A portion of these progression flows have multiple junior AFSCs progressing into a single senior AFSC as the scope of responsibility increases, necessitating explicit modeling of each to determine the correct accession level for each junior AFSC. AFSCs managed separately for visibility purposes, but having deterministic transitions from a single junior AFSC to a corresponding senior AFSC, are combined to reduce unnecessary noise in the projections. The starting inventory considered in the WRP instance of interest is the actual USAF enlisted inventory on 30 September 2021, which is the end of the fiscal year. The authorizations use the current and projected authorizations programmed

in the UMD recorded on 30 September 2021 (MPES-UMD Dataset, 2021).

3.3 Markov Decision Process Formulation

To effectively apply ADP and RL solution techniques, we first formulate the USAF WRP as a Markov decision process. We propose a finite-horizon formulation with a length of $T = 30$ years based on the enlisted maximum career length. Let

$$t \in \mathcal{T} = \{1, 2, ..., T\}. \tag{6}$$

3.3.1 State Variables

The state of the USAF personnel system S_t indicates the number of personnel with each combination of attributes. Let $S_{t,a,y} \in \mathbb{Z}_0^+$ be the number of personnel with AFSC $a \in \mathcal{A}$ and YOS $y \in \mathcal{Y}$ at time $t \in \mathcal{T}$, where $\mathcal{A}$ is the set of all skillsets and $\mathcal{Y} = \{0, 1, ..., Y\}$ is the set of all YOS with a maximum career length $Y = 30$. Let $S_t = (S_{t,a,y})_{a \in \mathcal{A}, y \in \mathcal{Y}}$ compactly indicate the state of the system at time t.

This problem requires the specification of several domain-specific parameters, represented within the starting state S_1. We define a set of manpower authorizations as

$$m_t = (m_{t,a,g})_{t \in \mathcal{T}, a \in \mathcal{A}, g \in \mathcal{G}}, \tag{7}$$

wherein $m_{t,a,g} \in \mathbb{Z}_0^+$ is the sum of the authorizations on the UMD for AFSC $a \in \mathcal{A}$ and grade $g \in \mathcal{G}$, with $\mathcal{G} = \{1, 2, ..., 6\}$, representing 6 enlisted grades. This approach combines the grades of E-1, E-2, and E-3 because of the structure of enlistment contracts that bring personnel in at different grades as well as E-8 and E-9 because of the small numbers of personnel in senior grades. Because manpower authorizations on the UMD are only programmed 5 years in advance, we use the authorizations

programmed for the latest year as an approximation of future demand. Moreover, the US Congress limits total end strength, constraining the total accessions for each annual timestep to A_t based on the expected aggregate departure rate for the following year and any desired change in the number of total personnel.

3.3.2 Decision Variables

The accessions decision at each time step t is defined as

$$x_t = (x_{t,a})_{a \in \mathcal{A}'}, \tag{8}$$

wherein $x_{t,a} \in \mathbb{Z}_0^+$ (i.e., the set of positive integers including zero) indicates the number of accessions for AFSC $a \in \mathcal{A}'$ and where $\mathcal{A}' \subset \mathcal{A}$ is the subset of AFSCs that use accessions to replenish their personnel instead of entirely relying on crosstraining or progression into the AFSC from a corresponding junior AFSC. For the USAF WRP instance of interest, $|\mathcal{A}'| = 186$.

A trivial and unhelpful solution to this problem would be to increase the total number of personnel in the force. Previous work included a penalty for overages in the cost function and allowed the algorithm to vary aggregate end strength slightly based on this tradeoff (Hoecherl et al., 2016). Such an approach is consistent with legal authorities granted to the Secretary of the Air Force, but it is inconsistent with the actual business processes and financial realities that govern the process to generate accessions. In practice, the annual total accession target, denoted A_t, is generated by a USAF personnel model that uses aggregate inventory by YOS and historical retention by YOS to estimate the number of personnel who will leave in the next year, modified for any desired end strength growth or decline. The accessions decisions for individual AFSCs are developed to ensure total end strength is satisfied. That is,

$$\sum_{a \in \mathcal{A}'} x_{t,a} = A_t \quad \forall t \in \mathcal{T}. \tag{9}$$

We also include the ability to select lower and upper constraints for each AFSC's accession decision, $\eta^-_{t,a}$ and $\eta^+_{t,a}$, letting

$$\eta^-_{t,a} \leq x_{t,a} \leq \eta^+_{t,a} \quad \forall a \in \mathcal{A}', t \in \mathcal{T}. \tag{10}$$

Together, these constraints restrict the potential actions x_t to the set $\mathcal{X}_t$. For the problem instance of interest, we set such constraints to 75% and 150% of the sustainment level. For operational applications, this would be modified to the actual constraints as recorded in Air Education and Training Command's Business Reporting and Intelligence Tool. Upper constraints represent limitations on the number of instructors, available dormitory space, or other AFSC-specific constraints. Lower constraints capture interrelationships between training pipelines where production of one specialty relies on corresponding training in another specialty, contractual obligations, or agreements with other services for shared training pipelines.

3.3.3 System Transition

A system transition function models a single time step stochastic transition from any given state to a potential future state. Let

$$S_{t+1} = S^M(S_t, x_t, \omega_{t+1}) \quad \forall t \in \mathcal{T}, \tag{11}$$

wherein $\omega_{t+1} \in \Omega$ represents the exogenous information discovered during the transition to the next time step, shown in Table 13, and Ω represents all possible outcomes.

To model this transition, we first develop a Kaplan Meier estimate of retention rates by AFSC and YOS, then use a binomial distribution to simulate future reten-

tion outcomes using those estimates. Cohorts at the maximum YOS deterministically transition out of the service whereas the selected number of new accessions transition into the service according to a stochastic YOS distribution based on historic arrival patterns. This YOS distribution accounts for both lengthy training pipelines and underlying processes wherein personnel who fail to complete one set of initial skills training are reclassified into another AFSC. In such a case, the reclassified member restarts the training process for their new AFSC. These individuals are only counted as part of the AFSC's manning once they have been awarded their AFSC and permanent party status. Students in these training pipelines and basic military training are accounted for separately in the total end strength using a separate account for student man-years.

Although rates of departure from and arrivals to the USAF are relatively stable, crossflows present a much more difficult phenomenon to model. Retraining policies that govern crossflows adjust both the number of quotas into different AFSCs as well as eligibility for personnel to retrain out of undermanned AFSCs after the service member's initial enlistment. This dynamic policy means that overly simple modeling with static transition probabilities can yield unrealistic behavior. We specify the transition probabilities based on particular state variable conditions at the beginning of each time step. We model the probability of transitioning out of an AFSC to any other AFSC, then model a second stochastic process to determine which AFSC gains the service member based on a constructed number of potential retraining quotas.

Potential Outcome	Probability Calculation	Distribution	AFSC-Specific Destination
Retain in AF	P(Retain $\mid a, y$)	Binomial	No
Remain in AFSC	P(Stay $\mid a, y$, Retain)	Binomial	Yes
Progress	P(Progress $\mid a, y$, Retain, Not Remain)	Binomial	Yes
Crosstrain Out	P(Cross Out $\mid a, y$, Retain, Not Remain, Not Progress)	Fully Determined	No
Gain to System	P(Gain in YOS $y \mid a, x_{t,a}$)	Multinomial	Yes
Crosstrain In	P(Cross to AFSC a $\mid$ Cross Out)	Uniform	Yes

Table 9. Potential State Transitions

All rates are measured based on historical observations using 5 years of Military

Personnel Data System data. The dynamic nature of the simulation of crossflows is an important step for modeling USAF personnel retention behavior. Past approaches either modeled retraining as distinct phenomena for each AFSC without modeling actual personnel flowing from one to the other (Hoecherl et al., 2016) or modeled fixed transition rates based on historical data. The rationale for the USAF administrative policies that created the conditions for those historical rates included nuanced perceptions of manning, overages, shortages, and political realities at the time, and computing those rates under different conditions implicitly assumes policies that would not be executed in the differing conditions in the modeled scenario.

3.3.4 Cost Function

Solving the WRP requires defining a cost function that well represents the preferences of the USAF. Although identifying the ideal state is straightforward (i.e., personnel match authorizations), evaluating the quality of other states is considerably more difficult. Past approaches focus on two methods to define rewards or penalties; both are inadequate to ensure desirable real-world outcomes. The first approach minimizes aggregate shortages and overages by AFSC. Although shortages are the primary concern, the USAF budget does not allocate resources for personnel that do not meet a funded authorization, so any overage in one AFSC directly results in a shortage elsewhere. However, this approach generates undesirable real-world outcomes, and its recommended solutions are trivial. If personnel within an AFSC are treated as exactly fungible, then the optimal solution is always to simply replace losses and increase or decrease accessions to grow or shrink the AFSC to any desired size, subject to accession constraints. However, personnel within an AFSC are not exactly fungible.

The second approach attempts to address this limitation by measuring personnel

inventories and authorizations by grade as well, which serves as a proxy for disparate competencies and experience. can be grouped by grade, which correspond to the grade of authorizations. Policies can be then developed to best minimize shortages by grade and AFSC. Increasing accessions to solve manning problems in one time period will result in a year group cohort that is larger on average than the steady-state level for the next 30 years. As the year group cohorts with fewer people eventually depart the system, accessions must be restricted below the steady-state level to prevent manning from rising above 100%, resulting in shortages in other AFSCs. This policy results in a pendulum effect, creating grade shortages far into the future to correct aggregate AFSC shortages. Hoecherl et al. (2016) consider an extended variant of the approach, developing accession policies that seek to minimize shortages and overages by both AFSC and grade, ensuring personnel can meet the workload associated with a required level of manpower authorizations.

However, this approach too does not capture the scope of the problem for two reasons. First, it ignores the ways that additional personnel in a higher or lower grade can partially compensate for a lack of needed personnel in a specific grade. Having the right number of total personnel is not sufficient for a good outcome, but it is necessary. Aggregate AFSC shortages still have relevance when defining which outcomes are good. Second, some AFSC grade structures are constructed in a way that is unsustainable. These AFSCs have too many or too few senior authorizations relative to the number of junior authorizations needed to grow the required senior personnel. If only considering grade shortages for an AFSC, an optimal solution for some of these AFSCs may be to simply accept that these positions cannot be filled and choose to prioritize feasible grade structures, consistently under-resourcing these AFSCs. To generate the cost functions for this research, we combine the two approaches to ensure both the aggregate disconnect and the effect of disproportionate

year group sizes are captured.

Notably, the state variables defined previously do not include grade. Static transitions by grade are difficult to model because the USAF modifies its grade structure and promotions policies based on several dynamic processes using significant subject matter expert input. This process is heavily influenced by the relative sizes of different YOS cohorts (i.e., year groups) and changes based on a number of quantitative and qualitative factors from year to year. Testing transition rates based on historical rates yielded simulated behavior that deviates substantially from historically observed grade structures and experience levels for individual grades. Given this complexity, a viable modeling approach must either replicate the additional complexity present in the promotions process or else translate the more parsimonious state vector into an inventory that includes grade. Because grade structures generally maintain the same approximate relationship between YOS and grade, we instead calculate an expected grade inventory for each AFSC, YOS combination. Let

$$S_{t,a,y,g} = S_{t,a,y} P(g|a,y) \quad \forall t \in \mathcal{T}, a \in \mathcal{A}, y \in \mathcal{Y}, g \in \mathcal{G}, \tag{12}$$

wherein $P(g|a,y)$ is the historically-observed probability of a person being in a given grade g given the person is in AFSC a and YOS y.

This approach allows us to use AFSC and YOS information to emulate realistic transition behavior within the system, computing expected grade distributions to compare these states to requirements (i.e., $m_{t,a,g}$) defined by AFSC and grade but not YOS. Each AFSC's grade authorizations $m_{t,a,g} \in \mathbb{Z}_0^+$ sum to the total AFSC authorizations, $m_{t,a}$:

$$m_{t,a} = \sum_{g \in \mathcal{G}} m_{t,a,g} \quad \forall t \in \mathcal{T}, a \in \mathcal{A}. \tag{13}$$

Because there are $|\mathcal{G}|$ grades, we weight AFSC shortages by a factor of $2|\mathcal{G}|$ to

reflect the general importance of the aggregate health. We also define the number of personnel in each AFSC as $S_{t,a} = \sum_{y \in \mathcal{Y}} S_{t,a,y}$ and the number of personnel in each AFSC and grade combination as $S_{t,a,g} = \sum_{y \in \mathcal{Y}} S_{t,a,y,g}$. This yields the following cost function:

$$C(S_t) = \sum_{a \in \mathcal{A}} \left[(2|\mathcal{G}|) \max \left(m_{t,a} - S_{t,a}, 0 \right) + \sum_{g \in \mathcal{G}} \max \left(m_{t,a,g} - S_{t,a,g}, 0 \right) \right]. \tag{14}$$

Although representing the general preferences of the USAF, this cost function based on shortages demonstrates a critical weakness: shortages impact AFSCs of different sizes with different levels of severity. For example, lacking 100 trained personnel in a community of tens of thousands of military police will result in less dramatic impacts to readiness than in a community of 400 specialized aircraft maintainers. For this reason, we consider a second cost function that uses manning percentage to measure how far below the total requirement an AFSC falls instead of a direct measure of shortages.

$$C(S_t) = \sum_{a \in \mathcal{A}} \left[(2|\mathcal{G}|) \max \left(1 - \frac{S_a}{m_a}, 0 \right) + \sum_{g \in \mathcal{G}} \max \left(1 - \frac{S_{a,g}}{m_{a,g}}, 0 \right) \right] \tag{15}$$

3.3.5 Objective Function

The objective of the Markov decision process is expressed as follows:

$$\min_{\pi \in \Pi} \left(\mathbb{E}^{\pi} \Big[\sum_{t=1}^{T} \gamma^{t-1} C(S_t) \Big] \right), \tag{16}$$

where γ is the discount factor and Π is the set of all possible policies. Recall that system transition occurs according to $S_{t+1} = S^M(S_t, x_t, \omega_{t+1})$, wherein the accession decision x_t is selected according to the policy $\pi \in \Pi$, expressed by the decision function

$X_t^\pi(S_t)$. For small enough problem instances, this policy is found by recursive selection of the optimal action according to the modified Bellman equation:

$$x_t^\pi = X_t^\pi(S_t) = \arg\min_{x_t \in \mathcal{X}_t} \Big(C(S_t) + \gamma \mathbb{E}\big[V(S_{t+1}|S_t, x_t)\big] \Big), \tag{17}$$

where $X_t^\pi(S_t)$ is the decision function, $V(S_{t+1})$ is the value of the next state at time $t+1$, and the expectation of the value of the next state at time $t+1$, $\mathbb{E}[V(S_{t+1}|S_t, x_t)]$, is the sum of the value of each potential future state weighted by the probability of transitioning to that state across all potential states at $t+1$. However, while this expectation can be directly computed for smaller problem instances, larger problem instances require an approximation of this expectation to achieve tractable computation times. For problems using Monte Carlo simulation to sample potential future states and approximate this expectation, we determine the approximate best action using the values of the parameter θ. The best known decision based on the current values of θ is the policy $X_t^\pi(S_t|\theta)$.

The discount factor impacts the relative value of different policies significantly and must be chosen with care. Values set too high place too much confidence in the authorization structure remaining static, valuing the ability to meet uncertain authorization levels in the future near equally to the ability to meet certain authorizations in the present. We utilize $\gamma = 0.8$ in our analysis, striking a balance between senior leaders' observed emphasis on solving problems in a short period of time while also reflecting the common wisdom that the long term impacts of personnel policies on national security are of great importance.

3.4 Algorithms

We first provide an overview of the benchmark algorithm currently in use by the USAF then describe two candidate algorithmic approaches we develop to improve

upon the policies generated by the benchmark.

3.4.1 Benchmark: Equilibrium Sustainment Model (Markov Chain)

As the benchmark for this research, we use the USAF's current policy baseline called the *sustainment model*. This model uses a Markov Chain for each AFSC, constructed with Kaplan Meier retention estimates by YOS. This approach does not consider grade requirements but is used to generate an equilibrium policy that would result in each AFSC being manned at 100% on average over the long term. This policy successfully addresses manning issues but only after a substantial delay as existing year groups that are larger or smaller than the steady-state distribution eventually depart the system.

Two primary issues negatively impact this sustainment model approach. The first weakness is the slowness to adapt to new mission requirements, especially for AFSCs that continually grow. Each time the number of authorizations increases, the new steady-state policy requirement increases, meaning that past accession policies admit too few personnel even compared to the steady-state policy. For example, a 16-year "get well plan" results in perpetual undermanning because the year group sizes are continually undersized for the new requirement. The second weakness of the current USAF policy is that it only adjusts accessions based on how they fill authorizations in their original AFSC. Future transitions to other AFSCs do not affect the policy calculation. This omission is especially problematic for AFSCs that progress upwards in a pyramid design, steadily expanding the service member's scope of responsibility. Such policies likely result in healthy AFSCs that rely primarily on accessions, but may result in negative secondary and tertiary effects in crossflow and progression AFSCs.

Because aggregate year group sizes vary, the same number of personnel do not

depart the USAF each year, and the required number of total accessions varies accordingly. In addition, the total number of accessions varies further based on desired changes in end strength for a given year. Changing end strength using accession levels is generally cheaper and less disruptive than requiring existing personnel to leave involuntarily. As this aggregate target changes, the sustainment level for accessions is simply scaled upwards or downwards proportionately for all AFSCs to comply with the end strength constraint. This current USAF approach serves as the benchmark.

We propose and test two algorithmic approaches to identify improved accession policies. The policies determined by these algorithms are compared to each other and the benchmark sustainment policy to examine the quality of such approaches. The sustainment policy cannot be used in its existing form due to the need to comply with end strength limits. For this reason, once the aggregate accession level is determined at each time step, the individual sustainment targets are proportionately scaled up or down to match the aggregate target. The current USAF sustainment model policy indicates an accession decision for each AFSC a at a given timestep t as $e_{t,a} \in \mathbb{Z}_0^+$.

3.4.2 Concave Adaptive Value Estimation (CAVE)

Godfrey and Powell (2001) developed Concave Adaptive Value Estimation (CAVE) to efficiently find optimal solutions for a single time step resource allocation problem with stochastic demand using a piecewise linear value function approximation. Godfrey and Powell (2002*a*) demonstrated this approach's effectiveness for large control problems with high-dimension action spaces.

Piecewise linear value function approximation allows for efficient updating of estimated gradients because the slope at every breakpoint is known to be monotonically decreasing. Subject to a learning rate to stabilize training with stochastic outcomes, observations of a lower gradient allow all higher gradients at higher values to be up-

dated simultaneously. The same procedure holds for observations of a higher gradient and updating lower gradients at lower values. This approach is particularly suitable for discrete functions, such as those measuring the value of personnel, where only whole people are observed in the system. By varying the length of the interval between breakpoints from large to small as the algorithm progresses, the algorithm very efficiently converges to high-quality solutions.

Godfrey and Powell (2001) and Godfrey and Powell (2002*a*) considered resource allocation problems wherein a resource was allocated or replenished to meet a single specific potential future demand. In this case, a single piecewise linear value function approximation could represent tradeoffs between the allocation of resources to different choices. Godfrey and Powell (2002*b*) extended this approach for multiperiod problems, but the algorithm still made decisions based on the single period use of a resource. Topaloglu and Powell (2003) proved that the CAVE approach converges to the optimal solution for the discrete newsvendor problem under certain assumptions. Kunnumkal and Topaloglu (2008) extended CAVE's use to multiperiod inventory problems with backlogged demands. More recently, Salas and Powell (2018) tested CAVE's performance using different stepsize rules contrasting a harmonic stepsize rule and a Bias-Adjusted Kalman Filter for an energy storage problem. However, each of these contributions developed CAVE variants for a problem structure that allocated a consumable resource. In the WRP, the inventory (i.e., personnel) are not consumed by demand. Instead, inventory can meet demand (i.e., authorizations) at multiple timesteps, and the length of the inventory's survival is not closely related to the utilization (i.e., inventory is not "used up" by meeting demand).

Several other contributions extended CAVE's application to WRPs. Song and Huang (2008) provided a related stochastic programming approach called the Successive Convex Approximation Method to solve a workforce capacity planning problem

with stochastic demands. However, their approach requires static transition rates and does not scale to the size of the USAF WRP, even with a simpler transition model. Hoecherl et al. (2016) applied the CAVE algorithm to solve a smaller form of the USAF WRP, one featuring both skillsets and grade, multiple time steps, and deterministic demand. Two key modifications enabled its successful application. First, the CAVE algorithm updated a direct lookahead policy based on the gradients of the individual decisions instead of the state variable. Second, the CAVE algorithm weighted the future observed gradients with an expectation of the probability of survival until the future demand, allowing an accurate estimate of the true gradient. A successful application of CAVE was possible because the USAF WRP exhibits a helpful problem structure; an accession decision's individual marginal impact to shortages at every future timestep is concave, holding all other decisions equal. Thus, the cumulative gradient of effects at all future timesteps is also concave.

Although this concave structure allows for efficient, simultaneous updating of gradient estimates, a problem arises. Generating good policies requires the algorithm to simulate far enough into the future to observe penalties or derive an alternative estimate of the value of a given state. Such an approach might require a lookahead horizon, defined as T_π, of 50 years for something like the WRP, since this approach must simulate effects 15-20 years away but still model policies at that time that do not become myopic and ignore future effects of accessions. Hoecherl et al. (2016) overcame this limitation by examining a single starting state and making a simplifying assumption that policies would revert back to the equilibrium sustainment policy after a specified number of years. In the WRP, each time step's decision interacts with other time steps' decisions due to the relatively long career length of the recruits resulting from any given accessions policy. For this reason, the direct lookahead approach can easily become trapped in a local optimum when modeling sequential decisions that

can fill the same demands, such as accessions policies over multiple years. This is also true for multiple decisions in the same time step that contain an interaction term, such as retraining and accession policies. Despite these limitations, the approach has been demonstrated to provide high-quality policies.

We propose a further modification of the CAVE algorithm, shown in Algorithm 3. This modification involves two major changes that should improve its solution quality compared to Hoecherl et al. (2016). First, the gradients sampled are based on the modified contribution function, which no longer penalizes overages, so the observed gradient is always positive. The constraint to accessions is not based on the overages, but on the actual end strength constraint, with the decisions being made based on the relative value of the AFSCs. Second, as the system model now captures transitions between AFSCs, the gradient for accessions in each AFSC is calculated based on the cumulative probability of a given accession filling a shortage in each of the 235 AFSCs at each time step, not just the original AFSC. This inclusion of cross-training allows for a much more realistic representation of actual business processes and behaviors because many AFSCs rely on cross-training as a source of personnel. The previous approach may have underestimated the value of accessions in skillsets that serve as a source for these lateral-entry AFSCs. Notably, the expectation may change based on whether particular AFSCs are overmanned or undermanned. As an approximation of this expectation, we calculate these probabilities based on the unrestricted crossflow-out rates from each AFSC. We also restrict the crossflow-in rates for this calculation for any AFSCs that rely primarily on accessions instead of retraining or a blend of retraining and accessions. These categories are provided by the research sponsor as a policy decision made with advice from the respective career field managers.

In our CAVE implementation, the piecewise linear value function approximation model is defined using the parameter tuple

Algorithm 1 CAVE Algorithm

Step 1: Initialization

Identify A_t, the aggregate constraint for accessions $\forall t \leq T_\pi$, where T_π is the desired length of the lookahead policy before reverting to equilibrium policy.

For each $x_{t,a}$, let $k_{t,a} = 2$, where $\nu^1_{t,a} = 0.0001$, $\nu^2_{t,a} = 0$, $u^1_{t,a} = 0, u^2_{t,a} = e_{t,a}$, where $e_{t,a}$ is the equilibrium policy s.t. $\sum_{a \in \mathcal{A}'} e_{t,a} = A_t \quad \forall t \in \{1, ..., T_\pi\}$.

Initialize parameters δ_n and α_n.

for n = 1 **to** N **do**

Step 2: Determine current policy $X^\pi_t(S_t \mid \theta)$

for $t \in \{1, ..., T_\pi\}$ **do**

Initialize policy with $x_{t,a} = 0$ by setting $k_{t,a} = 1 \quad \forall a \in \mathcal{A}'$

while $\sum_{a \in \mathcal{A}'} x_{t,a} < A_t$ **do**

Select AFSC a^+ with largest estimated gradient $\text{argmax}_{a \in \mathcal{A}'}(\nu^{k_{t,a}}_{t,a})$

Increase the accessions for decision x_{t,a^+} by setting $k_{t,a^+} = k_{t,a^+} + 1$.

end while

end for

Step 3: Collect Gradient Information

Simultaneously sample the gradients $\Delta^-_{t,a}(x_{t,a}, \omega)$ and $\Delta^+_{t,a}(x_{t,a}, \omega)$ over a finite time horizon with random outcomes $\omega \in \Omega \quad \forall t \leq T_\pi, a \in \mathcal{A}'$

Step 4: Define Smoothing Interval

Let $k^-_{t,a} = \min\{k_{t,a} \in \mathcal{K}_{t,a} : \nu^{k_{t,a}}_{t,a} \leq (1 - \alpha_n)\nu^{k_{t,a}+1}_{t,a} + \alpha_n \Delta^-_{t,a}(x_{t,a}, \omega)\}$.

Let $k^+_{t,a} = \max\{k_{t,a} \in \mathcal{K}_{t,a} : (1 - \alpha_n)\nu^{k_{t,a}-1}_{t,a} + \alpha_n \Delta^+_{t,a}(x_{t,a}, \omega) \leq \nu^{k_{t,a}}_{t,a}\}$.

Define the smoothing interval

$$U_{t,a} = \left[\max\{x_{t,a} - \delta_n, u^{k^-_{t,a}}_{t,a}, \eta^-_{t,a}\}, \min\{x_{t,a} + \delta_n, u^{k^+_{t,a}+1}_{t,a}, \eta^+_{t,a}\}\right).$$

Create new breakpoints at $x_{t,a}$ and the endpoints of $U_{t,a}$ as needed. Since a new breakpoint always divides an existing segment, the segment slopes on both sides of the new breakpoint are the same initially.

Step 5: Update θ based on current policy

For each segment in the interval $U_{t,a}$, update the slope according to $\nu^k_{t,a} = \alpha_n \Delta_{t,a} + (1 - \alpha_n)\nu^k_{t,a}$, where $\Delta_{t,a} = \Delta^-_{t,a}(x_{t,a}, \omega)$ if $u^k_{t,a} < x_{t,a}$ and $\Delta_{t,a} = \Delta^+_{t,a}(x_{t,a}, \omega)$ otherwise.

Adjust δ_{n+1} and α_{n+1} according to step size rules.

end for

End

$$\theta = (u_t, \nu_t)_{t \in \mathcal{T}_\pi}, \tag{18}$$

where $u_t = (u_{t,a})_{a \in \mathcal{A}'}$ and $\nu_t = (\nu_{t,a})_{a \in \mathcal{A}'}$ respectively represent the vectors of breakpoints and the gradient at each breakpoint for each AFSC a with an associated accession decision. The system transition function for the CAVE approach is then defined as $S_{t+1} = S^M(S_t, X_t^\pi(S_t|\theta), \omega_{t+1})$. Because these tuples of vectors show the gradients for multiple potential decisions for each AFSC a at time t, we use the variable $k_{t,a}$ to represent the selected breakpoint of the current decision, with the $k_{t,a}$th element of $u_{t,a}$ denoted $u_{t,a}^{k_{t,a}}$ defining the decision $x_{t,a}$.

In Step 1, we initialize each of these variables using the equilibrium policy. At each iteration n, we complete Step 2: determine current policy; Step 3: collect gradient information; Step 4: define smoothing interval; and Step 5: update θ based on current policy. In Step 2, we iteratively find the AFSC a^+ with the highest gradient at each timestep t and increment the associated breakpoint $k_{t,a}$ until the total number of accessions in timestep t meets the constraint A_t. In Step 3, we then sample the gradients using survival rates and simulated future outcomes to find the marginal effect on discounted cost below and above the current breakpoint, denoting these gradients as $\Delta_{t,a}^-(X_{t,a}, \omega)$ and $\Delta_{t,a}^+(x_{t,a}, \omega)$, respectively. In Step 4, the smoothing interval $U_{t,a}$ is updated to account for any constraints or potential concavity violations, then the algorithm inserts breakpoints above and below the selected decision at $k_{t,a}$ in $u_{t,a}$ and $\nu_{t,a}$. The piecewise linear value function approximation begins with a large smoothing interval as the new breakpoints are inserted based on high values of δ_n, then updates smaller intervals as the algorithm progresses and δ_n declines. This approach allows for large adjustments to decisions in early stages of training, then more granular adjustments as the algorithm progresses and the overall quality of the policy improves. Finally, in Step 5, the algorithm modifies the appropriate elements

of the estimate of the gradients, $\nu_{t,a}$, adjusted for a stepsize α_n. Training continues for N total iterations.

3.4.3 Parameterized Policy Generation with Deep Q-Networks

Deep Q-Networks (DQN) algorithms extend traditional Q-Learning to larger state spaces by mapping the value of the state action pairs with a neural network instead of tabulation, but many DQN implementations still optimize over a relatively small set of actions, such as the controls of an Atari console (Mnih et al., 2013). A primary difficulty for any algorithm attempting to solve large sequential decision-making problems is the computational demands for modeling the high-dimension state, action, and outcome spaces of these systems. We use several methods to address this challenge. First, the use of Monte Carlo simulation reduces the problems associated with the large outcome space. Second, by starting each simulation at the current state of the system and resetting after a specified number of time steps, the state space to be sampled is dramatically reduced to only states that may actually be visited during a finite length simulation. However, the action space remains a challenging problem.

DQN is a powerful technique when solving problems with small action spaces, but when the action space becomes too large, it faces significant limitations. Even with a high-quality value function approximation, simply searching the action space for a good policy is computationally demanding and potentially intractable, limiting its application. Because effective training typically requires iteratively solving problems many times, appropriately addressing the challenge of a large action space is a major concern for scaling deep reinforcement learning in general. This difficulty is present for the USAF WRP. Individual accession decisions have a large number of possible integer solutions and high dimensionality because of the large number of AFSCs.

Beyond simply applying more computation, different approaches have been de-

veloped to effectively scale deep Q-Networks or other value function approximation techniques to larger action spaces. However, each of these approaches must limit the action space to some smaller subset of the total space by sampling the action space pseudo-randomly and finding an approximate best action (Ho, 1999; Van de Wiele et al., 2020). Given the size of the action space for this problem and the large number of low-quality actions, randomly selecting a subset of actions would be an inefficient and likely ineffective method to explore other good state-action combinations. Another alternative is to develop parameterized policies that use known structure to solve the problem using a much smaller action space. If high-quality or optimal policies cannot be represented by the selected parameterization, this may significantly worsen solution quality, so the parameterization must be selected carefully.

Our second proposed algorithm, shown in Algorithm 2, is a DQN variant with a parameterized decision structure that USAF analysts have developed and used for personnel policy generation, but the structure has not been empirically tested to determine the appropriate parameter setting. This approach makes use of the existing equilibrium model and the knowledge that accessions should generally be higher for undermanned AFSCs and lower for overmanned AFSCs. We investigate policies of the following form. A discretized decision $d^p \in \mathcal{D}^p = \{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$ must be found regarding the proportion of shortages to fill for AFSCs with a shortage. We also describe these decisions with their corresponding index $d \in \{1, 2, ..., 6\}$. One benefit of discretizing this accession decision instead of treating it as a continuous variable is the ability to generate a separate output for the value of each action for a given state, making a search for the best action computationally efficient. AFSCs with a manning level between 95% and 105% simply receive the sustainment target automatically; these cutoffs are appropriate for future tuning as a parameter for this strategy. Donor AFSCs (i.e., those that are overmanned) receive a fair-share

proportion of their sustainment target once the other AFSCs receive what they need.

Given the broad variety of potential state outcomes, this parameterized approach may sometimes lead to absurdities. The following conditions handle those cases. If zero donor (overmanned) or needy (undermanned) AFSCs exist at a given timestep, the accession decision for all AFSCs defaults back to the sustainment target. If donor AFSCs have less total accessions than would be donated by the selected decision, the number of accessions to be transferred is set to the number available to be donated.

Algorithm 2 Baseline Parameterized Deep Q Network Algorithm

for n = 1 to N **Policy Improvement Loop**
  Initialize S_t as Starting Inventory S_1
  for $t \in \mathcal{T}$ **Policy Evaluation Loop**
    Record 1-dimensional vector of expected grade inventory of S_t as next row of S^{buffer}
    for $d^p \in \mathcal{D}^p$ **Policy Observation Loop**
      Determine $x_{t,a} \forall a \in \mathcal{A}'$ from d^p
      Observe transition to next state S_{t+1}
      Predict $Q(S_t, d^p \mid \theta^{\text{target}})$ and $\max_{d^p \in \mathcal{D}^p} Q(S_{t+1}, d^p \mid \theta^{\text{target}})$
      Set dth element of next row of v^{buffer} as $\hat{v}(d) =$
      $$(1-\alpha_n)Q(S_t, d^p \mid \theta^{target})+\alpha_n\Big(C(S_t)+\gamma\big(\max_{d^p \in \mathcal{D}^p} Q(S_{t+1}, d^p \mid \theta^{target})\big)\Big)$$
    end for
    Pursue ϵ-greedy state sampling strategy to transition to next state
  end for
  Select S^{sample} and v^{sample} as normalized random sample of 10% of S^{buffer} and v^{buffer}
  Update θ with single batch update with S^{sample} (input) and v^{sample} (output)
  if $n \mod N^{\Delta} = 0$ **then**
    $\theta^{\text{target}} = \theta$ (Update Target Network)
  end if
  Record $\max_{d^p \in \mathcal{D}^p} Q(S_1, d^p \mid \theta^{\text{target}})$ as estimated value of starting inventory
  if Range of last N^{Ω} estimates of $\max_{d^p \in \mathcal{D}^p} Q(S_1, d^p \mid \theta^{\text{target}}) <$ threshold V^{Ω} **then**
    End Training
  end if
end for

In addition to the problems with large action spaces, the potential for divergence

of DQN and other RL algorithms has been a subject of concern for years (Tsitsiklis and Van Roy, 1997). The combination of the deadly triad of function approximation, bootstrapping, and off-policy training, as described by Sutton and Barto (2018), and further explored by other reinforcement learning researchers (Van Hasselt et al., 2018), speaks to this problem. This triad can be devastating when used on deterministic systems, but the danger increases dramatically when bootstrapping off of stochastic outcomes, which may increase the odds of diverging if the function approximation happens to fit some amount of stochastic noise early in the training process. Much of the research on these algorithms focuses on deterministic problem sets; early testing suggests that stochasticity increases the potential for divergence and establishes the importance of finding effective ways to stabilize training.

In preliminary testing, we observed this divergent behavior during some training runs. To stabilize performance, we implemented a a target network and a modified form of the experience replay buffer as demonstrated by Mnih et al. (2015). The target network provides stability and efficient convergence traditionally observed in approximate policy iteration algorithms, where multiple observations are sampled with a given policy before updating (Alpaydin, 2014). The replay buffer effectively decorrelates observations from the simulation by storing previous experiences and only sampling a few observations from each run. Because samples are computationally costly to obtain, relatively small amounts of noise in the bootstrapped estimates of Q values can prevent the algorithm from converging in an acceptable number of iterations. We tested algorithm variants that record the predicted values of states and found improved stability, though this comes at the cost of rapidly updating the policy.

Algorithm 2 shows this baseline algorithmic approach. In our DQN implementation, the value function approximation is defined by the trained θ (i.e., the weights of

the neural network) which produces an estimate of the value of each parameterized action given a state and its corresponding features. Using similar notation to the CAVE approach, we iterate over $n = 1, 2, ..., N$ policy improvement loops to update θ and through the set of time steps $\mathcal{T}$ to generate the observations for these updates. While problems with longer trajectories that have thousands of time steps necessitate the use of a longer sequence length with many updates, the USAF WRP is primarily concerned with mapping the state-action values for states likely to be observed in the next several decades, observed as tens of time steps in our simulations. For this reason, each iteration simply observes one simulation length before performing an update. Iterating over $d^p \in \mathcal{D}^p$, we observe the next state S_{t+1} given each action and predict the corresponding current state value $Q(S_t, d^p \mid \theta^{\text{target}})$ and next state value $\max_{d^p \in \mathcal{D}^p} Q(S_{t+1}, d^p \mid \theta^{\text{target}})$. This information allows us to generate a new estimate of the value of the state-action pair $\hat{v}(d)$ adjusted for a learning rate α_n. These values are then added to the replay buffer in v^{buffer} while the corresponding features are recorded in S^{buffer}. We size these replay buffers such that they completely replace all observations every 10 iterations of n. When the simulation concludes, we sample the replay buffer to create a minibatch to update the trained network parameters θ, periodically updating the target network's parameters θ^{target} with the trained network every N^{Δ} updates. Finally, we establish a convergence criterion because the computation time can extend much longer than the alternative choice of algorithm, but stopping the algorithm while in the middle of a noisy training period can result in erratic estimated values and low quality solutions. If the last N^{Ω} estimates of the starting state vary less than V^{Ω}, then training stops, with N^{Ω} and V^{Ω} both being tunable hyperparameters. In our DQN implementation, the value function approximation θ is defined by the trained weights of the neural network which produce an estimate of the value of each parameterized action given a state and its corresponding

features.

3.5 Implementation, Results, and Policy Discussion

We seek to compare each of these solution approaches in terms of quality, computational effort, and robustness. While CAVE has relatively few hyperparameters compared to other algorithmic approaches, there are still several places to tune, as summarized in Table 10. Potentially the most important for this application is the length of the lookahead horizon. This length directly affects both the computational requirements, which increase as this quantity increases, and the quality of the solutions generated. Setting this horizon too short forces the algorithm to try to initiate all the necessary corrections for future years into a short timeframe before the policy reverts back to the equilibrium level. Solution quality faces a tradeoff with computational requirements, although the interactions between policies at different time steps may ameliorate or even reverse these effects on solution quality for shorter horizons. This is also true for the number of training iterations. The number of training iterations N was empirically tested at 100 and found to perform well. Two CAVE hyperparameters that affect the rate of convergence are the stepsize and initial update size (i.e., the gap between breakpoints). For both of these, we set the initial level relatively high and create a rule to steadily decrease the hyperparameter as training progresses. The stepsize rule is specifically developed and tested for the problem instance of interest, starting at 1 for the first 60 iterations, then linearly decreasing to 0.6. The reason for starting with the stepsize at 1 for an extended portion of training, effectively overwriting the current gradient with the observed gradient, is that this algorithm directly compares gradients between decisions. If one decision has recently been adjusted upwards, the algorithm will be updating a gradient that has not yet been updated from 0. Decisions that have already updated the gradient will appear

to be more valuable, creating a bias. This learning rate schedule is appropriate for the USAF WRP because the stochastic outcomes are not overly noisy. In general, the learning rate should be tailored for other problem instances. Finally, the initial interval starts at 8, then decreases by 50% every 15 iterations until reaching 1. This allows for rapid updates and fast movements early in the training, then smaller updates by iteration 45, and decreasing stepsizes starting at iteration 60.

Hyperparameter	Variable	Setting
Lookahead Horizon	T_π	5 years
Training Iterations	N	100
Stepsize Rule	α_n	$\begin{cases} 1 & \text{if } n \leq 60 \\ 1 - \frac{n-60}{N} & \text{otherwise} \end{cases}$
Initial Update Interval	δ_1	8

Table 10. CAVE Hyperparameter Settings

Table 11 reports the hyperparameter settings for our DQN algorithm. The DQN algorithm utilizes a neural network model for value function approximation. The architecture comprises 3 hidden layers with 600 neurons per layer as shown in Table 11. Heaton (2008) provides three rules as starting points for determining the number of neurons in hidden layers: the number of hidden layers should be between the size of the input and output layers, the number of hidden layers should be 2/3 the size of the input layer plus the output layer, and the number of hidden layers should be less than twice the size of the input layer. While the answer that meets all three recommendations is 942 neurons per hidden layer, Heaton (2008) makes these recommendations for neural networks in general, including shallower architectures. Because 3 layers is deeper than many feedforward neural networks and initial empirical testing showed that fewer neurons performed as well or better than this guideline, we reduce this number to 600. Because we need to produce many predictions in an iterative structure, we use Rectified Linear Unit (ReLU) activation functions to decrease the computational burden. The learning rate was empirically tested, and the algorithm

showed choppier (i.e., high variance) learning during preliminary testing at levels of 0.1 and 0.01 whereas a level of 0.005 resulted in a more consistent convergence. This architecture uses 6 outputs, so each potential action is sampled at every observation. This approach also uses an ϵ-greedy mechanism to select which observed next state the model transitions to and evaluates next. We select the batch size based on a single simulation of 50 years, run in parallel on 36 cores, resulting in 1800 observations being loaded into the replay buffer for each iteration. Finally, we require $N^{\Omega} = 40$ consecutive periods with the estimate of the value of the initial starting state varying no more than $V^{\Omega} = 5\%$ for the early stopping criterion. Decreasing this range parameter may increase solution quality, but preliminary testing showed increased computation times of 5-10 times longer.

Hyperparameter	Setting
Hidden Layers	3
Neurons per Hidden Layer	600
Activation Function	ReLU
Batch Size	1800
Neural Network Learning Rate	0.005
Stepsize (α_n)	Generalized Harmonic Stepsize, $a = 10$
Exploration Rate (ϵ)	0.2
Target Network Update Frequency (N^{Δ})	10
Stopping Criterion: Number of Estimates (N^{Ω})	40
Stopping Criterion: Maximum Range (V^{Ω})	5%

Table 11. DQN Hyperparameter Settings

To improve computational efficiency when generating observations, we ran multiple simulations in parallel for their complete length to observe the outcomes, then trained updates with a much larger batch size of 1800 and a proportionately higher learning rate of 0.005. Given this smaller number of higher-impact updates, we update the target network every 10 time steps and used a buffer sized such that each batch randomly sampled 10% of the buffer. To further stablize training, we also used 2-step bootstrapping (Sutton and Barto, 2018), which lets the algorithm view the

next two states and associated penalties instead of only one.

We compare the policies determined via CAVE and DQN approaches by running 48 simulations with a length of 30 years and recording the total discounted costs during each simulation. During these simulations, the CAVE approach dominated the other two approaches for both cost functions, as shown in Table 12. The computation times differed significantly, in part due to the different algorithmic approaches to training. Because the DQN algorithm seeks to train a value function approximation that is valid across multiple states, the algorithm trains for a longer period of time prior to simulating outcomes to measure the quality of the solution. Conversely, CAVE requires approximately 7.7 minutes to generate an accession policy, though it requires this time to generate each direct lookahead policy, so simulating over many years requires much more time. The DQN approach required more time upfront to train the neural network but was able to generate policies much more quickly during the simulation afterward. For operational use, the data used to train the DQN changes between each policy being generated, so the CAVE approach requires less computation. However, for algorithm testing purposes, the DQN approach requires less time to simulate many policies.

Cost Function	Model	Percent Discounted Cost Reduction versus Benchmark, 95% CI	Computation Time to Test (min)
Manning	CAVE	29.76 ± 1.04	7.7
	DQN	1.53 ± 1.14	314.9
Shortages	CAVE	17.38 ± 0.76	7.7
	DQN	4.45 ± 0.82	153.7

Table 12. Mean reduction in absolute aggregate prediction error on test dataset shows that CAVE outperforms both DQN and the benchmark for both potential cost functions.

The CAVE approach developed highly dynamic sets of policies, but the DQN approach selected the same parameter choice at each decision, set to 40% of existing shortages. While this DQN policy was consistent, the parameterized structure still

resulted in a relatively wide variety of accession decisions as AFSCs became better or worse manned over time. The level of 40% makes intuitive sense as a good setting and is remarkably close to the 35% setting chosen by subject matter experts. This policy selection may be due to the setting truly being the best possible setting at each time period for both cost functions, or it may reflect a lack of nuance to the learned value function approximation. The degree to which the CAVE algorithm improved solution quality compared to the DQN algorithm suggests that the policy parameterization approach, while a useful way to improve upon the benchmark, does not compete with higher quality optimization approaches such as our implementation of CAVE. Additional computation time may improve the policy quality slightly, but an alternative approach to searching the action space is necessary.

We tested to see if an approach that used a DQN and searched a small space around the given policy could develop an improved policy beyond the parameterized approach. This technique was able to develop statistically significant improvements to policies when searching around the benchmark but was unable to improve upon the parameterized policy developed here when limited to a computation time restriction of two orders of magnitude beyond the CAVE computation time.

Notably, the CAVE approach not only outperformed the other two approaches on average, it outperformed both competing approaches for every single simulation, as shown in Figure 28. This high-quality result is in addition to the lower computational burden required to generate a single CAVE policy. CAVE is the superlative performer with respect to solution quality, computational requirement, and robustness.

In addition to producing algorithms that can efficiently solve the workforce replenishment problem, a key insight for these approaches is determining how policies react to potential cost functions. Observing obviously incorrect policies can help inform cost function selection as a form of inverse reinforcement learning (Russell, 1998). We

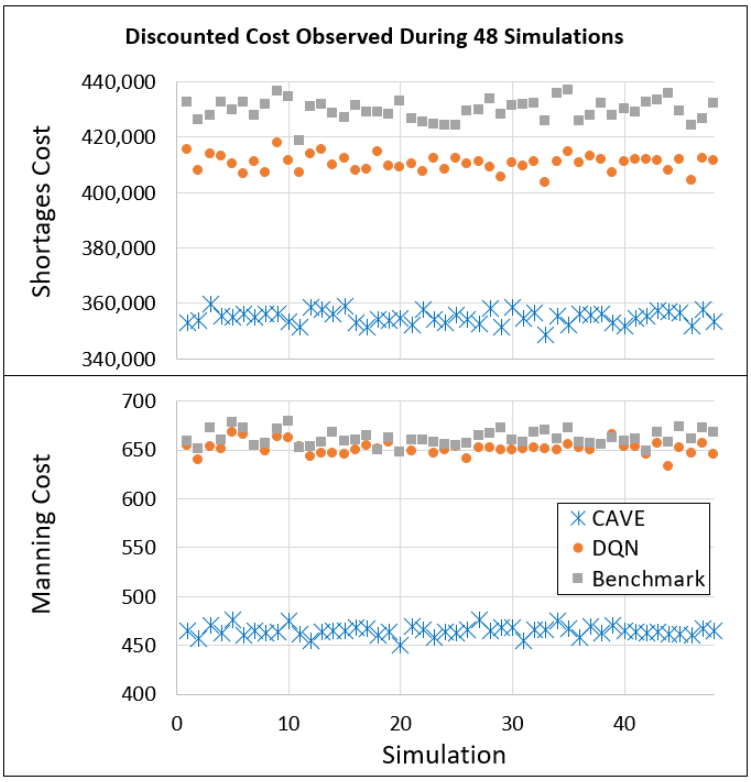

Figure 28. CAVE Consistently Outperforms both Benchmark and DQN Policies

see just such a case in Figure 29. For overmanned AFSCs, we see policies that make intuitive sense, where the applications of CAVE for each cost function reduce the accessions level compared to the benchmark policy. However, we see a strange occurrence for large, undermanned AFSCs. When finding good policies based on shortages (i.e., CAVE-S), we see the expected increase in accessions. However, when we use the manning cost function to choose policies (i.e., CAVE-M), we see the algorithm choose policies that actually reduce accessions, intentionally driving manning further down for these AFSCs. Upon further examination, this is actually a very good strategy to decrease the resulting penalties. These large AFSCs can be reduced by many accessions while only penalizing one AFSC, keeping a large number of AFSCs at or above 100% manning. This intentional undermanning is not inline with Congressional guidance to procure AFSC-specific talent, so this cost function is not suitable without

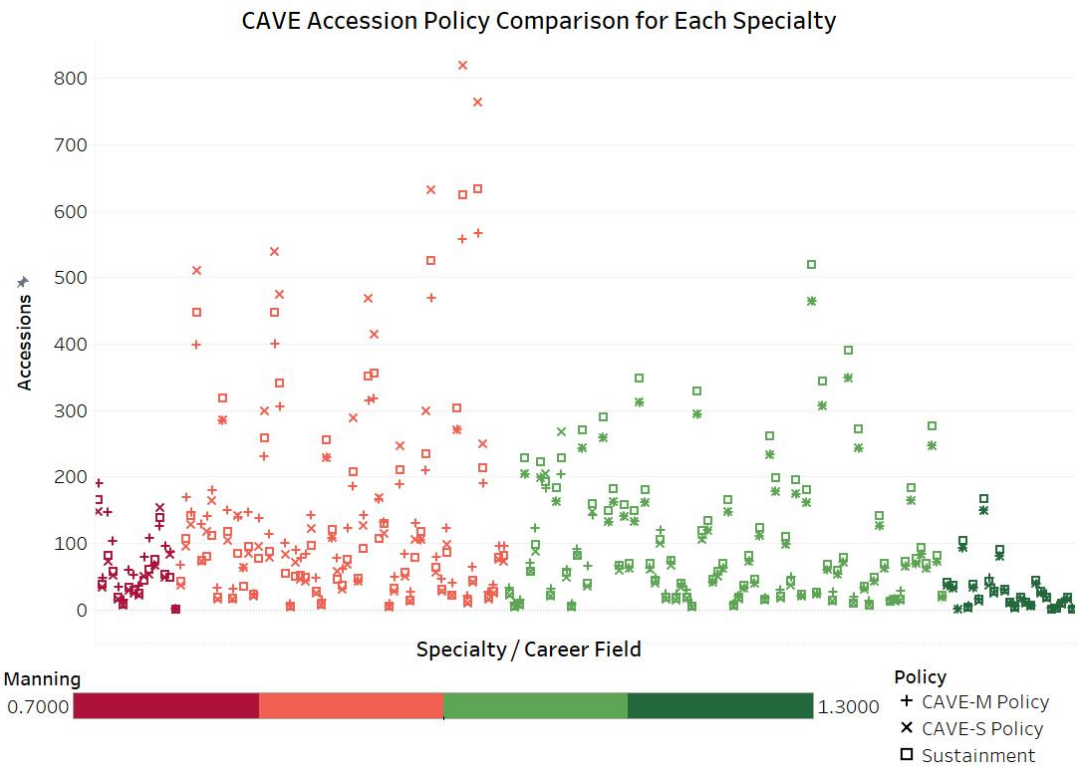

Figure 29. Policies Compared to Equilibrium Benchmark

weighting the AFSCs in some way to prevent this behavior, although a full weighting by AFSC size would yield the equivalent of the shortages cost function.

Figures 30 and 31 display the mean manning levels attained over time for selected AFSCs using the CAVE-M, CAVE-S, DQN, and benchmark sustainment policies. Examining the mean AFSC manning outcomes for AFSCs that rely only on accessions in Figure 30 yields some insight regarding how these solution approaches differ. In the top left panel, we observe that the manning levels generally converge close to 100% over time with the equilibrium policy. In the top right panel, the manning levels converge somewhat faster and remain close to 100% manning as stochastic outcomes that cause manning to drift are corrected by the DQN policy. In the bottom right panel, we see the CAVE approach when using the manning cost function. As previously observed in Figure 29, we see that the algorithm prioritizes some AFSCs

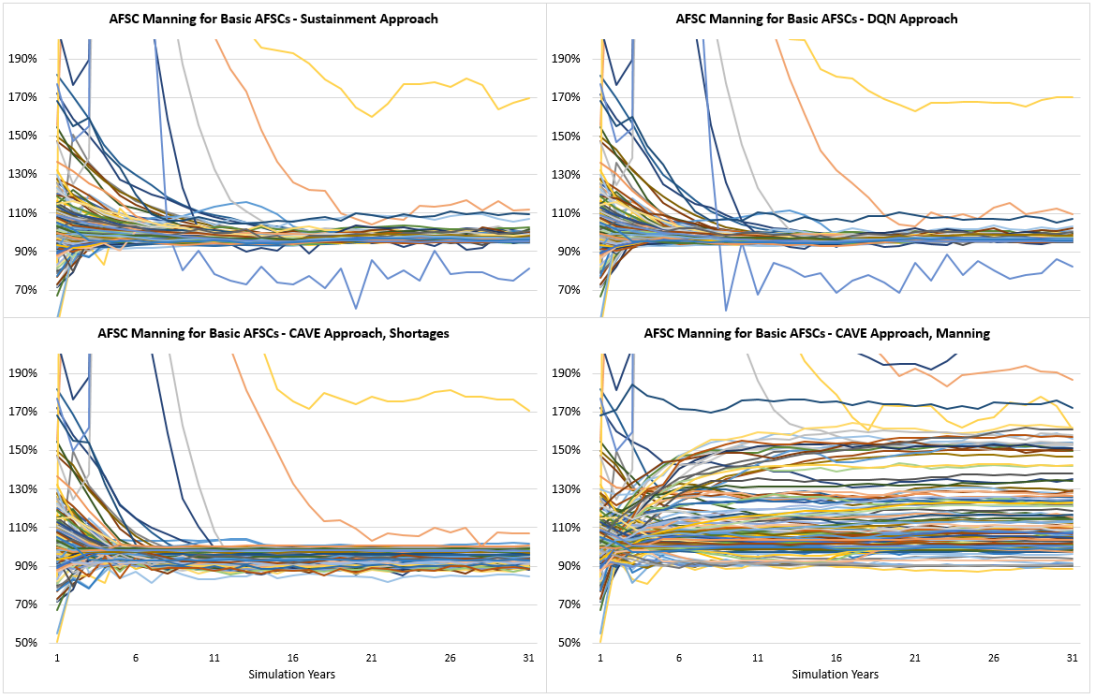

Figure 30. Mean AFSC Manning levels for basic AFSCs that rely entirely on new accessions

over others, resulting in a high variance of outcomes for AFSC manning. Finally, in the bottom left panel, we see that the CAVE approach using the shortages cost function acts more aggressively to close manning gaps early, but it displays more dispersed set of outcomes as some AFSCs are prioritized over others.

The shortages cost function also includes grade shortages and shortages in progression and lateral AFSCs, so quality of solution does not perfectly align with this measure when examining only the overall manning for AFSCs receiving accessions. We see one example of this in Figure 31, where the sustainment and DQN approaches result in increasing manning levels for lateral AFSCs. In the real world, high levels of manning would result in the modification of policies to avoid such outcomes because these overmanned AFSCs are causing undermanning elsewhere. The CAVE approach observes this outcome in measurements of the gradients for each accession and adjusts accession policies over time to avoid too many personnel working in AFSCs where

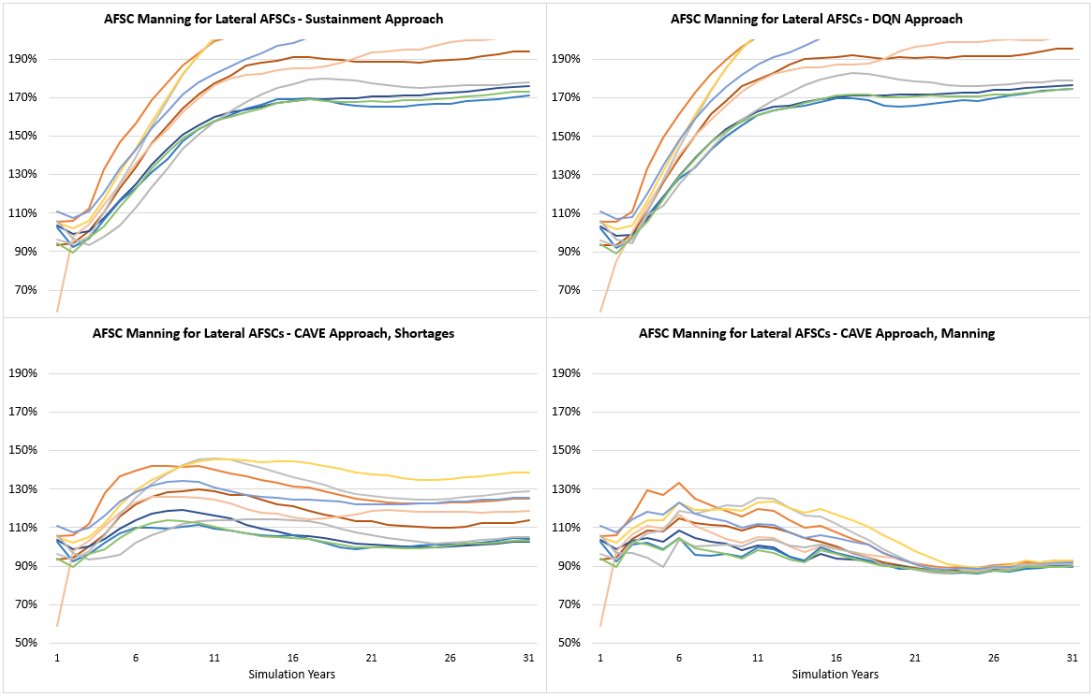

Figure 31. Mean AFSC Manning levels for lateral AFSCs that rely entirely on crosstraining

they are not needed.

3.6 Conclusions and Way Forward

This work develops a mathematical model of the USAF WRP and describes two solution approaches for determining enlisted accession policies. Both solution approaches perform better than the currently practiced benchmark equilibrium policy when tested using real USAF personnel data from 2015-2021. The CAVE approach outperformed the DQN approach in terms of solution quality and required computational resources. We recommend the use of CAVE for implementation by the research sponsor as a benchmark for future accessions planning and its inclusion as an addition to the current USAF Career Field Health modeling package. We tested two historically-accepted candidate cost functions and observed evidence suggesting that one was inappropriate for developing policies in its current form. This cost function testing indicates the robustness of the CAVE approach and its ability to determine high-quality policies for a range of potential cost functions.

Future work should expand the use of CAVE to other policies that exhibit concave structure for the underlying value function. The personnel policy space is considerably larger than just accessions, although these decisions remain of primary interest. One area that deserves further attention is the retraining policies that can supplement shortages in later years. The primary practical limitation of using algorithms to optimize this set of policies is the lack of a reliable set of constraints for these policies. In practice, USAF analysts generate recommendations and receive input from each AFSC's career field manager. Creating a reasonable simulation of future outcomes would require committing resources to recording which AFSCs can benefit from retraining and which are constrained to only cross-training personnel from other compatible AFSCs. Such AFSCs may have ratios of cross-training to direct accessions that need to be maintained as well as other restrictions. This set of constraints would require maintenance to generate consistently high-quality policies. Other potential

policies include those considering retraining out constraints and those targeting personnel to transition from the active duty force to the reserve components.

Future work should also expand the policy space to include the reserve components, including both the Air Force Reserve and the Air National Guard. Although some AFSCs may have a disparate impact on mission accomplishment, it is difficult to measure how much priority should be placed on these disparate impacts within the current problem construction. However, extending this work to include the reserve components would allow problem structures to directly address readiness, which is determined by a combination of personnel from different components, instead of just using AFSC and grade manning. This would enable the development of accession policies that would improve readiness beyond the current set of recommended policies.

One weakness of this approach is that the current approach to CAVE constructs a direct lookahead policy that can easily be trapped in a local optima. Adding a perturbation after convergence to a solution could improve solution quality at the expense of additional computational resources. Future work should also examine further refinements of CAVE under both static and dynamic requirements environments. While environments with dynamic requirements are inappropriate to use to directly train such algorithms without detailed validation of the requirements perturbation mechanism, selection of methodologies that perform well in such environments is entirely appropriate. Additionally, future work is needed to build and validate simulations of this dynamic requirements environment.

Although we use a simpler state space that does not allow for the complexity of some of the top-performing personnel retention models developed for the USAF (Hoecherl, Robbins, Borghetti and Hill, 2022; Pujats, 2020; Schofield et al., 2018), future work will explore including one or more economic parameters that can improve

the prediction quality in the short term without a large increase in the state space. In the long term, our confidence level in any economic prediction will be low, so we will return to an economically neutral forecast or a range of economic outcomes.

Although improved problem formulation and high-quality policy solutions are independently important, this cost also enables a significant business process realignment for manpower authorizations programming. Although authorizations change frequently, those changes result from decisions by senior USAF leaders and the US Congress. The future manning of these AFSCs and the warfighting ability generated by those personnel are vitally important planning considerations for choosing future manpower authorizations. Previous methods to simulate future outcomes required a lengthy, human-intensive process to generate realistic accessions policies, but this approach can provide high-quality policies rapidly enough to fit within current planning timelines without consuming scarce analytic resources. This approach offers the potential to tighten the relationship between the process to program authorizations and the corresponding personnel policies, informing higher quality decisions for both authorizations and personnel. When considering a candidate set of authorization changes, identifying constrained pipelines and shortages that cannot be closed enables senior leaders to select from three options:

1. Reduce the rate of required change for emerging requirements.

2. Find alternative offsets to allow human capital to be repurposed for the emerging requirement.

3. Apply required resources to relax the relevant constraint, allowing bottlenecks to be identified and removed during the planning process instead of waiting for the problem to manifest.

Such an approach would provide senior leaders and the US Congress the oppor-

tunity to make a decision about future human capital directly, rather than through the more roundabout process of making decisions about programmed levels of authorizations, with the hope that the personnel system will be able to deliver whatever has been programmed.

IV. SUPERCAVE: A Reinforcement Learning Approach for Integrating Workforce Replenishment Policies Across United States Air Force Regular and Reserve Components

4.1 Introduction

The United States Air Force (USAF) conducts operations through the use of complex systems and platforms requiring a high-skill workforce. The USAF's closed system, requiring senior personnel to be developed from junior personnel instead of hired from outside organizations, complicates the USAF's ability to ensure this workforce is appropriately recruited and trained. Furthermore, the US Congress and USAF senior leaders change the mix of required skillsets as missions and resourcing change over time (Hoecherl, Barger, Robbins and Zavislan, 2022). These factors cause decisions in one year to create significant long term consequences as the size of annual cohorts with varying levels of experience change based on accessions decisions at the time the cohort entered service. This decision-information structure necessitates solving a specific form of the closed workforce replenishment problem, using an approach appropriate for a problem with sequential decision-making under uncertainty.

In addition to the closed nature of the problem, workforce management is further complicated by the separation of personnel into multiple components of service, including the Regular Air Force (RegAF) consisting of active duty Airmen, the Air Force Reserve (AFR) consisting of reservists, and the Air National Guard (ANG) consisting of Airmen assigned to various states. Each of these components must compete for the same general pool of recruits, utilize many of the same training school resources for different skillsets, and rely on each other for specific mission sets in times of war, contingency operations, or emergencies. Additionally, many of the new personnel the AFR and ANG recruit are fully trained RegAF personnel departing active duty, attracted by increased stability or the lifestyle associated with the AFR

and ANG. While the components historically managed their accessions policies to recruit and train new personnel separately, the decreased size of the RegAF over time and increased costs of personnel necessitate a more holistic approach to maintain the effectiveness of the collective USAF personnel at an acceptable cost.

To address this problem, this research makes the following contributions to the workforce replenishment problem literature. First, we extend the benchmark equilibrium model for RegAF Air Force specialty codes (AFSCs) to the AFR and ANG, which do not currently have an enterprise-level model to provide as a benchmark. Second, we formulate this problem as a Markov decision process using realistic behavior developed from data stored in the USAF's Military Personnel Data System. This formulation includes existing accession policies as well as a policy lever to increase affiliations from the RegAF to the AFR and ANG over the baseline rate. Third, we demonstrate the efficiency of scaling a direct lookahead policy using Concave Adaptive Value Estimation (CAVE) to this larger problem set with 645 dimensions in the action space, over 3 times larger than previous applications. We test CAVE across multiple hyperparameters including lookahead horizon, training length, and stepsize rule, determining a superlative modeling structure for the problem instance of interest. Fourth, we develop and test a novel modification to the CAVE approach. Previous applications had no means to escape local optima created by the interactions between policies at different time steps. We propose the addition of a perturbation to developed policies with retraining to find improved policies, called Stochastic Use of Perturbations to Enhance Robustness of Concave Adaptive Value Estimation (SUPERCAVE). This approach develops improved solutions to the direct lookahead policy without the exponential increase in computing power typically required to exactly solve large, high-dimension sequential decision-making problems. We test this modification across two new hyperparameters, number of perturbations

and perturbation weighting scheme. Finally, we test the effect of including or excluding the additional affiliations policy lever on the quality of solutions available to provide additional insight to USAF decision-makers.

The remainder of this chapter is organized as follows. Section 4.2 describes the intercomponent USAF workforce replenishment problem. Section 4.3 reviews related work in the existing literature. Section 4.4 details our formulation of the problem as a Markov decision process, and Section 4.5 explains the optimization approaches we develop and test to find high-quality policies. Finally, Section 4.6 details the experimental designs and results, and Section 4.7 describes our conclusions and future work.

4.2 USAF Total Force Management

While many of the military recruiting challenges focus on identifying which personnel with specific characteristics, talents, and competencies should be recruited, the more basic question of how many personnel the USAF needs to enter each AFSC is a significant problem that must be solved prior to addressing these other pressing questions (Hoecherl, Barger, Robbins and Zavislan, 2022). When developing policies for determining the quantity of personnel to recruit, most private industry approaches to workforce replenishment prioritize meeting short term human capital needs because these organizations retain the ability to recruit more senior personnel at later stages of their career, correcting any problems created by earlier policies. For this reason, many of these approaches focus on immediate hiring decisions with shorter time horizons instead of considering decisions about the quantity of recruits with a long-term framing required in a closed workforce replenishment problem. Since many USAF skillsets require years of experience to fully mature, the ability to conduct operations in a given time period often depends on accession decisions made 3-10 years ago.

Most private sector problems are either relatively small compared to the military problem or decomposable into smaller problems, with workforce requirements determined at lower levels of the organization. Conversely, the military's plans to fund manpower positions, recorded as authorizations, require adjudication by the collective USAF corporate structure and approval by the US Congress, although commanders of major commands have the ability to make modifications to funded billets within certain constraints. Since the military provides a public good instead of pursuing a profit, the measurement of relative contributions from different allocations of human capital becomes much more difficult. This becomes relevant when considering that the military manages the total number of personnel, defined as end strength, to meet Congressional guidance and that the processes for recruiting new personnel or releasing existing personnel require significant coordination and involve multiple bureaucratic structures. Since end strength does not exceed the programmed level of manpower authorizations, excess personnel in one AFSC causes a shortage in another AFSC.

The RegAF, AFR, and ANG do not approach this problem in the same way. The RegAF manages skillsets and the associated accession decisions at an enterprise level because they have the ability to simply move personnel from one location to another to achieve the correct balance. The AFR and ANG rely on each wing to determine their training and recruiting requirements. Members of these components cannot be involuntarily relocated but may voluntarily move between locations. However, service members do not relocate at the scale needed to meet local imbalances given the lack of compelling incentives, their civilian employment, family considerations, and cultural affinity for remaining with one unit.

The AFR and ANG rely on a combination of personnel transitioning from the RegAF, called affiliations, and non-prior service accessions. However, the long term

trend of declining end strength in the RegAF has caused the number of affiliations to decrease over time, creating increased difficulties in achieving the correct skill mix (Hoecherl, Schulker, Hornberger and Walsh, 2022). While the AFR and ANG must still develop a more granular, location-specific accessions policy, this changing end strength dynamic necessitates enterprise modeling of skillsets first, integrated with RegAF modeling to allow a more holistic approach to managing personnel transitions through the different components. In addition, the USAF has traditionally managed affiliations from the RegAF to the AFR and ANG in an ad hoc manner. RegAF AFSC manning informs release of personnel, but the corresponding benefit to readiness or manning in the AFR and ANG is not considered. We propose the inclusion of intentional affiliations in addition to the current baseline volunteer rate.

The USAF's current modeling approach for RegAF personnel uses Kaplan Meier survival rates by completed years of service (YOS) to construct a Markov chain model for each AFSC. USAF analysts construct these rates based on longitudinal observations of personnel and their associated features in the Military Personnel Data System over a 5 year period. For basic AFSCs (i.e., those that only use crosstraining to correct manning problems), this model determines the equilibrium distribution of personnel by YOS sustained only by accessions. This approach also generates the equilibrium number of accessions, to maintain the AFSC at 100% manning in the aggregate, which provides a useful baseline for accession policies and resourcing decisions to determine training pipeline capacity. However, this approach has several weaknesses. First, it considers only retention within the original AFSC the airman enters, so the effects of accessions on other AFSCs that Airmen may transition to do not affect the required target. Some of these transitions are crossflows and may occur relatively infrequently, but some transitions reflect automatic AFSC changes as airmen progress from junior skill levels to more senior skill levels. In cases where progression AFSCs are too large

or small compared to their corresponding junior AFSCs, this imbalance can create overages or shortages, even in an equilibrium state. Second, this approach provides equilibrium policies, so any disproportionately large or small year group cohorts must age out of the system before it returns to full manning, potentially requiring more than a decade to fully restore AFSCs that are undermanned. Third, the USAF has only applied this enterprise level modeling to RegAF AFSCs. Because the AFR and ANG rely on many local decision-makers to choose accessions for their individual locations, past efforts have not generated enterprise-level models for career field health analysis. However, changing relative sizes between the RegAF, AFR, and ANG components necessitates a broader modeling approach to ensure adequate personnel. Senior leadership has expressed interest in using such a model to inform policy discussions and adjudication between decision-makers (Miller, 2017).

Given the rate at which authorizations change over time, an equilibrium policy is inappropriate. While USAF analysts generate highly customized policies to ensure AFSCs remain manned at appropriate levels, this policy construction is manual and very time-intensive, requiring multiple subject matter experts and detailed review. While some of the aspects of this review are not captured in the datasets available, improved benchmark policies are a valuable tool to reduce the required time to develop high-quality policies and improve the effects of anchoring bias toward low-quality policies.

4.3 Related Work

To develop good policies for the closed workforce replenishment problem, researchers have leveraged a number of approaches. Stochastic programming and goal programming have been applied to smaller workforce replenishment (WRP) problem instances, including for 33 broad specialties in the US Army (Gass et al., 1988) and

more detailed approaches to specialties in the US Army medical workforce (Bastian et al., 2015) and cyber workforce (Bastian et al., 2020). These approaches fail to scale to the size of the enlisted USAF problem while remaining computationally tractable, especially with transition rates that change based on the state variable to reflect the reality of USAF policies that guide retraining decisions. These dynamic transition rates require a dramatic increase in computational resources to apply goal programming or stochastic programming approaches. These approaches require policy or problem simplification to scale to the full USAF WRP.

The use of conventional dynamic programming to find optimal solutions to workforce problems exhibiting this structure cannot scale well to large problem instances due to the curse of dimensionality (Powell, 2011). Approximate dynamic programming provides a means to develop high-quality solutions to problems with this structure, including for the larger question of end strength management (Situ, 2018).

Because the workforce replenishment problem is a question of how to allocate a scarce resource (i.e., accessions) across different specialties, approximate dynamic programming algorithms designed to take advantage of structure in resource allocation problems are appropriate. Godfrey and Powell (2001) developed Concave Adaptive Value Estimation (CAVE) as a way to construct a piecewise linear value function approximation to solve the newsvendor problem, showing that it scales to high-dimension action spaces (Godfrey and Powell, 2002*a*). This approach leverages the concave nature of the value of additional accessions to efficiently update the value function approximation. The authors later extended this work to the multiperiod newsvendor problem (Godfrey and Powell, 2002*b*), and Topaloglu and Powell (2003) showed that CAVE converges to optimality under specified conditions. Researchers later extended CAVE to a multiperiod inventory control problem with backlogged demands (Kunnumkal and Topaloglu, 2008) and energy storage problems (Salas and

Powell, 2018).

Like the newsvendor and energy storage problems, the WRP exhibits concave structure, where each additional accession faces a decreasing probability of addressing an unmet demand. However, a difficulty arises when considering how to apply CAVE to the workforce replenishment problem. Unlike the newsvendor and energy storage problems, the workforce replenishment problem secures a resource that can meet multiple demands over time, has a stochastic survival rate, can meet different demands based on the length of survival, and is not consumed by this demand. There have been approximate algorithms that leverage concave structure in this problem to workforce planning for limited problem sizes (Song and Huang, 2008). Hoecherl et al. (2016) extended the CAVE approach to the workforce replenishment problem by constructing a direct lookahead policy with an assumed equilibrium policy beyond the lookahead horizon. Hoecherl and Robbins (2022) refined this further, modifying the problem structure to match USAF business processes and testing candidate contribution functions. However, this approach demonstrates a weakness: the algorithm leverages the same future simulation of shortages to inform simultaneous gradient updates for policies in multiple timesteps. Although this approach has been shown to converge to high-quality policies, this approach is not robust to the interactions between decisions in different time steps because accessions in multiple time steps can fill the same future needs.

4.4 Markov Decision Process Formulation and Simulation

To develop and assess high-quality policies, we first formulate the USAF WRP as a Markov decision process. This formulation uses a finite time horizon with annual time steps defined as

$$t \in \mathcal{T} = \{0, 1, ..., T\}, \tag{19}$$

where $\mathcal{T}$ is the set of all timesteps through the end of the finite horizon T.

4.4.1 State Variables

We define the state variable using the number of personnel $S_{t,c,a,y} \in \mathbb{Z}_0^+$ with each combination of component (c), AFSC (a), and YOS (y) at each time t.
The state at time t is then compactly defined as

$$S_t = (S_{t,c,a,y})_{c \in \mathcal{C}, a \in \mathcal{A}, y \in \mathcal{Y}}, \tag{20}$$

where $\mathcal{C}$ is the set of components, $\mathcal{A}$ is the set of all AFSCs or skillsets, and $\mathcal{Y} = \{0, 1, ..., Y\}$ is the set of all YOS with Y being the maximum career length.

4.4.2 Problem Parameters

The initial state S_0 includes fixed problem parameters that represent important and unchanging features of the problem. Included in this set of problem parameters is the sum of the programmed requirements $m_{t,c,a,g} \in \mathbb{Z}_0^+$ for each component $c \in \mathcal{C}$, AFSC $a \in \mathcal{A}$, and grade $g \in \mathcal{G}$, where $\mathcal{G} = \{1, 2, ..., G\}$, with G total grades, at time t. Let the authorizations at time t be compactly represented by

$$m_t = (m_{t,c,a,g})_{c \in \mathcal{C}, a \in \mathcal{A}, g \in \mathcal{G}}. \tag{21}$$

Importantly, authorizations do not specify the YOS of the required Airmen, relying on the grade variable to ensure the Airmen have the appropriate competencies and experience.

4.4.3 Decision Variables

At each time step, the USAF must decide a set of decisions d_a for accessions, a set of decisions d_r for personnel affiliating to the AFR, and a set of decisions d_g for personnel affiliating to the ANG, with the decision class defined as $d \in \mathcal{D} = \{d_a, d_r, d_g\}$, where $\mathcal{D}$ is the set of all decision classes. In addition, not every AFSC requires a corresponding decision of each class. While most AFSCs rely on accessions to replenish their personnel, some rely on service members progressing from a more junior AFSC, and some rely entirely on crosstraining from other AFSCs due to a need for maturity or general military experience. This categorization of AFSCs varies by component, and some AFSCs do not exist in certain components. For AFSCs that rely on accessions in each component c, we define the subset of AFSCs that require an accession decision as $\mathcal{A}'_{c,d_a} \subset \mathcal{A}$. In addition, we define the collection of AFSCs that require an affiliation decision as $\mathcal{A}'_{c,d_r} \subset \mathcal{A}$ for affiliations from component c to the AFR and $\mathcal{A}'_{c,d_g} \subset \mathcal{A}$ for affiliations from component c to the ANG. The decision at each time step is then defined as

$$x_t = (x_{t,c,a,d} \in \mathbb{Z}_0^+)_{c \in \mathcal{C}, a \in \mathcal{A}'_{c,d}, d \in \mathcal{D}}. \tag{22}$$

Whereas the affiliation decision is constrained only by the number of personnel in the AFSC at time t, the collective accession decisions are constrained by the need to manage total end strength within each component according to Congressional guidance. Although the services have some flexibility to deviate from these targets, the cost to do so even by relatively small margins is prohibitively expensive because of the high cost of personnel. Each year, an aggregate retention model predicts the total number of personnel who will retain for the next several years, generating a series of annual aggregate accession constraints for each component, denoted as $A_{t,c}$. The accessions for each individual AFSC must then sum to this total for each component

in the given year. This constraint is defined as

$$\sum_{a \in \mathcal{A}'_c} x_{t,c,a,d_a} = A_{t,c} \quad \forall c \in \mathcal{C}, t \in \mathcal{T}. \tag{23}$$

In addition to the aggregate constraint for accessions, the number of personnel accessed in each AFSC must also remain within pipeline constraints defined by Air Education and Training Command. Many training pipelines require specialized equipment, instructors, classroom or dormitory space, agreements with other services for shared pipelines, or significant coordination between complementary pipelines to ensure the correct training opportunities are available. Let

$$\eta^-_{t,c,a} \leq x_{t,c,a,d_a} \leq \eta^+_{t,c,a} \quad \forall c \in \mathcal{C}, a \in \mathcal{A}'_{c,d_a}, t \in \mathcal{T}, \tag{24}$$

wherein $\eta^-_{t,c,a}$ and $\eta^+_{t,c,a}$ represent the respective lower and upper pipeline constraints for each accession decision. These constraints limit possible actions at time t so that $x_t \in \mathcal{X}_t$, where $\mathcal{X}_t$ is the subset of feasible actions that meet both the aggregate and combined AFSC- and component-specific constraints.

4.4.4 System Transition

We model the transition during a single annual time step from a given state to a future state using a system transition function. This transition function uses the existing state, a selected action x_t, and the observation of the exogenous information discovered during the transition to simulate the composite result of each set of transitions shown in Table 13. Let

$$S_{t+1} = S^M(S_t, x_t, \omega_{t+1}), \tag{25}$$

wherein $\omega_{t+1} \in \Omega$ is the exogenous information discovered during the transition and

Ω represents all possible outcomes.

Potential Outcome	Probability Calculation	Distribution	AFSC-Specific Destination
Remain in AFSC	P(Stay $\mid c, a, y, \frac{S_{c,a}}{m_{c,a}}$)	Binomial	Yes
Progress	P(Progress $\mid c, a, y$, Not Remain)	Binomial	Yes
Crosstrain Out	P(Cross Out $\mid c, a, y$, Not Remain, Not Progress, $\frac{S_{c,a}}{m_{c,a}}$)	Binomial	No
Baseline Affiliation to AFR or ANG	P(Cross to $c' \mid c, a, y$, Not Remain, Not Progress, Not Crosstrain)	Binomial	Yes
Gain to System	P(Gain in YOS $y \mid c, a, X_{t,c,a}$)	Multinomial	Yes
RegAF Crosstrain In	P(Cross to AFSC $a' \mid c$, Cross Out, $\frac{S_{c,a'}}{m_{c,a'}}$)	Uniform	Yes
AFR, ANG Crosstrain In	P(Cross to AFSC $a' \mid c, a, y$, Cross Out)	Multinomial	Yes
Complete YOS	P($y+1 \mid c, a, y$, Remain or Cross or Progress or Affiliate)	Binomial	No
Depart System	P(Loss $\mid c, a, y$, Not Remain, Not Cross, Not Progress, Not Affiliate)=1	Fully Determined	No

Table 13. Potential State Transitions

The transition to the next state first determines the number of Airmen with a given component c, AFSC a, and YOS y who will remain in the same component and AFSC using a binomial distribution. While approaches have been developed to generate higher-quality estimates of retention behavior (Hoecherl, Robbins, Borghetti and Hill, 2022), we constrain our retention model to the information contained within the state variable to preserve the model's Markovian property. Next, the transition sequentially determines progression, crosstraining out, and affiliation transitions for the remaining personnel with each combination of features. Notably, the transition rate for remaining in or cross-training out of an AFSC is conditioned not only on the observed rate, but on the aggregate manning level of the AFSC, defined as $\frac{S_{t,c,a}}{m_{t,c,a}}$, where $S_{t,c,a} = \sum_{y \in \mathcal{Y}} S_{t,c,a,y}$ is the aggregate number of personnel in AFSC a and component c at time t and $m_{t,c,a} = \sum_{g \in \mathcal{G}} m_{t,c,a,g}$ is the aggregate number of authorizations for AFSC a and component c at time t. For personnel who have cross-trained out from AFSCs in the RegAF, the next transition is determined by a uniform random draw based on open retraining quotas for AFSCs. This approach reflects the existence of enterprise level policy development for the RegAF with significant involvement from USAF policy analysts and career field managers as well as the ability to move individuals to locations that match their new specialty. Conversely, AFR and ANG cross-training is not managed at the enterprise level, so base-level openings may allow cross-training out of specialties that are undermanned in the aggregate and into

specialties that are overmanned in the aggregate. Given this complexity, historical transition rates are treated as the most likely estimate of future behavior, and we generate a multinomial distribution of destination AFSCs using historically observed transition rates. Any personnel who started in a given component c, AFSC a, and YOS y at time step t but did not remain in the AFSC, crosstrain to a new AFSC, or affiliate to a new component deterministically transition out of the system as a loss.

While RegAF personnel who remain in the service reliably complete one YOS each year and progress to the next, the completion of each YOS used to calculate pay for the AFR and ANG depends on the service member's duty status during the year. For this reason, personnel transition to either the next YOS or remain in their current YOS according to a binomial distribution.

4.4.5 Cost Function

While the ideal state is for the inventory of personnel to perfectly match the authorized number of personnel by AFSC and grade, our state variable does not include grade. Modeling USAF personnel inventories with grade is difficult because promotion policies are modified each year both in the aggregate and by AFSC. Replicating these business processes requires both significantly increased computation as well as the development of complex rulesets to generate transition rates that replicate system behavior. A more stable approach is simply to model inventory with component, AFSC, and YOS as $S_{t,c,a,y} \in \mathbb{Z}_0^+$ and then calculate an expected grade inventory $S_{t,c,a,y,g} \in \mathbb{R}_0^+$, since the relationship between YOS and grade is relatively stable. Let

$$S_{t,c,a,y,g} = S_{t,c,a,y} P(g|c,a,y) \quad \forall c \in \mathcal{C}, a \in \mathcal{A}, y \in \mathcal{Y}, g \in \mathcal{G}, t \in \mathcal{T}, \tag{26}$$

wherein $P(g|c,a,y)$ indicates the historically observed probability of an airman being

in grade g given the airman is in component c, AFSC a, and YOS y.

Despite the simplicity of measuring whether inventory matches authorizations, measuring the relative goodness of states that do not perfectly match authorizations poses a more difficult problem. The first complicating factor is the comparison of inventory to authorizations in the aggregate and by grade. Meeting aggregate authorizations is an important consideration, as even having more junior or senior personnel than authorized is preferable to having no personnel to fill an authorization. However, personnel are not all equally capable of executing tasks and leading other personnel. For this reason, including an assessment of whether the correct number of personnel is available in each AFSC and grade is important. We use both, but weight aggregate measures by $2|\mathcal{G}|$ to emphasize the importance of the aggregate number of personnel in the AFSC and offset the larger number of measures for each grade.

In Chapter III, we observe that using the ratio of inventory to authorizations as a means to measure shortfalls is inappropriate because the resulting policies will sacrifice a few large AFSCs to preserve healthy manning in a large number of small AFSCs. However, using the actual number of shortages for each combination of features misses the large relative importance of shortages in small AFSCs who may be unable to adapt to missing personnel in the same way that a large AFSC can. For this reason, we develop a hybrid approach where the costs for manning shortfalls are weighted by a variable $\kappa_{c,m}$ such that manning shortfalls are 1/3rd the magnitude of the cost for shortages in each component based on the starting state S_0. Additionally, the RegAF is both much larger and acts as a donor for the other components. To ensure that the holistic approach developed here does not undercut the effectiveness of the RegAF, we further weight each component's costs by κ_c such that the RegAF contributes twice as much to the cost function as the other two components. This yields the following cost function:

$$C_t(S_t) = \sum_{c\in\mathcal{C}} \kappa_c \Bigg[\sum_{a\in\mathcal{A}} \big[2|\mathcal{G}| \max\big(m_{t,c,a} - S_{t,c,a}, 0\big) + \sum_{g\in\mathcal{G}} \max\big(m_{t,c,a,g} - S_{t,c,a,g}, 0\big)\big] + \kappa_{c,m} \sum_{a\in\mathcal{A}} \big[2|\mathcal{G}| \max\big(1 - \frac{S_{t,c,a}}{m_{t,c,a}}, 0\big) + \sum_{g\in\mathcal{G}} \max\big(1 - \frac{S_{t,c,a,g}}{m_{t,c,a,g}}, 0\big)\big] \Bigg]. \tag{27}$$

4.4.6 Objective Function

We define the objective of the Markov decision process as

$$\min_{\pi\in\Pi} \Big(\mathbb{E}^{\pi} \Big[\sum_{t\in\mathcal{T}} \gamma^{t-1} C_t(S_t) \Big] \Big). \tag{28}$$

where γ is the discount factor. The transition from S_t to S_{t+1} proceeds according to the transition function $S_{t+1} = S^M(S_t, x_t, \omega_{t+1})$. The decision x_t is chosen using the decision function $x_t = X_t^{\pi}(S_t \mid \theta)$ where θ is the set of estimated parameters for a value function approximation. The policy $\pi \in \Pi$, where Π is the collection of all possible policies, is the associated policy for a given θ.

4.4.7 Selected Parameters for the Intercomponent USAF WRP

We proceed by specifying particular parameter values for this generalized Markov decision process formulation to represent the specific system behavior of the USAF WRP. First, we use data from the USAF's Military Personnel Data System from September 2016 through September 2021 to measure the 5 years of transition rates and the starting personnel inventory as of September 2021, the beginning of fiscal year 2022 (MilPDS Dataset, 2021). In addition, we also record manpower authorizations from the Manpower Programming and Execution System - Unit Manpower Document for the next 5 years as of September 2021 (MPES-UMD Dataset, 2021). We treat the

authorizations for future years as maintaining the level of authorizations in the final year in this dataset, as this is the best estimate available for the magnitude of future authorizations.

It is important to strike a delicate balance when selecting a discount factor. We must set the discount factor low enough to reflect senior leaders' demonstrated urgency for addressing shortages in a timely manner and the real world uncertainty of future authorizations. Conversely, we must set the discount factor high enough to ensure that the algorithm does not generate policies that solve problems in the short term at the expense of poor outcomes in the future given the importance of long term impacts of personnel policies on national security. Setting $\gamma = 0.8$, we select a horizon length $T = 20$ years for this problem instance based on the selected discount factor, where the cumulative discount factor at the end of the time horizon $\gamma^T \approx 0.01$.

With 3 components (i.e., RegAF, AFR, and ANG), 236 AFSCs, and a maximum career length $Y = 40$ years, the dimensionality of the state variable is $21,240$. We show the dimensionality of each decision category for the USAF problem instance in Table 14. Baseline affiliation rates and affiliation decisions are only calculated for RegAF personnel because we do not model flows between the AFR and ANG or back to the RegAF. These additional flows are relatively small in magnitude but require significant additional computational burden to implement. Moreover, correct measurement of these flows necessitates additional data cleaning from an additional system, the Defense Civilian Personnel Data System.

We denote the equilibrium policy for each component and AFSC as $e_{c,a} \quad \forall c \in \mathcal{C}, a \in \mathcal{A}'_{c,d_a}$. In addition, we use $e_{t,c,a} \quad \forall t \in \mathcal{T}, c \in \mathcal{C}, a \in \mathcal{A}'_{c,d_a}$ to describe the modified equilibrium policy that complies with aggregate accession constraints. We set the AFSC-specific pipeline constraints for the RegAF at a default level based on a range above and below the equilibrium policy level, where $\eta^-_{t,c,a} = 0.75 e_{c,a}$ for the

Policy	Dimensions
RegAF Enlisted Accessions	186
AFR Enlisted Accessions	129
ANG Enlisted Accessions	144
RegAF Affiliations to AFR	186
RegAF Affiliations to ANG	186
Total	831

Table 14. Expanded Policy Set

lower bound and $\eta^{+}_{t,c,a} = 1.5e_{c,a}$ for the upper bound, respectively. For the AFR and ANG, we set the lower bound to $\eta^{-}_{t,c,a} = 0.35e_{c,a}$ to reflect the smaller population and lack of enterprise-level modeling of skillsets. However, operational applications would modify these to reflect the actual constraints recorded in Air Education and Training Command's Business Reporting and Intelligence Tool. Finally, when modeling grade, we combine E-1, E-2, and E-3 ranks because of enlistment contract structures that allow some personnel to enter directly as an E-3. We also combine E-8 and E-9 because of their low numbers and nuanced management practices, setting $G = 6$.

4.5 Optimization Approach

4.5.1 Baseline CAVE adapted to USAF WRP

In Chapter III, we demonstrate the relative efficacy of the CAVE approach for a smaller form of the USAF WRP compared to other reinforcement learning approaches, but for a smaller form of the problem with only the RegAF and the associated accession decisions for RegAF AFSCs. We adapt this implementation of CAVE to the larger intercomponent USAF WRP with three components and a second class of decisions.

Unlike model-free forms of value function approximation that must directly estimate the value of different states, CAVE instead constructs a direct lookahead policy for the next T_π years and estimates the gradient of the value function for decisions

during that period to find a high-quality solution. One limitation for direct lookahead policies is that they must be able to represent costs far enough into the future to avoid becoming myopic. One approach to address this concern is to extend T_π to the end of a finite horizon problem, but this requires significant computational resources for problems with long horizons. Alternately, one can either create a different model of future actions after the lookahead horizon T_π or develop an estimate of the value of the state at the end of the lookahead horizon. In the USAF WRP, we can use the equilibrium policy as a reasonable approximation of future policies.

A second limitation is that the gradient for a stochastic problem like the USAF WRP depends both on the future state outcomes such as manning or shortages as well as the survival of personnel to contribute to those outcomes. For example, the addition of one accession in time period t may solve a shortage in the same AFSC a at time $t + 7$, or the additional accession may solve a shortage in a different AFSC a' after crosstraining at time $t + 7$, or the additional accession may depart the service before 7 time periods have passed and fill no shortages. We overcome this limitation by calculating a survival probability for each decision to each respective component, AFSC, and YOS combination after t timesteps have passed. For accession decisions, this survival calculation simply weights the probability of survival to meet demands. For affiliation decisions, two survival probabilities must be calculated: the probability of meeting an authorization after transferring to the new component and the probability that the respective service member affiliating would have met an unfilled authorization by remaining in their original component. Because these transition probabilities can vary according to AFSC manning, there is no single set of survival probabilities for a given decision. We use the baseline, unrestricted transition probabilities for crossflows out of AFSCs as a close approximation and the restricted crossflow in rates for AFSCs who are intended to meet their authorizations with

accessions and only use retraining-in as a corrective measure. This prevents the algorithm from intentionally deciding to fill authorizations in one of these AFSCs with crossflows from another AFSC, counter to the intention of the career field manager and USAF policy analysts. This approach has been validated by Hoecherl et al. (2016) and Hoecherl and Robbins (2022), contributing to the development of high-quality policies.

In this CAVE variant, shown in Algorithm 3, we express the parameters of the piecewise linear value function approximation as

$$\theta = (u_t, \nu_t)_{t \in \mathcal{T}_\pi}, \tag{29}$$

wherein $u_t = (u_{t,c,a,d})_{t \leq T_\pi, c \in \mathcal{C}, d \in \mathcal{D}, a \in \mathcal{A}'_{c,d}}$, with $u_{t,c,a,d}$ being a vector of breakpoints for a specific decision, $\nu_t = (u_{t,c,a,d})_{\forall t \leq T_\pi, c \in \mathcal{C}, d \in \mathcal{D}, a \in \mathcal{A}'_{c,d}}$, with $\nu_{t,c,a,d}$ being a corresponding vector of gradients for a specific decision, and $\mathcal{T}_\pi = \{1, 2, ..., T_\pi\}$. The variable $k_{t,c,a,d}$ represents the selected breakpoint for a specific decision, where $x_{t,c,a,d} = u_{t,c,a,d}^{k_{t,c,a,d}}$.

In Step 1, the CAVE approach begins by defining each accessions policy using two breakpoints, 0 and the equilibrium policy $e_{t,c,a}$ constrained by $A_{t,c}$. CAVE selects a set of decisions on how to distribute a total number of accessions $A_{t,c}$ for a given $t \in \mathcal{T}_\pi$ and $c \in \mathcal{C}$ by iteratively identifying the AFSC $a^+ \in \mathcal{A}'_{c,d_a}$ with the highest gradient $\nu_{t,c,a,d_a}^{k_{t,c,a,d_a}}$, incrementing k_{t,c,a,d_a} to move to the next breakpoint for that decision until reaching the accessions constraint $\eta^+_{t,c,a}$. During initialization, we accomplish this by setting a small positive gradient at ν^1_{t,c,a,d_a} for accession decisions. For affiliation decisions, we start with 0 additional affiliations. Because the process to find a good policy for each $t \in \mathcal{T}_\pi$ and $c \in \mathcal{C}$ for these decision classes simply increases the decision for each AFSC a until we reach a $\nu_{t,c,a,d}^{k_{t,c,a,d}} \leq 0$, we set $\nu^1_{t,c,a,d} = 0$ during initialization.

In Steps 2 and 3, we select our decisions x_t at each time step t according to the decision function $X^\pi_t(S_t \mid \theta)$, where θ is defined as $(u_t, \nu_t)_{t \in \mathcal{T}_\pi}$. For accession decisions,

Algorithm 3 CAVE Algorithm

Step 1: Initialization

Identify $A_{t,c}$, the aggregate constraint for accessions $\forall t \leq T_\pi$, where T_π is the desired length of the lookahead policy before reverting to equilibrium policy.

for $t \in \mathcal{T}_\pi, c \in \mathcal{C}, d = d_a, a \in \mathcal{A}'_{c,d_a}$ **do**

  To model each accession decision x_{t,c,a,d_a}, let $k_{t,c,a,d_a} = 2$, $\nu^1_{t,c,a,d_a} = 0.0001$, $\nu^2_{t,c,a,d_a} = 0$, $u^1_{t,c,a,d_a} = 0, u^2_{t,c,a,d_a} = e_{t,c,a}$, where $e_{t,c,a}$ is the equilibrium policy s.t. $\sum_{a \in \mathcal{A}'_c} e_{t,c,a} = A_{t,c}$.

end for

for $t \in \mathcal{T}_\pi, c \in \mathcal{C}, d \in \{d_r, d_g\}, a \in \mathcal{A}'_{c,d}$ **do**

  To model each affiliation decision $x_{t,c,a,d}$, let $k_{t,c,a,d} = 1$, $\nu^1_{t,c,a,d} = 0$, $u^1_{t,c,a,d} = 0$.

end for

Initialize parameters $\delta_{n,d}$ and α_n.

for n = 1 **to** N **do**

  Step 2: Determine current policy $X^\pi_t(S_t \mid \theta)$

  for $c \in \mathcal{C}, t \in \mathcal{T}_\pi$ **do**

    Initialize policy with $x_{t,c,a,d_a} = 0$ by setting $k_{t,c,a,d_a} = 1 \quad \forall a \in \mathcal{A}'_{c,d}$

    while $\sum_{a \in \mathcal{A}'_c} x_{t,c,a,d_a} < A_{t,c}$ **do**

      Select AFSC a^+ with largest estimated gradient $\text{argmax}_{a \in \mathcal{A}'_{c,d}} (\nu^{k_{t,c,a,d_a}}_{t,c,a,d_a})$

      Increment the decision x_{t,c,a^+,d_a} by setting $k_{t,c,a^+,d_a} = k_{t,c,a^+,d_a} + 1$.

    end while

  end for

  Step 3: Identify Current Affiliations Policy

  for $c =$RegAF, $d \in \{d_r, d_g\}, t \in \mathcal{T}_\pi, a \in \mathcal{A}'_{c,d}$ **do**

    Initialize policy with $x_{t,c,a,d} = 0$ by setting $k_{t,c,a,d} = 1 \quad \forall a \in \mathcal{A}'_c$

    while $\nu^{k_{t,c,a,d}}_{t,c,a,d} > 0$ **do**

      Increment decision $x_{t,c,a,d}$ by setting $k_{t,c,a,d} = k_{t,c,a,d} + 1$

      Increment aggregate accessions for the donor component $A_{t,c}$ by 1

      Decrement aggregate accessions for the receiving component $A_{t,c'}$ by 1

    end while

  end for

Step 4: Collect Gradient Information

Simultaneously sample the gradients $\Delta^-_{t,c,a}(X_{t,c,a},\omega)$ and $\Delta^+_{t,c,a}(X_{t,c,a},\omega)$ over a finite time horizon with random outcomes $\omega \in \Omega \quad \forall t \leq T_\pi, a \in \mathcal{A}'_c$

Step 5: Define Smoothing Interval

Let $k^-_{t,c,a,d} =$
$\min\{k_{t,c,a,d} \in \mathcal{K}_{t,c,a,d} : \nu^{k_{t,c,a,d}}_{t,c,a,d} \leq (1-\alpha_n)\nu^{k_{t,c,a,d}+1}_{t,c,a,d} + \alpha_n \Delta^-_{t,c,a,d}(x_{t,c,a,d},\omega)\}$.

Let $k^+_{t,c,a,d} =$
$\max\{k_{t,c,a,d} \in \mathcal{K}_{t,c,a,d} : (1-\alpha_n)\nu^{k_{t,c,a,d}-1}_{t,c,a,d} + \alpha_n \Delta^+_{t,c,a,d}(X_{t,c,a,d},\omega) \leq \nu^{k_{t,c,a,d}}_{t,c,a,d}\}$.

Define the smoothing interval
$Q_{t,c,a,d} = \left[\max\{x_{t,c,a,d} - \delta_{n,d}, u^{k^-_{t,c,a,d}}_{t,c,a,d}, \eta^-_{t,c,a}\}, \min\{x_{t,c,a,d} + \delta_{n,d}, u^{k^+_{t,c,a,d}+1}_{t,c,a,d}, \eta^+_{t,c,a}\}\right)$.

Create new breakpoints at the endpoints of $Q_{t,c,a,d}$ as needed. Since a new breakpoint always divides an existing segment, the segment slopes on both sides of the new breakpoint are the same initially.

Step 6: Update Estimates

for each segment k in the interval $Q_{t,c,a,d}$ **do**
  if $u^{k_{t,c,a,d}}_{t,c,a,d} < x_{t,c,a,d}$ **then** $\Delta_{t,c,a,d} = \Delta^-_{t,c,a,d}(x_{t,c,a,d},\omega)$
  else $\Delta_{t,c,a,d} = \Delta^+_{t,c,a,d}(x_{t,c,a,d},\omega)$
  end if
  Update the slope $\nu^k_{t,c,a,d} = \alpha_n \Delta_{t,c,a,d} + (1-\alpha_n)\nu^k$.
end for
Adjust $\delta_{n+1,d}$ and α_{n+1} according to step size rules.
end for
End

we iteratively increase accessions for each component and time period $t \in \mathcal{T}_\pi$ in the AFSC with the highest gradient at the current breakpoint until the accessions constraint is met. For affiliation decisions, we increase the number of affiliations from each RegAF AFSC until the observed gradient at the current breakpoint falls to zero or below.

After the piecewise linear value function has been initialized and the current decision has been identified, the algorithmic approaches for accessions and affiliations are identical. In Step 4, the CAVE algorithm uses Monte Carlo simulation to simultaneously observe future outcomes and estimate both the negative gradient $\Delta^-_{t,c,a,d}(x_{t,c,a,d}, \omega)$ and the positive gradient $\Delta^+_{t,c,a,d}(x_{t,c,a,d}, \omega)$ for all decisions x_t where $t \in \mathcal{T}_\pi$.

In Step 5, we establish how wide the smoothing interval $Q_{t,c,a,d}$ must be with the existing breakpoints to avoid any concavity violations and establish these end points as $k^-_{t,c,a,d}$ and $k^+_{t,c,a,d}$. We next further modify the smoothing interval $Q_{t,c,a,d}$ by adding additional breakpoints based on the pipeline constraints $\eta^-_{t,c,a}$ and $\eta^+_{t,c,a}$ as well as the declining size of the interval width parameter $\delta_{n,d}$. In Step 6, we update the slopes below $x_{t,c,a,d}$ according to the observed gradient $\Delta^-_{t,c,a,d}(x_{t,c,a,d}, \omega)$ and the slopes above $x_{t,c,a,d}$ according to the corresponding observed gradient $\Delta^+_{t,c,a,d}(x_{t,c,a,d}, \omega)$, with the size of both updates adjusted for the stepsize parameter α_n.

4.5.2 SUPERCAVE

The CAVE variant converges to a high-quality solution, but the algorithm cannot escape local optima. To address this issue, we propose SUPERCAVE, which begins by finding a solution x^π_t using the CAVE algorithm, but then generates ϱ perturbed solutions around this solution. For each perturbation $p \in \{1, 2, ..., \varrho\}$, the algorithm randomly selects ξ AFSCs where the accession decision is increased in $t = 1$ as the

set of $\mathcal{A}^{-}_{c,d_a}$. The algorithm also selects ξ other AFSCs where the accession decision is decreased as the set of $\mathcal{A}^{+}_{c,d_a}$. Each set of AFSCs is split into pairs of a^{-}_{p} and a^{+}_{p}, where AFSC a^{-}_{p} has its corresponding accession decision decreased by ϵ accessions in time t while AFSC a^{+}_{p} has its accession decision increased by the corresponding amount.

We know that the number of personnel brought in through accessions decisions cannot be perfectly substituted by corresponding accessions at a later time period, but when the time periods are close these personnel do overlap heavily in which authorizations they can fill. For this reason, simply perturbing accessions in time t would likely result in any retraining undoing much of the perturbation because the AFSC has simply been under- or over-resourced based on the direction of the perturbation. To address this, we generate a corresponding perturbation of the same magnitude ϵ but opposite in direction at time $\tau \in \{2, 3, ..., T_\pi\}$. We also select a desired perturbation size σ, though this must be reduced to ϵ to avoid violating any pipeline constraints for either AFSC at time t or τ. Once all of these perturbations have been generated, we have ϱ perturbed policies in addition to the original trained policy.

Next, we retrain using a truncated form of the CAVE algorithm on these perturbed solutions. Because we know these perturbed solutions are already very close to high-quality solutions, we reduce the training length from N to N_r and set $\delta_{n,d} = 1$. We reinitialize each policy as we did with the original equilibrium policy as the baseline. Affiliation decisions are initialized as the unperturbed initial solution but are allowed to continue changing during the retraining process. Each final retrained set of decisions is recorded as x_p. To assess the relative goodness of each perturbed decision x_p, we simulate each for T timesteps and ζ replications and select the set of decisions x_p with the lowest mean discounted cost, unless the original unperturbed policy x^π

Algorithm 4 SUPERCAVE Algorithm

Step 1: CAVE Baseline

Identify $A_{t,c}$, the aggregate constraint for accessions $\forall t \leq T_\pi$, where T_π is the desired length of the lookahead policy before reverting to equilibrium policy.

Train CAVE algorithm, identify $x_t^\pi \quad \forall t = \mathcal{T}_\pi$

Step 2: Perturb accessions solutions

for $t \in \mathcal{T}_\pi, c \in \mathcal{C}$ **do**

  for $p = 1$ **to** ϱ **do**

    Select ξ AFSCs as $\mathcal{A}^-_{c,d_a} \in \mathcal{A}'_{c,d_a}$

    Select ξ AFSCs as $\mathcal{A}^+_{c,d_a} \in \mathcal{A}'_{c,d_a} : \mathcal{A}^-_{c,d_a} \cap \mathcal{A}^+_{c,d_a} = \emptyset$

    for each pair of AFSCs a_p^- and a_p^+ from $\mathcal{A}^-_{c,d_a}$ and $\mathcal{A}^+_{c,d_a}$ **do**

      Select future policy year $\tau \in \{2, 3, ..., T_\pi\}$ to perturb

      Identify perturbation size $\epsilon = \min\big(\sigma, x^\pi_{t,c,a_p^-,d_a} - \eta^-_{t,c,a_p^-}, \eta^+_{t,c,a_p^+} - x^\pi_{t,c,a_p^+,d_a}, x^\pi_{\tau,c,a_p^+,d_a} - \eta^-_{\tau,c,a_p^+}, \eta^+_{\tau,c,a_p^-} - x^\pi_{\tau,c,a_p^-,d_a}\big)$

      Set $x_{t,c,a_p^-,d_a,p} = x^\pi_{t,c,a_p^-,d_a} - \epsilon$

      Set $x_{t,c,a_p^+,d_a,p} = x^\pi_{t,c,a_p^+,d_a} + \epsilon$

    end for

  end for

end for

Step 3: Retrain perturbed solutions

for $p = 1$ **to** ϱ **do**

  for $t \in \mathcal{T}_\pi, c \in \mathcal{C}, d = d_a, a \in \mathcal{A}'_{c,d_a}$ **do**

    To model each perturbed accession decision $x_{t,c,a,d_a,p}$, let $k_{t,c,a,d_a} = 2$, $\nu^1_{t,c,a,d_a} = 0.0001$, $\nu^2_{t,c,a,d_a} = 0$, $u^1_{t,c,a,d_a} = 0, u^2_{t,c,a,d_a} = x_{t,c,a,d_a,p}$.

  end for

  for $t \in \mathcal{T}_\pi, c \in \mathcal{C}, d \in \{d_r, d_g\}, a \in \mathcal{A}'_{c,d}$ **do**

    To model each affiliation decision $x_{t,c,a,d}$, let $k_{t,c,a,d} = 2$, $\nu^1_{t,c,a,d_a} = 0.0001$, $\nu^2_{t,c,a,d_a} = 0, u^1_{t,c,a,d} = 0, u^2_{t,c,a,d} = x^\pi_{t,c,a,d_a}$.

  end for

  Initialize parameters $\delta_n = 1$ and α_n.

  Train with CAVE algorithm for $n = \{1, 2, ..., N_r = 15\}$ to find x_p.

end for

Step 4: Simulate to select superlative solution

Simulate each x_p for T timesteps and ζ replications to identify superlative policy

End

is the superlative performer.

4.6 Experimental Design and Results

Next, we test the relative performance of these algorithms. We run all experiments in MATLAB using MATLAB's parallel computing toolbox and using a GPU for certain matrix calculations. Each experiment is run locally with an Intel Xeon Gold 6240 CPU at 2.60 GHz with 36 cores and an NVIDIA Quadro RTX 8000 GPU. All computation times are reported for a full simulation length of 20 years with 35 replications run in parallel on the CPU, sharing GPU resources.

4.6.1 CAVE Performance

We first test the performance of CAVE with two candidate stepsize rules, two stepsize parameters for each stepsize rule, two training lengths, and three lookahead horizons as shown in Table 15 to find which hyperparameters deliver the best performance for the intercomponent USAF WRP. The stepsize rules test a deterministic stepsize rule, the Generalized Harmonic Stepsize, and a stochastic stepsize rule that accounts for the variance and distribution of the observed updates, the Bias Adjusted Kalman Filter. Both of these stepsize rules have been applied successfully to CAVE variants in past research. For each of these approaches, we test two settings for the stepsize decay parameter a, which determines how quickly the stepsize reduces as n progresses.

Hyperparameter	Settings
Stepsize Rule (α)	Generalized Harmonic Stepsize, Bias Adjusted Kalman Filter
Internal Stepsize Parameter (a)	5, 10
Training Length (N)	50, 100
Lookahead Horizon (T_π)	3, 5, 7

Table 15. CAVE Hyperparameter Testing

Previous applications of CAVE to the USAF WRP used a training length of $N =$

100, but preliminary testing suggested that a shorter training timeline might deliver similarly high-quality results. These previous applications also used a lookahead horizon of $T_\pi = 5$, in part because USAF accession plans tend to be developed for three to five year windows. While plans shorter than three years would not be as helpful for planners, we tested lookahead horizons of 3, 5, and 7 years. We hypothesized that shorter lengths would decrease problems with interactions between time periods but at the expense of being constrained to simpler policies.

Lookahead Horizon	Training Length (N)	Stepsize Schedule Rule	Stepsize Parameter (a)	Percent Improvement over Benchmark (95% CI)	Computation Time (hours)
3	50	BAKF	5	18.90 ± 0.45	6.7
3	50	BAKF	10	18.72 ± 0.52	6.7
3	50	GHS	5	21.98 ± 0.47	6.7
3	50	GHS	10	21.05 ± 0.44	6.7
3	100	BAKF	5	19.06 ± 0.47	13.5
3	100	BAKF	10	18.68 ± 0.43	13.4
3	**100**	**GHS**	**5**	**23.06 ± 0.44**	**13.3**
3	100	GHS	10	22.64 ± 0.47	13.5
5	50	BAKF	5	17.59 ± 0.58	9.5
5	50	BAKF	10	17.28 ± 0.47	9.5
5	50	GHS	5	19.97 ± 0.49	9.5
5	50	GHS	10	19.14 ± 0.38	9.5
5	100	BAKF	5	17.56 ± 0.43	18.8
5	100	BAKF	10	17.43 ± 0.40	18.9
5	100	GHS	5	20.61 ± 0.55	18.6
5	100	GHS	10	19.96 ± 0.49	18.8
7	50	BAKF	5	17.51 ± 0.54	12.3
7	50	BAKF	10	16.92 ± 0.39	13.0
7	50	GHS	5	19.42 ± 0.50	12.4
7	50	GHS	10	18.94 ± 0.37	12.3
7	100	BAKF	5	17.29 ± 0.40	25.4
7	100	BAKF	10	17.07 ± 0.44	25.5
7	100	GHS	5	19.82 ± 0.48	24.7
7	100	GHS	10	19.23 ± 0.44	25.2

Table 16. Policy performance comparison: shorter lookahead horizons and longer training times demonstrated the strongest performance.

We tested each combination of hyperparameters with 35 replications of a simulation over $T = 20$ years to observe the mean discounted cost for each algorithmic implementation. Table 16 shows very clear effects of each hyperparameter tested, with the Generalized Harmonic Stepsize rule with $a = 5$ outperforming every other

stepsize for all hyperparameter combinations, the shorter horizon length $T_\pi = 3$ outperforming all other horizons for every other hyperparameter combination, and the longer training length $N = 100$ outperforming the shorter training length for nearly every hyperparameter combination. The superlative combination of hyperparameters was found to use a Generalized Harmonic Stepsize with $a = 5$, $N = 100$, and $T_\pi = 3$.

The high performance of the short horizon length formulations was counter to initial hypotheses, but is a logical finding for two reasons. First, current manpower funding business practices prioritize near term authorizations. These practices often do not fully execute future authorizations with any required AFSC changes, resulting in demand signals that show significant change in the first year or two but little change in later years. This business practice is subject to change as these business processes improve, requiring further testing in the future. Second, we specifically design the SUPERCAVE approach to overcome interactions between time steps. With few changes in future years, the original CAVE approach may spread deviations from the equilibrium policy (i.e., "fixes") over a longer timeline, resulting in less responsive policies.

As expected, longer lookahead horizons and increased training length both clearly increased training time. While the effect of the stepsize rule on training time was less clear, the superlative stepsize rule requires fewer calculations than the Bias-adjusted Kalman Filter and demonstrates faster times for most combinations of lookahead horizon and training length.

While all other experiments used 35 replications using specified seeds, the sustainment results were replicated 350 times due to the lower computational burden of this approach. Table 16 shows the superlative result to be a statistically significant improvement over the equilibrium sustainment policy, with an estimate of the effect size as $23.06\% \pm 0.44\%$ with 95% confidence using a two sample t-test.

4.6.2 SUPERCAVE Improvement Over Baseline

We extend the CAVE algorithm using these superlative hyperparameter settings. While SUPERCAVE should provide larger performance gains for longer time horizons due to the greater number of interactions between policies, we test the shorter time horizon to ensure that the SUPERCAVE implementation can outperform the highest quality solutions generated by CAVE for the current USAF system. To investigate the performance of the SUPERCAVE algorithm, we design a test across a range of SUPERCAVE-specific hyperparameters, shown in Table 17.

Hyperparameter	Settings
Number of perturbations (ϱ)	5, 10, 20
Number of AFSCs (ξ), Size of perturbations (σ)	Small Setting: $\xi = 40$,$\sigma = 5$
	Large Setting: $\xi = \frac{\lvert\mathcal{A}'_{c,d_a}\rvert}{2}$,$\sigma = 10$
AFSC Sampling Approach	Uniform, Weighted

Table 17. SUPERCAVE Hyperparameter Testing

First, we test with the number of perturbations ϱ, where additional perturbations should improve solution quality but at the expense of computation time. Next we test the size of the perturbations, either testing a subset of 40 AFSCs each for positive and negative perturbations with a perturbation size $\sigma = 5$ accessions or perturbing half of the AFSCs in each direction with a perturbation size $\sigma = 10$ accessions. For the setting with each AFSC perturbed, all AFSCs are sampled each time, but the setting with $\xi = 40$ is also tested with both a uniform sampling mechanism as well as a weighted sampling mechanism based on the AFSC's distance from being 100% manned in the aggregate.

Table 18 shows improvements in mean performance for all but one of the experiment results which appears to be caused by noise in the stochastic outcomes. The highest performing model shows statistically significant improvements over the superlative CAVE model at the 95% confidence level using a paired t-test. Table 18

Perturbation Size	Number of Perturbations	AFSC Sampling Approach	Percent Improvement over CAVE (95% CI)	Computation Time (hours)
Small	5	Uniform	0.45 ± 0.49	31.9
Small	10	Uniform	0.06 ± 0.46	49.1
Small	20	Uniform	0.09 ± 0.54	84.9
Small	5	Weighted	0.32 ± 0.57	34.0
Small	10	Weighted	-0.05 ± 0.50	54.8
Small	20	Weighted	0.23 ± 0.49	101.2
Large	**5**	**Uniform**	**0.67 ± 0.42**	**30.9**
Large	10	Uniform	0.25 ± 0.55	47.6
Large	20	Uniform	0.42 ± 0.48	82.7

Table 18. SUPERCAVE policy performance comparison: Large perturbations improve performance, but the effect of the number of perturbations is lost in the noise.

also shows the computational time for one batch of 35 replications run in parallel. As expected, using weighted sampling results in slightly higher computation times, while the number of perturbations to retrain has large effects on the required computation time.

4.6.3 Affiliations Improvement Over Component-Centric Policy

We tested an alternative policy structure without the affiliation decision classes d_r and d_g to show the impact of managing these policies. We compared results using the superlative tested SUPERCAVE configuration with 5 perturbations and the large perturbation setting. This test showed a $14.98 \pm 0.58\%$ increase in mean costs when restricting affiliations to the baseline rate using a 95% confidence level and a paired t-test. This result suggests that the inclusion of policies to directly manage and optimize affiliations to the AFR and ANG can meaningfully improve the USAF's ability to maintain the required number of personnel across all components.

4.7 Conclusions and Future Work

This research first extends the benchmark equilibrium model for RegAF AFSC management to the AFR and ANG, providing the first approach to enterprise model-

ing of AFR and ANG skillsets at the enterprise level. This provides a policy baseline that can be used for resourcing training pipelines and an expected distribution of personnel that can provide significant insight regarding the composition of current personnel inventories to AFR and ANG policy analysts.

We then formulate the intercomponent USAF WRP as a Markov decision process and extend previous applications of CAVE to this larger problem, showing significant improvement over the benchmark equilibrium policy. We test CAVE's applications across a range of hyperparameters and find superlative settings that vary from previous applications of CAVE to the USAF WRP.

We next devise and test SUPERCAVE, a methodological improvement to the CAVE approach that demonstrates a statistically significant improvement versus CAVE. We show small but statistically significant improvements over the CAVE approach, demonstrating that this approach can deliver higher-quality solutions even for very short lookahead horizons. While improvements are relatively small in scale, the USAF spends billions of dollars on its personnel and deviations from the funded authorizations drive personnel utilization and talent management decisions at many different organizational levels, meaning that even small improvements in the USAF's ability to meet funded authorizations may have dramatic effects on mission effectiveness as well as airmen's quality of life and career satisfaction. Additionally, as future business processes change how the USAF funds future authorizations, the USAF may need to increase the use of more complex accession policies, where these benefits will increase. In the short term, SUPERCAVE can provide the highest-quality policy recommendations if adequate time and computational resources are available. If computation is a limiting factor, as may be the case if considering the effects of different future authorizations or pipeline constraints where rapid iteration is desirable, CAVE may provide policies that are useful for planning even if they accept some reduction

in solution quality. Selecting a lower number of iterations for training can extend this tradeoff further, maintaining most of the improvement over the benchmark while reducing the computation time substantially.

Finally, we validate the importance of beginning to optimize affiliations from the RegAF to the AFR and ANG. The USAF does not set specific targets for this policy lever, relying on individual volunteers and the concurrence of RegAF career field managers. However, we demonstrate the large potential benefits of directly managing affiliation targets from the RegAF to the other components. Because this application was constructed to assess the impact of additional affiliations above the volunteer rate, future applications of this approach should modify the policy structure to directly optimize the total number of affiliations. This structure will provide a useful target for USAF decision-makers when attempting to modify the volunteer rate, whereas the target number of "extra" personnel is not helpful without the baseline.

When operationalizing such an approach, the AFR and ANG will need to establish new processes to communicate with the disparate decision-makers at individual locations to determine how to inform accession-planning. While fully-centralized accessions are not compatible with current business processes and cultural expectations within the AFR and ANG, these policy baselines should be used as a starting point to inform local decision-makers as to the likely future consequences of their accession decisions and inform resourcing decisions to ensure AFR and ANG recruiters can find the right talent. Additionally, such a policy baseline and use of simulated results from the Markov decision process formulation can help the components negotiate when constrained training resources are fungible between components.

Future work should increase the number of replications to refine the estimated effect of SUPERCAVE's hyperparameters. While the increased quality of the large perturbation setting seems clear, increased testing on the individual effects of the

number of AFSCs to be perturbed and the size of the perturbation merits exploration. Additionally, the findings showing the highest performance with a small number of perturbations tested appears to be due to stochastic noise within the system. Future testing should seek to confirm this hypothesis or address why a smaller number of perturbations would provide higher-quality solutions.

Future work should also develop models for how authorizations may change over time. While such an approach is not appropriate for directly developing policy solutions, CAVE and SUPERCAVE hyperparameters should be further tested for robustness in an environment with changing authorizations. Current performance estimates assume that current projections of future authorization levels remain at the projected levels without further programmatic change, which would be a historical anomaly.

Finally, future work should test such approaches for more granular approaches that work to maintain specific skillsets. The US Space Force is currently experimenting with directly quantifying skillsets instead of relying on a career field designation to measure groups of skillsets. The CAVE and SUPERCAVE approaches are suitable for such a structure, although the computational complexity of such an approach increases as the granularity of the skillsets increases.

One area of concern when modeling more complex relationships is whether the approximation of the survival rate to potential future states is a good approximation, or whether solution quality could be limited by any difference between the true survival probability and the approximation. One approach to address this would be to observe these survival transitions during each simulation and update this approximation instead of relying on the original. This would potentially improve CAVE and SUPERCAVE's results on the current problem instance as well as enable their application to more complex state spaces where no initial approximation exists.

V. Conclusion

> "Rational decision-making requires a position of considerable political power. The sources of 'irrationality' are not simply muddled thinking or psychological quirks, but the regular intrusion of insistent lobbyists for some cause or interests, or inadequate bureaucratic structures or the divergent pull of opposing objectives."
>
> - Sir Lawrence Freedman
>
> Professor of War Studies, King's College
>
> The Evolution of Nuclear Strategy, 3rd Ed., p. 219

5.1 Summary of Research Contributions

While the potential scope of USAF talent management and workforce replenishment policies is large, this research improves the USAF's ability to manage these problems by answering the each of the following specific research questions.

Research Question 1: How can the USAF use MilPDS and publicly available data to accurately and precisely predict monthly retention behavior over a 12 month period?

In Chapter II, we show we can generate better predictions than the current benchmark Kaplan Meier model with both a feedforward neural network and by a feedforward neural network trained with a partially autoregressive feature. While the partially autoregressive neural network showed the superlative performance for the validation dataset, the traditional feedforward neural network showed the greatest performance on the test dataset. Importantly, to generate one high-quality model, many models needed to be trained and tested on a validation dataset. While the baseline neural network approach can be deployed to improve the quality of predictions, the partially autoregressive neural network (PARNet) model appears likely to outperform the baseline during periods of less volatile economic conditions and can

be used as a second estimate. While an ensemble of the various modeling approaches was not tested, the use of an ensemble to develop robust predictions of the likely range of outcomes may be operationally useful to help inform decision-making.

Research Question 2: How can the USAF improve the quality of accessions policies for the active duty force implemented by AFSC to reduce AFSC shortages and improve AFSC manning?

In Chapter III, we design, develop, and test novel approximate dynamic programming (ADP) and reinforcement learning (RL) algorithms that determine high-quality personnel accessions policies. We develop a direct lookahead policy modification of Concave Adaptive Value Estimation (CAVE) as well as a parameterized deep reinforcement learning approach to generate high-quality policies for decisions with high dimensionality while maintaining a low computational cost. We show that CAVE performs well for the USAF workforce replenishment problem (WRP) at a low computational cost and provide insight into cost function development by testing the effects on policy of two candidate cost functions.

While the primary use of this contribution will be to develop a baseline for accession policies across all AFSCs, this approach provides a standardized approach to examine the effect of existing policies and inform functional stakeholders for specific communities. This Markov decision process model and insights from this work have been used to inform the USAF operations research analyst career field management team's policy planning and coordination with AF/A1 (Hoecherl, 2022). The insights from this model have driven modifications to accession policies across multiple years to procure the required analytic talent to meet analytics and artificial intelligence initiatives directed by the Secretary of the Air Force.

Research Question 3: How can the USAF improve the quality of accessions policies across all components implemented by AFSC to reduce AFSC shortages and

improve AFSC manning? What policies that significantly impact AFSC manning need to be managed differently or start being managed? How do we ensure good solutions to those policies?

In Chapter IV, we design, develop, test, and compare multiple sequential decision-making approaches for determining high-quality personnel policies. This contribution extends the work proffered in Chapter III by considering a new, larger problem set, including RegAF, AF Reserve, and Air National Guard personnel. First, we extend the RegAF's benchmark equilibrium sustainment model to the AFR and ANG, then formulate this larger problem as a Markov decision process. We extend the CAVE approach to this larger problem and test performance across a range of hyperparameters. Finally, we develop and test a novel algorithm modification to the CAVE approach which leverages a perturbation and retraining process to improve solution quality at the expense of additional computation. Tests show statistically significant improvements over the baseline CAVE approach, which shows statistically significant improvements over the benchmark equilibrium policy. While the computational costs for implementing SUPERCAVE compared to CAVE are not trivial, relatively infrequent policy development of accession targets could support such an investment for improvements in solution quality for such a high-stakes set of policies.

Although the primary intent of such an algorithmic implementation is to provide high-quality accessions and affiliation policy baselines, this approach also has the potential to dramatically change the USAF's approach to making decisions about future human capital composition. Current approaches allow senior leaders to make decisions about future authorizations that are divorced from considerations of whether we can meet these authorizations with corresponding personnel. This results in many decisions that are not feasible within known policy constraints and frustration as senior leaders seek ways to procure the human capital needed for their various missions.

While some business processes would need to change to include AFSC-level detail in the USAF programming process, by projecting policies and personnel inventories this research can provide an alternative decision framework. With a considered set of future authorizations that are shown to be infeasible, senior leaders can instead choose to:

1. Reduce the rate of required change for emerging requirements.
2. Find alternative offsets to allow human capital to be repurposed for the emerging requirement.
3. Apply required resources to relax the relevant constraint, allowing bottlenecks to be identified and removed during the planning process instead of waiting for the problem to manifest.

With such a process, quick responses may be more important than fine policy adjustments, so using the superlative CAVE implementation with simulations over 5 years could provide such insights with less than 2 hours of computation on a comparable machine.

Additionally, this set of models and algorithms allows for an unprecedented level of integration with the AFR and ANG. One key to successful implementation of such an approach will be developing the relationships and business processes between enterprise-level modelers and the AFR and ANG commanders at each location that own the corresponding policies. This approach will be most effective if used to identify areas of concern with existing policy and inform local commanders, rather than centralize decision-making without an understanding of local conditions or commander's constraints. This can provide an avenue to identify the relative importance of different recruiting problems and inform resourcing decisions to overcome these limitations.

In addition to the individual accession and affiliation policies, this contribution has informed analysis of aggregate personnel behavior across components and the strategic consequences of such patterns, historically not observed within a specific model. Insights from examining such behavior were briefed at Operation Retrenchment Specter in December 2021, showing that aggregate retention within the RegAF had driven additional costs and a significant vulnerability during future conflicts with significant attrition. Based on this analysis, we developed a course of action demonstrating the need to begin managing affiliations directly and boosting the overall level of these affiliations. This course of action was rated as the top submission for overall quality at the wargame and is the subject of a follow-on paper (Hoecherl, Schulker, Hornberger and Walsh, 2022). Implementing this course of action would require data-informed policy development for affiliations, which do not currently have a data-informed target. The models in Chapter IV provide a defensible, integrated target, though future development may need to establish modifications to costs or constraints if a specific total affiliations target is established.

5.2 Future Work

Future retention modeling research should be conducted in four general directions. First, the selection of features used in Chapter II was informed by subject matter expertise of known relationships. Many other variables within the Military Personnel Data System may have significant explanatory power, though the addition of features will increase problems with imbalanced observations and statistical bias. The positive and negative effects of such feature selection is deserving of future study. In addition to personnel features, economic data may provide valuable information about the likelihood of personnel to depart without changing the distribution of retention observations, but requires multiple economic trends within the training

data to effectively measure. As more data is collected after the COVID-19 pandemic has passed, we can measure the effects of including this data either to supplement or to compete with the use of a partially autoregressive feature to predict trends in retention behavior.

A second area for future research is the development of loss functions that more closely respond to statistical bias. Because this approach leverages large minibatch sizes, one approach may be to create a bias-adjusted loss function that adjusts updates to the neural network based on the statistical bias measured across all predictions in a single minibatch update. Such an approach may provide higher quality predictions and an improved ability to assess model quality after initial training, though this approach may cause problems with training stability.

Because of the high level of noise in the quality of the models generated, many of the hyperparameters showed only weak relationships with model quality. Especially in combination with the development of new loss functions, further work to assess the effect of hyperparameters may provide additional insight, especially in more stable retention environments.

Finally, the level of variance in prediction quality is concerning from a practitioner's perspective. Barring further progress in some other area, constructing an ensemble approach to produce multiple predictions may improve robustness. One area to consider is the inclusion of the Random Forest models, which performed well in preliminary testing and displayed very consistent, robust predictions even though their superlative models did not produce predictions of the same quality as the superlative neural network models.

For the WRP, the most pressing future work is to reconfigure the Markov decision process formulation to directly optimize the number of affiliations. While the current approach was important for demonstrating the value of beginning to manage this

process directly, implementation will require a direct target.

While many of the advances from Chapter II are not suitable for a parsimonious model that must predict years into the future, economic data included in the starting state S_0 may be able to help provide high-quality, short-term retention predictions. While these effects should fade as the time progresses and confidence in economic conditions decreases, including economic features may increase the quality of policies generated to account for losses in the short term.

Although retraining policies are generally more difficult to model due to the high level of volition involved and the lack of natural experiments, retraining policies have a significant effect on the manning of many AFSCs. Further work to replicate the approach used for affiliations and extend this to transitions to other AFSCs within the same component may provide both better policy baselines for retraining as well as more refined accessions policies.

The effect of SUPERCAVE hyperparameters is difficult to measure because of the noise in stochastic outcomes. Further work to increase the sample size can refine our understanding of these effect sizes and the benefit of increasing the computational investment to further refine policies.

All current testing was conducted in a static authorizations environment, where future authorizations do not deviate from the projected plan. This assumption is inconsistent with USAF system behavior, where programmatic changes occur every year. Future testing of such approaches should examine performance in both static and dynamic authorizations environments to assess the robustness of different algorithmic approaches.

The USAF may increasingly need to measure and develop policies to procure more granular skillsets beyond the career field level of detail. Such an approach is already being explored by the USAF with its Multi-Capable Airmen initiative

and by the US Space Force, who are forgoing the use of career fields in favor of directly quantifying specific skills for personnel. We can easily adapt the CAVE and SUPERCAVE algorithmic approaches to such formulations, though at the expense of additional computation as the state space grows.

Finally, the CAVE and SUPERCAVE approaches rely on developing a gradient using both a direct observation of future outcomes as well as a survival function approximation. While the future outcomes are an exact measure of simulated future states, the survival function is currently built on an approximation of future transition rates because the actual transition rates are dynamic, so no single set of weights will be appropriate for all possible training scenarios. While the default setting is the most appropriate approximation for survival rates generally, the CAVE and SUPERCAVE algorithms can potentially produce a survival approximation specific to the starting state S_0 by simply observing the simulated survival and recording the actual transition rates. Given the stochastic transitions in the system, such an approach should leverage a stepsize and update its approximation by a small amount after each training iteration.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015), 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: *https://www.tensorflow.org/*

Alpaydin, E. (2014), *Introduction to Machine Learning*, 3rd edn, The MIT Press, Cambridge, MA.

Asch, B. J. (2019), *Setting Military Compensation to Support Recruitment, Retention, and Performance*, RAND Corporation.

Barger, J. (2017), Personal communication. Deputy Division Chief, Human Resources Data, Analytics, and Decision Support Division, HQ USAF/A1XD.

Bastian, N. D., Lunday, B. J., Fisher, C. B. and Hall, A. O. (2020), 'Models and methods for workforce planning under uncertainty: Optimizing US Army cyber branch readiness and manning', *Omega* **92**, 102171.

Bastian, N. D., McMurry, P., Fulton, L. V., Griffin, P. M., Cui, S., Hanson, T. and Srinivas, S. (2015), 'The AMEDD uses goal programming to optimize workforce planning decisions', *Interfaces* **45**(4), 305–324.

Bernanke, B. S., Boivin, J. and Eliasz, P. (2005), 'Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach', *The Quarterly journal of economics* **120**(1), 387–422.

Betts, R. K. (1995), *Military readiness: concepts, choices, consequences*, Brookings Inst Press.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Chakraborty, K., Mehrotra, K., Mohan, C. K. and Ranka, S. (1992), 'Forecasting the behavior of multivariate time series using neural networks', *Neural Networks* **5**(6), 961–970.

Chollet, F. (2021), *Deep learning with Python*, Manning Publications Company, Shelter Island, NY.

Clevert, D.-A., Unterthiner, T. and Hochreiter, S. (2015), 'Fast and accurate deep network learning by exponential linear units (elus)', *arXiv preprint arXiv:1511.07289* .

Gal, Y. and Ghahramani, Z. (2016), Dropout as a bayesian approximation: Representing model uncertainty in deep learning, *in* 'international conference on machine learning', PMLR, pp. 1050–1059.

Gass, S. I., Collins, R. W., Meinhardt, C. W., Lemon, D. M. and Gillette, M. D. (1988), 'OR practice—The army manpower long-range planning system', *Operations Research* **36**(1), 5–17.

Géron, A. (2019), *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, Sebastopol, CA.

Godfrey, G. A. and Powell, W. B. (2001), 'An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution', *Management Science* **47**(8), 1101–1112.

Godfrey, G. A. and Powell, W. B. (2002*a*), 'An adaptive dynamic programming algorithm for dynamic fleet management, i: Single period travel times', *Transportation Science* **36**(1), 21–39.

Godfrey, G. A. and Powell, W. B. (2002*b*), 'An adaptive dynamic programming algorithm for dynamic fleet management, ii: Multiperiod travel times', *Transportation Science* **36**(1), 40–54.

Harrison, T. (2014), 'Rethinking readiness', *Strategic Studies Quarterly* **8**(3), 38–68.

Heaton, J. (2008), *Introduction to neural networks with Java*, Heaton Research, Inc., St Louis, MO.

Ho, T. K. (1995), Random decision forests, *in* 'Proceedings of 3rd international conference on document analysis and recognition', Vol. 1, IEEE, pp. 278–282.

Ho, Y.-C. (1999), 'An explanation of ordinal optimization: Soft computing for hard problems', *Information Sciences* **113**(3-4), 169–192.

Hoecherl, J. C. (2022), 'Operations research analyst (15A) sustainment: System behavior, problems, and solutions', *Air Force Institute of Technology Technical Report* .

Hoecherl, J. C., Barger, J. C., Robbins, M. J. and Zavislan, S. M. (2022), 'The fundamentals of United States Air Force human capital analytics', *Air Force Institute of Technology Technical Report* .

Hoecherl, J. C. and Robbins, M. J. (2022), 'Reinforcement learning approaches to improve United States Air Force accession policies', *Air Force Institute of Technology Technical Report* .

Hoecherl, J. C., Robbins, M. J., Borghetti, B. J. and Hill, R. R. (2022), 'Partially autoregressive machine learning: development and testing of methods to predict United States Air Force retention', *Computers & Industrial Engineering* **171**.

Hoecherl, J. C., Robbins, M. J., Hill, R. R. and Ahner, D. K. (2016), Approximate dynamic programming algorithms for United States Air Force officer sustainment, *in* '2016 Winter Simulation Conference (WSC)', IEEE, pp. 3075–3086.

Hoecherl, J. C., Schulker, D., Hornberger, Z. and Walsh, M. (2022), 'Antifragile Air Force: Systematic policy changes to build talent for the high-end fight', *Air and Space Operations Review* .

Hornik, K., Stinchcombe, M. and White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural networks* **2**(5), 359–366.

Hosek, J. and Wadsworth, S. M. (2013), 'Economic conditions of military families', *The Future of Children* pp. 41–59.

Ioffe, S. and Szegedy, C. (2015), 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', *arXiv preprint arXiv:1502.03167* .

Joffrion, J. L. and Wozny, N. (2015), 'Military retention incentives: Evidence from the air force selective reenlistment bonus', *Upjohn Institute Working Paper* (15-226).

Kennedy, B. (2018), 'Most americans trust the military and scientists to act in the public's interest. pew research center'.

Klambauer, G., Unterthiner, T., Mayr, A. and Hochreiter, S. (2017), Self-normalizing neural networks, *in* 'Proceedings of the 31st international conference on neural information processing systems', pp. 972–981.

Kunnumkal, S. and Topaloglu, H. (2008), 'Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems', *Operations Research* **56**(3), 646–664.

Lim, N. (2015), Air force commander's guide to diversity and inclusion, Technical report, RAND PROJECT AIR FORCE SANTA MONICA CA.

McCulloch, W. S. and Pitts, W. (1943), 'A logical calculus of the ideas immanent in nervous activity', *The bulletin of mathematical biophysics* **5**(4), 115–133.

Meeker, W. Q. and Escobar, L. A. (2014), *Statistical methods for reliability data*, John Wiley & Sons.

Menard, S. (2001), *Applied Logistic Regression Analysis (Quantitative Applications in the Social Sciences)*, 2nd edn, Sage Publications, Inc., California.

Miller, M. (2017), 'Briefing feedback from Chief of Air Force Reserve'. Pentagon, Washington D.C.

MilPDS Dataset (2021). Military Personnel Database (MILPDS) Extracts, HQ USAF/A1XD HR Data Analytics Division. Updated Oct. 08, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M. (2013), 'Playing Atari with deep reinforcement learning', *arXiv preprint arXiv:1312.5602* .

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al. (2015), 'Human-level control through deep reinforcement learning', *Nature* **518**(7540), 529–533.

MPES-UMD Dataset (2021). Manpower Programming Execution System (MPES) Unit Manpower Document (UMD) Extracts, HQ USAF/A1XD HR Data Analytics Division. Updated Oct. 08, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Powell, W. (2011), *Approximate Dynamic Programming: Solving the curses of dimensionality*, Vol. 703, 2nd edn, John Wiley & Sons, Hoboken, NJ.

Pujats, T. S. (2020), Forecasting attrition by AFSC for the United States Air Force, Technical report, Air Force Institute of Technology Technical Report.

Recruiting and Retention of Military Personnel (2007), Technical report, North Atlantic Treaty Organization, Research Task Group HFM-107.

Rostker, B. D. and Yeh, K. (2006), *I want you!: the evolution of the All-Volunteer Force*, Rand Corporation.

Russell, S. (1998), Learning agents for uncertain environments, *in* 'Proceedings of the eleventh annual Conference on Computational Learning Theory', pp. 101–103.

Salas, D. F. and Powell, W. B. (2018), 'Benchmarking a scalable approximate dynamic programming algorithm for stochastic control of grid-level energy storage', *INFORMS Journal on Computing* **30**(1), 106–123.

Schofield, J. A., Zens, C. L., Hill, R. R. and Robbins, M. J. (2018), 'Utilizing reliability modeling to analyze United States Air Force officer retention', *Computers & Industrial Engineering* **117**, 171–180.

Siebold, G. L. (2006), 'Military group cohesion', *Military life: The psychology of serving in peace and combat* **1**, 185–201.

Sims, C. A. (1980), 'Macroeconomics and reality', *Econometrica: journal of the Econometric Society* pp. 1–48.

Situ, J. X. (2018), An approximate dynamic programming approach to analyzing military personnel end-strength planning, PhD thesis, George Mason University.

Smith, L. N. (2018), 'A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay', *arXiv preprint arXiv:1803.09820* .

Smith, T. D., Asch, B. J. and Mattock, M. G. (2020), An updated look at military and civilian pay levels and recruit quality, Technical report, RAND National Defense Research Inst, Santa Monica, California.

Song, H. and Huang, H.-C. (2008), 'A successive convex approximation method for multistage workforce capacity planning problem with turnover', *European Journal of Operational Research* **188**(1), 29–48.

Sutton, R. S. and Barto, A. G. (2018), *Reinforcement learning: An introduction*, 2nd edn, MIT press, Cambridge, MA.

Taskaya-Temizel, T. and Casey, M. C. (2005), 'A comparative study of autoregressive neural network hybrids', *Neural Networks* **18**(5-6), 781–789.

Topaloglu, H. and Powell, W. B. (2003), 'An algorithm for approximating piecewise linear concave functions from sample gradients', *Operations Research Letters* **31**(1), 66–76.

Triebe, O., Laptev, N. and Rajagopal, R. (2019), 'Ar-net: A simple auto-regressive neural network for time-series', *arXiv preprint arXiv:1911.12436* .

Tsitsiklis, J. N. and Van Roy, B. (1997), 'An analysis of temporal-difference learning with function approximation', *IEEE Transactions on Automatic Control* **42**(5), 674–690.

Van de Wiele, T., Warde-Farley, D., Mnih, A. and Mnih, V. (2020), 'Q-learning in enormous action spaces via amortized approximate maximization', *arXiv preprint arXiv:2001.08116* .

Van Hasselt, H., Doron, Y., Strub, F., Hessel, M., Sonnerat, N. and Modayil, J. (2018), 'Deep reinforcement learning and the deadly triad', *arXiv preprint arXiv:1812.02648* .

Wooldridge, J. M. (2016), *Introductory econometrics: A modern approach*, Cengage Learning, Boston, Massachusetts.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
19–08–2022	Doctoral Dissertation	August 2019 — August 2022

4. TITLE AND SUBTITLE

Retention Prediction and Policy Optimization for United States Air Force Personnel Management

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Hoecherl, Joseph C., Maj, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-DS-22-S-062

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

HQ USAF/A1XD (HAF)
Mr. Douglas Boerman
1040 Air Force Pentagon Room 5B349
Washington, DC 20330
douglas.boerman@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

AF/A1XD

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approval for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Effective personnel management policies in the United States Air Force (USAF) require methods to predict the number of personnel who will remain in the USAF as well as to replenish personnel with different skillsets. To improve retention predictions, we develop and test traditional machine learning models as well as partially autoregressive forms, outperforming the benchmark on a test dataset by 62.8% and 34.8% for the neural network and the partially autoregressive neural network, respectively. We formulate the workforce replenishment problem as a Markov decision process for active duty enlisted personnel, then extend this formulation to include the Air Reserve Components. We develop and test an adaptation of Concave Adaptive Value Estimation (CAVE) on the active duty problem, finding that CAVE reduces costs from the benchmark policy by 29.76% and 17.38% for the two cost functions tested. We test CAVE across a range of hyperparameters for the larger intercomponent problem, reducing costs by 23.06% from the benchmark, then develop the Stochastic Use of Perturbations to Enhance Robustness of CAVE (SUPERCAVE) algorithm, reducing costs by another 0.67%. Resulting algorithms and methods are directly applicable to contemporary USAF personnel business practices, enabling more accurate, less time-intensive, and data-informed policy targets for current processes.

15. SUBJECT TERMS

machine learning, sequential decision-making, computational stochastic optimization, neural network, approximate dynamic programming, human capital analytics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Jeffery D. Weir, Ph.D., AFIT/ENS
U	U	U	UU	177	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x4523; jeffery.weir@afit.edu