ABSTRACT

SEMMEL, AUSTIN DENNIS. The Readiness–Utilization Trade Space in U.S. Army Aviation: Policy, Decision Behavior, and Information Quality. (Under the direction of Hans Sebastian Heese, Brandon M. McConnell, and Benjamin Rachunok.)

This dissertation investigates how information, readiness metrics, and policy design jointly shape maintenance and utilization decisions in U.S. Army aviation. Focusing on AH-64 Apache units, the work examines how OR and usage interact as coupled outcomes of maintenance capacity and policy design.

The first phase characterizes observed decision behavior. Using a generalized additive model (GAM), we quantify the relationship between OR, phase maintenance proximity, and daily aircraft utilization. Results show that while units continue to fly aircraft despite degraded readiness ratings, they systematically avoid flying aircraft approaching or emerging from phase maintenance. These patterns reveal implicit prioritization rules not captured by aggregate readiness metrics and indicate a measurable gap between doctrinal guidance and observed unit behavior.

Building on these findings, the second phase develops a data-driven framework to compare the impact of unit-level decision-making on efficiency outcomes. Units are evaluated along a Pareto frontier defined by OR and flying hours per aircraft, and a self-organizing map identifies latent decision-making profiles associated with distinct efficiency profiles. The results show that units operating under similar policy environments can achieve different performance outcomes on the frontier and that differences in decision behavior help explain these differences.

The final phase embeds these behaviors within a controlled simulation environment to assess the operational value of prognostic information. A decision-tree policy, optimized using a heterogeneous island-model genetic algorithm, is evaluated under varying levels of remaining useful life (RUL) prediction accuracy. A factorial experimental design isolates the causal effects of information quality and policy adaptation. The findings show that performance gains are driven primarily by signal quality and exhibit diminishing returns beyond moderate prognostic accuracy.

Collectively, this work demonstrates that readiness outcomes emerge from a joint system of metrics, policies, and information quality. These results provide a principled basis for evaluating legacy readiness measures and prioritizing investments in data-informed maintenance capabilities.

The Readiness–Utilization Trade Space in U.S. Army Aviation: Policy, Decision Behavior, and Information Quality

by
Austin Dennis Semmel

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Operations Research

Raleigh, North Carolina
2026

APPROVED BY:

Hans Sebastian Heese Co-chair of Advisory Committee	Brandon M. McConnell Co-chair of Advisory Committee
Benjamin Rachunok Co-chair of Advisory Committee	Shu-Cherng Fang

Mike Powell

DEDICATION

For Morgan Kathryn Semmel, who has carried me through every difficult season across ten years of marriage. I would not be here without her, and this work belongs to her as much as it does to me.

BIOGRAPHY

Austin Dennis Semmel was born in Chester, Pennsylvania, to Cheryl and Dennis Semmel in 1993. He has one younger brother, Brandon Semmel. He attended Parkland High School and graduated from the United States Military Academy in 2015 with a Bachelor of Science in Economics and was commissioned as a Finance Officer. In 2020, he transitioned to the Operations Research and Systems Analysis functional area. His assignments include the 82nd Airborne Division; 1st Special Warfare Training Group of the United States Army John F. Kennedy Special Warfare Center and School; United States Army Special Operations Command Headquarters; 8th Military Police Brigade; 25th Infantry Division; and the United States Military Academy at West Point in the Department of Mathematical Sciences. He holds a Master of Statistics (2019) and a Master of Operations Research (2024), both from North Carolina State University. Outside of work, he enjoys trail running, is a certified sommelier, and plays guitar and ukulele. He married Morgan Kathryn Semmel in 2016; they have one son together, Griffin Harvey Semmel, born in 2024.

ACKNOWLEDGEMENTS

First and foremost, I owe a tremendous debt of gratitude to my three co-advisors. Dr. Brandon McConnell saw potential in me and brought me into the NC State family with the audacious idea of extending my ACS-funded Master's program into a doctoral pursuit. The very possibility of this dissertation exists because of him. Dr. Sebastian Heese challenged my ability to critically evaluate information and elevated my writing in ways I could not have achieved alone; his nuance and attention to narrative clarity have left a lasting mark on my workflow. Dr. Benjamin Rachunok brought an inspirational methodological perspective and an uncanny ability to discern whether an idea was worth pursuing. More than once, a single offhand remark of his developed into multiple pages and plots in this dissertation, and I am deeply thankful for his support and guidance.

I am also grateful to the remaining members of my committee. Dr. Shu-Cherng Fang's network flows course helped me navigate a conditional pass on my qualifying exams. What had been a shallow understanding of linear programming became a genuine strength under his instruction. His courses in SVMs and neural networks further stabilized the analytical foundation that made the more complex extensions in this work manageable. COL Mike Powell has been an exceptional mentor at USMA and someone I could always turn to when I needed to bridge the gap between complex statistical concepts and practical understanding. As a former aviator, he understood the operational nuance in these issues and served as a reliable sounding board — if I could not convince him, I knew I was off base.

COL Thomas Dirienzo is the inspiration behind this entire dissertation. Conversations with him in Hawaii from 2019 to 2021, while I served in 8$^{\text{th}}$ Military Police Brigade and 25$^{\text{th}}$ Infantry Division, framed my understanding of the problem and sharpened my research questions. His guidance and mentorship allowed me to hit the ground running when I arrived at NC State.

I would also like to recognize several instructors whose teaching shaped my trajectory. David Haaf at Parkland High School first cultivated my love of statistics and calculus. At West Point, Benjamin Hung sparked my interest in mathematical modeling in MA103, and COL Victor Trujillo deepened my understanding of calculus. COL(Ret) Nicholas Clark inspired me as a cadet in his MA206 course, COL(Ret) Michael Phillips strengthened my foundation in applied statistics, and Alexander Knight sent me to a business analytics course in 2017 that set this entire journey in motion. COL Francis Murphy and COL (Ret) David Lyle deepened my analytical thinking through their econometrics courses, COL(Ret) Andrew Glen exercised great patience in teaching me mathematical statistics, COL David Beskow introduced me to systems engineering and the R programming language, and COL Riley

Post instilled lessons in macroeconomics and leadership that I treasure to this day. At NC State, Dr. Justin Post has been an inspirational leader in the field of statistics and a model for the kind of educator I aspire to be. Dr. Yahya Fathi's linear programming and Dr. Negash Medhin's non-linear courses stretched my understanding of linear algebra and have been foundational in my understanding of constrained optimization. Collectively, these instructors fostered in me a deep curiosity and love for mathematics that has served me through every stage of my development.

My high school cross country coaches, Loretta and Steve Dodson, and track coach Doug Finley instilled in me a sense of competitiveness and drive that has pushed me through many challenges since. LTC(Ret) Martin Wennblom and his family served as my sponsors at West Point and provided steady support during some of the most difficult moments of my cadet years.

I owe special thanks to my coworker, officemate, and friend Dr. Rachel Gidaro for listening to me talk through half-formed ideas and for countless whiteboard sessions that sharpened this work. LTC Jon Paynter in the USMA Department of Mathematical Sciences shared many engaging conversations about these topics and helped spark the ideas that became Chapter 4. I am also grateful to my fellow students in the OR program at NC State for their friendship, camaraderie, and willingness to navigate the PhD process together. In particular, I want to thank Robert Smith, Jabari Myles, Jake Benhart, Pierce Secola, Laura Mora, Cam Lisy, Mat Fukuzawa, Russ Nelson, Will Kirschenman, Erik Rosenstram, and Veronica Pacheco.

To the Soldiers, NCOs, and Officers I have had the privilege of serving alongside in the 82$^{\text{nd}}$ Sustainment Brigade, 1$^{\text{st}}$ Battalion 1$^{\text{st}}$ SWTG(A), USASOC Headquarters, 8$^{\text{th}}$ MP Brigade, 25$^{\text{th}}$ Infantry Division, and the USMA Department of Mathematical Sciences; you have all inspired me to reach this point, and I am honored to have served with each of you.

I am profoundly thankful for my family. My parents, Dennis and Cheryl Semmel, and my brother Brandon have offered unwavering support throughout this journey. My father- and mother-in-law, Michael and Michelle Mummey, and my sisters-in-law, Megan and Madison Mummey, have provided a bountiful encouragement and the occasional child support. My grandparents Stanley and Marilyn Roth nurtured in me a love of learning, and I carry the memory of Delores and Lamar Semmel with me always. I am also grateful to my extended family, especially Jaime and Karen Betancourt and Bruce and Veronica Redline, for their steadfast support over the years.

Finally, I would like to thank the Omar N. Bradley Foundation for their gracious backing and fellowship offer in 2025.

Disclosures:

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Army, Department of Defense, the United States Military Academy at West Point, or the U.S. Government.

Statement on Use of Generative AI:

The author used generative AI tools (ChatGPT, Google Gemini, Claude.ai, and Claude Code) throughout the development of Chapters 1, 3, 4, and 5 of this dissertation. These tools assisted with coding in R and Python (designing and debugging functions), LaTeX formatting (including tables), repository setup for Chapter 4, and iterative prose revision (both drafting and reviewing for clarity of intent). All AI-generated outputs were reviewed and validated by the author, who takes sole responsibility for all content presented in this document.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Authorship Statement

Austin Dennis Semmel is the sole author of Chapters 1 and 5 and the primary author of Chapters 2, 3, and 4. For Chapters 2 through 4, the author designed the research questions, developed all models and code, conducted the analysis, and wrote the manuscripts. Advisory committee co-chairs Hans Sebastian Heese, Brandon M. McConnell, and Benjamin Rachunok provided guidance on methodology, interpretation, and manuscript revision. Committee members Shu-Cherng Fang and Mike Powell also contributed to manuscript revision.

Use of Generative Artificial Intelligence

The author used generative AI tools (ChatGPT, Google Gemini, Claude.ai, and Claude Code) throughout the development of Chapters 1, 3, 4, and 5 of this dissertation. These tools assisted with coding in R and Python (designing and debugging functions), LaTeX formatting (including tables), repository setup for Chapter 4, and iterative prose revision (both drafting and reviewing for clarity of intent). All AI-generated outputs were reviewed and validated by the author, who takes sole responsibility for all content presented in this document. Generative AI was not used in the development of Chapter 2.

Chapter 1

Operational Readiness: An Eighty-Year Conversation

"The traditional R.A.F. habit of using 70–75% serviceability as a yardstick for measuring a squadron's efficiency is thoroughly misleading."

—C.H. Waddington, *O.R. in World War 2*, 1973

describing findings from the RAF Coastal Command O.R.S., 1943

1.1 The Eighty-Year Conversation

The tension between readiness metrics and operational outcomes is not new. The concept dates back to the origins of operations research itself. In 1943, C.H. Waddington and the RAF Coastal Command Operational Research Section discovered that the serviceability target of 75% was a poor proxy for squadron efficiency [3]. Their analysis revealed that readiness and utilization were coupled through maintenance processes in non-obvious ways. Unscheduled repairs increased sharply after scheduled maintenance events, and units could artificially inflate serviceability rates by simply flying less.

Waddington's work remained classified for thirty years [4]. By the time it was published in 1973, peacetime doctrine had already moved on. The critique resurfaced periodically in the decades that followed [5–10], but its clearest contemporary expression came from General David Goldfein, then USAF Chief of Staff, in 2019 [11]:

> "The fastest way for me as chief and for us as the Air Force to increase the [mission capable] rate is to stop flying. If I gave that airplane to maintenance, bought them the parts, they're going to get that MC rate high. But I'm not going to have people trained, I'm not going to have folks airborne, so my overall performance is going to go down."

The US Army formalized readiness reporting through AR 220-1 [12], which introduced the Green/Amber/Red rating system in 1963 and first used the phrase "Operational Readiness" in 1978. The current 75% Fully Mission Capable (FMC) target was established in AR 700-138 in 1985 [13]. This abstraction made Operational Readiness (OR) reportable and comparable across units but divorced the metric from how readiness was actually generated. Today, the US Army faces the exact same tension that Waddington identified, yet it relies on a similar scalar metric to manage it.

1.2 The Policy Puzzle

If the 75% OR threshold functions as intended, we would expect to observe a negative relationship between degraded readiness and subsequent flight dispatch. As this dissertation

will show, observed patterns in AH-64 Apache operations suggest the opposite. Units are often *more* likely to fly when OR falls below 75%.

This creates a fundamental disconnect. If the primary metric used to manage the fleet does not actually constrain unit behavior, it loses its diagnostic value. We are left with a system where units may meet mission demands through high utilization while appearing to fail readiness standards, or conversely, conserve aircraft to meet readiness standards while under-training pilots.

As the Army invests heavily in predictive maintenance and advanced sensors, this gap matters. The AH-64 Apache fleet alone represents a $13.79 billion investment of over 2,400 aircraft [14]. We risk layering sophisticated information systems on top of a management paradigm that fundamentally misinterprets the relationship between maintenance and usage.

1.3 The OR-Usage Framework

To resolve this puzzle, this dissertation proposes that readiness cannot be evaluated in a vacuum as a scalar value. Instead, we develop a framework that interprets readiness in two dimensions: availability (OR) and utilization (flying hours per aircraft, FHPA).

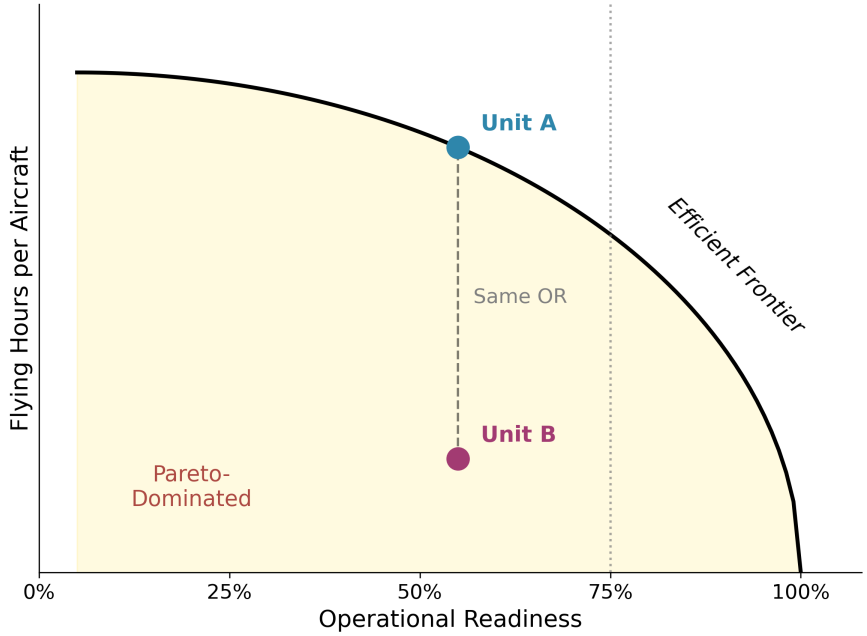

Figure 1.1: Conceptual OR-Usage tradespace. Units A and B both report the same OR, but occupy different positions: Unit A operates near the efficient frontier while Unit B is Pareto-dominated. Scalar OR cannot distinguish these cases. The dashed line indicates the R-1 readiness rating threshold for US Army aviation equipment (75%).

The central argument of this work is that OR and utilization are *joint outcomes* of a coupled maintenance-usage system. As illustrated in Figure 1.1, units operate along an efficiency frontier. Unit A (high usage, moderate OR) and Unit B (low usage, moderate OR) may report identical readiness scores, yet they possess vastly different operational capabilities. By reframing the problem in this two-dimensional tradespace, we can move from checking compliance ("Is the unit at 75%?") to analyzing efficiency ("Where is the unit relative to the Pareto frontier?").

1.4 Research Questions and Contributions

This dissertation operationalizes the OR-Usage framework through three sequential inquiries. The progression moves from diagnosis to classification and finally to prescription.

RQ1 (Diagnosis): Does OR restrict flight behavior as policy assumes? (Chapter 2)

We first test the validity of the current policy assumption using Generalized Additive Models (GAMs). The dataset contains 314,575 FMC aircraft-day observations. After exclusion of units with incomplete reporting, 265,472 observations from 423 AH-64 Apaches enter the final model. The analysis examines whether low OR or proximity to phase maintenance actually deters flight dispatch.

RQ2 (Classification): Do unit decision-making profiles inform position on the Pareto frontier? (Chapter 3)

If OR does not bind flight behavior, units may still differ in how they navigate the tradespace between utilization and readiness. We use Bayesian logistic regression and Self-Organizing Maps (SOMs) to characterize unit-level decision patterns and examine how these profiles inform position relative to the Pareto frontier. A minimum improving distance (MID) metric provides each unit a specific vector toward a better-performing peer. Robustness checks based on profile similarity (Gower-Jaccard, Intersection over Union), temporal traces, and Dynamic Time Warping assess whether MID targets are behaviorally plausible and whether each unit's time-averaged frontier position faithfully represents its month-to-month trajectory.

RQ3 (Valuation): What is the marginal value of prognostic accuracy for fleet outcomes? (Chapter 4)

RQ1 and RQ2 characterize a system where readiness targets do not constrain behavior and units vary in how they navigate the tradespace. RQ3 asks what commanders gain from better prognostic information and how that gain changes as accuracy improves. Chapter 4 uses a stochastic simulation calibrated to AH-64 maintenance doctrine and optimizes interpretable policies via genetic algorithm. Prognostic accuracy is parameterized by the coefficient of variation (CV) of remaining useful life (RUL) estimates. A factorial experimental design isolates the causal mechanism that links accuracy to operational outcomes and maps the shape of the value-of-information curve.

Summary of Contributions

The central claim across these investigations is that readiness and utilization are joint outcomes of a coupled maintenance-usage system, not separable indicators of unit performance. Empirical analysis shows that commonly used readiness targets do not function as binding behavioral constraints, and that units with similar readiness levels can occupy fundamentally different positions in the tradespace based on how they allocate available resources. A stochastic simulation establishes that the operational value of prognostic information is governed by its coefficient of variation, exhibits diminishing returns, and is bounded by existing policy structures. Together, these results reframe readiness as a tradespace that links measurement, behavior, and information quality.

1.5 Background: The Coupled Maintenance-Usage System

1.5.1 Operational Tempo and Phase Maintenance

Army doctrine explicitly recognizes the coupling between OR and utilization. ATP 3-04.7 states:

> "Units reporting high OR/RTL rates while not supporting high operational requirements may mask the ability to regenerate combat power. High bank time without corresponding combat or training flight hour execution demonstrates aircraft under-utilization and reduced combat presence in sustained operations.

> Units executing high flight hours against strong OR/RTL rates while sustaining or improving bank time ensure flexibility, predictability, and combat power generation."

This coupling emerges from fleet structure. Army Regulation 95-1 [15] requires pilots to fly 70 hours every six months, but pilots are not tied to unique aircraft. Utilization decisions therefore affect the fleet rather than individual pilot-aircraft pairings. Operational Tempo (OPTEMPO) determines how quickly aircraft accumulate hours toward phase maintenance. When an aircraft exceeds the phase threshold, it becomes immediately ineligible for flight and must undergo a comprehensive set of maintenance tasks.

1.5.2 Maintenance Definitions and Practices

Army aviation equipment falls into one of several readiness categories based on its current maintenance and supply status. Table 1.1 lists these status codes as defined in AR 700-138.

The Army's unit status reporting system translates FMC percentages into readiness ratings. Table 1.2 shows the thresholds; aircraft units face a stricter standard than ground equipment, with 75% FMC required for an R-1 rating.

The timelines in Table 1.3 represent upper limits; actual maintenance times vary with parts availability and workload.

1.5.3 Phase Maintenance

Phase maintenance is a comprehensive, scheduled inspection required at fixed flight-hour intervals (every 250 and 500 hours for AH-64 aircraft), as shown in Table 1.3. When an aircraft exceeds this threshold, it becomes immediately ineligible for flight until the inspection is complete.

The Department of the Army advocates for a "sliding scale" approach (ATP 3-04.7) that evenly spaces aircraft hours until phase maintenance. Figure 1.2 depicts this process: the straight line represents ideal hours remaining, with deviations in either direction considered undesirable.

1.5.4 Decision Authority

Flight decisions span multiple echelons: company commanders approve low-risk missions, battalion commanders approve moderate-risk missions, and brigade commanders approve high-risk missions [15]. Maintenance, however, is mostly managed at the battalion level [1].

Table 1.1: Description of possible maintenance status codes (AR 700-138)

Status Code	Full Status Code	Description of Status Code
FMC	Fully Mission Capable	Aircraft has no limitations; can accomplish all mission types based on total weapon system readiness.
PMCM	Partially Mission Capable: Maintenance	Aircraft is limited in its capabilities; can accomplish some mission sets, potentially in a diminished state.
PMCS	Partially Mission Capable: Supply	Aircraft is capable of performing certain missions but other types require a new piece of equipment.
NMCS	Non-Mission Capable: Supply	Aircraft is unable to fly because it requires a part that is not currently available.
DADE	Department of the Army Directed Event	Mandatory recall; usually requires aircraft to return to a depot-level maintenance center
NMCM	Non-Mission Capable: Maintenance	Aircraft are incapable of performing any of their assigned missions because of maintenance requirements.
FIELD	Subcategory of NMCM	Aircraft repair is handled by a field maintenance team.
SUST	Subcategory of NMCM	Aircraft repair is handled by a depot-level team.

This separation means the echelon making flight decisions differs from the echelon managing maintenance capacity.

1.5.5 Maintenance Paradigms

The Army currently operates using preventive and corrective maintenance [16]. Predictive maintenance (PdM), which uses condition monitoring and data analytics to forecast failures, is emerging but lacks unified implementation doctrine. Chapter 4 addresses the value of prognostic information in this context.

Table 1.2: Army readiness level requirements by FMC percentage

R-Level	R-1	R-2	R-3	R-4
Equipment other than aircraft	100–90%	89–70%	69–60%	less than 60%
Aircraft	100–75%	74–60%	59–50%	less than 50%

Adopted from Table 5-3 in AR 220-1

Table 1.3: Maintenance goals (phase/periodic inspections)

Aircraft Type	Phase	Flying Hours	Goals in Working Days
AH-64D/E	Minor	250 hours	11 days
	Major	500 hours	44 days
CH-47F	Minor	200 hours	28 days
	Major	400 hours	39 days
UH-60A/L/M	Minor	480 hours	30 days
	Major	960 hours	35 days

Adopted from Table 1-1 in Army Techniques Publication 3-04.7

1.6 Data

Chapters 2 and 3 use empirical data provided by the Engineer Research and Development Center (ERDC), which consolidates Army aviation flight records and maintenance status logs. The dataset covers October 2019 through May 2022 and contains 314,575 FMC aircraft-day observations from 423 unique aircraft across Active Duty units; 265,472 observations remain after exclusion of incomplete records. Chapter 4 uses a stochastic simulation informed by this data. Figure 1.3 shows data availability over the study period.

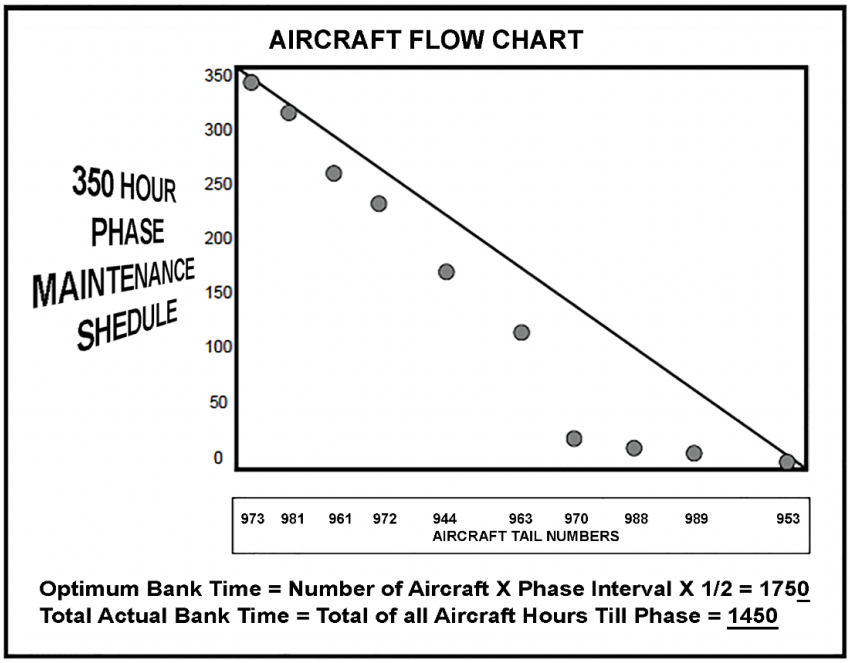

Figure 4-3. Aircraft flowchart with less than optimum

Figure 1.2: US Army aviation units currently employ a "sliding scale" of bank time to uniformly phase aircraft into maintenance (adopted from ATP 3-04.7).

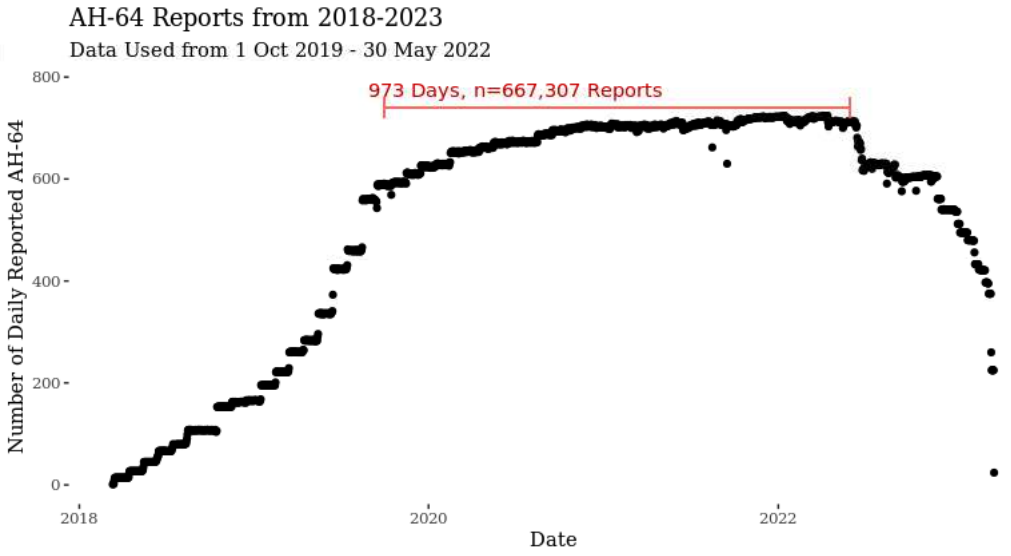

Figure 1.3: Reports in the ERDC maintenance logbook stabilized in FY2020.

1.7 Definitions and Acronyms

Appendix A consolidates the acronyms, systems of record, and doctrinal publications referenced throughout this dissertation. Table A.2 provides a full glossary with chapter-level usage; Table A.1 lists the Army data systems; and Table A.3 lists the doctrinal publications cited.

Chapter 2

Evaluating the Implementation of Operational Readiness and Maintenance Policies in US Army Aviation

Abstract

Published online in the *Journal of Defense Modeling and Simulation* [2]

This study examines AH-64 Apache dispatch decisions to assess the implementation of Operational Readiness (OR) and maintenance policies in the US Army. Current policies are designed to promote a ready and flexible force that is prepared to respond to global force projection requirements. The Army dictates a 75% OR target for aviation equipment and urges units to utilize aircraft uniformly to distribute maintenance capacity and prevent backlog. Given these objectives, we would expect a reduced OR rating to compel fewer sorties and uniformly distributed flying hours over the phase maintenance horizon. However, using a Generalized Additive Model (GAM), findings indicate that diminished OR does not deter flight operations. Moreover, aircraft are more likely to be grounded when approaching scheduled phase maintenance. Further analysis exposes a significant interaction effect; units place greater weight on an aircraft's hours until phase maintenance in the presence of low OR, highlighting a potential risk aversion in decision-making. Interestingly, control variables (the day of the week and reporting period proximity) highly associate with flight decisions. The findings suggest that current aviation readiness metrics may have an unintended influence on units' resource allocation. Future research should investigate unit-specific decision-making frameworks to improve aviation maintenance and OR efficiency.

2.1 Introduction and Motivation

The US Army maintains a globally-deployable ground force, capable of rapid power projection in order to deter, fight, and win its nation's wars. In order to do so, it seeks to efficiently allocate resources in a manner that sustainably maximizes the Army's lethality over time. A lethal force requires a combination of trained soldiers and functional weapon systems. However, an inherent trade-off exists between training and equipment serviceability. The more a unit trains, the more its equipment is used, fails and requires maintenance. On a daily basis, commanders must balance their internal and external training and mission requirements with a maintenance plan designed for sustained operations. While armor and mechanized infantry units certainly boast complex pieces of equipment, few Army units feel the sting of a maintenance mishap or the dread of a supply backlog more than an attack reconnaissance battalion of Apache helicopters. Thus, the Army designs policies and regulations that prescribe minimum levels of expected equipment serviceability and training to aid commanders in navigating these trade-offs.

This paper aims to investigate the behavior of units operating under current policies and uncover decision-making patterns surrounding the practical implementation of their flying hour programs. Army commanders are evaluated based on their ability to effectively train their units while maintaining readiness levels prescribed by the Headquarters, Department of the Army [12]. The Army refers to a unit's equipment serviceability as its "Operational Readiness" (OR) [12, paragraph 5-5]. Broadly speaking, each Active Duty AH-64 Apache pilot is required to fly 70 hours semi-annually with a unit goal of achieving at least 75% OR at all times [17]. Thus, OR and training are naturally competing objectives. Commanders and their representatives are tasked with making daily decisions on aircraft utilization while considering a unit's mission requirements, OR, maintenance schedules, and other factors (such as the weather and budget). This decision-making landscape is the subject of our investigation. We analyze the decision to fly a capable AH-64 Apache helicopter on a given day using a generalized additive model (GAM) to investigate the impact of these factors on a unit's aircraft deployment decisions.

One of the covariates in our model is a unit's current OR rating. As an example, if one helicopter is Fully Mission-Capable (FMC) to a unit for 24 out of 30 days in a reporting period (reported as 576 of 720 hours), then this piece of equipment maintained an 80% OR rating for that specific reporting period. This process is averaged across all equipment items of that type in a given unit. The Army establishes thresholds denoted as R-levels to measure a unit's ability to maintain their on-hand equipment; per Table 5-3 in HQDA [12], a unit achieves level R-1, if its OR rating is above 75%, R-2, if it is above 60%, R-3, if it is at least

50%, and R-4 otherwise.

Strict standards dictate if a helicopter is FMC or assigned to another maintenance category at a given time. A piece of equipment receives the maintenance status FMC if it is fully operational, configured in a safe and proper manner as designated by the US Army, and able to perform its combat mission without endangering the lives of the crew or operators. In contrast, an aircraft designated as Not Mission-Capable (NMC) compromises unit effectiveness by limiting both training and operational capabilities [13]. As such, Aviation Company and Battalion commanders have a strong incentive to ensure their units maintain an R-1 status when possible.

Due to helicopters' complex maintenance needs, achieving R-1 status is not always feasible. They are technical pieces of equipment that require regular maintenance intervals and also experience individual sub-component failures that demand immediate unplanned maintenance in order to regain FMC status. Planned (or, scheduled) maintenance intervals are called Phase Maintenance Cycles and occur at prescribed times according to the number of flying hours since the last phase maintenance occurred. Phase maintenance is a labor-intensive process that requires the thorough disassembly and inspection of an aircraft. These intervals vary based on the type of helicopter; Apaches must enter into major phase maintenance after the 500^{th} cumulative flying hour. Phase maintenance goals also vary by aircraft with AH-64D/E aircraft being expected to complete their 500-hour phase maintenance in no more than 44 days [1]. Each aircraft type has its own dedicated maintenance team. Based on the data, maintenance timelines are generally more rapid in practice, and the goals in the table functionally serve as upper bounds on phase maintenance timelines. Our goal is to evaluate how aviation units make decisions across different aircraft and fleet conditions.

2.2 Hypothesis Development

The objective of our study is to evaluate the effectiveness of existing Army policies on aircraft deployment decisions at the unit level. Specifically, we focus on the requirements regarding Operational Readiness and Phase Maintenance. We review these policies and derive related hypotheses in the following.

2.2.1 Operational Readiness

OR was first introduced to the US Army in doctrine in 1978 and, in 1985, AR 700-138 established that the "objective of aircraft readiness is to achieve a 75% FMC goal at all times" [6, p. 6]. In 1998, the Office of the Deputy Chief of Staff for Logistics requested that

the Army's Operations Research Center conduct a study on the need for an OR reporting system which then branched into a study of the history of OR in the Army [6]. They concluded that the 75% benchmark was created without "any analytical/engineering design criteria" and is not "linked to unit resources or capabilities" [6, p. 3]. Instead, they proposed a new method to monitor OR using control charts.

In the 2020s, the Army commissioned multiple Government Accountability Office (GAO) reports on readiness. Following a study using FY17–19 data, the GAO noted that the "services reported a variety of challenges related to air domain force elements including [...] the effects of trained pilot shortages on the Army's AH-64 attack helicopter" [18, p. 13]. It remains unclear how pervasive this shortage is across various units, and whether the Army employs a prioritization strategy to allocate limited resources, potentially leading to uneven effects on OR across the force. Importantly, newly minted pilots require additional training and can stress a unit's OR upon their initial arrival [19]. The FY17-19 study focused primarily on the human aspect of readiness, while future studies would investigate readiness from a systems and technology perspective [20, 21].

The GAO concluded a report on predictive maintenance for weapon systems that recommends the Army to efficiently improve its weapon system availability by identifying targets of opportunity, such as aircraft that are highly likely to experience failure in the near future [20]. In order to do so, the GAO suggested "reducing unplanned and unneeded maintenance" [20, p. 1] The Army differentiates between *planned* maintenance, which is preventive or predictive, and *unplanned* maintenance, which is reactive. There has been a strategic push towards predictive maintenance with an underlying goal of improving OR. Since 2002, many aviation units have started implementing predictive maintenance with mixed results due to the lack of standardized reporting metrics to properly evaluate its impact [20]. To convert unplanned maintenance occurrences into planned efforts, units now utilize sensors on aircraft that flag abnormal readings. Still, no standardized predictive maintenance doctrine exists, leading to varied implementations and effects [20]. In a recent success story, the 244^{th} Expeditionary Combat Aviation Brigade reports heightened maintenance efficiency with the adoption of the 'Griffin' AI-based predictive platform [22].

Given the challenges posed by the absence of a unified predictive maintenance doctrine, it is valuable to explore how recent literature addresses various maintenance strategies and their associated impact on aircraft readiness. Lipina [23] finds Air Force aircraft availability to be closely tied to the capacity of qualified maintenance manpower. MacKenzie et al. [24] demonstrate through simulation how a 10% change in maintenance capacity can significantly impact readiness. Choo et al. [25] further extend this concept, linking sortie generation directly to usage. They argue that for a given level of maintenance capacity, a unit can

determine an upper bound on long-run fleet availability, which is highly useful for planning and wargaming purposes.

Despite the advancements in maintenance strategy research, a significant gap exists in the literature regarding overarching policy evaluation. This is evident in studies such as Ritschel et al. [26], which admittedly overlook the impact of maintenance capacity on flight hours. McLean and Reiman [27] illustrate how small adjustments in the spare parts order process can significantly enhance readiness and cost efficiency, pointing towards the potential for policy-level interventions.

Informed by the history, doctrine, and literature above, and because Army commanders are held to the standards we have outlined, we hypothesize that as OR rates fall below R-1 status, units are deterred from continuing operations at the same operational tempo.

Hypothesis 1: *The probability of flying a Fully Mission-Capable (FMC) aircraft positively associates with increased operational readiness (OR).*

2.2.2 Phase Maintenance Interval Management

In order to balance OR and utilization, the Army has recommended the usage of *bank hours*, which is simply the number of hours available to be flown, per aircraft, until phase maintenance. The Army advocates for using a bank hour flow chart for efficient maintenance scheduling. Proper phase maintenance interval management is shown in Figure 2.1 via the *circle* unit. Contrary to this consistent expected flow into maintenance is the backlogged *triangle* unit, which exemplifies poor management practices (three aircraft all simultaneously about to enter phase maintenance). Both example units have eight aircraft (A1–A8). An efficient unit will induct aircraft into phase maintenance over smooth, uniform intervals, which ultimately prevents backlog and surges in spare part requirements while simultaneously offering a greater degree of predictability and capability [1, paragraph 4-59]. Thus, the Commander at the appropriate echelon will specify goals for their unit's OR and total fleet bank time in a training cycle [1, paragraph 1-29]. It further warns Commanders that "reporting high OR rates while not supporting high operational requirements" may "mask the ability to regenerate combat power" [1, paragraph 1-28].

The shift towards predictive and preventive maintenance strategies is also central to current research. Gavranis and Kozanidis [28] introduce the concept of *residual flight time*—akin to what we refer to here as *hours until phase maintenance*—proposing algorithms to maximize fleet availability. This idea of strategic scheduling is further developed by Barde et al. [29], who focus on minimizing equipment downtime through a sequential reinforcement learning model. Öhman et al. [30] discuss *frontlog-scheduling* and emphasizes the optimization

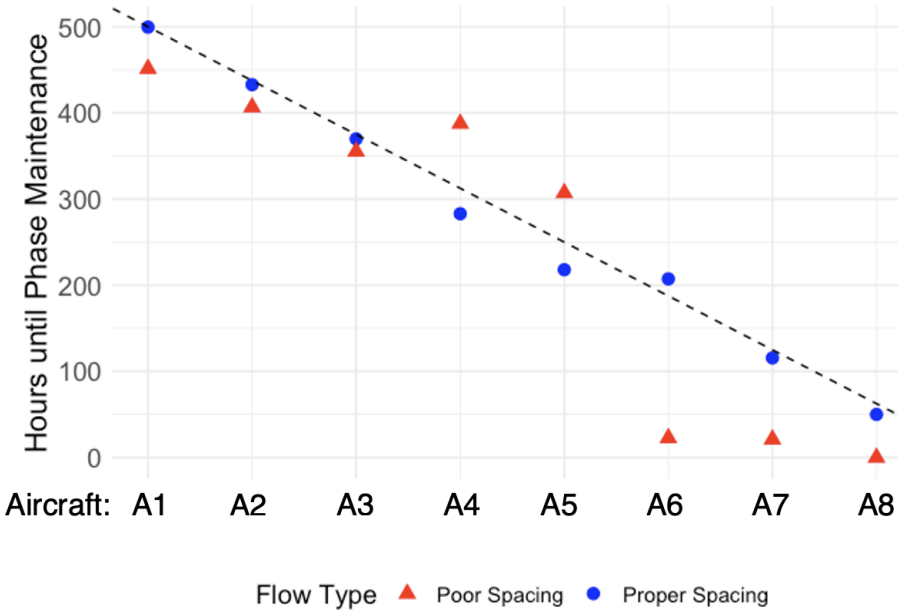

Figure 2.1: Examples of proper vs poor phase maintenance interval management (adapted from [1], ATP 3-04.7, Figures 4-2 and 4-3)

of maintenance schedules. These studies collectively highlight the importance of systematic maintenance planning in enhancing military aviation readiness.

A study by Colonel (retired) Bradley Pippin on *Allocating Flight Hours to Army Helicopters* reveals how manual flight hour allocation at the battalion level leads to decreased deployability, particularly when units hold back aircraft with low hours until phase maintenance [7]. Pippin suggests a policy adjustment in Army aviation reporting metrics by introducing a new availability metric that accounts for the deployability of aircraft in relation to maintenance schedules. He points out that "an aircraft with only one flight hour remaining until phase maintenance may be FMC, but is not available [for deployment]" [7, p.37].

The maintenance interval management doctrine discussed in ATP 3-04.7 aims to prevent maintenance bottlenecks by expecting units to distribute aircraft usage evenly throughout each maintenance cycle in order to specifically avoid the worst-case scenario of multiple aircraft with low hours until phase maintenance as highlighted by Pippin. Given the requirement to employ an aircraft evenly throughout its maintenance cycle, we hypothesize that the decision to deploy an aircraft should not be affected by its remaining hours to phase maintenance.

Hypothesis 2: *The probability of flying a FMC aircraft is independent of the remaining number of hours until required phase maintenance.*

2.3 Data and Methodology

We model the decision to fly a fully mission-capable aircraft via a GAM given a set of covariates relevant to a decision-maker.

2.3.1 Data Origin and Filtering Criteria

We commence our analysis with a dataset from the Army's Engineering Research and Development Center. This dataset includes information such as the date, aircraft model, aircraft serial number, unit hierarchy, a detailed breakdown of maintenance status by hour, and recorded flying time in hours.

We apply filtering criteria to analyze AH-64 Apache helicopter data from 1 October 2019 to 30 May 2022. In order to avoid crossing an additional fiscal year (which begins on 1 October each year), we exclude three months of data from the summer of 2019. The dataset does not continue past June 2022. This time frame comprises 973 days, encompassing 642,232 daily status reports from 797 unique aircraft as recorded on Department of the Army Form 1352. We focus on Active Duty, non-training units, reducing the dataset to 417,678 observations across 61 companies within 21 battalions. The final filtered data includes 423 unique aircraft serial numbers, each of which, coupled with a date, creates a single observation that we refer to as an *aircraft day*. We concentrate on the decision to deploy a FMC aircraft, and so filter out any observations without at least one FMC hour in the aircraft day (8.2% of data), reducing the dataset to 314,575 observations. We also impute a zero for units that do not report any flying hours for a given day. The unit with the highest observed missing flight data is missing 41% of its fleet's total potential observations. This unit had a seven-months-long period from August 2021 until February 2022 in which they did not report at least seven of their 24 aircraft. However, this time period is known to correspond to a deployment for that unit. Removing observations that contain incomplete data, such as the unit described above, reduces the dataset further. This step involved the removal of six companies and two battalions, one of which did not have available data until 2021. Flight occurs on 16.7% of the days with an average sortie length of 3.3 hours. Lastly, there are 7,165 instances in which an aircraft flew on a given day in which no FMC hours are observed—these hours were originally presumed to be comprised mainly of partially mission-capable hours. Surprisingly, there were also 1,665 occurrences (2.9% of flying days) in which an aircraft logged 24 hours of NMC time in a day and also flew. Of these occurrences, 83% spent all 24 hours in a FIELD state, suggesting that field maintenance teams routinely conduct flight tests while the aircraft are in their hands. These are all fringe cases and are ultimately dropped from

the modeling portion of the paper as we are only interested in the decision to fly when an aircraft observes FMC hours in a day.

2.3.2 Variables: Description, Operationalization, and Controls

Based on our hypothesized effects, the key independent variables in our model influencing the decision to fly a mission-capable aircraft are unit operational readiness and the remaining flying hours until phase maintenance is required. The day of the week plus the remaining number of days in the reporting period are controls that could plausibly affect the decision to fly. We further account for the year, month, and battalion to control for seasonality and training cycles. The response variable is binary: whether a FMC aircraft flies on a given day or not. First, we compare the bivariate relationships between each of the covariates and our response.

Operational Readiness OR is a unit's monthly average equipment availability rating, measured as its percentage of time in FMC status. The reporting period begins on the 16^{th} of the current month and closes on the 15^{th} of the following month. From Figure 2.2, the probability of aircraft flying on a given day decreases roughly linearly from 60% to 100%. There appears to be an inflection point at about 60% OR, suggesting that the overall relationship might not be linear. Only 15.4% of observations fall into R-3 or R-4; accordingly, the confidence interval as we approach the tail end of OR widens substantially.

Hours until Phase Maintenance Throughout this study, when referencing phase maintenance, we are referring to *major phase maintenance*, which Apaches are required to undergo after every 500 flying hours. For example, if an aircraft has flown 50 hours since its last phase maintenance, it would have 450 flying hours left until its next one. Per Figure 2.3, the probability of flying is lowest just before (after) aircraft enter (exit) phase maintenance.

We impute the number of hours that an aircraft has until phase maintenance by searching for each aircraft's top five longest consecutive streaks of NMC hours using Algorithm 1 in Appendix B.2. The logic of this algorithm is the following: first, we calculate the total flying hours over the dataset's duration and estimate the number of 500-hour phase maintenance cycles. We then focus on identifying the top five longest periods of NMC status for each aircraft to search for candidate phase maintenance cycles.

For each identified NMC period, we assess the flying hours accumulated prior to the period. This step is crucial in estimating the likelihood of the NMC period in question representing a phase maintenance cycle. We consider both the duration of downtime and the flying hours leading up to the NMC period. In making this estimation, it is important

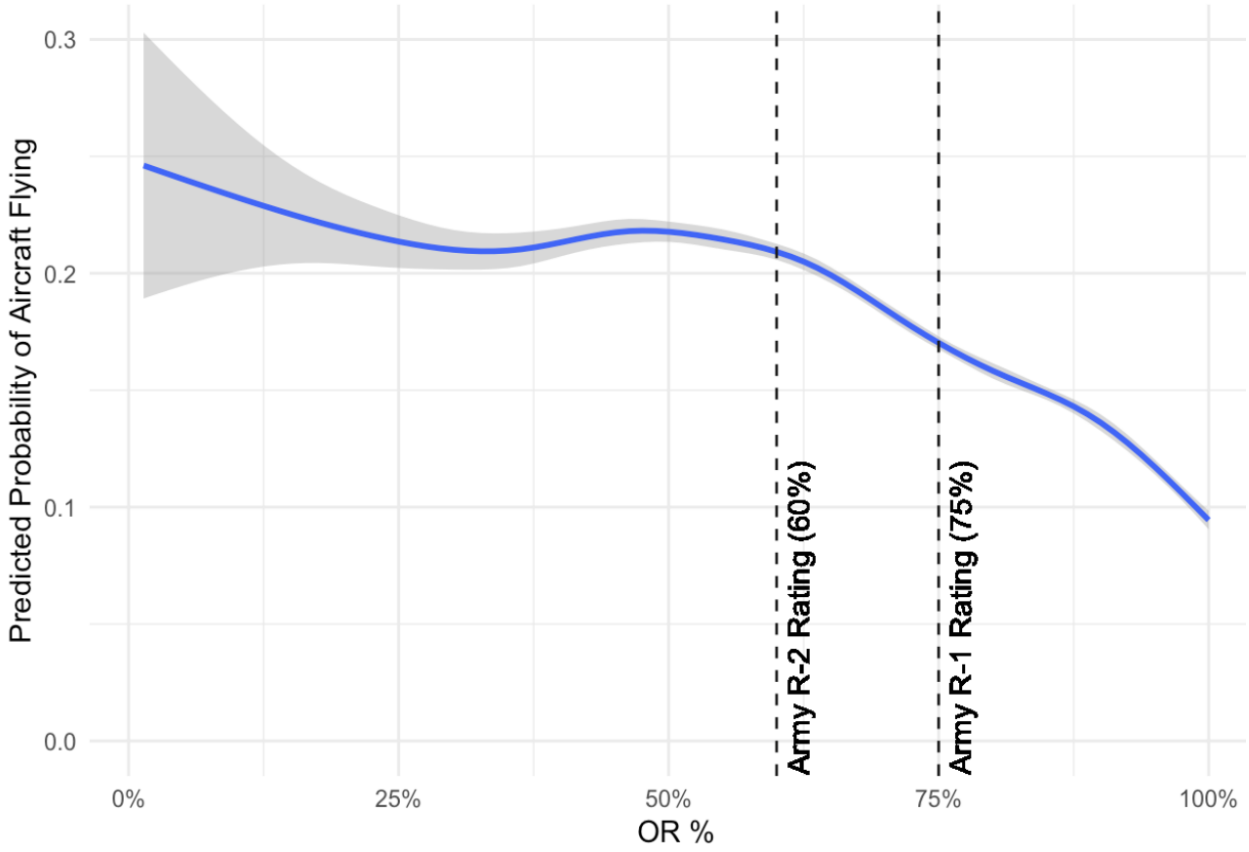

Figure 2.2: Flying vs OR: fitted GLM via LOESS (95% CI)

to recognize that extremely short durations of NMC, such as three days, are highly unlikely to correspond to a full phase maintenance period due to the insufficiency of time for all required maintenance tasks. Finally, we allocate the maintenance gaps to the most probable maintenance blocks following approximately 500 flying hours. In doing so, we also account for the possibility of early phasing by units as recommended in Paragraph 4-46 of HQDA [1], acknowledging the operational variability in maintenance scheduling across different battalions. Once we determine the date on which an aircraft enters into phase maintenance, we decrement the number of bank hours remaining on the aircraft each time it flies to keep a running tally of the number of hours until phase maintenance is required.

Control Variables We control for *Days until Reporting Period Closure*. Units are required to report their monthly average OR rating on the 15^{th} of each month–the process then resets on the following day. Thus, on any given day, we determine the number of days until the reporting period ends. We did not discretize this covariate as there is no distinct block in which doing so made practical sense. It is treated as continuous from zero to thirty

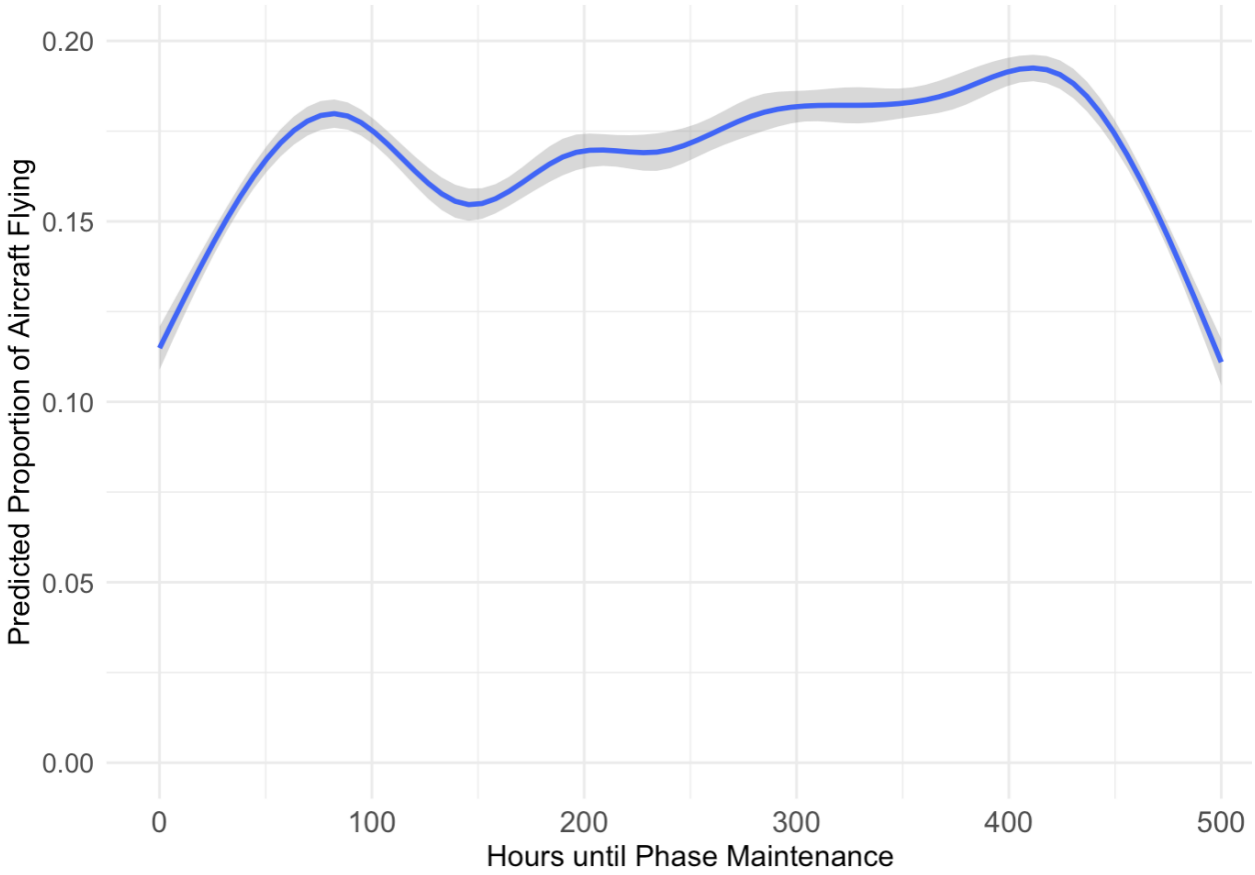

Figure 2.3: Flying vs hours until phase maintenance: fitted GLM via LOESS (95% CI)

days. We also control for the *Day of the Week*, which we treat as a categorical variable, with Wednesday serving as the reference value. We further include *Month* (reference value: January) and *Year* (reference value: 2019) as categorical variables to control for seasonality, and an aircraft's *Battalion* as a random effect.

2.3.3 Generalized Additive Model

We employ a GAM to investigate the determinants influencing the log-likelihood of a FMC aircraft being dispatched for flight on a given day. Let Y_{ij} represent the binary outcome of flight for aircraft j on date i, where $Y_{ij} = 1$ indicates flight and $Y_{ij} = 0$ indicates grounding

of aircraft j on date i; then, the log-likelihood of flight can be expressed as

$$\begin{aligned}\log\left(\frac{\hat{P}(Y_{ij}=1)}{1-\hat{P}(Y_{ij}=1)}\right) &= \beta_0 + f_1 \times \text{Operational Readiness}_{ij} \\ &\quad + f_2 \times \text{Hours until phase maintenance}_{ij} \\ &\quad + f_3 \times \text{Days remaining in reporting period}_i \\ &\quad + f_4 \times \text{Battalion}_j \text{ (Random Effect)} \\ &\quad + \beta_1^{\text{d}} \times \text{Day of the week}_i \\ &\quad + \beta_2^{\text{m}} \times \text{Month}_i + \beta_3^{\text{z}} \times \text{Year}_i + \varepsilon_{ij}.\end{aligned} \tag{2.1}$$

The smooth functions f_1, f_2, and f_3 represent continuous covariates and are estimated non-parametrically. Each function can be articulated as:

$$f_i(x) = \sum_{k=1}^{K_i} a_{ik} b_{ik}(x), \tag{2.2}$$

where the $b_{ik}(x)$ denote the basis functions, and the a_{ik} are the coefficients estimated for the i-th smooth function. K_i represents the number of basis functions utilized for the i-th smooth term. The function f_4 is associated with the smoothing spline for the random effect related to an aircraft's battalion. In this framework, the β coefficients resemble those in a logistic regression model, signifying fixed effects for each day of the week d, each month m, and each year z (2019-2022), plus an intercept. In contrast, the a_{ik} coefficients correspond to the non-linear effects captured by the smooth functions. The term ε_{ij} encapsulates the random error or the unexplained variation in the data for aircraft j on date i.

The spline fitting process minimizes a penalized likelihood criterion, where the smoothing parameter applied to reduce overfitting is adjusted based on cross-validation. Credible intervals for the fitted splines are calculated using the estimated covariance matrix of the coefficients, assuming a Gaussian distribution for these coefficients, as detailed in Wood [31].

2.4 Results and Discussion

GAMs are flexible, non-parametric models that allow non-linear fits to a response if such fits are warranted. Splines and their credible intervals in Figures 2.6–2.8 show how OR, maintenance requirements, and reporting periods relate to fight decisions. These illustrations, combined with model results in Table 2.1, allow us to assess the effect of each covariate on the decision to fly a FMC aircraft. The continuous nature and visualization of the splines aid

in model interpretability and allow us to identify trends in the data without relying strictly on fixed effects.

By selecting the optimal prediction acceptance threshold via cross-validation, we are able to achieve a maximum F1 score of 0.38 and 63.5% accuracy, as shown in Figure 2.4. F1 gradually decreases as the model moves towards a standard 50-50 weighting on flying and non-flying days. F1, which balances sensitivity and precision, is the conventional metric for classification models and is well-suited for evaluating imbalanced data such as ours. Explained deviance is adapted from Faraway [32, p. 32]; F1, sensitivity, and precision are adapted from Dalianis [33]. The number of knots chosen for each spline is selected via cross-validation. Table B.8 in Appendix B.3 shows the adequacy of fit at the given knot level of four, five, and five for OR, hours until phase, and days until report, respectively.

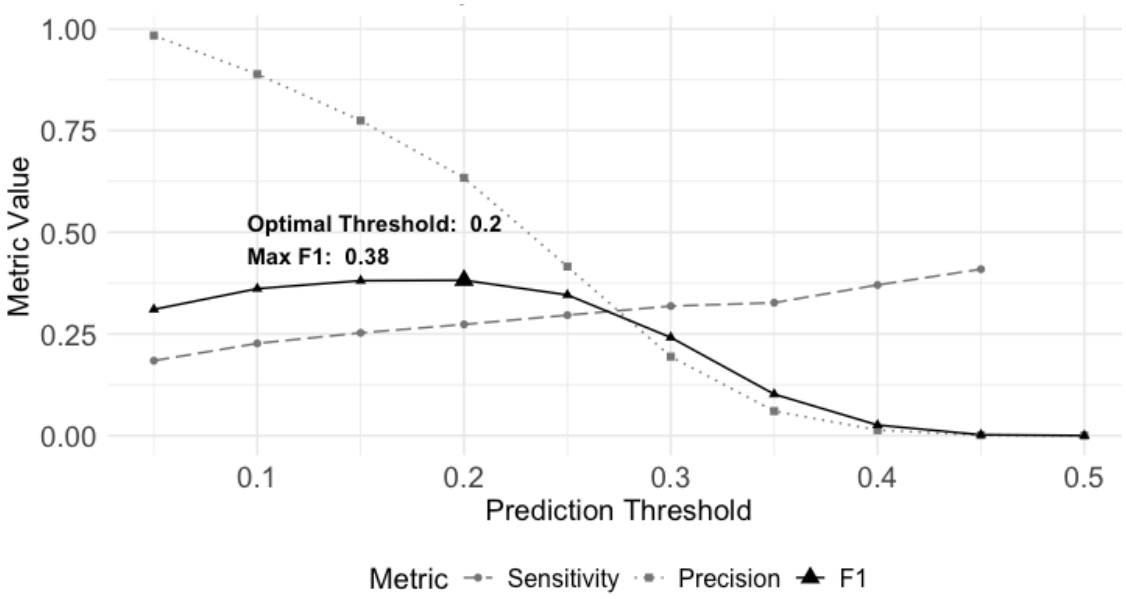

Figure 2.4: Final model (D) performance metrics across prediction acceptance thresholds using 100 iterations of 5-fold cross-validation

2.4.1 Analysis and Test of Hypotheses

We summarize the results of the regression in Table 2.1. The final model (Model D) achieves an Adjusted R^2 of 0.11 and explained deviance of 9.7%.

We evaluate Hypotheses 1 and 2 using a combination of the results presented in Table 2.1 and a component-wise visual inspection of the smoothing splines and their respective signif-

icance regions in Figures 2.6, 2.7 and 2.8. Here, we use the term *credible interval* to refer to these significance regions, which is the Bayesian analog to frequentist *confidence intervals* [34]. We follow conventional practice by using credible intervals of the smoothing splines to perform inference (cf., Hastie and Tibshirani [35] and Wood [36], sections 6.8 and 6.10). Building on the work from Marra and Wood [37], who derive credible intervals for individual smoothing splines using the covariance matrix V_f of each spline's coefficients, Wood [38] shows these intervals have Wald-like test statistics with frequentist coverage properties, allowing for the creation of posterior $1 - \alpha$ credible intervals. Following the suggestion by Ruppert et al. [39, section 6.8], we use visual inspections of the first derivative to identify areas with higher rates of change between flying and our covariates.

Table B.4 in Appendix B.1 presents odds ratios for each covariate at different levels, with all other variables held at their medians. The odds ratios provide context to the models and illustrate the practical implications of the model. The results indicate that units are almost 9% less likely to fly an aircraft with 10 hours until phase compared to 400 hours, so they are unlikely to be following the spirit of the phase interval management doctrine.

Interestingly, as shown in Figure 2.5, the day of the week has a strong impact on the decision to fly a FMC aircraft. It is the strongest determinant, in terms of magnitude, of the odds of flight on a given day in the model. While we observe a statistically significant drop in the odds of flight on the weekend of approximately 80.0%, we also find a significant decrease in the odds on Mondays (21.3%), Thursdays (19.4%), and Fridays (62.9%) compared to Wednesdays, all else constant. As expected, units are more likely to fly in the spring and fall months, which corresponds with traditional training cycles. Units are least likely to fly in 2020 and gradually increase sortie occurrences in both 2021 and 2022 but still do not return to 2019 levels.

The battalion random effect is highly statistically significant, pointing towards location-specific idiosyncrasies in determining flight operations. The number of days remaining in the reporting period has a statistically significant yet practically slight effect on flight operations. Specifically, when comparing 1 day to 15 days until the reporting period closure, the odds ratio increases from 0.465 to 0.478, reflecting a 2.8% increase. Further extending the timeframe to 25 days elevates the odds ratio to 0.523, which constitutes a 9.4% rise from 15 days and a 12.5% increase from the 1-day scenario.

Hypothesis 1: Operational Readiness Table 2.1 and Figure 2.7 provide convincing evidence against Hypothesis 1. In fact, units are generally *more* likely to fly when their OR rating is in the R-2, R-3, or R-4 levels compared to R-1, as evidenced by the sharp first derivative between OR levels of 0.8 to 1.0. The spline for OR is statistically significant,

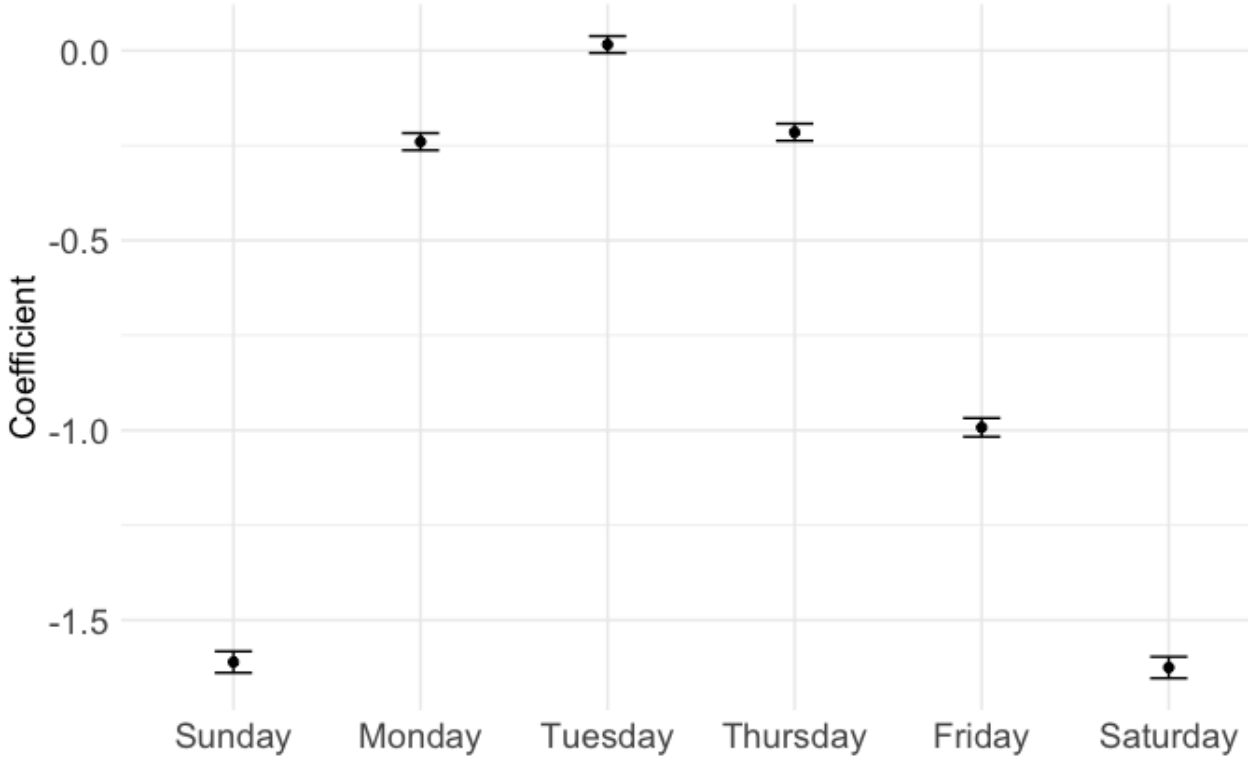

Figure 2.5: Estimated fixed effect of day of the week (full model D)

95% Confidence Interval

with an effective degrees of freedom (EDF) of 2.997, a Chi-squared statistic of 5992.7, and a p-value below 0.001, indicating a strong non-linear effect on the response variable. Our initial hypothesis that a diminished OR rating acts as a deterrent to flying is thus rejected. In Table B.4, the odds ratio for a flight occurrence at 60% OR is 0.563, which is a 64.1% increase from the odds ratio at 90% OR (0.343). Similarly, there is a 14.4% increase from the odds ratio at 75% OR (0.492), with all other variables held constant at their median values.

Hypothesis 2: Hours until Phase Maintenance We also reject **H2**. Evidence in Figure 2.8 and Table 2.1 suggests that units are significantly less likely to fly a FMC aircraft that is close to phase maintenance (either recently departed or soon to enter). The first derivative is visibly increased in the 0-50 hours and the 450-500 hours intervals, indicating a potential change in deployment decisions for aircraft in these intervals. Holding all other covariates at their median levels, an isolated increase in *Hours until Phase* from 10 to 400 hours corresponds to a 19.9% rise in the odds ratio, from 0.442 to 0.530. Additionally, a

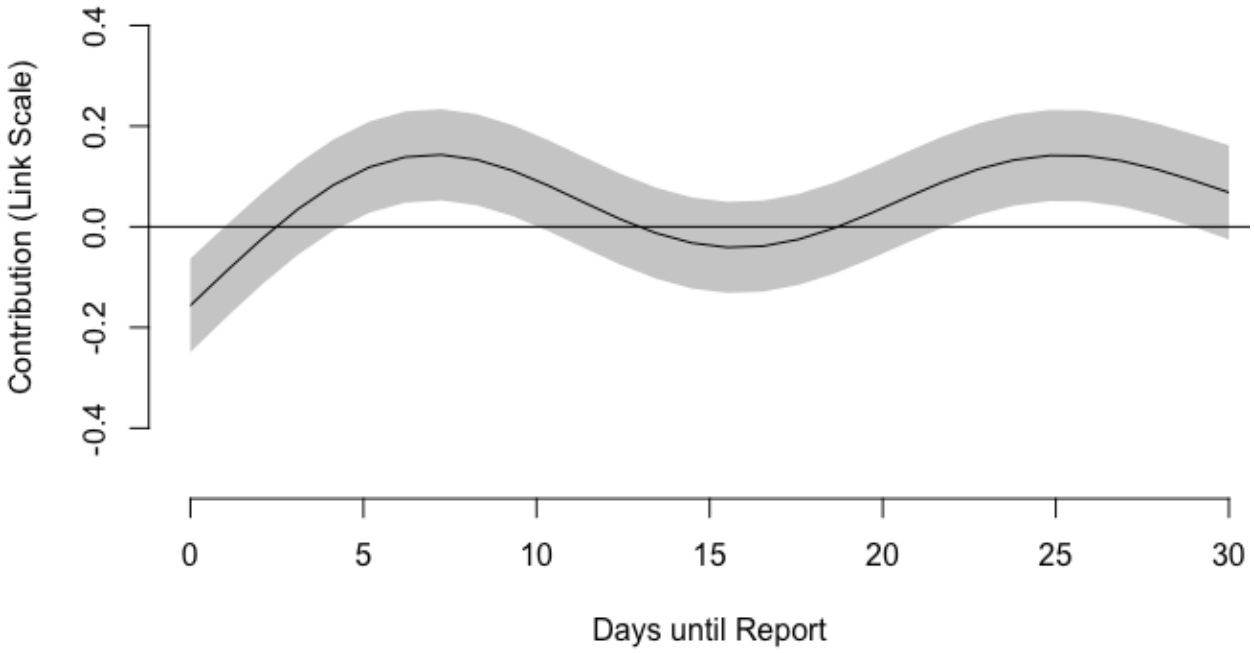

Figure 2.6: Days until report spline with 95% credible interval (full model D)

change from 250 to 400 hours yields a 7.3% increase in the odds ratio, from 0.494 to 0.530. Units clearly do not place equal emphasis on flying each aircraft regardless of its remaining hours until major phase maintenance.

Given that the Army explicitly states in its doctrine that the goal for equipment readiness is 75% FMC [13, 1-19] and that units should seek well-spaced flow into phase maintenance (such as in Figure 2.1), our findings suggest a disconnect between policy objectives and their practical implementation.

Contingency tables outlining the frequency of observations in various OR and hours until phase levels can be found in Tables B.1 and B.2 in Appendix B.1. From the contingency tables, only 45.3% of observations meeting our filtering criteria fall into R-1 status–the majority of them fall below it. Looking at unfiltered, aggregated battalion-level data in Table B.2, only 41.3% of battalion days achieve R-1 status. Again, units are more likely to fly in R-2 or R-3/R-4 compared to R-1 (cf. Figure 2.2. One potential reason behind this could be cyclical training patterns in which high flying levels and, thus, reduced OR, are associated for periods of time. Falling below R-1 status does not preclude units from further operations. Certainly, the mission clearly does (and should) take precedence, but it brings into question the efficacy of an OR-centric policy. If a unit is meeting its current mission requirements, what levels of OR might be acceptable? Currently, in doctrine, there is no

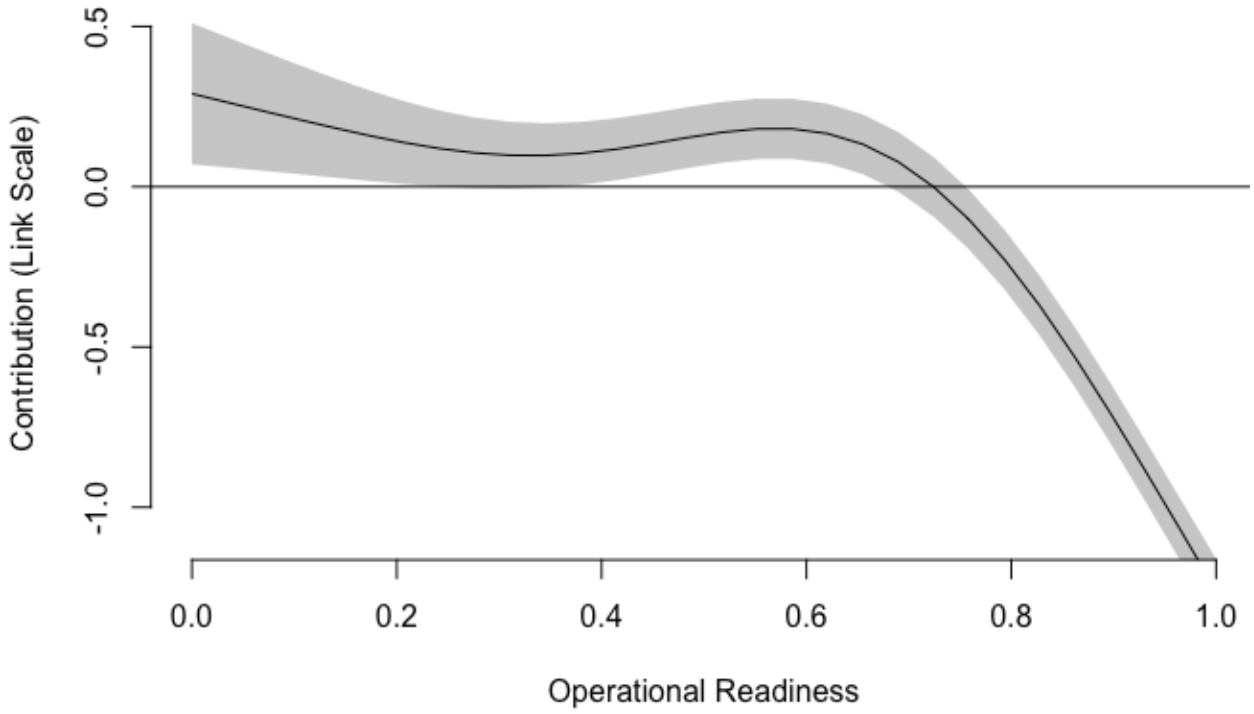

Figure 2.7: OR spline with 95% credible interval (full model D)

official link between operational intensity and OR.

In addition, we would expect maintenance officers to adjust flying hours on aircraft judiciously across a fleet to ensure the unit achieves the appropriate balance of 'bank hours' that evenly spaces aircraft into phase maintenance. It seems, however, that units are more likely to resort to adjusting flight patterns of aircraft that have either just left phase maintenance or are about to enter it. This behavior, while perhaps rational under certain circumstances, suggests a lack of long-range planning endemic to the entire force.

2.4.2 Post Hoc Analysis of an Interaction Effect

Our observations in relation to Hypotheses 1 and 2 suggest that other factors, possibly behavioral, may influence flight operations. A unit's OR status plausibly affects its flight patterns. Despite the extensive research on optimizing maintenance scheduling and spare parts logistics in military supply networks [40–44], one aspect often remains overlooked: the human element in executing these systems. Behavioral factors play a role in the effectiveness of any operational strategy, influencing outcomes in sometimes unpredictable ways. In order to further test flight operation decision-making at the unit level, we develop a model that takes into account potential differences between units operating at different OR levels. We

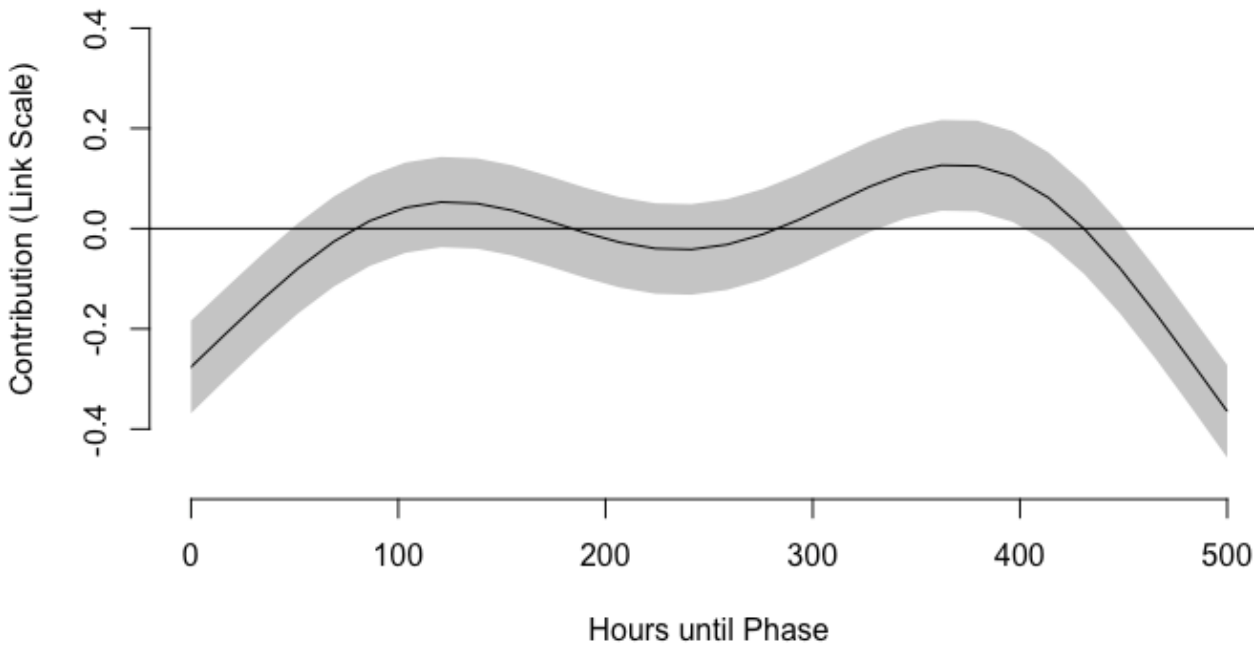

Figure 2.8: Hours until phase maintenance spline with 95% credible interval (full model D)

fit a tensor product as the interaction between OR and hours until phase. An aircraft's battalion is still treated as a random effect, serving as a proxy for latent location-specific factors like weather, mission requirements, geography, budget, and command preferences.

Again, let Y_{ij} represent the binary outcome of flight for aircraft j on date i. The formulation is consistent with Equation 2.1, with the addition of the term f_5, which represents the tensor product spline $f_1 \otimes f_2$ and models the interaction between OR and hours until phase maintenance.

A Chi-squared test of significance from an ANOVA comparing Model D and Model E indicates a statistically significant improvement in fit. Full model results are presented in Table B.5 in Appendix B.1; splines for hours until phase, OR, and days until reporting period closure are illustrated in Figures B.1-B.3 in Appendix B.1. The optimal knot selection remains the same as in Model 1 from Figure B.5. The adjusted R^2 increases to 0.113 and explained deviance to 9.81%. While interpretability suffers from the inherent additional complexity due to the inclusion of a tensor product, the interaction between hours until phase and OR is statistically and practically significant. Rather than compare odds ratios from spline coefficients, we instead can directly view predicted probabilities of flight across various scenarios, as shown in Table 2.2.

The non-linear, disproportionate decrease in the probability of flight occurrence across OR levels indicates a potential interaction effect present in the decision-making process. While the overall impact of low OR still dominates–a unit at 90% OR is not more likely to fly an aircraft than at 75% OR, regardless of the hours until phase maintenance remaining–the interaction can be summarized succinctly from Table 2.2. We observe that, as hours to phase maintenance decrease from 400 to 100, for an average unit, there is a 9.7% decrease in the probability of flying an aircraft if the OR rating is at 75% compared to a 6.4% decrease at 90% OR. Similarly, as hours to phase maintenance decreases from 250 hours to 10 hours, there is a 7.7% decrease in the probability of flying an aircraft at 75% OR versus a 12.0% decrease at 90% OR. We conclude that the marginal effect of hours until phase on aircraft deployment decisions clearly is not independent of OR. Our observations suggest the possible presence of risk-averse behavior: units striving to meet readiness ratings could opt for a conservative approach, holding back aircraft close to phase maintenance in order to reduce risk. The results are consistent with the notion from Lehman et al. [45] that low-performing organizations may avoid risky behavior.

Interestingly, the effect of days until the reporting period, as shown in Figure B.3, is diminished and is modeled as a nearly linear fit with an EDF of 1.54. In the second model, holding all other covariates at their median levels, a unit with 30 days remaining in the reporting period exhibits a 1.8% increase in the predicted probability of flying compared to when there is just one day remaining. The effect of the day of the week is still also statistically and practically significant. The findings related to Hypotheses 1 and 2 are consistent with the previous model. A deeper discussion of the post hoc model assumptions and diagnostics, specifically with regard to concurvity, can be found in Appendix B.3.

2.5 Conclusion

Evidence from this study suggests a disconnect between policy objectives and their practical implementation. Given the doctrine surrounding Army aviation, we would expect to find flight operations reduce in a unit when OR falls below R-1 and that the time until phase maintenance for an aircraft does not associate with usage. Instead, we find that diminished OR levels do not seem to preclude units from continued flight operations, and the remaining number of hours until phase maintenance has a significant impact on a unit's decision to dispatch a FMC Apache.

Our results suggest that units flying according to mission requirements may exercise limited discretion in their day-to-day operations, aiming instead to meet the overarching needs of the Command. Further, units with reduced OR are significantly less likely to fly

an aircraft with low time until phase maintenance, possibly indicating risk-averse behavior and a potential lack of long-range planning in accordance with the recommended phase maintenance interval management doctrine. Control variables such as the day of the week and the number of remaining days in the reporting period are significant across all models. These behavioral factors, which on the surface may not seem linked to flight decisions, show an unexpectedly high correlation with aircraft flight occurrences. This study thus provides potential evidence that current policies overly focus on OR and consequently do not place enough emphasis on flying (and thus training).

Future research should focus on the decision-making frameworks of different units, as such patterns could dictate outcomes that vary in efficiency and effectiveness. These findings suggest that OR alone may not be a sufficient statistic to assess aviation readiness or unit performance. A more holistic approach that also accounts for the human elements present in the system could provide the Army and aviation community with a deeper understanding of a unit's true potential for sustained and reliable combat power projection.

Table 2.1: Consolidated model summary

	Model			
Term	A	B	C	D
Spline EDF (Chi-Squared Test Stat.)				
OR	—	2.993	—	2.993
	—	**(6045.7)**	—	**(5961)**
Hours until Phase	—	—	3.994	3.993
	—	—	**(1269.2)**	**(1180.4)**
Days until Report	3.989	3.989	3.97	3.989
	(502.2)	**(507.6)**	**(504.3)**	**(508.5)**
Battalion†	17.918	17.882	17.914	17.876
	(3993.4)	**(2848.8)**	**(3791.8)**	**(2719.9)**
Coefficient Estimates (Std. Error)				
Sunday	**-1.72**	**-1.61**	**-1.72**	**-1.61**
	(0.014)	(0.014)	(0.014)	(0.014)
Monday	**-0.31**	**-0.24**	**-0.31**	**-0.24**
	(0.011)	(0.012)	(0.011)	(0.011)
Tuesday	-0.007	0.008	0.016	0.019
	0.014	(0.011)	(0.011)	(0.011)
Thursday	**-0.23**	**-0.22**	**-0.23**	**-0.22**
	(0.011)	(0.011)	(0.11)	(0.011)
Friday	**-1.05**	**-0.99**	**-1.05**	**-0.99**
	(0.012)	(0.013)	(0.012)	(0.013)
Saturday	**-1.73**	**-1.62**	**-1.73**	**-1.62**
	(0.014)	(0.014)	(0.014)	(0.014)
Month	✓	✓	✓	✓
Year	✓	✓	✓	✓
Model Metrics				
Adjusted R^2	0.097	0.11	0.0997	0.112
Explained Deviance	8.3%	9.49%	8.56%	9.71%
Count				265,472

Note: **Bold** indicates $p - Value < 0.001$; EDF: Estimated Degrees of Freedom; † random effect; ✓control variable included but not reported

Table 2.2: Predicted probability of flying for varying hours until phase and OR percentages

Hours until Phase	**OR**		
	60%	**75%**	**90%**
10 hrs.	0.536	0.489	0.330
250 hrs.	0.642	0.567	0.450
400 hrs.	0.635	0.586	0.394

Chapter 3

A Framework for Analyzing Operational Efficiency in US Army Aviation: Unsupervised Clustering of Flight Dispatch Decisions

Abstract

US Army aviation policy uses Operational Readiness (OR) as the primary mechanism for moderating aircraft utilization, yet prior evidence shows that readiness targets do not seem to meaningfully restrain flying behavior. Substantial variation in utilization-readiness outcomes exists across units operating under the same policy environment. This paper develops a diagnostic framework to characterize that heterogeneity by examining how units make flight dispatch decisions and how those patterns relate to realized outcomes.

Using longitudinal AH-64 Apache data from 19 battalions (2019–2022), we employ a three-stage methodology: (i) estimate unit-level decision profiles via Bayesian logistic regression that capture how each unit's propensity to fly varies with readiness status, proximity to phase maintenance, and temporal factors such as day of the week and days remaining in the monthly reporting cycle; (ii) cluster units on posterior coefficient distributions using a self-organizing map; and (iii) map clusters onto an empirical Pareto frontier defined by flying hours per aircraft (FHPA) and OR. This frontier captures the inherent tension between utilization and readiness by treating both as joint outcomes of observed decisions rather than optimization targets. The frontier does not prescribe how units should operate. Rather, it provides a structured way to assess how decision-making patterns relate to realized outcomes.

We find that distinct decision-making profiles correspond to different regions of the FHPA–OR trade space. Units with similar readiness can exhibit markedly different flying levels and occupy different regions of the efficiency frontier. The self-organizing map's topology-preserving structure enables a "minimum improving distance" metric that quantifies the smallest peer-anchored adjustment required for a dominated unit to resemble more efficient peers. The framework provides guidance that respects commander agency.

3.1 Introduction, Motivation, and Paper Contributions

The US Army operates under a strategic mandate to allocate resources in a manner that sustainably maximizes lethality over time [46]. Military aviation units face a fundamental tension in pursuit of this mandate. Flying aircraft builds pilot proficiency and operational capability, but accelerates wear and increases maintenance burden. Restricting flying preserves equipment availability, but degrades training and readiness for combat operations. Commanders must continuously balance aircraft utilization against maintenance capacity under limited resources, and may cycle between periods of readiness recovery and intensive use as mission demands evolve.

US Army aviation policy attempts to manage the utilization–maintenance tension through monthly reporting of Operational Readiness (OR), a scalar measure of aircraft availability that determines readiness classification, subject to command evaluation. Doctrine treats OR as a regulatory signal: declining OR is intended to indicate increasing maintenance risk and, on average, to moderate utilization without prohibiting flying when mission demands require it [1, 12, 15]. Prior empirical evidence, however, shows that OR does not meaningfully deter utilization [2]. Substantial heterogeneity exists across battalions operating under the same policy environment. Some units generate high flying hours while sustaining readiness comparable to peers, whereas others achieve similar readiness with markedly lower utilization. These differences cannot be explained by readiness targets alone. Instead, they suggest variation in how units translate observed conditions into flying decisions under a common policy. This paper characterizes that heterogeneity by estimating decision profiles for each unit from longitudinal data. A decision profile summarizes how a unit's propensity to fly varies with observed operational conditions. We cluster units on these profiles and use the results to generate peer-anchored guidance.

In Army aviation, the utilization–maintenance trade space can be represented using flying hours per aircraft (FHPA) and OR, where OR summarizes aircraft availability over a reporting period and anchors readiness classification. Doctrine defines 75% OR as the threshold for the highest readiness category (R-1) [12, 17]. While FHPA and OR are useful for reporting, they do not explain how units arrive at a given combination, nor why units in the same policy environment achieve different outcomes. The available data omit potential determinants of frontier position (e.g., mission demand, weather, maintenance capacity). We therefore reframe FHPA and OR as joint outcomes of observed flying decisions rather than optimization targets. We use the FHPA–OR plane as a diagnostic lens to analyze systematic differences in how units navigate this trade space. These decision profiles are mapped to an empirical Pareto frontier to assess whether differences in decision-making patterns are

associated with a unit's position relative to the frontier. We operationalize efficiency as the distance to the Pareto frontier, where units closer to the frontier achieve better utilization–readiness combinations relative to peers. For Pareto-dominated units, we identify how their patterns differ from those of more efficient peers.

3.1.1 Novel Contributions

In this paper, we introduce a novel framework that integrates Bayesian logistic regression with self-organizing maps (SOMs) to analyze flight dispatch decisions across US Army aviation units. Our specific contributions are:

Conceptual Contribution: Joint Outcomes Framing. Rather than treating utilization (FHPA) and readiness (OR) as optimization targets, we reframe them as joint outcomes of observed flying decisions. This two-dimensional framing facilitates more granular comparisons than readiness thresholds alone, enables construction of an empirical Pareto frontier, and supports a diagnostic approach that characterizes how decision patterns differ across units.

Methodological Contribution I: Diagnostic Pipeline. We introduce a framework that links operational decision patterns to efficiency outcomes through a three-stage process: (i) clustering units based on their decision-making profiles derived from the coefficients of our Bayesian logistic regression model; (ii) mapping these clusters to the Pareto frontier to assess whether decision profiles inform a unit's position in the trade space; and (iii) calculating the minimum adjustments needed for units in dominated clusters to adopt behaviors associated with more efficient outcomes.

Methodological Contribution II: Minimum Improving Distance. The SOM is not itself a decision model but rather a clustering vehicle that groups units based on their estimated logit coefficients, which constitute each unit's behavioral profile. We extend this unsupervised approach by introducing a "minimum improving distance" (MID) that quantifies the smallest adjustment in these coefficients required to resemble more efficient peers. Because the SOM preserves topological structure, adjacent clusters represent similar behavioral profiles. Movement to an adjacent cluster therefore represents an operationally realistic behavioral adjustment rather than an abstract optimization target. MID identifies which operational variables a unit would need to adjust, and by how much, to adopt the decision-making characteristics of more efficient peers.

3.2 Background and Literature Review

Choice Theory: The Pareto Frontier Commanders make operational choices about how to balance utilization and readiness based on their specific missions, constraints, and risk tolerances, and those choices are often justifiable given local operating conditions. A Pareto frontier provides a diagnostic lens for comparing the outcomes of those choices across units. It characterizes the set of utilization–readiness combinations for which no unit can improve along one dimension without sacrificing performance along the other. In this context, a unit is Pareto-dominated if another unit achieves higher utilization with equal or higher readiness, or higher readiness with equal utilization. This construction provides a relative efficiency benchmark for units operating under a common policy environment. The use of Pareto frontiers to evaluate trade-offs and relative efficiency is well established in choice theory and operations research [47–49]. Here, the frontier does not prescribe how units should operate; it provides a structured way to assess how observed decision-making patterns relate to aggregate utilization and readiness outcomes.

Applications of Pareto Frontiers and DEA for Efficiency Evaluation Pareto frontiers are traditionally used to assess efficiency by mapping trade-offs between inputs and outputs [50, 51]. Anderson et al. [52] use Bayesian probabilistic modeling alongside Pareto frontier constraints to find military force structures that remain robust under uncertainty. More generally, Data Envelopment Analysis (DEA), introduced by Charnes et al. [53], provides a method for comparing decision-making units against a best-practice frontier. While DEA provides efficiency scores relative to a frontier, it assumes decision-makers can identify and adjust specific inputs to reach the frontier. In our setting, the data do not reveal which operational levers commanders can manipulate or what constraints they face. We observe decisions but not the factors that produced them, so the problem is not structured in a way that DEA can address. Clustering techniques and frontier analysis have been integrated across multiple domains. For instance, Ganhadeiro et al. [54] combine SOMs with DEA to assess energy distribution efficiency in Brazil, and Kanmani et al. [55] use SOMs to assess environmental sustainability across countries. The manner in which US Army aviation units navigate the utilization–readiness trade space through systematic differences in operational decision-making has received comparatively little empirical attention.

Operational Background: Phase Maintenance One operational consideration shaping flying decisions within this trade space is proximity to scheduled phase maintenance. Army rotary-wing aircraft undergo scheduled phase maintenance at fixed flight-hour inter-

vals, during which aircraft are unavailable for flight until inspection and repair are complete [1]. For AH-64 Apache helicopters, major phase maintenance occurs at 500 cumulative flight hours and can require up to 44 days to complete [1, Table 1-1], during which the aircraft is removed from the pool of Fully Mission-Capable assets. Therefore, proximity to major phase maintenance represents an aircraft's operational state and can influence a unit's propensity to allocate additional flying hours to that aircraft. Commanders must navigate the distribution of remaining hours across the fleet in order to sustain a uniform flow of aircraft into this resource-intensive maintenance process.

Related Literature on Flight and Maintenance Planning The flight and maintenance planning literature has largely adopted a prescriptive, centralized perspective, in which a single decision-maker selects flight and maintenance schedules (often under multiple objectives and uncertainty) to achieve operational objectives. This work includes linear and nonlinear mixed-integer formulations for phase maintenance flow and availability maximization [28, 56, 57], long-horizon planning models with feasible maintenance windows, sustainability constraints, or workload smoothing [44, 58, 59], and computational extensions using heuristics or decomposition [60, 61]. Simulation-based studies similarly evaluate alternative fleet management policies under uncertainty by comparing outcomes such as availability, utilization, and schedule adherence, without modeling decentralized decision behavior [62, 63].

Across this literature, three features predominate: (i) units (when represented at all) appear as organizational groupings or resource pools accounted for by a central decision-maker, rather than as decision-making entities themselves; (ii) readiness and utilization are treated as optimization objectives or policy outputs rather than as emergent outcomes of decentralized unit-level decisions; and (iii) phase maintenance is modeled as a scheduling constraint rather than as operational state information that shapes flying decisions. In summary, the prescriptive FMP literature asks what schedule is optimal for a single unit under centralized coordination; we ask what behavioral patterns distinguish efficient from inefficient units across a fleet operating under a common policy environment. Table C.17 in Appendix C.9 organizes this body of work along these and related dimensions.

A complementary diagnostic literature analyzes realized maintenance and operational data to characterize performance variation. Data envelopment analysis has been used to assess maintenance efficiency across time periods or squadrons [64, 65], and recent statistical work examines the empirical relationship between operational readiness and flying behavior using unit-level effects [2]. These approaches focus on realized outcomes rather than on the decision processes that generate them.

Here, we adopt a diagnostic foundation while generating prescriptive guidance [66, 67].

We model units as behavioral entities and offer what we term *peer-anchored prescription*: guidance derived from how efficient units actually behave, rather than from optimization or policy mandate. Flying decisions are characterized as functions of operational state, including readiness and proximity to phase maintenance, and clustered to reveal usage patterns across units. Mapping these behavioral profiles onto an empirically observed utilization–readiness Pareto frontier associates differences in decision sensitivities with efficient and dominated outcomes. Detailed scheduling optimization, prognostics-driven condition-based maintenance, and life-cycle policy design are outside the scope of this analysis.

3.3 Data

We use data from the Army's Engineering Research and Development Center that includes daily status reports from 1 October 2019 to 30 May 2022 on US Army AH-64 Apache helicopters.[1] The dataset includes 314,575 observations from 423 unique aircraft after filtering on Active Duty and non-training units and excluding incomplete records. For each aircraft-day, the data record the aircraft's assigned unit, flight hours, downtime, and the reported cause of any downtime.

The purpose of this study is not to analyze individual aircraft operations directly, but to characterize unit-level decision patterns. We interpret unit behavior through aggregate aircraft-day dispatch patterns, not through direct observation of commander intent. To do so, we extract information from an initial explanatory model that relates flight occurrence to relevant state variables faced by units. The primary independent variables in this first-stage model are OR and the number of hours remaining until required phase maintenance. OR is measured as a unit's monthly rolling average equipment availability rating. Hours until phase maintenance represent the remaining flying time before an aircraft enters major scheduled maintenance, which for the AH-64 occurs at the 500^{th} flight hour and can require up to 44 days to complete [1, Table 1-1]. Following Semmel et al. [2], we impute remaining hours to phase maintenance. Control variables include the number of days until the reporting period closes, the day of the week, the month, and the year. On average, flight occurs on 16.7% of FMC aircraft-days, with a mean sortie length of 3.3 hours.

[1]Data cannot be published directly due to distribution restrictions; access may be granted by the issuing agency upon formal request and approval.

3.4 Methodology and Diagnostics

3.4.1 Study Design Overview

This study employs a three-stage approach to identify patterns in flight dispatch decisions and relate these to operational efficiency. We define a unit's *decision-making profile* as the set of regression coefficients (β values) from the logistic model in Equation 3.1 that characterize how specific factors influence unit flight decisions. Our methodology consists of two modeling stages followed by an optimization stage:

Stage I: Bayesian Logistic Regression. We perform Bayesian logistic regression to estimate the decision-making profile for each battalion. This captures how units respond to various operational and environmental circumstances. The Bayesian approach quantifies uncertainty in each coefficient, which propagates into the clustering stage and allows us to assess the robustness of cluster assignments.

Stage II: Unsupervised Clustering via SOM. We apply a SOM to cluster battalions based on the posterior distribution of their estimated β coefficients. We organize coefficients into five data layers (Table 3.1), each containing the full set of coefficients for a single decision dimension (e.g., the OR layer contains two coefficients: High and Low, with R-2 as the reference). We weight the SOM data layers using the kohonen package [68], which supports user-specified layer weights. We derive these weights from permuted variable importance measures from a Random Forest model. Because the Random Forest is fit to the same binary flight-occurrence outcome as in Stage I, we can use its importance measures as scaling weights for the SOM. This ensures that more influential dimensions carry proportionally greater weight in the clustering, rather than treating all coefficient sets as equally informative.

Stage III: Pareto Mapping and Minimum Improving Distance Calculation. We map the resulting clusters to the FHPA–OR Pareto frontier to assess whether distinct decision-making profiles correspond to systematically different efficiency outcomes. The SOM's topology-preserving structure enables the computation of a *minimum improving distance*, which quantifies the smallest change in a unit's decision-making profile required to shift it from a given cluster to a more efficient neighboring cluster. This calculation is performed at the layer level to allow adjustments to be interpreted in terms of specific decision dimensions, with interaction effects constrained to remain consistent with their associated main effects (see Appendix C.1 for the full formulation).

Figure 3.1 illustrates the empirical FHPA–OR Pareto frontier, which serves as the analytical canvas on which decision-profile clusters are later mapped.

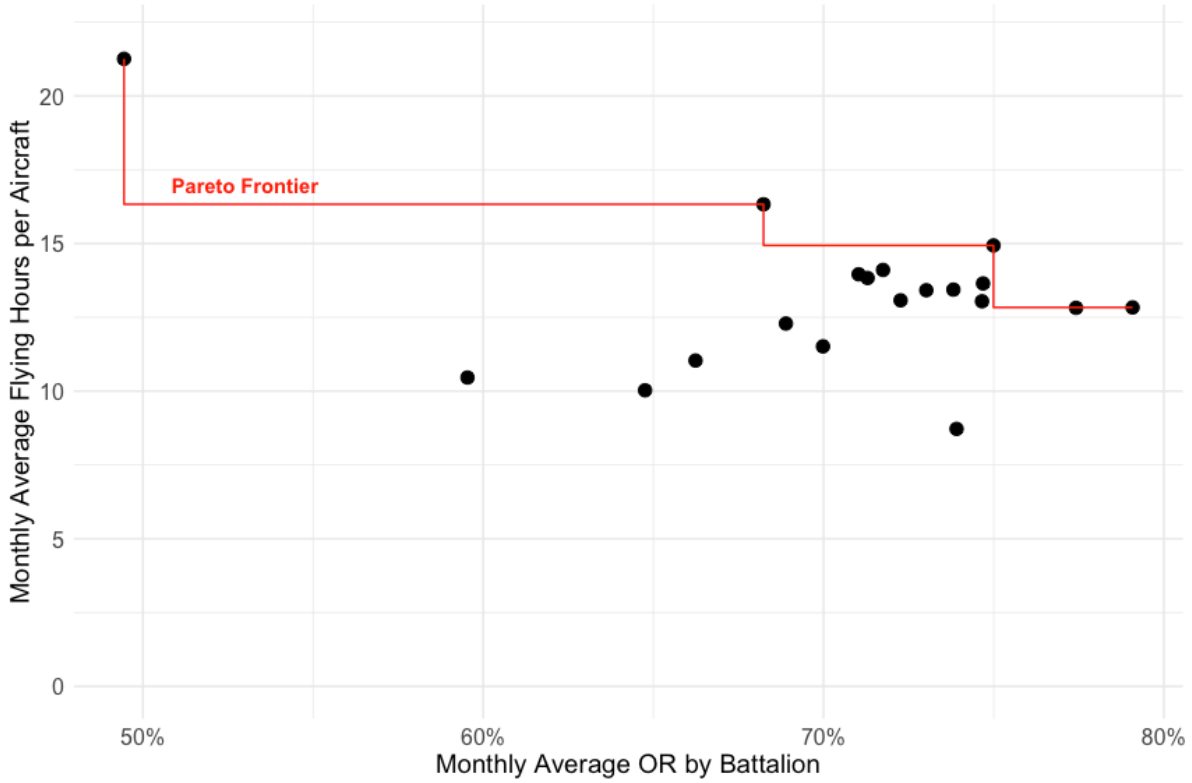

Figure 3.1: Pareto Frontier: Average Monthly FHPA vs OR

3.4.2 Stage I: Bayesian Logistic Regression

We estimate a separate Bayesian logistic regression model for each of the 19 battalions in the dataset; no hierarchical structure or partial pooling is imposed. Pooling would shrink estimates toward a common mean, which risks obscuring the behavioral heterogeneity we aim to characterize. Each model estimates the log-odds of flight occurring for an FMC aircraft on a given day, consistent with the formulation in Semmel et al. [2]. Following common practice, we choose a weakly informative prior distribution of $\beta \sim \text{Normal}(\mu = 0, \sigma = 10)$ [69, p. 55]. We discretize hours until phase and OR into High / Medium / Low categories because both variables exhibit non-linear relationships with flight occurrence [2]: hours until phase shows a quadratic pattern, while the log-odds of flight decreases as OR rises. A continuous specification would smooth over this structure. Each coefficient represents the expected change in the log-odds of flight associated with a one-unit change in the corresponding

operational factor, conditional on the other factors in the model. For OR, we follow the Army guidelines described in Table 5-3 of HQDA [12]:

R1: 100% – 75% Operational Readiness (High)
R2: 74% – 60% (Medium)
R3: 59% – 50% (Low)
R4: below 50% (Low)

For modeling purposes, we combine R3 and R4 into a single "Low" category, with R2 (Medium) as the reference. These categories follow doctrine directly [12]; results should be interpreted as differences across readiness levels rather than as claims about smooth marginal effects. We discretize hours remaining until phase maintenance into High, Medium, and Low categories using empirically derived thresholds at 400 and 100 flight hours. These cutpoints correspond to the roots of a generalized additive model that capture interpretable transitions in the relationship between proximity to phase maintenance and flight occurrence; full details are provided in Appendix C.4. Days until reporting period ends ($\beta_{(3)}$) is the number of days remaining in the monthly reporting period and is included as a continuous predictor. Day of the week, month, and year are treated as factors.

Let Y_{ijb} represent the binary outcome of flight for aircraft j on date i from battalion b, where $Y_{ijb} = 1$ indicates flight and $Y_{ijb} = 0$ indicates grounding. We model $Y_{ijb} \sim \text{Bernoulli}(P_{ijb})$, where the log-odds of flight are

$$\begin{aligned}\text{logit}(P_{ijb}) = \beta_0 + \boldsymbol{\beta}_{(1)}^\top \mathbf{OR} + \boldsymbol{\beta}_{(2)}^\top \mathbf{Hrs} + \beta_{(3)}\text{Days} + \boldsymbol{\beta}_{(4)}^\top \mathbf{DoW} \\ + \boldsymbol{\beta}_{(5)}^\top (\mathbf{OR} \otimes \mathbf{Hrs}) + \boldsymbol{\beta}_{(6)}^\top \mathbf{Month} + \boldsymbol{\beta}_{(7)}^\top \mathbf{Year},\end{aligned} \tag{3.1}$$

with **DoW** specifying the day of the week. Table 3.1 summarizes the notation and layer structure used throughout this paper. Month and year are included as controls in the regression model to account for temporal effects, but are excluded from the SOM because they do not reflect persistent unit decision-making behavior.

Table 3.1: Coefficient Structure and SOM Layer Weights ($\alpha_{(\ell)}$)

Layer (ℓ)	Coefficient Vector	Dimension	Components	$\alpha_{(\ell)}$
1	$\boldsymbol{\beta}_{(1)}$	2	OR: High / Low (Ref: Medium)	0.205
2	$\boldsymbol{\beta}_{(2)}$	2	Hours to phase: High / Low (Ref: Medium)	0.123
3	$\beta_{(3)}$	1	Days until reporting period ends	0.259
4	$\boldsymbol{\beta}_{(4)}$	6	DoW: Mon / Tue / Thu / Fri / Sat / Sun (Ref: Wed)	0.361
5	$\boldsymbol{\beta}_{(5)}$	4	OR × Hours interactions	0.052
–	$\boldsymbol{\beta}_{(6)}$	11	Month (excluded from SOM)	–
–	$\boldsymbol{\beta}_{(7)}$	3	Year (excluded from SOM)	–

Individual components within a layer are indexed as $\beta_{(\ell_i)}$: layer ℓ, component i. For example, $\beta_{(1_1)}$ and $\beta_{(1_2)}$ denote OR High and OR Low, the two components of the layer 1 coefficient vector, $\boldsymbol{\beta}_{(1)}$.

Posterior inference is performed via Hamiltonian Monte Carlo in Stan [70]. From each battalion's model, we draw $n = 500$ samples from the posterior distribution of each of the 15 coefficients (see Table 3.1). This produces a 500×15 matrix of posterior draws per battalion. Across all 19 battalions, this produces 9,500 total observations (19×500), each characterized by 15 coefficient values. Figure 3.2 illustrates one example of how coefficient distributions can differ across units.

3.4.3 Stage II: Unsupervised Clustering via the Self-Organizing Map Algorithm

We perform unsupervised clustering on the full posterior distribution of the estimated regression coefficients to capture uncertainty in each unit's decision-making profile. For each battalion, the clustering input consists of posterior draws of the $\boldsymbol{\beta}$ coefficient vectors from the Stage I Bayesian logistic regression. This approach allows units to be compared based on both the magnitude and uncertainty of their estimated responses to operational conditions. The SOM is trained on posterior draws rather than point estimates specifically to assess the robustness of cluster assignments to estimation uncertainty.

The clustering proceeds in three steps. First, we group the regression coefficients into *data layers*, where each layer consists of the subset of $\boldsymbol{\beta}$ coefficients associated with a single decision dimension, as defined in Table 3.1. Second, we weight these layers using relative importance measures obtained from a Random Forest model, which quantify each dimension's influence on flying decisions. Third, we apply a SOM to cluster units based on the result-

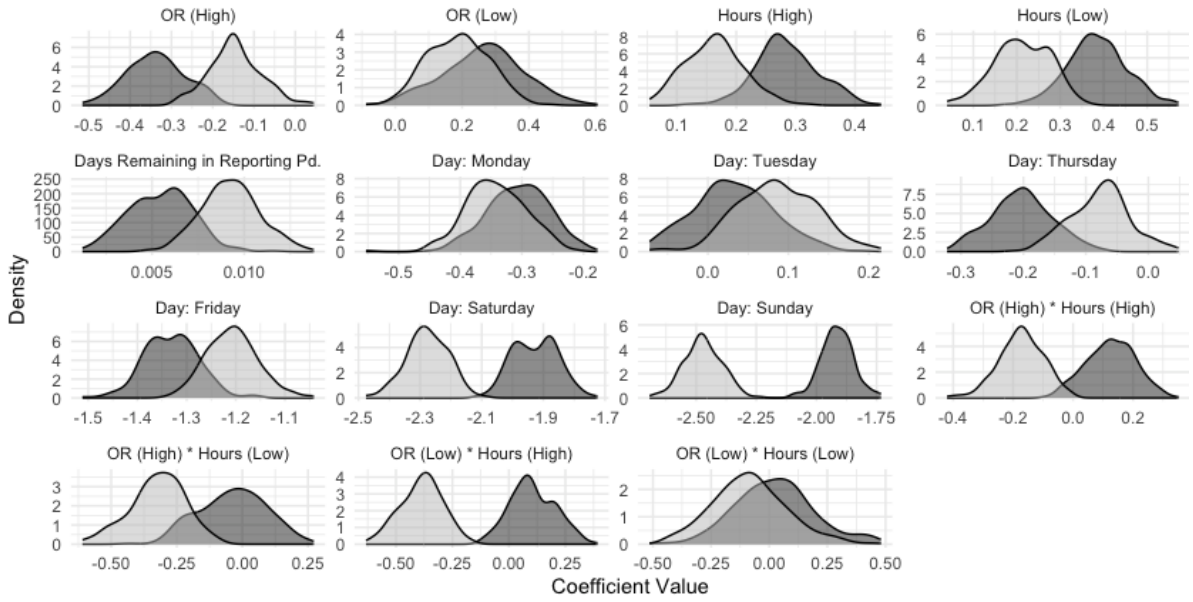

Figure 3.2: Coefficient density comparison for two high-efficiency battalions (D and H, both Cluster 2). Different operational emphases can produce similar positioning relative to the efficiency frontier.

ing weighted coefficient structure. Month and year controls ($\boldsymbol{\beta}_{(6)}$, $\boldsymbol{\beta}_{(7)}$) are excluded from the clustering to focus on persistent decision-making behavior rather than circumstantial temporal effects.

Model Layer Weights via Random Forest Permuted Variable Importance We use permuted variable importance as outlined in Nicodemus et al. [71] to calculate the weights of each of the five data layers. Since the GINI impurity metric can be biased towards factor variables with more categories (e.g., three categories for OR and hours to phase versus seven for day of the week), we elect to use permuted importance instead [72]. The weights are scaled such that their sum equals one. These weights prevent the SOM from conflating estimation uncertainty with behavioral significance. In an unweighted Euclidean space, the clustering algorithm prioritizes coefficients with the largest variance. For Bayesian posteriors, high variance often reflects statistical noise rather than operational emphasis. The RF importance weights force the SOM to measure distance based on predictive power rather than posterior spread. The Random Forest serves purely as a structural scaling device, not as a competing behavioral model. The day of the week is the most influential, with a weight of 0.361, followed by days until report (0.259), OR (0.205), hours to phase (0.123), and interaction terms between OR and hours to phase (0.052).

SOM Application and Grid-structure Tuning We fit the SOM for 5,000 iterations over each grid structure. Our selection of $(x, y) = (3, 2)$ is based on the traditional elbow method [73], adapted by utilizing the sum of the squared quantization error as the distance metric [74, p. 49, 129]. Larger grids risk overfitting: with only 19 battalions, a 12-node grid averages fewer than two units per cluster, while a 16-node grid would leave most clusters empty. We select 6 nodes to balance fit and interpretability within the bias-variance tradeoff. The SOM arranges these nodes in a hexagonal lattice structure, where each node connects to its immediate neighbors. This structure preserves topological relationships, so behaviorally similar clusters occupy adjacent positions on the lattice. Table 3.2 shows the quantization error by grid structure.

Table 3.2: Quantization Error for Different Grid Structures

x	y	Nodes	Quant. Error
2	2	4	1.086
3	**2**	**6**	**0.426**
2	3	6	0.687
4	2	8	0.723
2	4	8	0.446
3	3	9	0.587
4	3	12	0.288
3	4	12	0.346
4	4	16	0.262

3.4.4 Stage III: Pareto Mapping and Minimum Improving Distance

This stage measures how dominated unit profiles differ from frontier unit profiles. The efficiency frontier is constructed empirically from unit-level FHPA–OR aggregates computed over the full observation period, with the non-dominated envelope defining the realized utilization–readiness trade space. Distance to the frontier is the Euclidean distance to the nearest frontier point after (z-score) standardization of each dimension (FHPA–OR) to ensure equal weighting. Because unit positions reflect data aggregated over the span of the study, temporal robustness is examined separately in Section 3.5. The lattice structure constrains which clusters a unit can realistically move toward. A unit in cluster c cannot directly adopt the profile of a distant cluster without first resembling intermediate neighbors.

Variable Definitions and Layer Structure Each unit's decision-making profile is represented by its posterior mean coefficients, while cluster boundaries are defined by the SOM prototype vectors. Perturbation analysis computes the minimum change to a unit's posterior mean coefficients, by layer, required to cross the decision boundary into a more efficient cluster. Let $\boldsymbol{\beta}$ represent the posterior mean coefficient values for each unit from Equation 3.1, partitioned into $L = 5$ analytical layers:

$$\boldsymbol{\beta} = \big(\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \beta_{(3)}, \boldsymbol{\beta}_{(4)}, \boldsymbol{\beta}_{(5)}\big),$$

where $\boldsymbol{\beta}_{(1)}$ = OR layer, $\boldsymbol{\beta}_{(2)}$ = Hours until Phase layer, $\beta_{(3)}$ = Days until Report layer, $\boldsymbol{\beta}_{(4)}$ = Day of Week layer, and $\boldsymbol{\beta}_{(5)}$ = Interaction layer. When referencing individual components within a layer, we use $\beta_{(\ell_i)}$ to denote component i of layer ℓ; for example, $\beta_{(1_1)}$ and $\beta_{(1_2)}$ represent the OR High and Low coefficients within the coefficient vector $\boldsymbol{\beta}_{(1)}$. Month and Year controls ($\boldsymbol{\beta}_{(6)}$, $\boldsymbol{\beta}_{(7)}$) are excluded from the SOM clustering to focus on persistent operational behaviors. We use importance weights $\boldsymbol{\alpha} = (\alpha_{(1)}, \alpha_{(2)}, \alpha_{(3)}, \alpha_{(4)}, \alpha_{(5)}) = (0.205, 0.123, 0.259, 0.361, 0.052)$ for each layer, derived from the Random Forest model in Section 3.4.3 (see Table 3.1).

Minimum Perturbation for Independent Layers For layers 1–4, we solve an unconstrained optimization problem. Let $\mathbf{p}^c_{(\ell)}$ and $\mathbf{p}^n_{(\ell)}$ denote the SOM prototype vectors for layer ℓ in clusters c (current) and n (neighbor), where $c > n$ indicates movement toward a more efficient cluster (clusters are indexed by efficiency ordering, with cluster centroids at increasing distances from the Pareto frontier; see Section 3.5). The decision boundary normal vector is

$$\mathbf{u}_{(\ell)} = \mathbf{p}^n_{(\ell)} - \mathbf{p}^c_{(\ell)}.$$

The unnormalized signed projection of the unit's position $\boldsymbol{\beta}_{(\ell)}$ relative to the decision boundary is

$$d_{(\ell)} = \mathbf{u}^\top_{(\ell)}\boldsymbol{\beta}_{(\ell)} - \tfrac{1}{2}\big(\|\mathbf{p}^n_{(\ell)}\|^2 - \|\mathbf{p}^c_{(\ell)}\|^2\big). \tag{3.2}$$

Geometrically, $d_{(\ell)} < 0$ indicates the unit's coefficient vector lies on the current-cluster side of the hyperplane bisecting the two prototypes, while $d_{(\ell)} \geq 0$ indicates it has crossed to the neighbor-cluster side.

The minimum perturbation required to cross the boundary is

$$\Delta\boldsymbol{\beta}_{(\ell)} = -\frac{d_{(\ell)}}{\|\mathbf{u}_{(\ell)}\|^2}\,\mathbf{u}_{(\ell)}. \tag{3.3}$$

The *minimum improving distance* (MID) for layer ℓ is the $\alpha_{(\ell)}$-weighted Euclidean norm of this perturbation:

$$\text{MID}_{(\ell)} = \alpha_{(\ell)} \cdot \|\Delta\boldsymbol{\beta}_{(\ell)}\|_2 = \alpha_{(\ell)} \cdot \sqrt{\sum_i \left(\Delta\beta_{(\ell_i)}\right)^2}, \tag{3.4}$$

where $\alpha_{(\ell)}$ is the layer importance weight from Table 3.1. MID quantifies the smallest weighted change in a unit's estimated decision-response pattern that would move it into a cluster associated with a more efficient frontier position. In operational terms, MID identifies the path of least resistance for a unit to shift its behavioral profile toward the prototype of a more efficient cluster. Because these adjustments come from observed peer behavior, they describe how efficient units actually operate rather than how units should operate in some theoretically optimal sense.

The SOM assigns units based on the weighted sum of per-layer distances. The full decision boundary between two clusters satisfies $\sum_\ell \alpha_\ell d_\ell = 0$. A single-layer crossing is sufficient for reclassification only when the remaining layers already favor the target cluster on net. MID identifies the layer that requires the smallest behavioral adjustment, but crossing that layer's boundary alone does not guarantee reclassification. Robust recommendations require concordance between the MID-identified target and the cluster most similar to the unit's full profile. The profile concordance analysis in Appendix C.8 addresses this gap.

Interaction Layer: Constrained Optimization The interaction layer requires special consideration to ensure consistency among the four interaction coefficients. Interaction terms are not treated as independent decision variables. To preserve practical implementability, perturbations to the interaction layer are constrained to be consistent with the perturbations applied to their corresponding main-effect layers. Without this constraint, the optimization could recommend contradictory adjustments, such as simultaneously increasing and decreasing sensitivity to OR, or produce interaction effects that conflict with the underlying main effects (e.g., when two negative perturbations yield a positive product). The constraint ensures that interaction effects amplify or attenuate changes already implied by the primary variables, rather than introduce independent structure that commanders cannot operationalize.

The full constrained optimization formulation is provided in Appendix C.1. It includes auxiliary variables and exclusivity constraints that ensure no coefficient is simultaneously increased and decreased.

3.5 Results and Discussion

The SOM clustering identifies six distinct behavioral profiles. Cluster 1 occupies the efficiency frontier as measured by OR and FHPA under the observed policy environment; Clusters 2 through 6 fall at progressively greater distances from it, as measured by mean distance to the frontier (Table 3.3). The primary differentiators are sensitivity to OR status and phase maintenance timing. Case studies illustrate how the minimum improving distance metric translates these behavioral differences into relative recommendations for underperforming units.

Stage II Results: SOM Clustering and Operational Profiles Units cluster into six distinct behavioral profiles based on their estimated decision-making patterns. Cluster membership is summarized by the proportion of posterior draws mapping to each SOM node; concentration of mass on a single node indicates a robust behavioral profile, while dispersion across nodes reflects estimation uncertainty rather than evidence that the unit operates some fraction of the time under different cluster prototypes. Under the SOM clustering, 18 of 19 battalions map to their modal cluster in at least 96.8% of posterior draws (mean 98.1%; minimum 72.4%; see Appendix C.2 for details on the outlier unit). Figure 3.3 displays the prototype profiles associated with units from each cluster.

Behavioral heterogeneity across units is systematic and measurable. Two units with identical monthly OR can occupy different clusters and exhibit markedly different flying patterns. OR alone cannot distinguish these behaviors.

Stage III Results: Minimum Improving Distance Analysis Cluster assignment helps explain where a unit lands on the FHPA–OR frontier. Figure 3.4 maps the observed assignments for each unit to this space. Clusters 2 and 3, nearest the frontier, are closely grouped, whereas Cluster 5 shows greater dispersion. Table 3.3 quantifies these differences. Average distance to the frontier increases from 0.47 (Cluster 2) to 2.35 (Cluster 6). Admittedly, unobserved factors such as mission requirements, weather, and maintenance capacity likely also influence frontier position. Still, this correspondence raises a natural question: what behavioral adjustments would move a unit from its current cluster toward a more efficient one?

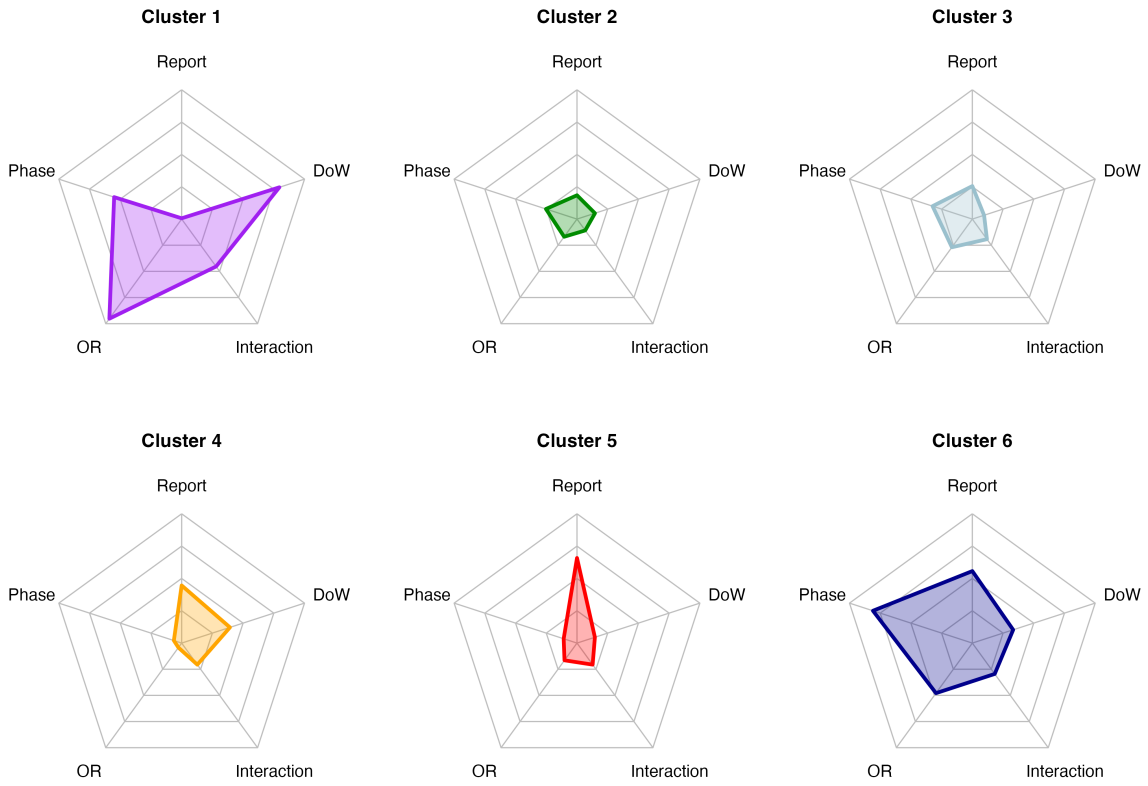

Figure 3.3: (color online) Characteristic profiles of each cluster across five operational dimensions. Each axis displays the mean of the absolute SOM prototype coefficients for that dimension. Larger coefficients indicate stronger sensitivity to conditions in that layer, not better outcomes; the mapping to the Pareto frontier reveals which profiles correspond to which efficiency regions.

Table 3.3: Summary of Normalized Distances to Pareto Frontier by Cluster. Cluster 1 lies on the frontier and thus has no measurable distance.

Cluster	Avg. Dist.	Median Dist.
1	–	–
2	0.474	0.395
3	0.619	0.574
4	0.968	1.04
5	1.16	1.16
6	2.35	2.35

The minimum improving distance for each unit can be computed as a function of the distance from that unit's position in the coefficient space to the centroids of neighboring

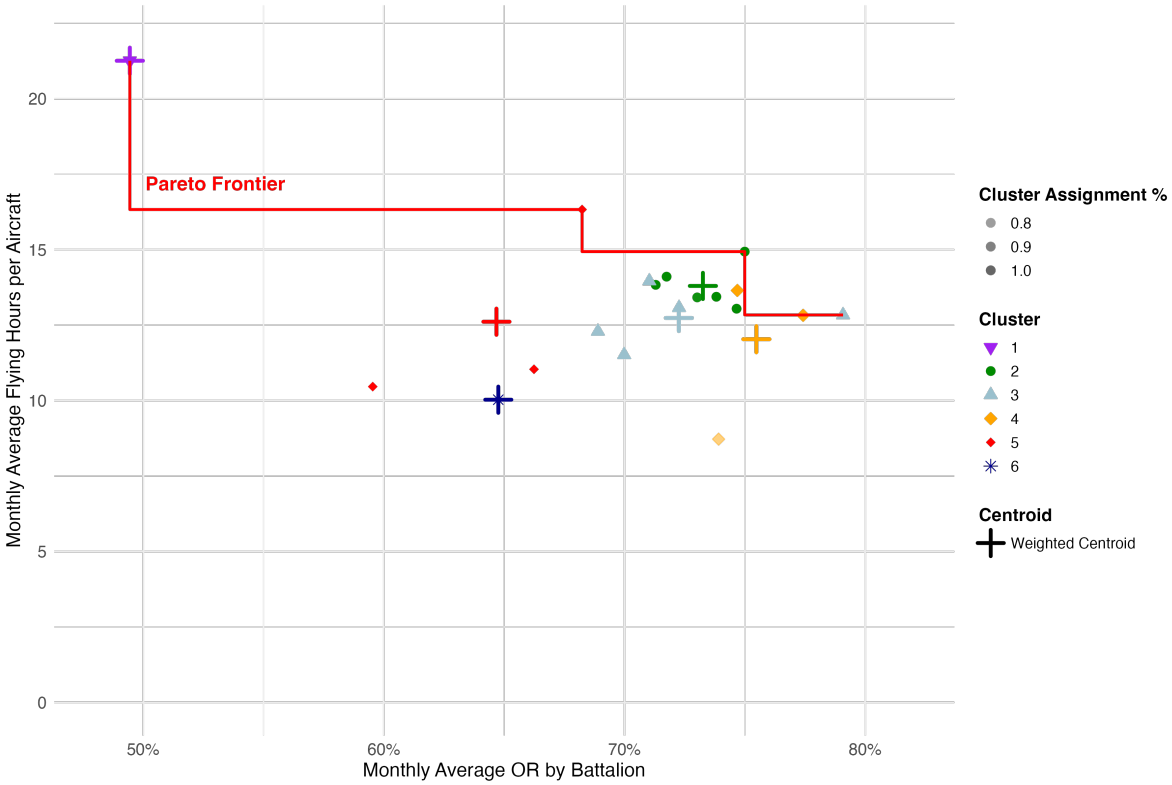

Figure 3.4: (color online) Mapping of Unit Cluster Assignment to Pareto Frontier

clusters with better efficiency. Minimum improving distance applies only to Clusters 3–6. Cluster 1 is a single unit with an atypical profile, and Cluster 2 units are largely non-dominated and consistently perform near the frontier over time (Section 3.5.2 examines this temporal consistency in detail). Neither group stands to benefit from the analysis. Figure 3.5 shows the minimum total perturbations by layer for each unit in Clusters 3–6. Lower perturbation values identify the decision dimensions where the smallest behavioral adjustments would move a unit toward a more efficient cluster.

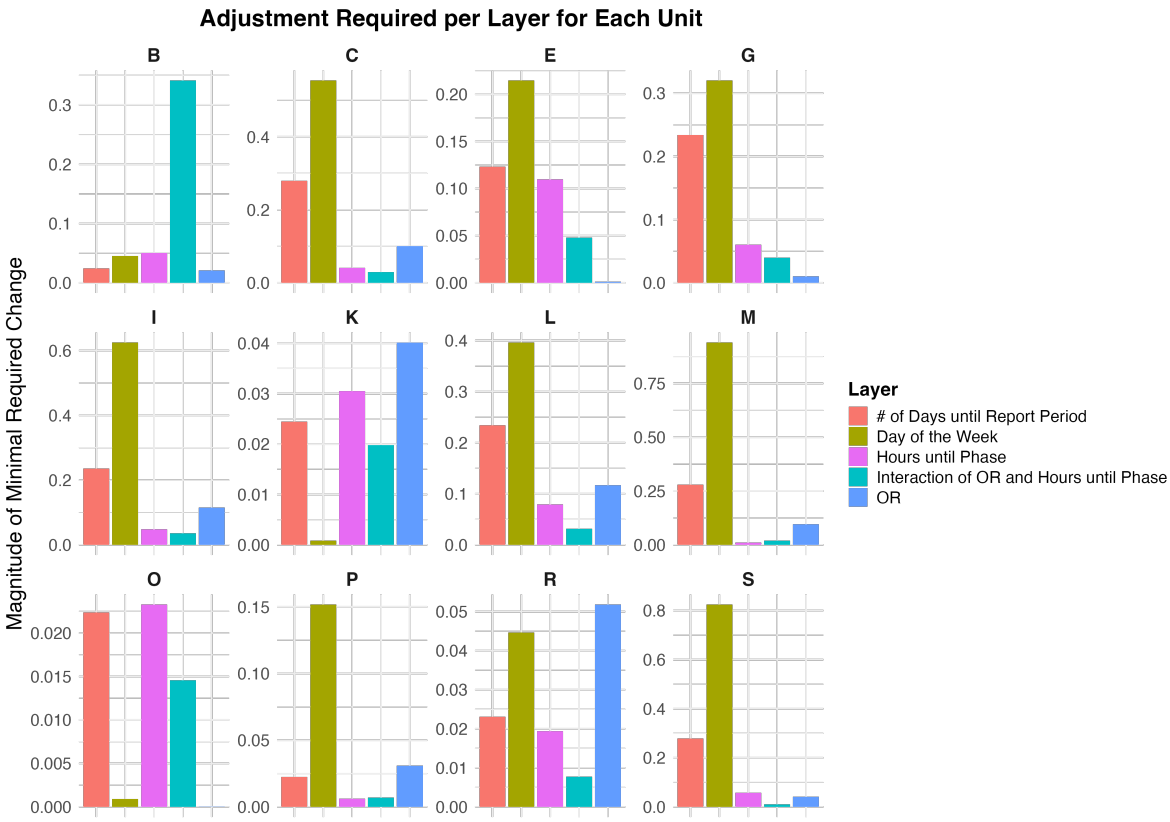

Figure 3.5: (color online) Minimum total perturbations by layer for each unit, with values weighted by the α layer importance factors. Lower values indicate that smaller adjustments to that layer would be sufficient to shift the unit into a more efficient neighboring cluster.

MID identifies the most efficient entry point: the single layer requiring minimum perturbation. A low MID does not guarantee that the unit's full behavioral profile resembles the target. To check whether the nearest boundary also represents the most similar destination, we compute profile concordance metrics against each candidate target (Appendix C.8). The Generalized Jaccard Index [GJ; 75] measures element-wise overlap between a unit's behavioral profile vector and a cluster prototype. Values range from 0 (no overlap) to 1 (identical profiles). GJ agrees with Intersection over Union [IoU; 76, 77] on the radar chart for all 18 non-frontier units. As an independent check, Quantization Error (QE) uses the same weighted distance in the full 15-dimensional coefficient space that guided grid selection (Table 3.2; 78, p. 53). Among the 12 non-frontier units in Clusters 3–6, 7 are already closer to the MID-recommended target than to their own cluster prototype by this measure. For 7 of the 9 units with multiple improving neighbors, all concordance metrics confirm the MID target. For Unit E, the metrics diverge. Three cases illustrate how the diagnostics translate

into unit-level recommendations.

3.5.1 Case Study: From Analysis to Peer-Anchored Recommendations

To translate our decision support framework into practical recommendations, we present a case study examining three units (B, E, and R). Table 3.4 summarizes baseline readiness and flight activity for these units. Table 3.5 shows how flight rates vary with OR levels and phase proximity.

Table 3.4: Baseline readiness and flight activity metrics for Units B, E, and R. The unit of observation is one FMC aircraft on one day. OR is a unit-level daily measure shared by all aircraft in the unit. Hours to phase is aircraft-specific: each aircraft's remaining flight hours until phase maintenance on that day. Flight Rate is the percentage of FMC aircraft-days on which any flight occurred. Monthly FHPA equals the mean daily flying hours per aircraft multiplied by 30.

	Unit B	**Unit E**	**Unit R**
Cluster	3	6	3
Readiness			
Avg OR (%)	78.8	64.9	70.0
OR SD (%)	8.3	19.5	14.1
Flight Activity			
Flight Rate (%)	14.5	11.7	12.0
Avg Sortie (hrs)	3.4	2.9	3.3
Monthly FHPA	14.9	10.2	12.0
Avg Hrs to Phase	248	203	212

Table 3.5: Conditional flight rates (%) for Units B, E, and R. Each value is the percentage of FMC aircraft-days on which any flight occurred, restricted to the indicated condition. Left panel: conditioned on the unit's daily OR level, a unit-level measure shared by all aircraft on a given day. Right panel: conditioned on the individual aircraft's hours remaining until phase maintenance, an aircraft-specific measure. Unit E maintains a nearly consistent flight rate regardless of phase proximity, while efficient peers modulate flight decisions based on aircraft state.

	By Unit OR			**By Hours to Phase**		
Unit	**$\geq$ 75%**	**60–75%**	**$<$ 60%**	**$>$ 400 hrs**	**100–400 hrs**	**$<$ 100 hrs**
B	12.6	19.3	19.6	14.0	15.0	13.9
E	5.8	14.5	15.8	11.2	11.4	12.4
R	9.2	14.9	10.5	12.6	12.6	10.4

The MID coefficients, flying hours estimates, and concordance metric values in each case study illustrate relative magnitude and diagnostic context. They are not precise implementation targets, nor are they guarantees. MID identifies the nearest decision boundary. The concordance metrics assess whether that boundary leads to a structurally similar destination. Together they inform commander judgment rather than prescribe specific outcomes.

Unit B: OR Adjustment Unit B (Cluster 3) exhibits decision patterns that differ from Cluster 2 primarily in OR management. Cluster 2 profiles reflect a 1.8% to 2.7% higher flying probability, driven by greater propensity to fly under both high OR (+0.0176) and low OR (+0.0271) conditions. Given an average sortie length of 3.4 hours, this represents 0.64–0.96 additional flying hours per month across the fleet. More efficient peers fly more frequently across all OR conditions while maintaining comparable average readiness levels.

Unit E: Target Discordance MID identifies Cluster 2 as the nearest single-layer boundary for Unit E (Cluster 6), with an OR perturbation of $+0.0018/-0.0012$. GJ and IoU both favor Cluster 3 (GJ = 0.55 vs. 0.51 for Cluster 2), while QE favors Cluster 4 (QE = 1.75 vs. 1.81 at Cluster 2). Unit E is the sole unit in Cluster 6 and the only unit with three improving neighbors. Each metric family selects a different target. If Unit E targeted Cluster 3 instead, the minimum single-layer MID would run through the Hours layer (0.002 vs. 0.001 for Cluster 2 via OR). Compared to Cluster 3 peers, Unit E differs in how it flies aircraft into phase (+0.0030 for high hours, +0.0046 for low hours), approximately 0.15 additional flying hours per month. Unit E is the only unit where MID and the concordance metrics fully diverge. Unit B shows partial divergence (GJ and IoU favor Cluster 1 while MID and QE favor Cluster 2), but all similarity values are low at both targets. Unit E illustrates the intended use of this framework as a diagnostic lens rather than an automated prescription.

When the metrics diverge, the framework surfaces the trade-off for commander evaluation rather than selecting a single answer.

Unit R: OR and Hours to Phase (Interaction) Adjustment Compared to more efficient peers in Cluster 2, Unit R (Cluster 3) differs in how it schedules aircraft based on the unit's current OR levels and each aircraft's hours remaining until phase maintenance. Specifically, more efficient peers tend to prioritize flying aircraft closest to phase maintenance during high OR periods ($\geq 75\%$) and prioritize flying aircraft farthest from phase maintenance during low OR periods ($<60\%$). This pattern pushes aircraft through phase when readiness can absorb the downtime and preserves flexibility during periods of constrained availability. In practice, the adjustment is not about total flight volume but about aircraft selection. When OR is high, the commander would prioritize aircraft nearest to phase for flight. This accelerates phase turnover while the unit has readiness to spare. When OR drops below 60%, priority shifts to aircraft farthest from phase to avoid additional phase inductions when the unit can least afford the downtime. Table 3.5 shows the gap: Unit R currently flies phase-urgent aircraft (< 100 hours) at 10.4% compared to 12.6% for phase-distant aircraft (> 400 hours). Cluster 2 peers reverse this ratio.

3.5.2 Temporal Robustness of SOM Clustering

The Pareto frontier in Figure 3.4 displays time-averaged positions, which may obscure month-to-month variation in individual unit performance. A unit represented by a single averaged point could have spent every month in a tight vicinity of that position, or it may have varied substantially around it. The Stage II SOM clustering (Section 3.4.3) uses the posterior distribution of coefficients estimated from data pooled across the full study period. A natural concern is whether these clusters reflect persistent decision-making patterns or are artifacts of averaging over heterogeneous periods (e.g., if a unit's behavior changed substantially due to deployment or a major training rotation). To address this concern, we employ two complementary approaches: (1) Time Above Pareto (TAP), which traces each unit's frontier position over the study period, and (2) Dynamic Time Warping (DTW), which clusters units by the shape of their movement through the FHPA–OR space.

Time Above Pareto (TAP) TAP quantifies the proportion of monthly observations each unit spends above a series of constructed Pareto frontiers, defined at various FHPA percentiles with OR fixed at the 75% R-1 threshold. Figure 3.6 illustrates these frontiers for Cluster 3. Table 3.6 shows that TAP generally decreases as we move down the cluster efficiency rankings (Cluster 1, a single-unit outlier, is the exception), and this pattern holds

across all FHPA thresholds. Figure 3.7 shows month-to-month positions for Cluster 2 units relative to the Pareto frontier (OR = 0.75, 75^{th} percentile FHPA). Cluster 2 exhibits the highest TAP values across all thresholds. Trace plots for all clusters appear in Appendix C.3.

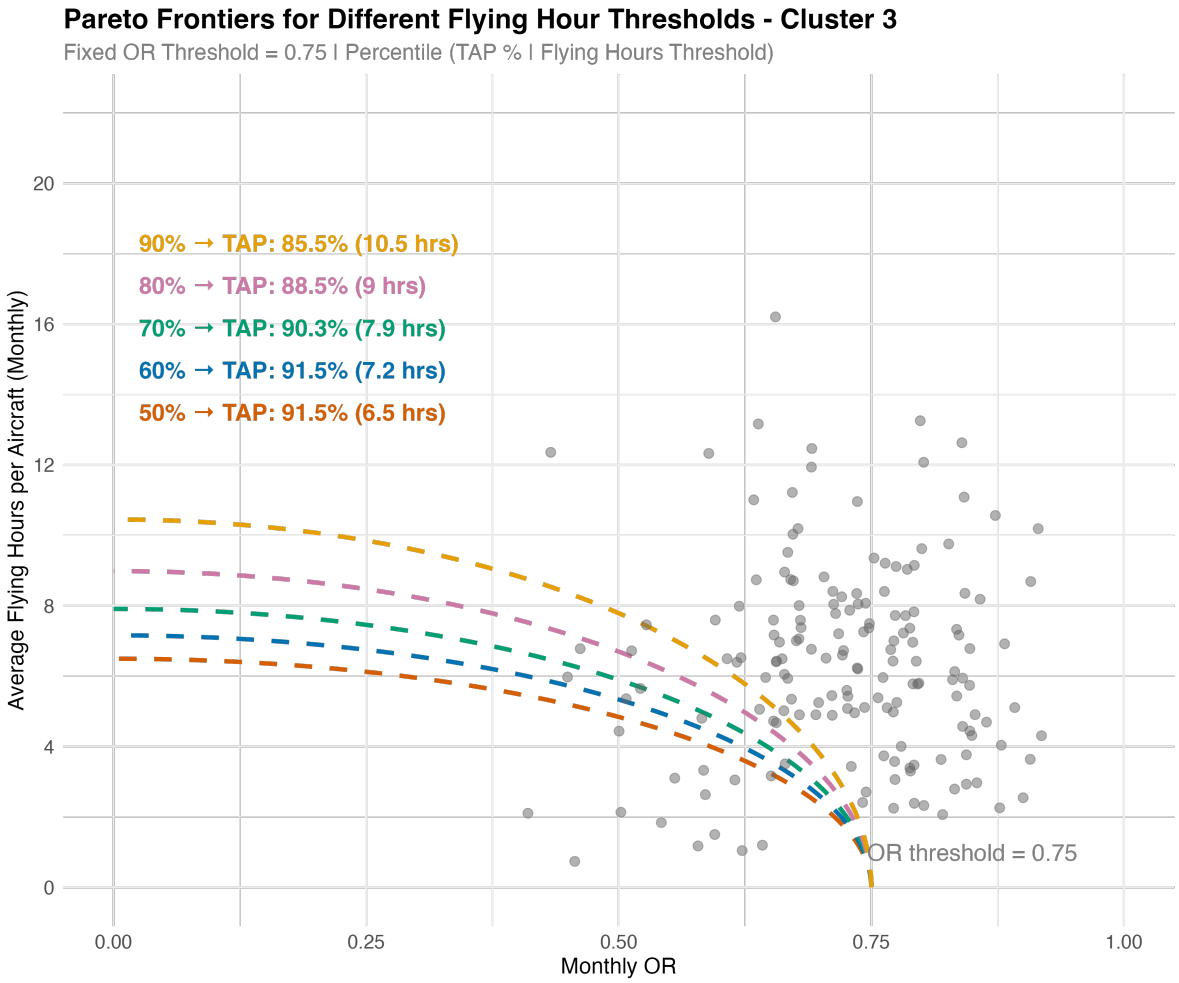

Figure 3.6: Pareto Frontiers at Various FHPA Percentiles for Cluster 3. Each frontier corresponds to a different FHPA threshold (50%–90%), with OR fixed at 0.75. TAP values indicate the proportion of monthly observations above each frontier.

Table 3.6: TAP Sensitivity to Flying Hours Threshold (Fixed OR Threshold = 0.75)

Cluster	FHPA Percentile				
	50%	**60%**	**70%**	**80%**	**90%**
1*	93.9	93.9	84.8	84.8	78.8
2	98.5	98.5	96.0	94.9	94.4
3	91.5	91.5	88.5	87.9	85.5
4	87.9	86.9	86.9	85.9	83.8
5	78.8	77.8	76.8	74.7	66.7
6*	81.8	81.8	75.8	69.7	66.7

*Clusters 1 and 6 each contain a single unit; TAP values reflect that unit's trajectory rather than a cluster-level pattern.

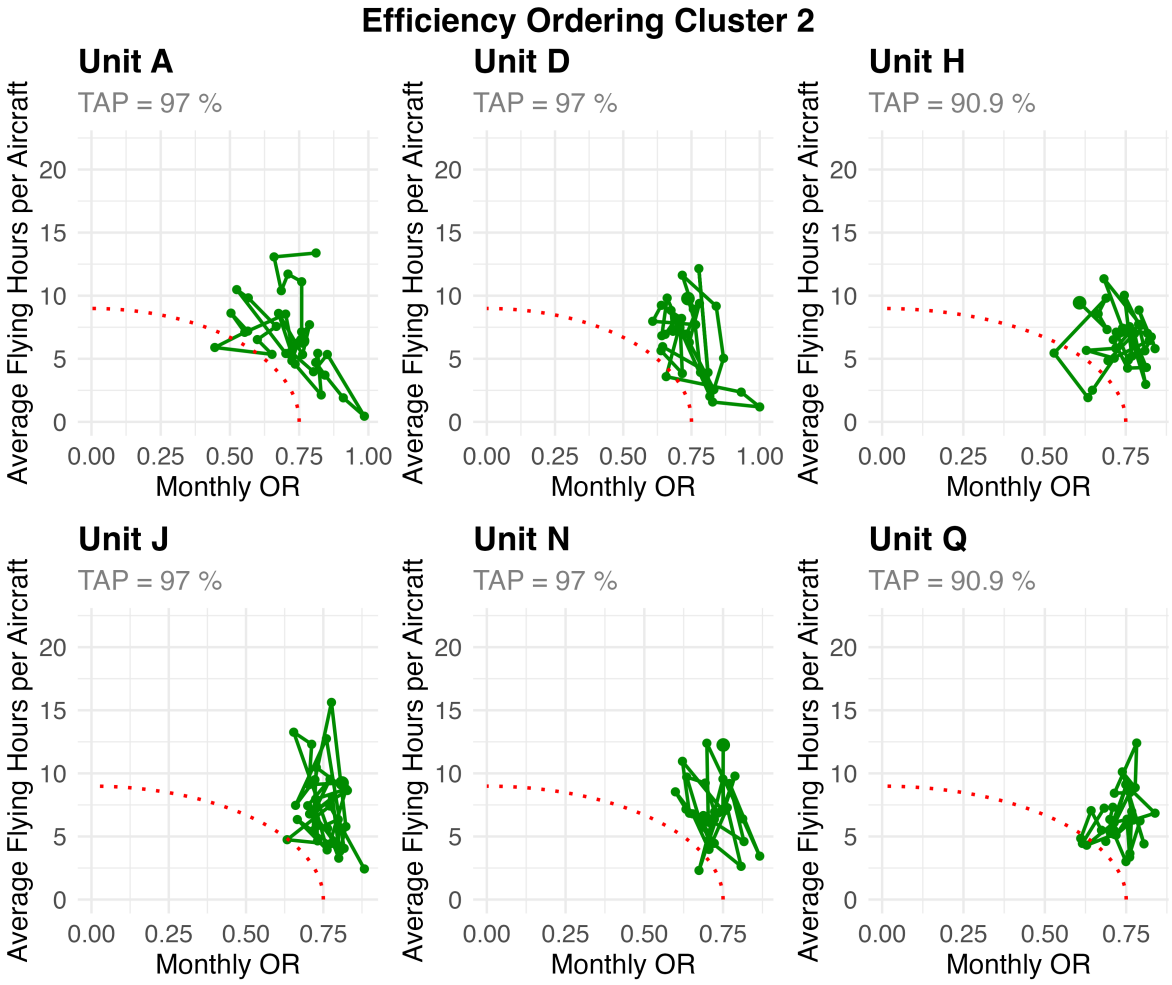

Figure 3.7: Cluster 2 Traces: Monthly OR vs FHPA positions relative to the Pareto frontier (OR = 0.75, 75^{th} percentile FHPA).

Dynamic Time Warping (DTW) DTW clusters units by the shape of their movement through the FHPA–OR space over time [79]. We compute pairwise DTW distances between unit traces and apply two clustering methods, hierarchical [80] and spectral [81]. To quantify agreement between these temporal groupings and the SOM-based clusters, we compute Normalized Mutual Information (NMI) [82], which is well-suited for comparing clusters with imbalanced sizes [83, 84]. Between SOM and DTW with hierarchical clustering, we obtain $NMI = 0.416$; between SOM and DTW with spectral clustering, we obtain $NMI = 0.450$. These values suggest moderate concordance by informal analogy with κ benchmarks [85], though no consensus thresholds exist for NMI in the clustering literature.

Together, these results provide evidence that the SOM clusters are robust to temporal considerations. TAP decreases with efficiency rankings, and DTW-based groupings show moderate concordance with coefficient-based clusters.

3.6 Conclusion, Limitations, and Future Work

The current OR-based reporting structure does not explain why units operating under the same policy achieve different utilization-readiness outcomes. By reframing FHPA and OR as joint outcomes of observed decisions, our three-stage framework identifies the decision-making profiles that distinguish efficient units from less efficient ones and maps them to positions on the efficiency frontier.

Our framework links observable decision-making patterns to efficiency outcomes and translates those patterns into actionable, peer-anchored guidance. The "minimum improving distance" metric quantifies which operational variables a dominated unit would need to adjust, and by how much, to resemble more efficient peers. This guidance is behaviorally grounded, in that it shows commanders how efficient units behave under the same policy and similar circumstances, without prescribing optimal schedules or mandating policy changes. The implicit message is not "you must do this," but "this is what units outperforming you actually do."

Limitations There are practical limitations to our model. Most significantly, we lack mission-specific and deployment status data for units. Events such as Combat Training Center (CTC) rotations or deployments lead to increased flight activity and introduce usage patterns that our model cannot capture. Additionally, the clusters we identify represent average behaviors within each group.

Scope and Interpretation One might argue that OR was never intended to measure efficiency, or that utilization is inherently mission-driven. Both points are valid. OR was designed for readiness classification, not efficiency measurement; interpreting it alongside utilization data extends rather than criticizes its original purpose. However, the existence of systematic behavioral variation within the same policy environment suggests that unit-level decisions, not just mission demands, shape outcomes. Understanding these patterns does not require redefining OR; it requires interpreting OR in context. Frontier position is conditional on unobserved factors; units closer to the frontier may face systematically different operational demands than those farther away.

Recommendations The following recommendations concern measurement, not doctrine. Current Army aviation metrics focus primarily on OR—not on operational tempo. A unit flying significantly more while maintaining the same OR level is demonstrating higher efficiency, yet no official policies exist to measure these differences. We recommend that the Army consider metrics beyond OR that allow commanders to understand their position relative to the Pareto frontier.

Framework Transferability An analyst applying this framework to a different fleet would (1) estimate unit-level decision profiles using the Stage I regression structure, (2) cluster units by behavioral similarity, (3) map clusters to the relevant efficiency frontier, and (4) compute minimum improving distances for dominated units. In principle, the framework is agnostic to airframe type; the primary requirement is longitudinal data linking equipment status, utilization decisions, and maintenance timing. Treating readiness and utilization as joint outcomes opens the door to more meaningful evaluation of unit performance without redefining readiness itself. Alternative frontier constructions (e.g., mission-capable subcategories or maintenance man-hours where available) represent natural extensions of this approach.

Acknowledgements

This research was made possible due to the gracious support from the Omar N. Bradley Fellowship. A special thank you to LTC Thomas Dirienzo from the US Army; Dr. Kyle Miller from Carnegie Mellon University; Mr. Gregory Jinks and Mr. John Holdcraft from the US Army's Program Executive Office–Aviation; and Ms. LaKenya Walker from the US Army Engineer Research and Development Center.

Chapter 4

The Marginal Value of Prediction Accuracy in Capacity-Constrained Fleet Maintenance under Stochastic Demand

Abstract

Predictive maintenance strategies aim to preempt reactive failures using evolving health estimates, yet the marginal value of improving prediction accuracy in capacity-coupled fleets remains poorly understood. We study a stochastic fleet management problem in which noisy, real-valued estimates of remaining useful life (RUL) are used to allocate utilization and maintenance under stochastic demand. This paper quantifies the marginal value of RUL prediction accuracy in such systems. We ground this analysis in US Army rotorcraft operations, where structured maintenance doctrine provides a transparent environment for isolating information-driven decisions. We use a stochastic simulator calibrated to doctrine to optimize interpretable decision-tree policies across five accuracy levels, measured by the coefficient of variation (CV) of the RUL signal. A blocked factorial design isolates the causal effects of prediction accuracy and policy adaptation.

We find that the value-of-accuracy relationship is strongly concave. Approximately 79% of achievable improvement is captured by CV=25%, and 98% by CV=10%. Factorial results show that prediction accuracy is the dominant driver of performance. Policies trained under noisier prognostic information perform nearly as well as those trained under more precise signals when evaluated at the same CV. The contribution is decision-theoretic: organizations should identify the level of prediction accuracy at which operational returns saturate and design maintenance policies to adapt to that information, rather than treat prognostics as an overlay on usage-based systems. Within capacity-constrained fleet operations, CV determines whether the fleet is constrained by information or by maintenance capacity.

4.1 Introduction

Fleet operators face a fundamental tradeoff between utilization and maintenance under uncertain equipment health [86–89]. When demand is stochastic and maintenance capacity is constrained, this tradeoff becomes a dynamic scheduling problem in which decisions depend on imperfect health information. A common prognostic output is remaining useful life (RUL), the estimated operating time before an asset requires maintenance. The resulting challenge is a value-of-information (VOI) question: how does uncertainty in RUL estimates propagate through maintenance decisions to affect fleet-level outcomes? Prior studies show that condition monitoring improves decisions [90–92], but they rarely isolate how information quality alone affects performance. This paper focuses on that marginal effect. Quantifying the marginal value of accuracy requires a framework that bounds the range of possible outcomes and isolates information quality as the causal factor.

The theoretical bounds of this analysis are defined by two benchmark policies. A usage-based policy that ignores RUL entirely and intervenes only at fixed utilization thresholds or upon failure establishes the lower bound. A policy with perfect health information (CV=0%) establishes the upper bound, where $\text{CV} = \sigma/\mu$ is the coefficient of variation of the RUL estimate. Prediction accuracy, as used throughout this paper, refers to the precision of the RUL estimate available to the planner. Improvements may come from better sensors, better prognostic models, or better signal processing; the framework is agnostic about the source. The difference defines the *prognostic potential*: the maximum operational improvement achievable through perfect health information. Existing analyses typically compare "with prognostics" to "without" while simultaneously changing decision rules or operating assumptions. Because information quality, policies, and operating environments interact, those designs cannot determine how much accuracy itself drives performance gains. This requires a framework that holds the operating environment fixed, allows policies to adapt to each accuracy level, and isolates the marginal benefit of improved prognostics.

This study provides that framework. We address the "how good is good enough" question for RUL prediction by estimating the marginal improvement in operational outcomes associated with reductions in prognostic uncertainty. The policy class and operating environment are held fixed while the policy itself adapts to different levels of RUL accuracy. We ground the analysis in US Army rotorcraft operations, where structured maintenance doctrine provides a transparent and constrained planning environment. Capacity coupling is central to the problem: each aircraft's maintenance induction consumes a shared slot, changes fleet-wide availability, and shifts the priority ranking for every remaining aircraft. Although calibrated to military aviation, the model represents a general fleet problem with

stochastic demand, capacity-limited maintenance, and imperfect health signals.

Fleet performance is assessed through two complementary metrics. Operational readiness (OR) is the proportion of a unit's fleet that is mission capable at a given time, with a doctrinal target of 75% [12]. Mission success (MS) captures whether available aircraft satisfy stochastic daily demand. Together, OR and MS operationalize the utilization-maintenance tradeoff. OR measures fleet availability and MS measures operational throughput. Figure 4.1 illustrates this tension using unit-level readiness and utilization data from Semmel et al. [2], which reports monthly OR and FHP achievement rates for US Army rotorcraft units over a multi-year observation period. Units that fly more aggressively achieve higher mission throughput but lower readiness.

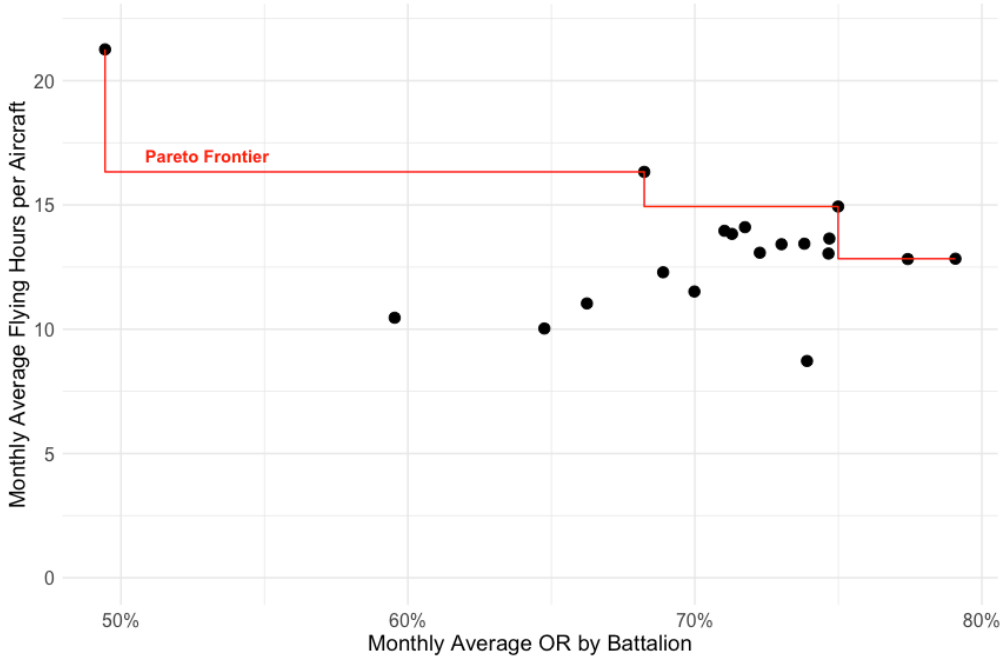

Figure 4.1: Operational readiness vs. Flying Hour Program achievement tradeoff for US Army rotorcraft units [2]. Units face a fundamental tension: flying more hours to meet mission requirements degrades aircraft and reduces readiness.

We organize the study around two research questions:

1. **What is the marginal value of prediction accuracy in fleet operations?** We seek the shape of the value-of-accuracy relationship across MS and OR jointly, from no prognostic information to perfect foresight.

2. **Does the improvement arise from better predictions, from better-adapted**

policies, or from their interaction? We isolate these effects through a blocked factorial design and test whether the finding generalizes across adverse operating conditions.

This paper contributes a simulation-optimization framework for partially observed fleet systems with capacity coupling, paired with a blocked factorial design that isolates prediction accuracy from policy structure. Both components transfer beyond the military context. The simulation-optimization approach applies to any setting with stochastic demand, imperfect health signals, and binding maintenance constraints. The factorial design provides a template for causal inference in value-of-information studies where policies adapt to information quality. Together, they reframe the central question from "is predictive maintenance beneficial?" to "at what level of prediction accuracy do operational returns saturate?"

4.2 Related Work

4.2.1 Flight and Maintenance Planning

The Flight and Maintenance Planning (FMP) problem integrates asset scheduling (which aircraft to fly when) with maintenance scheduling (which aircraft to maintain when) to maximize fleet availability and related operational objectives under capacity constraints. FMP-style problems arise in commercial aviation [93], military operations [58, 61, 94], and in the scheduling and maintenance of heavy equipment fleets characterized by utilization-driven maintenance and binding capacity constraints. FMP is difficult for three reasons. First, maintenance triggers depend on cumulative utilization and couple today's scheduling decisions to future maintenance obligations. Second, constrained maintenance resources must be allocated across competing demands. Third, asset failures introduce stochastic downtime with uncertain duration.

Early work in military aviation explicitly recognized this coupling between flight hour allocation and downstream maintenance burden. Pippin [7] formulates the Army flight hour allocation problem as an optimization model designed to stabilize the flow of aircraft into phase maintenance. His work shows that naive maximization of short-term flying hours can compromise deployability despite high reported readiness. This work establishes a foundational FMP insight: future fleet-level availability depends not only on aggregate utilization, but on how utilization is distributed across assets relative to maintenance thresholds.

The standard FMP formulation later formalized these ideas using mixed-integer programming to optimize scheduling over a rolling horizon [56]. Recent extensions address

increasingly realistic operational constraints: Verhoeff et al. [44] introduce token-based capacity modeling validated against Dutch military rotorcraft; Marlow and Dell [58] extend the horizon to life-of-type (12+ years) with hierarchical "phase flow staircase" constraints; Altner et al. [61] integrate 16 criteria including heterogeneous maintenance types and tolerance bands. Commercial aviation applications address routing constraints alongside maintenance [93], while military applications address heterogeneous usage-based and time-based tasks [94].

We calibrate our simulation to US Army aviation maintenance doctrine. This provides a structured and operationally grounded environment for evaluation. The Army employs a three-tier maintenance hierarchy: (1) *preventive maintenance* at fixed intervals; (2) *reactive maintenance* triggered by component failures; and (3) *phase maintenance*, comprehensive overhauls required after accumulating specified flying hours [1]. Fleet availability is measured as the percentage of time assets are Fully Mission-Capable (FMC), with a doctrinal target of 75% [13]. Army doctrine, therefore, provides a transparent and reproducible basis for simulation parameters and performance metrics.

This setting is operationally demanding for several reasons. Unplanned downtime has direct and cascading effects on mission execution, scheduled maintenance flow, and future fleet availability. Maintenance capacity is constrained by personnel, facilities, and supply chains, which limits the system's ability to absorb shocks. Operational demand is autocorrelated due to training cycles and deployments rather than independent day-to-day tasking, and operating conditions vary substantially across time and mission profiles.

Government Accountability Office reports document persistent challenges in achieving readiness targets and recommend increased integration of predictive maintenance technologies. These reports also note the absence of standardized frameworks for implementation and evaluation [20]. In response, the Army's Prognostic and Predictive Maintenance (PPMx) initiative aims to shift maintenance from a reactive response to a proactive, condition-based decision-making approach. However, as Parker and Bellocchio [95] argue, the value of predictive maintenance technologies depends fundamentally on the decision framework into which they are embedded: requirements for data fidelity, model accuracy, and training all flow from how prognostic information is used. This makes Army rotorcraft operations a setting in which the value of prognostic information is both measurable and consequential, and where the consequences of inaccurate or poorly integrated predictions are nontrivial.

4.2.2 Predictive Maintenance and RUL as Information for Decisions

RUL, or more generally prognostic information, affects fleet outcomes primarily through the decisions it informs rather than through prediction accuracy alone. In contrast to fixed-interval preventive maintenance, where actions are taken on calendar or utilization schedules regardless of condition, RUL-based approaches allow maintenance timing and asset utilization to adapt to inferred health state. The operational value of prognostics, therefore, lies in how information is used to decide when to intervene, which assets to prioritize, and how to allocate constrained maintenance resources across a fleet.

Existing work incorporates RUL information into maintenance decision-making through a small number of recurring policy structures that appear across fleet-management settings. *Alarm-based rules* trigger maintenance when predicted RUL falls below a threshold. This trades earlier intervention against the risk of failure [89]. *Prioritization heuristics* use relative health information to decide which assets to operate and which to maintain. They direct scarce resources toward those most at risk. *Index policies* go a step further by assigning each asset a priority score based on its current condition and anticipated future costs. These scores yield simple decision rules that approximate optimal scheduling under limited maintenance capacity [96]. In aviation maintenance applications, these general policy classes are implemented by mapping aircraft-level prognostic signals into maintenance and utilization priorities, which enables planners to allocate limited maintenance capacity across competing aircraft [97].

A consistent theme across this literature is the asymmetric impact of prognostic errors. Premature maintenance triggered by overly conservative signals consumes component life and maintenance capacity but typically results in relatively short, predictable downtime. In contrast, missed or delayed interventions can lead to unscheduled failures that impose longer and more variable ground times, disrupt maintenance queues, and propagate through tightly coupled scheduling systems. In a capacity-constrained system, the shadow price of a reactive failure is significantly higher than that of a premature preventive intervention because of queueing effects and stochastic ground times. As emphasized in both doctrinal discussions and recent optimization-based scheduling studies, unscheduled maintenance events are disproportionately disruptive relative to planned interventions [61]. This asymmetry suggests that prediction accuracy influences fleet performance through its interaction with maintenance decisions and affects both reactive failure risk and the MS-OR tradeoff. The relevant question becomes *how* prediction uncertainty propagates through decision policies.

4.2.3 Decision Sensitivity and Research Gap

Few studies explicitly link prognostic signal quality to operational outcomes through decision policies. Building on the theory of indexable resource allocation problems [96], recent work examines how predictive signals can be incorporated into maintenance prioritization rules. Paynter [98] develops a POMDP-based index policy for Army helicopters and identifies a cautionary result: unless predictive signals achieve sufficiently low false positive rates, preemptive maintenance consumes scarce capacity and can degrade OR despite increased preventive intervention. Complementarily, Kamariotis et al. [99] introduce a decision-oriented metric that varies prediction accuracy through controlled noise injection and show diminishing returns to accuracy improvements beyond a moderate threshold. Together, these studies suggest that prediction accuracy matters for operational outcomes and that its value is nonlinear.

However, these frameworks leave critical questions unanswered. Paynter [98] models health using a coarse, binary signal rather than continuous RUL, which collapses asymmetric forecast errors into threshold-crossing events. His LP-based formulation fixes policy structure and allows only parameter tuning as accuracy changes, rather than permitting decision rules to adapt structurally across accuracy levels. Kamariotis et al. [99] study single-component replacement under fixed heuristics and abstract away fleet-level prioritization under stochastic demand and maintenance capacity constraints.

While index-based policies are well motivated theoretically [96] and operationally applied in aviation contexts [97], we lack decision-relevant guidance for how accurate RUL must be to move fleet outcomes when policies adapt and capacity is scarce. Modern FMP models treat prognostic information as exogenous; the prognostics literature focuses on prediction metrics with limited connection to operational consequences. No existing framework quantifies the marginal fleet-level value of improved prediction accuracy or isolates the mechanism by which information quality generates operational gains.

A parallel methodological gap exists in solution approaches. Deep reinforcement learning (DRL) has been applied to maintenance decision-making at the component and single-aircraft level. Lee and Mitici [100] train a DRL agent to time engine replacements from probabilistic RUL estimates, and Hu et al. [101] optimize sequential maintenance actions for one aircraft across a long horizon but acknowledge that extending to a multi-aircraft fleet would render the state-action space intractable. At fleet scale, Tseremoglou and Santos [67] schedule maintenance for 34 commercial aircraft but decompose the problem into a component-level POMDP and a fleet-level scheduling layer solved separately. Kosanoglu et al. [102] apply DRL-guided simulated annealing and use reinforcement learning to direct the search pro-

cedure rather than to make maintenance decisions directly. The closest existing work to the present problem is Vos et al. [103], who apply Q-learning to jointly optimize mission assignment and maintenance scheduling for a fleet of aircraft. Their framework outperforms heuristic baselines for two and three aircraft but degrades substantially at four, where the Q-table covers only 30.5% of the state-action space and the performance advantage narrows from roughly 10% to 4%. The scalability challenge documented by Vos et al. [103] arises in a setting with full observability, deterministic demand, and no maintenance capacity coupling. The present formulation adds partial observability, stochastic demand, and binding capacity constraints to a larger fleet. The effective dimensionality grows combinatorially in fleet size and compounds across daily decisions over the full planning period. To the authors' knowledge, no published reinforcement learning formulation jointly optimizes utilization and maintenance across a capacity-coupled fleet under stochastic demand, partial observability, and binding slot constraints over a year-long horizon. For this reason, the optimization approach adopted in Section 4.4 avoids value-function learning entirely and instead searches the policy space directly.

Our approach builds on the foundational decision-theoretic frameworks of Raiffa and Schlaifer [104] and Howard [105], which established that information has value only to the extent that it alters decisions. Classical pre-posterior analysis and the "value of clairvoyance" provide theoretical foundations, but capacity couplings and stochastic dynamics in fleet management render analytical derivation intractable. We therefore extend these principles using simulation-based optimization to quantify the value of imperfect information in a high-dimensional, constrained setting.

This paper addresses that gap by treating prediction accuracy as an explicit experimental factor within a coupled utilization–maintenance system. We quantify how incremental improvements in RUL precision translate into changes in MS and OR under stochastic demand and capacity constraints, while explicitly allowing decision policies to adapt to the information quality.

4.3 Problem Formulation

A centralized planner makes two linked decisions each day. The first is which FMC aircraft to fly, and the second is which non-flying FMC aircraft to induct into preventive maintenance. Decisions are made using a noisy RUL signal and phase-cycle information, under binding slot and token constraints. The planner does not observe true component health, future demand realizations, or maintenance durations before they occur, so all scheduling is performed under partial observability. Aircraft health evolves deterministically with utilization, and

maintenance events return aircraft to service with reset health. The objective is to maximize a weighted combination of mission success and operational readiness over a 365-day horizon.

We first define the problem in abstract terms to establish methodological generality, then present the US Army rotorcraft context.

4.3.1 General Problem Formulation

State space The fleet consists of N assets indexed by $i \in \{1, \ldots, N\}$ operating over a discrete horizon $t = 1, \ldots, T$. Each asset is characterized by a state tuple $s_{i,t} = (\text{RUL}_{i,t}, u_{i,t}, v_{i,t})$, where $\text{RUL}_{i,t} \in \mathbb{R}_+$ denotes true remaining useful life, a latent quantity estimated by the prognostic system, $u_{i,t} \in \mathbb{R}_+$ denotes accumulated utilization, and $v_{i,t} \in \{0, 1\}$ denotes availability status. The state tuple may include additional variables such as maintenance cycle position depending on the application.

Action space At each decision epoch, the planner selects an action $a_{i,t} \in \{\text{fly}, \text{maintain}, \text{hold}\}$ for each available asset. Flying consumes health and utilization capacity while contributing to operational throughput. Maintenance transitions assets to unavailable status for a stochastic duration and resets health or utilization counters upon completion.

Transitions True RUL evolves deterministically with utilization: $\text{RUL}_{i,t+1} = \text{RUL}_{i,t} - \delta_{i,t}$, where $\delta_{i,t}$ is the utilization increment on day t. Maintenance events reset RUL to a maximum value $\text{RUL}_{\max}$ or reset utilization counters to zero. Availability transitions are governed by maintenance triggers (RUL depletion, utilization thresholds) and maintenance completion times drawn from type-specific distributions.

Partial observability The planner does not observe true RUL $\text{RUL}_{i,t}$ directly. Instead, a noisy signal $\widehat{\text{RUL}}_{i,t}$ is observed after each utilization event. The observation model is $\widehat{\text{RUL}}_{i,t} \mid \text{RUL}_{i,t} \sim F(\text{RUL}_{i,t}, \sigma)$, where σ parameterizes observation noise. This formulation captures the partial observability inherent in prognostic systems.

Capacity constraints Maintenance capacity is limited by (i) concurrent processing slots and (ii) a budget of maintenance actions per period. When capacity is exhausted, assets requiring maintenance remain unavailable until resources are released.

Objective The planner maximizes a weighted combination of operational throughput (fraction of demand satisfied) and asset availability (fraction of fleet available) over the planning horizon.

Informational benchmarks To anchor the value-of-information analysis, we define three reference points along the information spectrum. At one extreme, a *zero-information* baseline uses only utilization state (hours since phase maintenance) for flight assignment and performs no preventive maintenance; aircraft fly until failure triggers reactive maintenance. This baseline exploits utilization information but not health information, so the normalization below isolates the value of prognostic health signals specifically. At the other extreme, a *perfect-information* ceiling gives the planner access to true RUL (CV=0%), which eliminates all prognostic uncertainty within the policy class studied. Between these bounds, *noisy-information* policies observe $\widehat{\text{RUL}}$ with CV $\in \{0.50, 0.25, 0.10, 0.05\}$.

The difference in performance between the perfect-information ceiling and the zero-information baseline defines the *prognostic potential*: the maximum operational improvement achievable through health information alone. The value of prediction accuracy at a given noise level is the fraction of this potential that is captured:

$$\text{Value Captured(CV)} = \frac{J(\pi^*_{\text{CV}}) - J(\pi^*_{\text{zero}})}{J(\pi^*_{\text{perfect}}) - J(\pi^*_{\text{zero}})}. \tag{4.1}$$

This normalization makes the diminishing-returns finding directly interpretable and comparable across operational contexts.

4.3.2 Application to US Army Rotorcraft Operations

To evaluate this formulation in a realistic setting with validated parameters, we ground the analysis in US Army AH-64 rotorcraft operations. Structured maintenance doctrine provides transparent parameter values and performance benchmarks. Table 4.1 summarizes the correspondence between abstract variables and doctrinal terms.

Table 4.1: Correspondence between general formulation variables and US Army rotorcraft doctrine. The left column defines the abstract variable; the right column provides the doctrinal term used throughout the application.

Abstract Variable	**Army Doctrine Term**
$\text{RUL}_{i,t}$ (true RUL)	RUL measured in flying hours
$u_{i,t}$ (utilization)	Phase cycle hours (0–500)
$v_{i,t}$ (availability)	FMC/NMC status
$\widehat{\text{RUL}}_{i,t}$ (observed signal)	Observed RUL from prognostics
Capacity slots	Routine (2) + Phase (1) maintenance slots
Budget tokens	Annual maintenance budget (K)

System Dynamics and Maintenance Representation

The fleet consists of $N = 8$ aircraft indexed by $i \in \{1, \ldots, N\}$ operating over a discrete daily horizon $t = 1, \ldots, 365$. At any point in time, each aircraft is classified as either Fully Mission-Capable (FMC) or Not Mission-Capable (NMC). An aircraft is considered available (FMC) only when fully operational and properly configured for its assigned mission. Aircraft that do not meet these standards are unavailable (NMC) and cannot be tasked [13].

While Army doctrine includes additional readiness and maintenance subcategories (e.g. PMC, NMCS, NMCM, depot-level events), the model abstracts these distinctions and represents availability using only the FMC/NMC classification. In addition to availability status, each aircraft is characterized by accumulated flying hours within the phase maintenance cycle and a health state represented by true remaining useful life $\text{RUL}_{i,t}$, measured in flying hours.

Deterministic degradation True health evolves deterministically with utilization. If aircraft i flies $h_{i,t}$ hours on day t (the daily utilization increment $\delta_{i,t}$ from the general formulation), then

$$\text{RUL}_{i,t+1} = \text{RUL}_{i,t} - h_{i,t},$$

and $\text{RUL}_{i,t}$ is unchanged on days when the aircraft does not fly. Under this assumption, each aircraft's true RUL degrades at the same rate per flying hour. Any differences in true RUL trajectories therefore arise from utilization decisions alone. Differences in perceived health arise additionally from observation noise.

This assumption has an important implication for interpreting results. Because degradation is deterministic, all uncertainty in RUL estimates comes from observation noise rather than from stochastic variation in the degradation process itself. Perfect prediction (CV=0%) therefore eliminates RUL uncertainty entirely. In practice, component degradation rates vary across units due to manufacturing variation, operating conditions, and environmental exposure. If degradation were stochastic, even perfect prediction could not eliminate RUL uncertainty because the prognostic CV would have a floor set by process variability. The deterministic assumption therefore isolates the value of observation accuracy specifically and provides an upper bound on the operational gains achievable through improved prediction alone.

Although degradation is deterministic, the planner cannot exploit this structure to bypass the prognostic signal. After each maintenance event, true RUL is drawn from a Uniform(25, 300) distribution that the planner does not observe. The policy has access to utilization features (hours since phase maintenance) but not the initial RUL draw, so cumulative flight

hours alone do not determine remaining life. The noisy RUL signal is the planner's sole source of health information.

Maintenance events and state resets Maintenance actions transition aircraft into NMC status for a stochastic duration (defined in Table 4.2) and apply state resets upon completion. All maintenance types reset the aircraft's component health by re-initializing $\text{RUL}_{i,t}$. Phase maintenance additionally resets the phase-cycle counters associated with the 250-hour and 500-hour utilization thresholds. Preventive Maintenance Daily (PMD) checks defined by Army maintenance doctrine are not modeled explicitly, as they do not affect next-day availability.

Maintenance categories The simulator represents four maintenance categories: preventive maintenance (short planned interventions), reactive maintenance (unplanned corrective actions after failure), and minor and major phase maintenance (longer scheduled events at 250 and 500 flying hours). This structure preserves the key planning tradeoffs without modeling every maintenance task. Each maintenance action consumes tokens from a fixed annual budget that represents the unit's constrained capacity to perform work over the fiscal year. When tokens are exhausted, additional maintenance cannot be initiated until the next period. Table 4.2 summarizes the triggers, duration models, token costs, and slot classes used in the simulator.

Table 4.2: Maintenance categories and modeled resource requirements. Each row defines a maintenance type by its trigger condition, stochastic duration, token cost, and slot class. Preventive and reactive events share two routine slots; phase events use a single dedicated phase slot.

Type	Trigger	Duration (days)	Tokens	Slot Class
Preventive	$\widehat{\text{RUL}}_{i,t} < \tau$	Uniform[1,4]	1	Routine
Reactive	$\text{RUL}_{i,t} \leq 0$	LogNormal(μ=2.3, σ=0.43)	1	Routine
Minor Phase	250 flying hours	Uniform[8,14]	3	Phase
Major Phase	500 flying hours	Uniform[38,50]	10	Phase

Capacity constraints Maintenance capacity is constrained by (i) slot availability and (ii) an annual token budget. The system includes two *Routine* maintenance slots that can be used for preventive or reactive maintenance and one *Phase* maintenance slot reserved for minor and major phase events. An aircraft can occupy at most one slot at a time, and if no

appropriate slot is available when maintenance is required, the aircraft remains NMC until capacity becomes available. Budget feasibility is enforced through an annual allocation of K maintenance tokens released quarterly in equal increments; tokens may be used at any point within the year but do not carry over across simulation years. Preventive and reactive maintenance consume one token per event, while minor and major phase maintenance consume three and ten tokens, respectively.

Stochastic Environment

The system evolves in a stochastic operating environment defined by three exogenous processes that are independent of policy decisions and common across all policy evaluations: daily mission demand, uncertainty in prognostic information about RUL, and maintenance duration. These processes define the operating environment faced by all policies and ensure that performance differences arise from decision-making rather than from changes in underlying conditions.

Mission demand Daily mission demand is modeled as a stationary discrete-time Markov chain with state space $D_t \in \{0, 1, \ldots, 7\}$, where D_t denotes the number of aircraft required on day t. The transition structure exhibits strong inertia and favors transitions between nearby demand levels. This design reflects the inertia associated with training schedules and operational tempo, which tend to evolve gradually rather than change abruptly from day to day. While the expected demand sets the average utilization rate, it is the variance and autocorrelation of demand that stress maintenance capacity. High-demand periods that persist over multiple days can deplete maintenance capacity and create backlogs, whereas independent daily demand would allow capacity to recover between peaks. The Markov structure captures this stress mechanism. We adopt a specification that yields realistic autocorrelation and an expected baseline demand of two aircraft per day. This corresponds to 25% of the fleet, slightly above the long-run Army-wide average of 16.7% reported in Semmel et al. [2], which represents a moderately demanding but realistic operating tempo. Sensitivity scenarios modify the transition matrix to represent higher or more volatile operational tempo while preserving the same modeling structure.

RUL-scaled observation noise All maintenance actions reset component health. After completion of any maintenance event, an aircraft's true remaining useful life $\mathrm{RUL}_{i,t}$ is drawn from a Uniform$(25, 300)$ distribution (hours) and subsequently decrements deterministically (one-for-one with flying hours, as described in Section 4.3.2) until the next maintenance event.

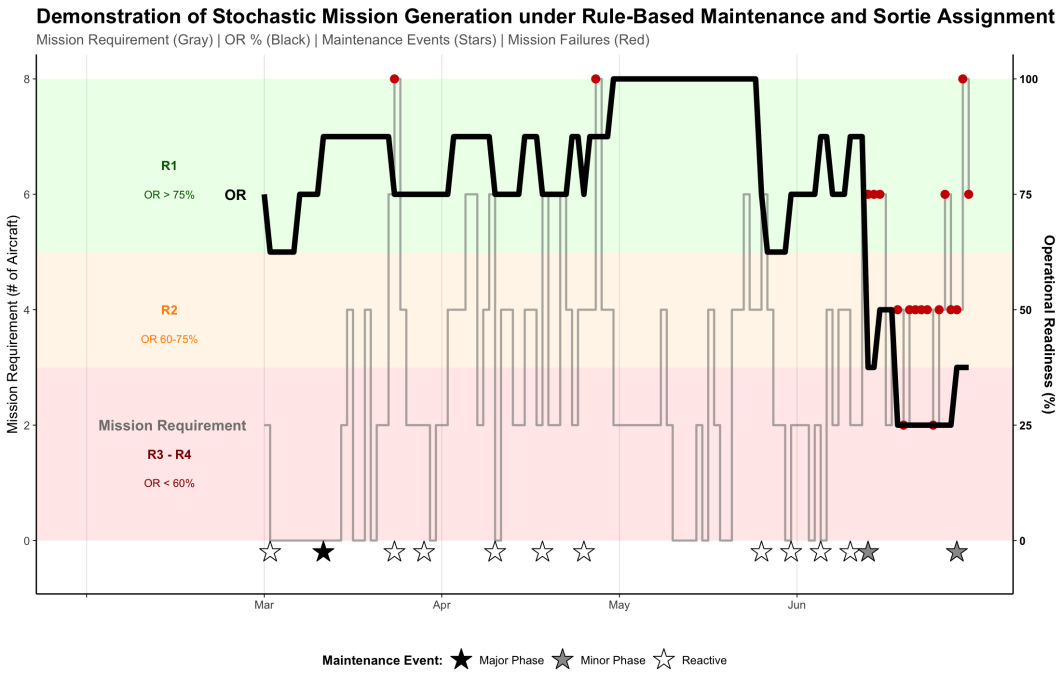

Figure 4.2: Sample simulation trajectory illustrating the interaction between stochastic mission demand, maintenance events, and operational readiness. Gray bars indicate daily mission demand (left axis); the black line tracks operational readiness percentage (right axis). Background shading denotes doctrinal R-level thresholds: R1 (green, OR $>$ 75%), R2 (orange, 60–75%), and R3–R4 (red, $<$ 60%). Stars along the bottom mark maintenance events (filled = major phase, gray = minor phase, open = reactive). Red dots indicate mission failures. The figure shows how reactive failures (open stars) drive OR into lower R-levels. Consecutive mission failures accumulate when demand spikes coincide with degraded availability.

The policy does not observe true RUL directly. Instead, after each flight, a fresh estimate $\widehat{\text{RUL}}_{i,t}$ is sampled from a conditional observation distribution centered on the current true value. We require that this estimate be (i) unbiased, with $\mathbb{E}[\widehat{\text{RUL}}_{i,t}] = \text{RUL}_{i,t}$, and (ii) have relative uncertainty controlled by the coefficient of variation, $\text{CV} = \sigma/\mu$.

The Gamma distribution is well-suited to this specification. Its support on $(0, \infty)$ ensures predictions remain positive, and for any Gamma(α, θ), the coefficient of variation equals $1/\sqrt{\alpha}$, regardless of θ, which allows relative uncertainty to be held fixed across aircraft and over time. The Gamma distribution is also the maximum entropy distribution for a random variable with fixed mean and fixed geometric mean, which makes it a neutral choice that avoids introducing hidden biases into the noise structure. This allows independent control of relative precision through α and centering through θ. We set shape $\alpha = 1/\text{CV}^2$ to achieve the desired relative precision, and scale $\theta_{i,t} = \text{RUL}_{i,t} \cdot \text{CV}^2$, which is re-computed at each

observation to center the distribution on the current true value. The resulting non-stationary parameterization (scale varies with true RUL, shape fixed) tracks the evolving health state at constant relative precision. This yields the conditional observation model:

$$\widehat{\mathrm{RUL}}_{i,t} \mid \mathrm{RUL}_{i,t} \sim \mathrm{Gamma}\left(\frac{1}{\mathrm{CV}^2},\ \mathrm{RUL}_{i,t}\,\mathrm{CV}^2\right). \tag{4.2}$$

Appendix D.9 illustrates the resulting observation distributions at selected true RUL values. This construction yields unbiased predictions, $\mathbb{E}[\widehat{\mathrm{RUL}}_{i,t}] = \mathrm{RUL}_{i,t}$, with standard deviation $\sigma = \mathrm{RUL}_{i,t} \times \mathrm{CV}$; for example, CV=50% means the standard deviation equals half the true RUL. Because absolute uncertainty scales with true RUL, prediction intervals naturally tighten as aircraft approach failure. The Gamma model is consistent with reliability-based RUL formulations that assume monotonic degradation driven primarily by operating time and provide interpretable confidence bounds [86, 87, 106]. Observations are re-sampled independently after each flight; each sortie produces a fresh, conditionally independent assessment of RUL rather than propagating a single noisy estimate forward. This memoryless observation structure is a deliberate modeling choice. It isolates the value of single-observation accuracy, which is the operationally relevant quantity when health monitoring systems report point estimates rather than posterior distributions. While historical trends can refine an asset's health narrative, the maintenance decision in a high-tempo environment ultimately collapses into a threshold comparison. As emphasized by Blechertas et al. [107], CBM must eventually provide actionable information to a practitioner who is compelled to act; at this stage, maintenance becomes a thresholding problem where decisions are driven by systematic policy rather than intuition alone. Whether the governing metric is a simple scalar or a complex risk-weighted sum, it still functions as a decision threshold that triggers an intervention. A belief-state model that aggregates observations over time could reduce effective noise through temporal filtering; we revisit this extension in Section 4.7.2.

Maintenance durations Maintenance durations are stochastic and depend on the type of maintenance performed. Preventive, reactive, minor phase, and major phase maintenance durations follow the distributions summarized in Table 4.2. Mean durations for minor and major phase maintenance (11 and 44 days, respectively) are calibrated to doctrinal planning factors in HQDA [1].

The use of right-skewed distributions for unscheduled maintenance durations is consistent with prior aircraft maintenance modeling studies, which represent repair times using lognormal distributions [108, 109]. Scheduled maintenance activities are modeled using bounded distributions that reflect planned work with limited variability. This distinction aligns with

assessments that reactive (unscheduled) maintenance generally requires greater effort and longer completion times than scheduled preventive actions [20]. Duration uncertainty affects aircraft availability but does not alter the underlying evolution of aircraft health.
Together, these processes define the stochastic operating environment in which maintenance and utilization decisions are made under imperfect information and uncertain demand. Figure 4.2 illustrates a sample trajectory from the simulation.

Decision Structure

The fleet management problem isolates how the quality of health information changes the decisions a planner makes and the outcomes those decisions deliver, given capacity-constrained maintenance and stochastic demand.

The planner observes a noisy, continuous RUL signal for each asset and must rank assets for use or maintenance under binding capacity constraints. Errors in this signal have asymmetric operational consequences. Acting too early consumes maintenance capacity and usable component life, while acting too late leads to reactive maintenance events that are longer and more variable. These unplanned events increase maintenance congestion and reduce scheduling flexibility. When capacity is fixed, policies that limit reactive failures preserve slack in maintenance resources and reduce the risk that simultaneous failures overwhelm available slots and degrade system performance.

Policy representation A policy π outputs an ordering (ranking) over FMC aircraft using the per-aircraft feature vector defined in Section 4.4.1. This ranking is the policy's sole control signal. A fixed adjudication procedure converts the ranking into executable actions under the realized operating conditions of the day. Comparisons across policies therefore reflect differences in prioritization logic, not execution rules.

We adopt a ranking-based policy to retain interpretability and to align the decision mechanism with the VOI objective. The policy structure and optimization procedure are detailed in Section 4.4.

Mandatory vs. discretionary events Reactive maintenance is triggered automatically when an aircraft fails ($\text{RUL}_{i,t} \leq 0$), and phase maintenance is triggered when utilization thresholds are reached. These events are non-discretionary: the aircraft enters maintenance when capacity and tokens become available and remains NMC until then. The policy controls only two decisions: (i) which FMC aircraft fly each day, and (ii) which non-flying FMC aircraft, if any, are inducted into preventive maintenance when eligible.

Preventive maintenance as an alarm-based rule To incorporate prognostic information into maintenance decisions, we use an alarm-based eligibility rule that is common in RUL-driven maintenance models. Following de Pater et al. [89], an FMC aircraft becomes eligible for preventive maintenance when its observed prognostic signal falls below a threshold τ, i.e., $\widehat{\mathrm{RUL}}_{i,t} < \tau$. The threshold τ parameterizes risk tolerance. Higher thresholds initiate maintenance earlier and reduce failure exposure at the expense of unused component life and routine capacity consumption. Lower thresholds defer intervention and increase exposure to reactive failure. Because the signal is noisy, the effectiveness of a given threshold depends on prediction accuracy, which determines how often the alarm triggers prematurely or fails to trigger before a true failure.

In the experimental design, τ is treated as a configuration parameter (values specified in Section 4.5) to reflect that different units may rationally select different alarm settings even under the same operating environment.

Performance Metrics

The broader framework motivates readiness and utilization as joint outcomes of a coupled system. In the simulation, utilization is not a free variable. Stochastic demand arrives daily and the fleet either meets it or does not. Mission Success (MS) captures this directly: the fraction of days on which available aircraft satisfy demand. Operational Readiness (OR) captures the availability side: the mean fraction of the fleet that is FMC on any given day. Together, MS and OR operationalize the utilization-readiness tradeoff within a demand-driven system.

Mission Success (MS) MS measures the proportion of days on which sufficient aircraft are available to meet mission demand:

$$\mathrm{MS} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\{C_t \geq D_t\}\,, \tag{4.3}$$

where D_t is the number of aircraft required on day t, T is the simulation horizon, and C_t denotes the number of aircraft that complete the mission on day t among those tasked. A day contributes to MS only if at least D_t aircraft both launch and return mission-capable. If any tasked aircraft fails during the mission such that fewer than D_t aircraft complete, the day is recorded as a mission failure.

Operational Readiness (OR) OR is defined as the average fleet availability over the planning horizon:

$$\text{OR} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{I}\{\text{aircraft } i \text{ is FMC on day } t\}\,, \tag{4.4}$$

where N is the fleet size and T is the length of the simulation horizon. This quantity is equivalent to the mean daily FMC rate across the year-long simulation.

Army doctrine reports OR as a rolling monthly average, assessed over a monthly reporting window [12]. The simulation computes OR as a single annual average, which serves as the time-averaged availability measure used for policy comparison. Because demand is stationary and policies are time-invariant, monthly and annual averages converge to the same expectation; the annual window reduces sampling noise.

Preference evaluation Aviation units exhibit distinct mission-readiness preference profiles [110]. To examine performance across this range, results are summarized using a weighted combination of MS and OR:

$$\mathit{Utility} = w_{\text{ms}} \cdot \text{MS} + (1 - w_{\text{ms}}) \cdot \text{OR}, \tag{4.5}$$

where $w_{\text{ms}} \in [0, 1]$ denotes the relative emphasis placed on mission execution versus fleet availability. This expression is used solely as a reporting device to assess outcome consistency under different preference profiles.

Results are reported for two representative weight settings. The first, $w_{\text{ms}} = 0.7$, corresponds to a mission-focused unit. The second, $w_{\text{ms}} = 0.3$, corresponds to a readiness-focused unit. These cases serve as illustrative points within a continuum of possible unit preferences.

Reactive failures are recorded and reported separately as a diagnostic outcome. Although not incorporated into the summary measure, failure frequency provides additional context on how policies balance utilization, maintenance timing, and operational risk.

4.4 Methodology

The problem defined in Section 4.3 involves partial observability, shared maintenance constraints across aircraft, and long-horizon effects that make analytical solution intractable. We therefore adopt a simulation-based optimization approach in which candidate policies are evaluated over full 365-day operational trajectories in a stochastic fleet simulator.

Model parameters are drawn from operational data when available and from Army doctrine when data are unavailable [1, 12, 13, 15]. Subject matter expert judgment is applied only for parameters specified by neither source. Common random number seeds are shared

across policies within each Monte Carlo replication so that observed performance differences reflect decision logic rather than variation in realized randomness.

4.4.1 Feature Representation

The feature representation specifies what information is available to the policy when prioritizing aircraft for utilization and maintenance. Together, these features form a continuous encoding of aircraft condition, maintenance proximity, and synchronization with the fleet that defines the policy's information set at each decision epoch. The objective is to capture the core information planners use when balancing utilization, maintenance timing, and failure risk in the presence of imperfect prognostics.

The selected features correspond to quantities routinely tracked in unit-level planning and maintenance management. Observed remaining useful life ($\widehat{\mathrm{RUL}}_{i,t}$) represents the prognostic health signal; hours to minor and major phase represent proximity to utilization-driven maintenance thresholds; and bank hour deviation represents doctrinal phase flow management aimed at maintaining uniform spacing across phase maintenance events [1]. As a result, learned policies can be articulated directly in operational terms. For example, a policy may prioritize aircraft that are behind phase flow with high remaining RUL.

The policy observes only per-aircraft features and produces a ranking based solely on individual aircraft characteristics. Fleet-level constraints and state variables (including realized mission demand, maintenance queue depth, remaining maintenance tokens, and slot availability) are not inputs to the policy. Instead, they are handled by a fixed adjudication procedure that applies the ranking subject to resource constraints and stochastic manifestations (Section 4.3.2). Performance differences therefore arise from how policies act on imperfect information, not from execution logic.

Per-Aircraft Features

Each FMC aircraft is characterized by a small set of decision-relevant features that summarize its current health, maintenance obligations, and position within the fleet maintenance cycle. These features are observed at the start of each day and updated as aircraft are flown or inducted into maintenance. The features encode three planning considerations. Observed RUL captures condition-based failure risk. Hours to minor and major phase capture proximity to utilization-driven maintenance obligations. Bank hour deviation captures fleet-level synchronization of phase maintenance.

1. **Observed RUL** ($\widehat{\mathrm{RUL}}_{i,t}$): A noisy prognostic estimate of remaining component life, measured in flying hours. Lower values indicate elevated failure risk and greater ur-

gency for preventive maintenance. Observations are re-sampled after each flight according to the stochastic noise model described in Section 4.3.2.

2. **Hours to Minor Phase Maintenance** ($250 - h_{\text{minor},i,t}$): Remaining flying hours until the aircraft reaches the 250-hour minor phase maintenance threshold. Minor phase events require dedicated phase maintenance capacity and induce moderate downtime, temporarily removing the aircraft from the available pool.

3. **Hours to Major Phase Maintenance** ($500 - h_{\text{major},i,t}$): Remaining flying hours until the 500-hour major phase maintenance threshold. Major phase events are long and resource-intensive and, therefore, result in extended aircraft unavailability.

4. **Bank Hour Deviation** ($\delta_{\text{bank},i,t}$): A signed measure of whether the aircraft is ahead of or behind its ideal position in the phase maintenance sequence. This feature operationalizes doctrinal guidance on uniform phase spacing and captures the extent to which utilization has deviated from the planned maintenance flow [1].

The feature set is deliberately minimal. Other potentially relevant quantities (e.g., forecasted demand, queue congestion, remaining budget) are excluded to preserve a clear mapping between information quality and decision outcomes.

Bank Hour Deviation as a Fleet Synchronization Metric

Army doctrine defines *bank hours* as the number of flying hours available until phase maintenance [1]. Effective phase management requires aircraft to enter phase maintenance at smooth, evenly spaced intervals to avoid surges in workload, spare part demand, and aircraft unavailability [1, paragraph 4-59]. Units routinely monitor bank hour flow charts, and commanders specify fleet-level bank hour targets as part of training cycle planning [1, paragraph 1-29].

Phase flow objectives have been addressed explicitly in prior optimization-based formulations (e.g. Pippin [7], Gavranis and Kozanidis [28], Marlow and Dell [58], Altner et al. [61]), where phase queue positions and residual flight times are modeled as decision variables and deviations from ideal spacing are penalized within centralized MILP frameworks. In contrast, the present framework does not optimize phase flow directly. Instead, it embeds synchronization information into a per-aircraft feature. This allows coordination to emerge from local ranking decisions under stochastic execution rather than from global queue optimization.

Although bank hour deviation is computed using information about all aircraft in the fleet, the resulting value is expressed as an aircraft-specific attribute. Each aircraft receives

its own deviation signal, which the policy uses when ranking aircraft. Fleet-level aggregation occurs only during feature construction and not within the policy itself.

Bank hour deviation is defined as the signed difference between an aircraft's actual bank hours and its ideal position under uniform phase spacing:

$$\delta_{\text{bank},i,t} = \underbrace{(500 - h_{\text{major},i,t})}_{\text{actual bank hours}} - \underbrace{\frac{500 \cdot (N - r_i + 1)}{N}}_{\text{ideal bank hours for rank } r_i}, \tag{4.6}$$

where $h_{\text{major},i,t}$ denotes cumulative hours since the last major phase event and $r_i \in \{1, \ldots, N\}$ is aircraft i's rank when aircraft are ordered by remaining bank hours in descending order.

Positive values indicate that the aircraft has more remaining bank hours than its ideal position; restoring uniform spacing would require additional utilization of that aircraft. Negative values indicate fewer remaining bank hours than ideal; because additional utilization would exacerbate phase bunching, uniform spacing is restored either through increased utilization of other aircraft (which causes rank reordering) or, for the lowest-ranked aircraft, through progression into major phase maintenance.

For an eight-aircraft fleet, ideal uniform spacing ranges from 500 hours (rank 1) to 62.5 hours (rank 8). Bank hour deviation measures how each aircraft's actual position compares to this ideal. Consider two cases. First, an aircraft at rank 8 with 20 hours remaining has $\delta_{\text{bank}} = -42.5$: it is closer to phase than uniform spacing would place it. In isolation, this is unremarkable; some aircraft must always be nearest to phase. The deviation becomes informative when multiple aircraft share large negative values. Multiple large negative deviations signal bunching. Second, an aircraft at rank 5 with 300 hours remaining has $\delta_{\text{bank}} = +50$ (ideal is 250): it is farther from phase than ideal. It could absorb additional flying hours to help rebalance the fleet. The policy uses these deviations to distribute utilization in a way that maintains even phase flow. Figure 4.3 illustrates ideal versus uneven phase configurations.

Encoding phase flow as a per-aircraft feature rather than an optimization constraint allows phase synchronization, prognostic risk, and maintenance proximity to be traded off within a single ranking-based policy. The remainder of this section defines the policy class and optimization procedure.

4.4.2 Policy Representation: Decision Trees for Aircraft Ranking

The policy must perform two functions simultaneously: (i) assess relative maintenance urgency across the fleet based on observed health and utilization features, and (ii) rank all FMC aircraft to prioritize flight assignments and maintenance induction when resources are

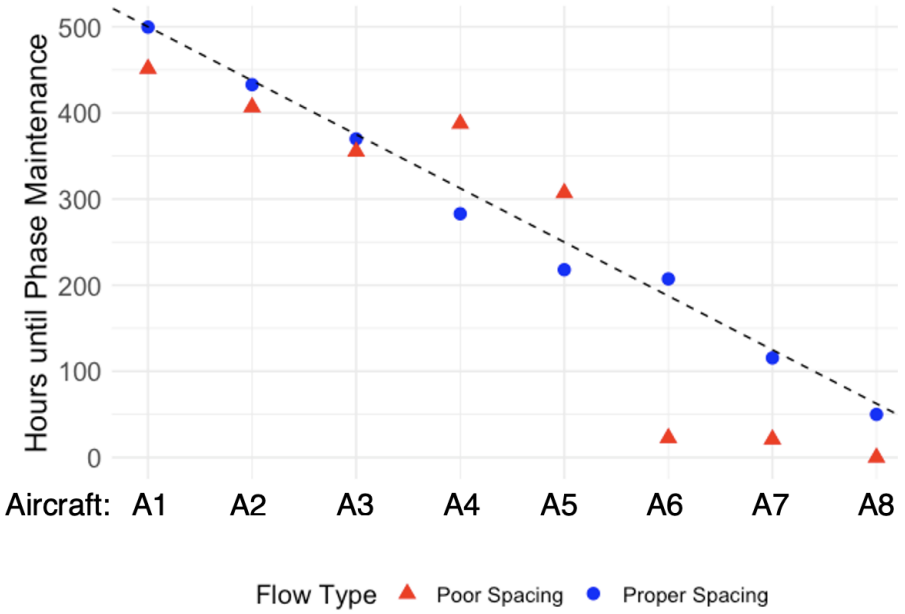

Figure 4.3: Proper versus poor phase maintenance interval management. Blue circles show proper spacing: aircraft hours-to-phase decrease uniformly across the fleet so phase entries are staggered. Red triangles show poor spacing: several aircraft cluster near zero hours, and simultaneous phase entries reduce fleet readiness. The dashed line marks ideal uniform spacing. Adapted from ATP 3-04.7, Figures 4-2 and 4-3 [1].

constrained. We represent this operating policy as a binary decision tree that maps per-aircraft features to priority buckets. This structure serves both functions within a single interpretable representation. Hereafter, "policy" refers to this learned decision tree.

Decision tree structure The policy uses a depth-3 binary decision tree with 7 internal (split) nodes and 8 leaf nodes. Each internal node compares one of the four per-aircraft features (observed RUL, hours to minor phase, hours to major phase, or bank hour deviation) to a threshold. Aircraft are routed left or right based on the comparison, ultimately arriving at one of 8 leaf nodes. Each leaf corresponds to a priority bucket $b \in \{1, \ldots, 8\}$. Intuitively, bucket 8 contains the healthiest aircraft (those best suited to fly) while bucket 1 contains aircraft in the worst condition, which are prioritized for preventive or mandatory maintenance rather than flight. The adjudication procedure (Algorithm 1) reflects this: higher bucket numbers receive priority for flight assignment, while lower bucket numbers receive priority for maintenance induction.

Chromosome encoding The decision tree is encoded as a 15-gene chromosome. Seven discrete feature indices ($\Omega_i \in \{0, 1, 2, 3\}$) select the split feature at each internal node. Seven

continuous thresholds ($\omega_i \in [0, 1]$) define normalized decision boundaries, scaled to each feature's operational range during evaluation. A single discrete gene selects the tiebreaker feature for within-bucket ordering.

This encoding creates a mixed discrete-continuous optimization problem that is computationally intractable for exhaustive search. The discrete component yields $4^7 = 16{,}384$ distinct tree structures, while each structure admits a 7-dimensional continuous threshold space. To prevent infeasible branches, thresholds are constrained to be path-consistent: if an aircraft has already been routed left through a split on RUL ≤ 200, any subsequent split on RUL along that path must use a threshold < 200. Figure 4.4 illustrates this structure.

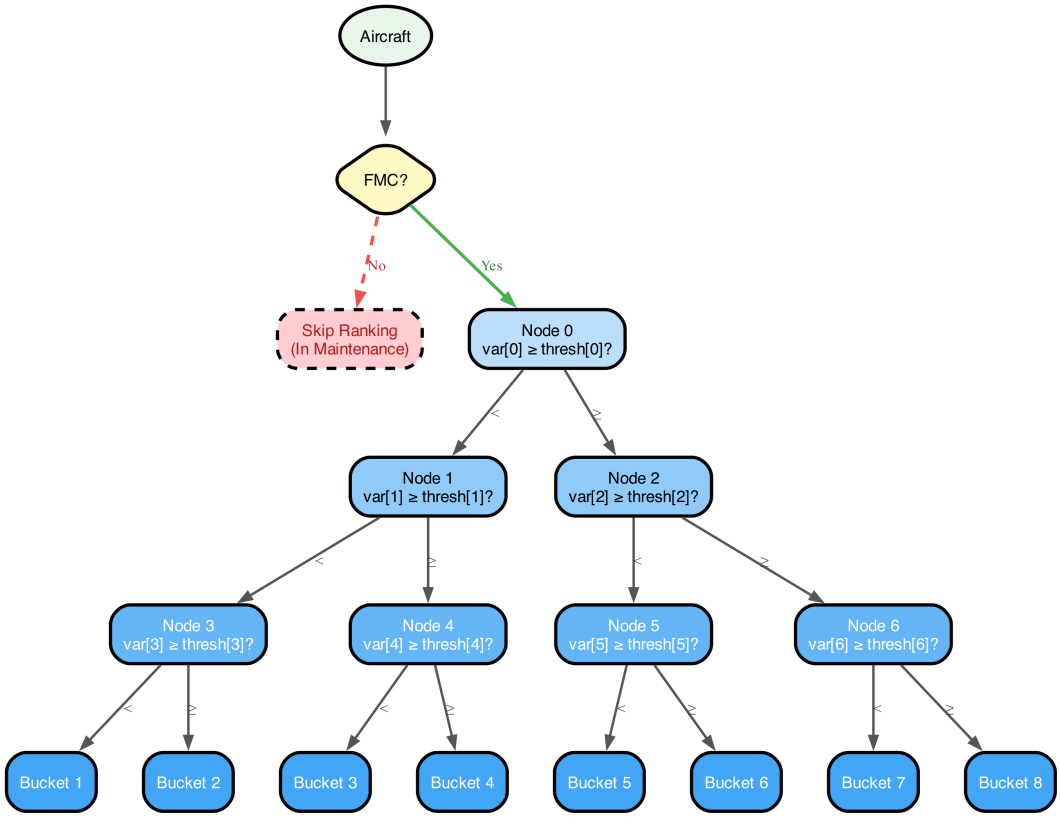

Figure 4.4: Decision tree policy structure. The 15-gene chromosome encodes 7 feature indices (which feature to split on at each internal node), 7 thresholds (decision boundaries), and 1 tiebreaker feature.

Interpretability Decision trees produce auditable decision logic (e.g., "if hours-to-major < 100 and observed RUL < 50, assign to bucket 1").

4.4.3 Policy Execution via Daily Adjudication

The decision tree outputs a priority bucket for each FMC aircraft, but does not directly specify actions. A fixed adjudication procedure converts the ranking into executable actions under the realized operating conditions of each day. This adjudication procedure is identical across all policies and all experimental scenarios. The only variation is the ranking produced by the policy. This invariance ensures that observed performance differences arise from ranking logic, not execution mechanics.

The adjudication procedure operates through three sequential sweeps, each constrained by slot availability and remaining maintenance tokens. Algorithm 1 formalizes this process.

Algorithm 1 Daily adjudication procedure

Require: Fleet state $\mathcal{F}$, demand D_t, policy π, alarm threshold τ
Ensure: Action assignment for each aircraft

Sweep 1: Mandatory Maintenance (NMC Aircraft)
for each NMC aircraft i awaiting maintenance **do**
if reactive or phase maintenance required **and** slot available **and** tokens remain **then**
Assign i to appropriate maintenance slot
Decrement token budget
end if
end for

Sweep 2: Flight Assignment (FMC Aircraft)
Compute bucket $b_i = \pi(\mathbf{x}_i)$ for each FMC aircraft i
Sort FMC aircraft by bucket (descending), breaking ties by tiebreaker feature
$n_{\text{assigned}} \leftarrow 0$
for each FMC aircraft i in sorted order **do**
if $n_{\text{assigned}} < D_t$ **then**
Assign i to fly
$n_{\text{assigned}} \leftarrow n_{\text{assigned}} + 1$
end if
end for

Sweep 3: Preventive Maintenance (Non-Flying FMC Aircraft)
Sort non-flying FMC aircraft by bucket (ascending)
for each non-flying FMC aircraft i in sorted order **do**
if $\widehat{\text{RUL}}_i < \tau$ **and** routine slot available **and** tokens remain **then**
Assign i to preventive maintenance
Decrement token budget
end if
end for

The three-sweep structure ensures that: (1) mandatory maintenance obligations are satisfied first; (2) mission demand is met using the highest-priority available aircraft; and (3) preventive maintenance is allocated to eligible aircraft in priority order, subject to the alarm threshold τ and remaining capacity. The policy controls *ranking*; the adjudication procedure enforces *constraints*.

4.4.4 Genetic Algorithm Design

As discussed in Section 4.2.3, existing DRL formulations for fleet maintenance face intractable state-action spaces when capacity coupling and partial observability are introduced. A genetic algorithm (GA) circumvents the credit assignment problem entirely. Rather than learning a value function over individual state-action pairs, the GA evaluates each candidate policy by its aggregate fitness across the full simulation horizon. Credit need not propagate backward through individual decisions because the optimization objective is the trajectory-level outcome. The mixed discrete-continuous search space rules out enumeration but poses no difficulty for evolutionary search. The same optimization procedure applies at every level of prognostic precision, so differences in discovered policies reflect changes in information quality rather than artifacts of the learning algorithm. This controlled stability is essential for isolating the value of information.

Among simulation-based metaheuristics, genetic algorithms are well suited to decision tree induction [111]. The chromosome representation accommodates mixed discrete-continuous variables without special treatment [112, 113]. Population-based search is well suited to the multimodal nature of the landscape, where many structurally distinct policies achieve comparable performance [114]. Subsequent design choices (including the use of an island model to preserve population diversity [114, 115], heterogeneous island parameters [116], adaptive mutation [117], and controlled migration [118]) are motivated by robustness considerations. Rather than accelerating convergence, these mechanisms are intended to delay premature convergence and encourage sustained exploration.

Chromosome encoding and operators The mixed discrete-continuous chromosome requires gene-type-specific operators (Table 4.3). Following established practices for decision tree induction via genetic algorithms [111, 119], we apply uniform crossover to discrete genes and blend crossover (BLX-$\alpha = 0.5$) to continuous genes. Mutation rates decay exponentially over generations (Figure 4.5):

$$\sigma(g) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \cdot \exp\left(-\frac{6.13 \cdot g}{G}\right), \tag{4.7}$$

where g is the current generation and G is the maximum number of generations [117].

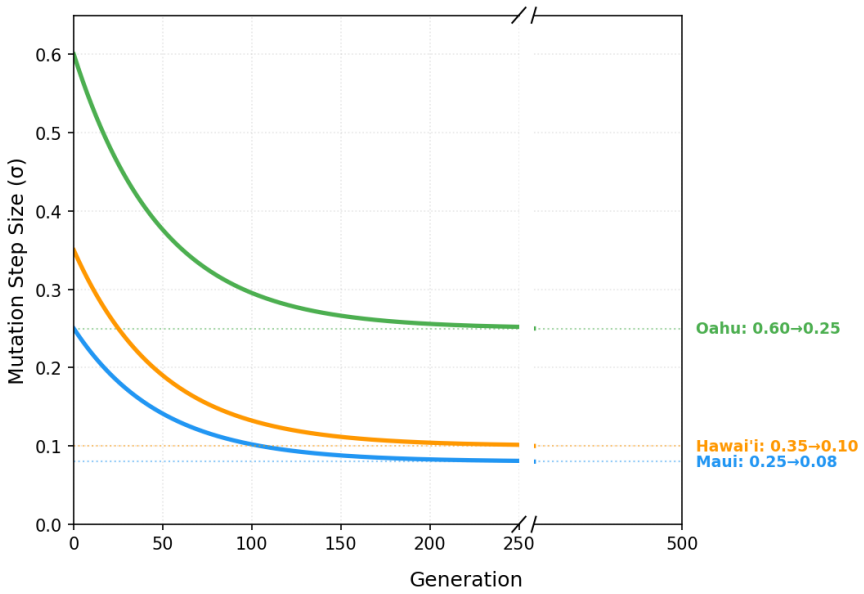

Figure 4.5: Adaptive mutation decay. Mutation rate (vertical axis) decreases exponentially over generations (horizontal axis). Early generations favor broad exploration; later generations favor local refinement.

Table 4.3 summarizes the operators applied to each gene type.

Table 4.3: Mixed-integer gene encoding and operators. Each row specifies a gene type, its role in the decision tree, its domain, and the crossover and mutation operators applied during evolution.

Gene Type	Role	Domain	Crossover	Mutation
Feature index	Split feature	$\{0, 1, 2, 3\}$	Uniform	Random reassignment
Threshold	Decision boundary	$[0, 1]$	BLX-0.5	Adaptive Gaussian
Tiebreaker	Within-bucket sort	$\{0, 1, 2, 3\}$	Uniform	Random reassignment

Island model The island model maintains three subpopulations connected by a feed-forward ring migration topology. Each island operates under distinct selection pressure, mutation intensity, and crossover rate to balance exploration and exploitation, while migration propagates promising candidate solutions across populations.

Table 4.4 summarizes the configuration of each island. The heterogeneous structure is designed to preserve diversity across candidate policies and reduce sensitivity to local optima, rather than to accelerate convergence. In preliminary experiments without islands, populations typically converged within 20–50 generations. With the island model, improvements often persisted for several hundred generations. This illustrates the role of heterogeneity as a robustness mechanism rather than a performance shortcut.

Table 4.4: Island model configuration. Each row specifies an island's role, population size, tournament size, mutation rate range, and crossover rate.

Island	Role	Population	Tournament k	Mutation Rate	Crossover Rate
Oahu	Explorer	20	2	70%→30%	50%
Maui	Refiner	20	3	20%→5%	85%
Hawai'i	Validator	25	3	35%→10%	70%

Islands exchange individuals through asymmetric feed-forward migration every 40 generations (Figure 4.6).

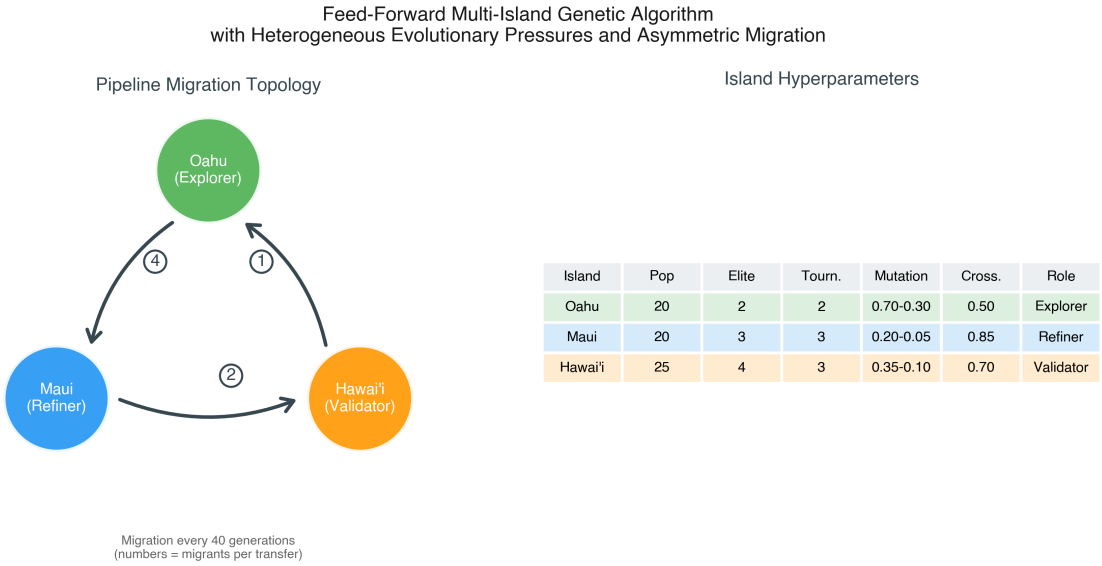

Figure 4.6: Heterogeneous island model architecture. Three islands with distinct evolutionary pressures are connected by a feed-forward ring migration topology. Exploratory solutions flow from Oahu through Maui to Hawai'i in a unidirectional feed-forward topology.

Selection, fitness, and convergence Each generation, offspring are produced by selecting parents via tournament selection: k individuals are sampled uniformly at random without replacement from the current population, and the one with highest fitness becomes a parent. Tournament size k varies by island (Table 4.4). Smaller tournament sizes reduce selection pressure and encourage exploration. Elitism preserves approximately the top 10% of individuals (2–4 candidates, depending on island population size) unchanged into the next generation; the remaining slots are filled by offspring produced through crossover and mutation [120].

Each chromosome is evaluated over 50 independent 365-day simulation episodes during training. Fitness is the utility defined in Equation (4.5), with $w_{\text{ms}} \in \{0.3, 0.7\}$ reflecting the commander preference profiles described in Section 4.3.2. Training evaluations use independent random seeds for each chromosome. Common random numbers are reserved for final policy evaluation (Section 4.5.4), where policies trained separately are compared on identical stochastic realizations. Within training, elitism preserves the best chromosome from any prior generation regardless of noise in current evaluations, and tournament selection compares candidates within the same generation where relative fitness ranking is more stable than absolute estimates.

Optimization terminates when no improvement exceeding 0.001 in best fitness is observed for 121 consecutive generations, or at a hard cap of 500 generations. This patience threshold corresponds to three complete migration cycles (40 generations each) plus one additional generation, so a beneficial solution discovered on any island can propagate through all three before termination. The criterion ensures that policies have stabilized sufficiently for comparative evaluation across accuracy levels, without claiming global optimality. All optimization runs in this study satisfied the early stopping criterion. Full hyperparameter specifications are provided in Appendix D.4.

Simulation-optimization procedure The framework comprises two nested loops: an outer optimization loop that evolves decision-tree policies, and an inner simulation loop that evaluates each candidate policy via Monte Carlo trajectory evaluation. Algorithm 2 summarizes this structure.

The policy π is a decision tree encoded as a chromosome of 15 genes (7 feature indices, 7 thresholds, 1 tiebreaker). The genetic algorithm searches over this encoding and evolves the tree structure and decision boundaries. Each candidate policy is evaluated by running E independent simulation episodes and averaging the resulting mission success and operational readiness outcomes. The fitness function is therefore a simulation-based estimate of expected operational performance.

For each prediction accuracy level CV, a separate genetic algorithm run discovers an effective policy for that information environment. Each policy therefore adapts structurally to its information environment rather than being tuned from a common baseline.

Algorithm 2 Simulation-optimization for policy discovery

Require: Accuracy level CV, fitness weight w_{ms}, population size P, episodes E
Ensure: Best policy π^*_{CV} (decision tree optimized for accuracy level CV)

Phase 1: Initialization
Initialize population of P random decision-tree chromosomes
best_fitness $\leftarrow -\infty$

Phase 2: Evolutionary Optimization (Outer Loop)
for generation $g = 1$ to $G_{\max}$ **do**

// Simulation-Based Fitness Evaluation (Inner Loop)
for each chromosome c in population **(parallel) do**
Decode chromosome c into decision tree π_c
for episode $e = 1$ to E **do**
Initialize fleet: N aircraft at staggered health states
for day $t = 1$ to T **do**
Process maintenance completions
Get mission demand D_t from Markov chain
$\mathbf{b} \leftarrow \pi_c(\mathbf{x})$ {Policy ranks aircraft into priority buckets}
Assign flights to top-ranked FMC aircraft
Induct eligible aircraft into preventive maintenance
for each flying aircraft i **do**
$\text{RUL}_{i,t+1} \leftarrow \text{RUL}_{i,t} - h_{i,t}$ {Deterministic degradation}
$\widehat{\text{RUL}}_{i,t+1} \sim \text{Gamma}(1/\text{CV}^2, \text{RUL}_{i,t+1} \cdot \text{CV}^2)$ {Fresh noise}
end for
Check failure triggers (RUL ≤ 0) and phase thresholds
end for
Record episode outcomes: $(\text{MS}_e, \text{OR}_e)$
end for
fitness$(c) \leftarrow w_{\text{ms}} \cdot \overline{\text{MS}} + (1 - w_{\text{ms}}) \cdot \overline{\text{OR}}$ {Average over E episodes}
end for

// Genetic Operators
Update best chromosome if max(fitness) > best_fitness
Preserve elite individuals (top 10%)
Generate offspring via tournament selection, crossover, adaptive mutation
Check convergence; terminate if no improvement for 121 generations
end for

Phase 3: Output
return Best chromosome decoded as policy π^*_{CV}

The procedure is repeated independently for each CV level in the experimental design (CV $\in \{0\%, 5\%, 10\%, 25\%, 50\%\}$). This yields a library of policies, each optimized for its respective information environment. Common random number seeds ensure that stochastic elements (demand realizations, maintenance durations) are identical across policy comparisons and isolate the effect of decision logic from environmental variation.

Section 4.5 describes the experimental environments and parameter ranges used to evaluate these policies.

4.5 Experimental Design

The preceding sections defined the policy representation and optimization architecture. This section describes the inferential framework used to support causal claims about the value of prediction accuracy. We explicitly differentiate between two distinct components: the *training grid*, which generates candidate policies across operating assumptions, and a *factorial experiment*, which isolates causal mechanisms by manipulating training and evaluation RUL accuracy independently.

The training grid produces a library of policies; the factorial experiment tests competing causal hypotheses about why those policies perform as they do. A 2×2 blocked factorial structure isolates the effect of prediction accuracy from the effect of the learned policy.

The baseline configuration represents a company-sized fleet of eight AH-64 aircraft operating over a fiscal-year horizon with daily decision epochs. The annual horizon aligns with the maintenance budget cycle and is sufficiently long to capture phase maintenance events, reactive failures, and the downstream consequences of interacting maintenance and utilization decisions. The daily decision cadence provides a simple representation of unit-level planning without introducing unnecessary sub-daily scheduling complexity.

4.5.1 Experimental Factors

Policy training is conducted over a structured grid of operating assumptions that reflect differences in prognostic information quality, commander preferences, and risk tolerance. These factors define the environments under which candidate policies are learned. Importantly, this training grid is exploratory in nature; it characterizes how policies adapt to different informational and operational settings rather than supporting causal claims on its own.

The first factor is prediction accuracy (CV), the coefficient of variation of the Gamma observation noise model (Equation 4.2). Levels range from perfect information (CV= 0%)

to highly uncertain prognostics (CV= 50%); Appendix D.3 maps these to decision-relevant probabilities. The second factor is commander preference (w_{ms}), the weight on mission success in the fitness function. Lower values ($w_{\text{ms}} = 0.3$) favor readiness; higher values ($w_{\text{ms}} = 0.7$) favor mission completion. The third factor is the alarm threshold (τ), the observed RUL threshold below which aircraft become eligible for preventive maintenance. Lower thresholds delay action until failure is imminent; higher thresholds trigger earlier intervention. Crossing τ with CV in the training grid allows us to observe whether the optimal alarm setting shifts with prediction accuracy, rather than fixing a single threshold that could confound the accuracy-performance relationship.

Table 4.5 summarizes the full training grid and evaluation structure.

Table 4.5: Experimental design summary. The training grid crosses five prediction accuracy levels with two preference weights and three alarm thresholds. Each of the 30 resulting configurations is evaluated over 10,000 replications under four operational scenarios.

Factor	**Levels**	**Values**
Prognostic accuracy (CV)	5	0%, 5%, 10%, 25%, 50%
Commander preference (w_{ms})	2	0.3 (readiness-focused), 0.7 (mission-focused)
Alarm threshold (τ)	3	25h, 50h, 100h
Operational scenarios	4	Standard, High Optempo, High Variance, Resource Constrained
Training grid	30	$5 \times 2 \times 3$ configurations
Factorial design	24	4×6 (treatments $\times$ blocks)
Training episodes per chromosome	50	Independent seeds per chromosome
Evaluation replications per cell	10,000	Common random numbers across policies

The resulting $5 \times 2 \times 3$ training grid yields a library of 30 policies, each optimized under a distinct combination of informational quality and operating preferences. This library serves as the input to the blocked factorial experiment described in Section 4.6.3, which isolates the causal role of prediction accuracy during policy evaluation.

4.5.2 Benchmark Policies

GA-optimized policies are compared against three benchmark policies that represent common practice and do not exploit RUL information. Fixed-Interval 25h (FI25) triggers preventive maintenance every 25 flying hours since the last maintenance event. This conservative policy maintains high availability at the cost of frequent interventions. Fixed-Interval 50h (FI50) triggers preventive maintenance every 50 flying hours and balances availability

against maintenance frequency. The Heuristic (Reactive) benchmark performs no preventive maintenance. Aircraft fly until failure triggers reactive maintenance, and flight assignment prioritizes aircraft closest to major phase maintenance to smooth phase flow. Because the Heuristic ignores RUL entirely, its performance is invariant to prediction accuracy and serves as the zero-information baseline in Equation 4.1. These benchmarks span the utilization–maintenance tradeoff without using prognostic information and serve as baselines against which to measure the value of RUL-informed decision-making.

4.5.3 Operational Scenarios

To assess robustness, policies are evaluated under four operational scenarios that stress different aspects of fleet management. The *Standard Baseline* scenario uses expected demand of 2.0 aircraft per day and an annual token budget of $K = 120$ and represents normal garrison operations. The *High Optempo* scenario increases expected demand to 3.0 aircraft per day (+50%) and represents sustained operational pressure during training cycles or deployments. The *High Variance* scenario preserves mean demand but increases transition entropy by 32%, which makes demand less predictable. This tests whether information value increases when operational tempo fluctuates unpredictably. Finally, the *Resource Constrained* scenario reduces the token budget to $K = 100$. This increases token exhaustion risk from approximately 5% to 25% (Figure D.1 in Appendix D.1) and tests whether RUL-informed policies make more efficient use of limited maintenance capacity. Full transition matrix specifications are provided in Appendix D.2.

4.5.4 Statistical Evaluation Protocol

Final policy evaluation uses 10,000 independent Monte Carlo replications per policy. Common random number seeds are used across policies within each replication to reduce variance in pairwise comparisons.

Differences in mean outcomes are assessed using Tukey's Honest Significant Difference (HSD) procedure to control family-wise error rates across multiple comparisons [121]. Reported confidence intervals reflect this adjustment.

4.5.5 Implementation and Reproducibility

Parameter selection hierarchy Simulation parameters follow a precedence rule: operational data are used when available, Army doctrine is used when data are unavailable [1, 12, 13], and subject matter expert judgment is applied only for parameters not specified

by either source. This hierarchy ensures operational realism while maintaining transparency about abstraction choices. Appendix D.1 documents the rationale for key simulation parameters.

Computational cost The simulation and optimization framework is implemented in Python. Each GA run uses a three-island model with a combined population of 65 chromosomes and a budget of 500 generations. Each fitness evaluation simulates 50 independent 365-day episodes. Runs typically converge between generations 200 and 450. A single configuration trains in approximately 2–4 hours on a standard multicore workstation. The full 30-configuration training grid ($5 \times 2 \times 3$) therefore requires roughly 60–120 hours of sequential computation, though configurations are independent and were run in parallel. Final policy evaluation (10,000 replications per policy across 4 scenarios) adds approximately 8–12 hours.

Reproducibility The complete simulation codebase, including genetic algorithm implementation, policy configurations, evaluation scripts, and analysis notebooks, is available as supplementary material. Random seeds and configuration files for all reported experiments are included to enable exact replication.

4.6 Results

4.6.1 Baseline Policy Behavior

RUL-informed policies substantially outperform the benchmark policies, which do not use prognostic information. Fixed-Interval benchmarks trigger preventive maintenance every 25 or 50 flying hours, while the doctrinally motivated heuristic performs no preventive maintenance and instead prioritizes aircraft closest to major phase for flight assignment to smooth phase flow. Detailed numerical comparisons across all metrics and accuracy levels are reported in Appendix D.5.

GA-optimized policies dominate all benchmarks across the full CV range (Figure 4.7). Reactive failures drop from approximately 16 per aircraft-year under the heuristic to 3–5 under GA policies, and as prediction accuracy improves, policies shift outward on the MS–OR frontier. Even imperfect RUL information yields operational improvements over fixed-interval or reactive strategies.

Together, these results establish that incorporating RUL information fundamentally shifts the utilization–maintenance balance and delivers simultaneous improvements in readiness and mission success relative to non-prognostic policies. RUL-informed policies sustain this

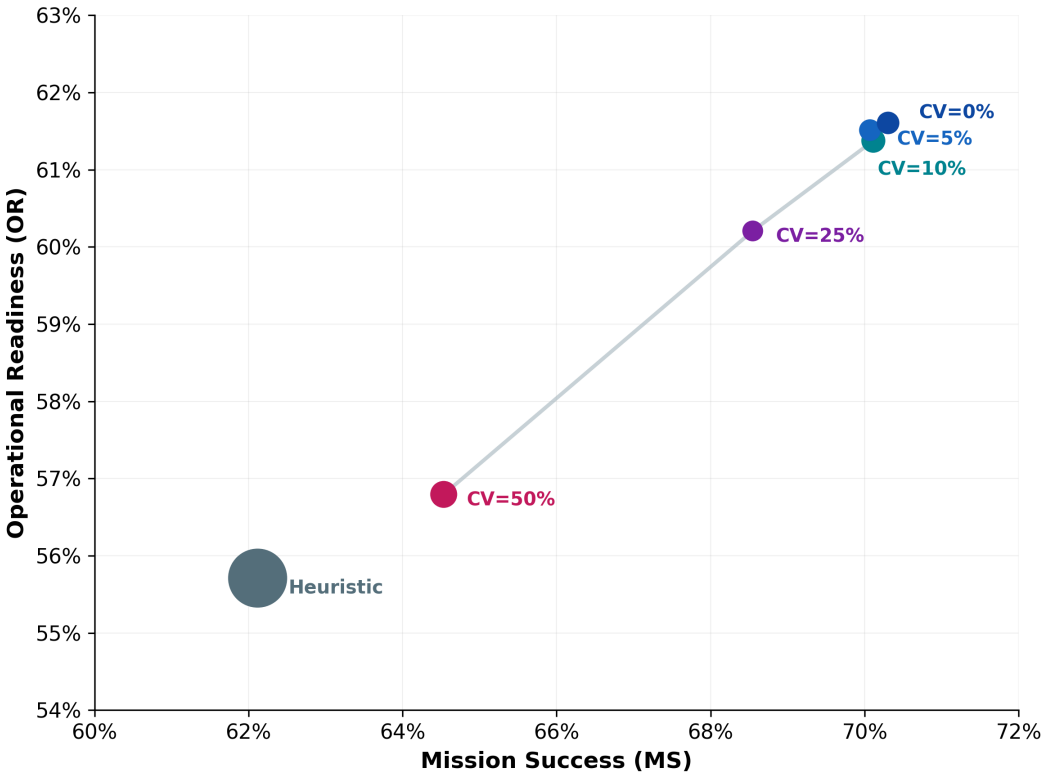

Figure 4.7: GA policy performance in MS–OR space by prediction accuracy level. Marker size is proportional to reactive failures per aircraft-year. As CV decreases, policies shift toward the upper-right (better on both metrics) while reactive failures decrease. Fixed-interval benchmarks are omitted; including them would distort the scale.

improvement by operating with a smaller bank hour buffer of approximately 100–130 fleet-average hours until major phase, versus 200 hours under benchmark policies (Figure D.8 in Appendix D.5). Prognostic information enables the policy to tolerate lower margins because it can distinguish at-risk aircraft from healthy ones. What would otherwise be a riskier utilization posture becomes a more efficient one.

4.6.2 The Value of Prognostic Accuracy

Under standard conditions, the gap between the perfect-information ceiling and the reactive heuristic baseline is approximately 8 percentage points MS and 6 percentage points OR. The relationship between prediction accuracy and operational performance is strongly concave. As accuracy improves, the binding constraint shifts from information quality to maintenance capacity, and further gains in prediction precision yield progressively smaller operational returns (Section 4.7). The largest improvement occurs when moving from CV=50% to

CV=25%, with progressively smaller gains at lower CV levels. Detailed CV effect plots by preference profile are provided in Appendix D.5.

These findings hold for the readiness-focused preference profile ($w_{\text{ms}} = 0.3$) as well. Under that profile, approximately 89% of achievable MS improvement and 91% of achievable OR improvement are captured by CV=25%, with full saturation by CV=10%. The CV effect plot for that configuration is provided in Appendix D.10. Moderate prediction accuracy captures most of the achievable operational benefit regardless of preference weight, with diminishing returns as accuracy improves past CV≈25%.

The largest gains occur in the first accuracy improvements. Moving from the heuristic baseline to CV=50% yields approximately +5 percentage points MS and +4 percentage points OR, and further improvement to CV=25% adds another +4 percentage points MS and +4 percentage points OR. Beyond CV=25%, returns diminish sharply, with the final steps from CV=10% to CV=0% contributing only ≈1 percentage point MS and ≈1 percentage point OR. Figure 4.8 quantifies the cumulative value captured (Equation 4.1) at each accuracy level: approximately 79% of achievable improvement is captured by CV=25%, and 98% by CV=10%. The CV=5% and CV=10% results are nearly identical, with small reversals in some configurations where CV=5% slightly underperforms CV=10%. This convergence is itself evidence of saturation. Across all training configurations, the performance gap between CV=5% and CV=10% is less than 0.5 percentage points at each alarm threshold (Appendix D.8), smaller than the evaluation standard errors reported in Tables 4.6 and 4.7. The fitness landscape flattens as accuracy improves past CV=10%. Many structurally distinct decision trees achieve near-equivalent performance because the information constraint no longer limits which aircraft can be correctly prioritized. When further accuracy improvements do not change which policies perform well, the information has diminishing decision relevance.

This diminishing-returns pattern holds across all three metrics (MS, OR, and reactive failures) and has practical implications: predictions at CV≈25% capture most of the achievable benefit, while investments pushing accuracy beyond CV≈10% yield proportionally smaller operational improvements. The primary mechanism underlying both MS and OR gains is reactive failure reduction. GA policies reduce reactive failures by 65–88% compared to the heuristic baseline and convert unpredictable, long-duration events into planned, shorter preventive maintenance. Fewer reactive failures mean more aircraft available for missions (higher OR) and more days meeting fleet requirements (higher MS). Supporting figures showing improvement by threshold and preference profile are provided in Appendix D.5.

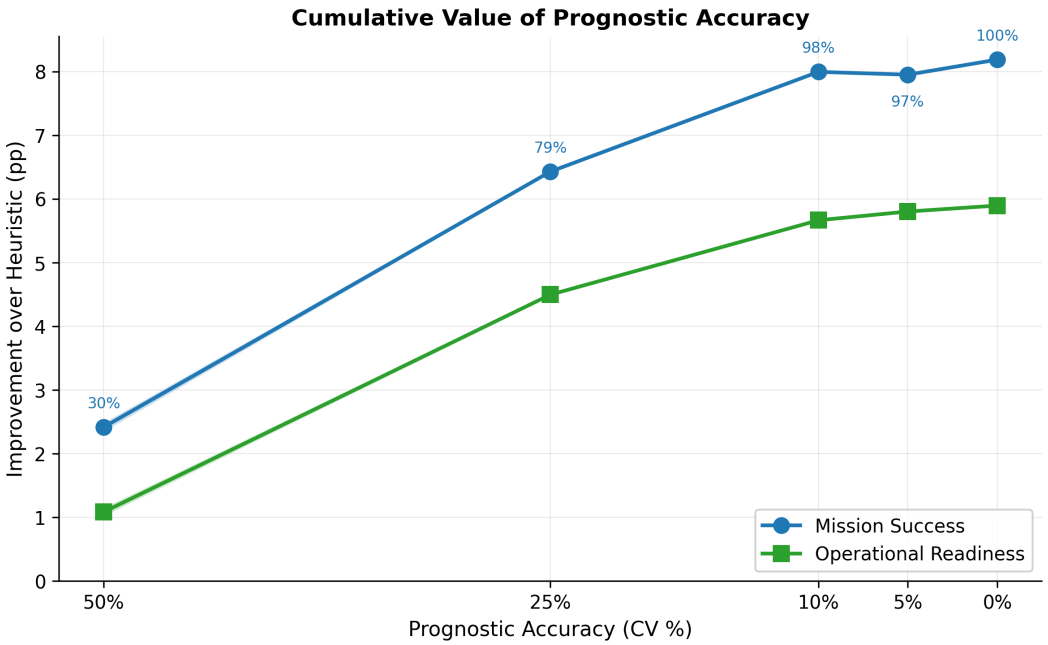

Figure 4.8: Percentage of maximum improvement captured at each prediction accuracy level (averaged across alarm thresholds for the mission-focused objective). At CV=25%, approximately 79% of the achievable MS improvement has been captured. Beyond CV=10%, nearly all value (98%) has been realized.

4.6.3 Causal Decomposition: Prediction Accuracy vs. Policy Effect

The preceding results establish that GA-optimized policies outperform benchmarks, but they leave open a causal question. Does performance improve because better predictions enable better decisions, because the learned policy is intrinsically superior, or because the two interact? We isolate these effects using a blocked factorial design [122]. Causal statements in this section refer to mechanisms within the simulated system under controlled experimental manipulation, not to empirical causality inferred from observational data.

Experimental design The design crosses two factors at two levels each: Factor A (Policy), determined by training CV $\in \{0\%, 50\%\}$, and Factor B (Prediction), the RUL accuracy during testing, CV $\in \{0\%, 50\%\}$.

This yields four treatment combinations where superscripts denote factor levels ($-$ = CV=0%, $+$ = CV=50%): $\mu_{A^-B^-}$ (both at 0%), $\mu_{A^-B^+}$ (policy at 0%, prediction at 50%), $\mu_{A^+B^-}$ (policy at 50%, prediction at 0%), and $\mu_{A^+B^+}$ (both at 50%). Extreme accuracy levels provide maximum contrast for the factorial decomposition. Section 4.6.2 established the accuracy-

performance curve. This design identifies which factor (prediction accuracy or policy structure) accounts for the observed improvement. To verify that the inference remains consistent across experimental configurations, we block by alarm threshold (τ) and fitness weight (w_{ms}) for a total of 6 blocks. Each of the 24 cells (6 blocks $\times$ 4 treatments) is evaluated over 10,000 replications.

Statistical model The factorial structure supports clean decomposition of observed performance into main effects and interaction:

$$\text{Prediction effect (B)} = \frac{\mu_{A^-B^-} + \mu_{A^+B^-}}{2} - \frac{\mu_{A^-B^+} + \mu_{A^+B^+}}{2} \tag{4.8}$$

$$\text{Policy effect (A)} = \frac{\mu_{A^-B^-} + \mu_{A^-B^+}}{2} - \frac{\mu_{A^+B^-} + \mu_{A^+B^+}}{2} \tag{4.9}$$

$$\text{A} \times \text{B interaction} = \frac{(\mu_{A^-B^-} - \mu_{A^-B^+}) - (\mu_{A^+B^-} - \mu_{A^+B^+})}{2} \tag{4.10}$$

Main effects and interactions Figure 4.9 shows the interaction plots for Mission Success across all six blocks. Plots for OR and reactive failures can be found in Appendix D.10. Near-parallel lines indicate additive effects (prediction accuracy dominates with minimal A×B interaction), while non-parallel lines indicate interaction.

The prediction effect (B) is large and consistent (+4.9 to +7.0 percentage points MS across blocks), while the policy effect (A) is comparatively small (−0.2 percentage points on average). A policy trained at CV=50% performs nearly as well as one trained at CV=0%, provided it receives accurate RUL information. Conversely, even a policy trained at CV=0% degrades substantially when tested with noisy predictions. Interaction plots for OR and reactive failures, which show consistent prediction-dominance patterns, are provided in Appendix D.10. Table 4.6 summarizes the effect decomposition across metrics.

Table 4.6: Factorial effect magnitudes (percentage points) averaged across six blocks (standard errors in parentheses). Within the decision-tree policy class studied, prediction accuracy (B) is the dominant driver of performance. Policy training conditions (A) contribute minimally. Reactive column shows reduction in failures from noisy to accurate conditions.

Effect	**MS**	**OR**	**Reactive ↓**
Prediction (B)	+5.6 (0.5)	+5.3 (0.4)	−1.1 (0.1)
Policy (A)	−0.2 (1.3)	+0.3 (0.9)	−0.4 (0.5)
A×B	<1.0*	<1.0	<0.1

*Exception: the readiness-focused, τ=100h block shows +3.2 percentage point A×B interaction for MS

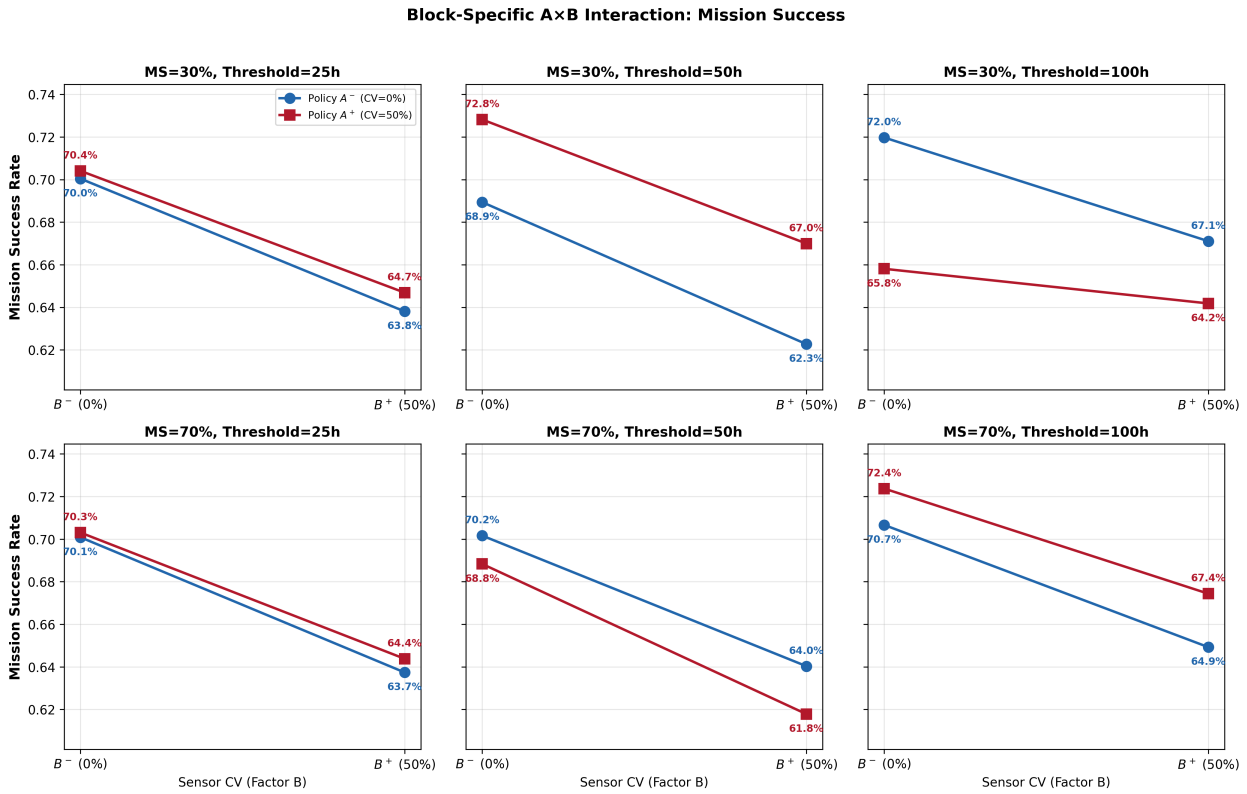

Figure 4.9: Factorial interaction plots for Mission Success. Near-parallel lines across most blocks indicate that the prediction effect dominates policy structure within the decision-tree class studied, with minimal A×B interaction. The ms30/100h block shows the largest interaction (+3.2 percentage points).

Because the A×B interaction magnitude varies across blocks, we apply Tukey's Honest Significant Difference (HSD) procedure at $\alpha = 0.05$ to identify where interaction effects are practically meaningful (full results in Appendix D.7). Only the readiness-focused, τ=100h block shows a significant synergy effect (+3.2 percentage points); in the remaining 5 blocks, interaction is <1 percentage point. In that configuration, both the conservative threshold and the readiness-focused objective penalize over- and under-maintenance, so precise intervention timing matters. Observation noise degrades the policy's ability to distinguish true urgency, and the narrow margin for error amplifies the cost of that confusion.

4.6.4 Robustness Across Operating Conditions

To assess whether the standard-scenario findings generalize, we evaluate policies under the three stress scenarios defined in Section 4.5.3, which perturb demand level, demand variance, and maintenance capacity.

Table 4.7 reports mean MS and OR by CV level across all four scenarios. Each cell averages across the six training configurations (two preference weights × three alarm thresholds) and 10,000 replications per configuration. At CV≤25%, GA-optimized policies outperform the heuristic benchmark under all four scenarios. The concave value-of-accuracy pattern persists under stress: most improvement is captured by CV=25%, with diminishing returns beyond CV=10%. The largest absolute advantage occurs under High Variance, where GA policies at CV=25% improve MS by 4.3 percentage points and OR by 2.6 percentage points over the heuristic. High Variance also produces the largest reactive failure reduction (89% fewer at CV=25%), consistent with the hypothesis that RUL-informed scheduling becomes more valuable when demand is volatile.

At CV=50%, however, GA policies underperform the heuristic in three of four stress scenarios. Under Resource Constrained conditions, CV=50% yields 59.2% MS and 52.4% OR versus the heuristic's 62.0% and 55.6%. High Variance and High Optempo show similar reversals. Noisy prognostics can be actively harmful under stress: the policy acts on misleading health signals. It schedules preventive maintenance for aircraft that do not need it and misses those that do. This finding reinforces the paper's central thesis. Information quality, not merely information availability, determines whether prognostic-informed policies improve outcomes. The reversal disappears by CV=25%, further supporting the saturation argument. Scenario-specific CV effect plots are provided in Appendix D.10.

This reversal has a practical implication. There exists a prediction quality threshold below which prognostic-based scheduling helps and above which it can hurt under stress. In our fleet configuration, that boundary falls between CV=25% and CV=50%. The specific value will depend on fleet size, demand intensity, and maintenance capacity, but the existence of a crossover point is general. Organizations should evaluate their prognostic pipeline against their operational environment before replacing a usage-based policy. If prediction quality falls in a regime where noisy signals misallocate scarce resources, the simpler heuristic may be preferable.

At moderate accuracy (CV≤25%), the relative advantage of RUL-based policies persists under all tested adverse conditions. Operational context affects the magnitude of improvement but not its direction. RUL-informed policies provide the largest advantage when operations are most challenging, a desirable property for units that cannot control their operational tempo or resource allocation. The optimal alarm threshold τ does not shift systematically with CV. One might expect that higher CV would favor a more conservative (higher) threshold to compensate for noisy readings, or a less conservative (lower) threshold to avoid false-positive interventions. Neither pattern emerges. Performance differences across thresholds are small (1–3 percentage points) at each accuracy level. The likely expla-

Table 4.7: Cross-scenario performance summary. Mean MS (%) and OR (%) by prediction accuracy level, averaged across six training configurations (standard errors of the six configuration means in parentheses). Heuristic baseline shown for reference. Bold values indicate GA underperformance relative to the heuristic.

	Standard		**High Optempo**		**High Variance**		**Resource Constr.**	
CV	MS	OR	MS	OR	MS	OR	MS	OR
Heuristic	62.1	55.7	34.3	40.0	58.0	55.5	62.0	55.6
50%	64.9 (0.8)	57.8 (1.0)	35.2 (0.5)	**38.8** (0.4)	**57.7** (1.0)	**54.1** (0.8)	**59.2** (0.9)	**52.4** (1.1)
25%	69.0 (0.4)	60.7 (0.6)	37.0 (0.2)	40.1 (0.2)	62.3 (0.5)	58.1 (0.4)	62.8 (0.5)	55.0 (0.6)
10%	70.8 (0.5)	62.5 (0.7)	37.7 (0.4)	40.8 (0.2)	65.1 (0.6)	60.6 (0.5)	64.9 (0.5)	57.0 (0.7)
5%	70.3 (0.1)	61.6 (0.1)	37.5 (0.1)	40.6 (0.1)	64.4 (0.1)	60.2 (0.1)	64.5 (0.1)	56.3 (0.2)
0%	70.3 (0.4)	61.9 (0.5)	37.7 (0.3)	40.8 (0.2)	64.8 (0.6)	60.4 (0.4)	64.8 (0.5)	56.8 (0.6)

nation is that the GA-optimized decision tree absorbs threshold variation through its learned split points. A policy trained at $\tau = 100$h can effectively implement a more conservative posture through its feature thresholds, and vice versa. The choice of τ is relatively robust to prediction accuracy.

4.7 Discussion

4.7.1 Mechanisms and Managerial Insights

CV is the decision-relevant metric Within the studied fleet configuration, the coefficient of variation governs the decision relevance of health predictions. Policies trained and evaluated under identical CV behave equivalently regardless of training conditions, policy class, or alarm threshold. The factorial analysis confirms that, within the decision-tree class, prediction accuracy dominates and policy adaptation is second-order. The practical investment question is therefore not "how accurate is the prediction?" but "what is the CV of the prediction model?"

The concave return on accuracy reflects a shift in which constraint binds. At high CV levels the information bottleneck dominates. When CV is 50%, the prediction intervals for different aircraft overlap substantially, and sorting becomes unreliable. At CV of 25%, the distributions separate enough for the decision tree to discriminate effectively. Accuracy improvements in this range relax the information constraint and yield large operational gains. Once accuracy reaches moderate levels, however, the bottleneck shifts from information to capacity. The planner can now distinguish which aircraft need attention, but slot availability limits how many can be maintained simultaneously. Further accuracy improvements provide finer discrimination but cannot relax the capacity constraint. The system transitions

from information-poor to resource-poor. Once slots are filled efficiently, more precise health estimates add little value. The value of information is ultimately bounded by the physical constraints of the maintenance system rather than by the limits of prediction technology. Moreover, the modeling simplifications adopted here (single-component degradation, stationary demand, perfect parts availability) generally bias results *against* finding strong CV effects, so the concavity result would likely strengthen under more realistic conditions (see Section 4.7.2).

Figure 4.1 illustrates the aggregate outcomes of a complex operational ecosystem, where readiness and utilization reflect dozens of uncontrolled factors. The simulation holds those factors constant and isolates the specific causal pathway of prediction accuracy. It cannot reproduce the historical scatter, nor does it attempt to. Its value is the opposite. By controlling for the operational noise that obscures any single mechanism in observed data, the model identifies the saturation point of one factor that would otherwise be undetectable.

Our results suggest three guiding principles for prognostic investment. First, the bottleneck shift implies that the value-of-accuracy relationship is concave and bounded. Early improvements in accuracy yield large operational gains. Later improvements yield progressively less. The exact saturation point varies by operational context, but the shape is general. Second, managers should prioritize prediction accuracy over algorithmic sophistication. Investment in signal processing, physics-of-failure model refinement, or sensor calibration will yield greater returns than complex optimization applied to noisy data. The factorial analysis shows that prediction accuracy drives operational improvement. Within the decision-tree policy class studied, accurate predictions improve outcomes even with simple decision rules. As CV degrades, reactive failures increase and consume more maintenance capacity (Table 4.8). The heuristic achieves the highest phase slot utilization (92.2%) because it never preempts failures with preventive maintenance. It cycles aircraft through phase less frequently and holds higher average bank hours than any GA policy (Figure D.8). Third, using prognostic information matters more than tuning precisely. Policies trained under mismatched noise conditions still outperform policies that ignore RUL entirely. Decision rules learned under one noise environment transfer effectively to other noise environments. Organizations need not perfectly recalibrate decision software every time prediction performance changes.

The accuracy ceiling and what lies beyond The results reveal a practical ceiling (Figure 4.8): at CV=10%, 98% of achievable improvement has been captured. The remaining 2% falls within the variability introduced by stochastic GA training (different runs at the same CV level produce policies that differ by 1–2 percentage points) and offers no practical

Table 4.8: Maintenance slot utilization by prediction accuracy level. Columns show reactive failure rate (events per aircraft-year), routine slot utilization (preventive plus reactive maintenance as a fraction of the two available routine slots), and phase slot utilization (fraction of the single phase slot). The heuristic performs no preventive maintenance and has the highest phase slot utilization (92.2%).

Policy	Reactive Events/Yr	Preventive/Reactive Slot Utilization	Phase Slot Utilization
GA, CV=0%	3.3	18.8%	78.5%
GA, CV=10%	3.9	19.7%	78.0%
GA, CV=25%	3.3	20.2%	75.5%
GA, CV=50%	4.9	24.7%	66.9%
Heuristic	15.8	21.6%	92.2%

return on further accuracy investment. Each fleet must identify its own saturation point. Once that threshold is achievable, deploy prognostic-informed policies rather than wait for better predictions. The binding constraint beyond the saturation point is not measurement accuracy. It is doctrine. Fleets governed by usage-based inspection intervals and fixed maintenance schedules cannot benefit from prognostic information regardless of prediction quality. The policy must adapt. The organization must enable that adaptation.

Organizations should monitor CV as an operational metric alongside traditional readiness measures. If CV drifts from 10% to 20%, the mission impact is modest. If it drifts from 25% to 40%, performance degrades substantially. Tracking CV provides early warning of decision quality degradation before it manifests in readiness shortfalls.

Although calibrated to Army rotorcraft, the framework applies to any setting with high-value assets, capacity-constrained repair facilities, and noisy health signals. Commercial aviation faces analogous tradeoffs between scheduled and unscheduled maintenance under similar capacity constraints. The specific CV threshold will vary across domains, but the saturation mechanism is general: once information quality is sufficient to fill maintenance capacity efficiently, further accuracy yields diminishing returns. Organizations should identify their saturation point rather than pursue maximal accuracy.

4.7.2 Limitations and Future Work

The findings hold under modeling simplifications that define both boundary conditions and research extensions. The model assumes a single limiting component per aircraft, but real aircraft have multiple subsystems with heterogeneous prediction accuracy. Future work should investigate whether coordinating maintenance across interacting failure modes amplifies or diminishes the value of accuracy investment. Parts availability is assumed perfect, yet sup-

ply chain delays in practice increase the cost of reactive failures. Incorporating stochastic lead times would test whether the VOI curve shifts toward earlier intervention. Demand is stationary, whereas real operational tempo varies with deployment cycles and training schedules. Time-varying demand models would reveal whether proactive strategies become more or less valuable during surge periods. Finally, the results guide relative prioritization of accuracy levels rather than budget allocations. Translating operational improvements into dollar values requires organization-specific cost accounting.

The capacity-constraint mechanism that drives these results binds for small fleets. Larger fleets with more scheduling flexibility would see the VOI curve flatten earlier, though the qualitative pattern of diminishing returns should persist. Methodologically, a multi-objective evolutionary algorithm could characterize the full Pareto frontier at each accuracy level rather than the individual weighted-sum solutions used here.

The finding that policy sophistication contributes less than prediction accuracy holds within the decision-tree policy class studied. Depth-3 trees have eight leaf nodes and four features, a constrained function approximator. Whether more expressive policy representations (e.g., neural networks optimized via DRL) could compensate for degraded prediction accuracy by learning more complex decision boundaries remains an open question. However, the factorial results suggest that the information loss at high CV is substantial. If a policy cannot distinguish aircraft with 50 hours remaining from those with 150 hours remaining, as occurs at CV of 50%, no amount of policy sophistication can recover that lost discrimination. The information bottleneck appears to bind before the policy bottleneck in this setting.

The factorial result also depends on the memoryless observation structure. The model re-samples RUL observations independently after each flight, so each maintenance decision rests on a single noisy reading. A belief-state POMDP that aggregates multiple observations over time could, in principle, reduce effective observation noise through temporal filtering and partially compensate for prediction degradation. The current design forecloses this possibility deliberately. It isolates the value of single-observation accuracy, which is the operationally relevant quantity when health monitoring systems report point estimates rather than posterior distributions. Extending the model to allow belief-state updating would test whether temporal filtering shifts the saturation point.

As noted in Section 4.7.1, these simplifications are likely to bias results *against* finding strong CV effects. Single-component models understate coordination benefits, stationary demand understates surge-period value, and perfect parts availability understates reactive failure costs. The core finding that CV governs decision relevance would likely strengthen, not weaken, under more complex conditions.

4.8 Conclusion

This paper quantified the marginal operational value of prediction accuracy in fleet maintenance by studying how information quality interacts with capacity-constrained decision policies in a partially observed stochastic system. We optimized decision-tree policies across five prediction accuracy levels and obtained clear answers to the research questions posed in Section 4.1: the value-of-accuracy curve is strongly concave with saturation near CV=25%, and the dominant mechanism is preemption of reactive failures rather than policy sophistication. The concavity reflects a bottleneck shift. At high CV, the information constraint dominates and accuracy improvements yield large gains. Once predictions reach moderate precision, the binding constraint shifts from information to maintenance capacity, and further accuracy improvements cannot relax the physical slot constraint.

The marginal value of prediction accuracy is strongly concave. Moving from CV=50% to CV=25% provides larger operational gains than moving from CV=10% to CV=5%. At CV=25%, approximately 79% of achievable improvement has been captured. At CV=10%, that figure reaches 98%. Perfect prognostics are not required. Under standard conditions, even noisy predictions at CV=50% outperform benchmark policies that ignore RUL entirely, though this advantage can reverse under operational stress (Table 4.7). Improved accuracy shifts the mission success–operational readiness Pareto frontier outward. Units gain on both dimensions simultaneously because fewer unplanned maintenance events free capacity for scheduled work.

This performance gain is driven primarily by the preemption of reactive failures. As prediction accuracy improves, the decision policy identifies at-risk aircraft earlier and schedules maintenance before failure occurs. The factorial analysis confirms that prediction accuracy, not policy sophistication, is the dominant driver. Within the decision-tree policy class studied, accurate predictions improve outcomes even with simple decision rules. This mechanism remains dominant across all tested operational contexts. GA-optimized policies outperform benchmarks under high optempo, high-variance demand, and resource-constrained environments. The relative advantage is largest when operations are most challenging.

The contribution is not a predictive-maintenance endorsement but a result with both practical and theoretical implications. Within capacity-constrained fleet operations, CV governs decision relevance. If an organization knows its prognostic CV, it knows whether further investment in prediction accuracy, modeling effort, or policy redesign will yield operational benefit.

Managerial implications These findings imply three shifts in how organizations should approach prognostic investment. First, within the fleet configuration studied, target CV of approximately 25% rather than perfect accuracy, since most operational value is captured at moderate precision. Second, deploy RUL-informed policies even with imperfect predictions, because using prognostic information dominates ignoring it. Third, monitor CV as a leading indicator of decision quality and treat it with the same operational importance as readiness itself.

Methodological outlook The simulation-optimization framework developed here extends naturally to other fleet settings with binding capacity constraints and imperfect health information. The factorial design provides a template for isolating prediction accuracy from policy structure in any value-of-information study where policies can adapt to the quality of available information.

Supplementary Materials

The complete simulation codebase, including all genetic algorithm implementations, policy configurations, evaluation scripts, and analysis notebooks, is available at `https://github.com/defense031/preventive_maintenance_us_army_aviation_genetic_algorithm` (archived at `https://doi.org/10.5281/zenodo.18653364`). The repository includes full documentation for replication and extension to other platforms. A `requirements.txt` file specifying all Python dependencies is included in the repository.

Chapter 5

Conclusion: Continuing an Eighty-Year Conversation

5.1 Introduction

This dissertation began with a deceptively simple question: How do Army aviation units actually make decisions about flying their aircraft? The answer reveals a fundamental tension between how we *measure* readiness and how readiness is actually *produced*. Waddington's observation about RAF Coastal Command in Chapter 1 could have been written yesterday.

Current policy treats OR as an input to decision-making: when availability falls below 75%, units should reduce utilization to preserve readiness. This dissertation shows the opposite. OR is an emergent property of the coupled system, not a control signal that governs it. The central thesis is that OR and aircraft utilization are joint outputs of a coupled maintenance-usage system shaped by capacity constraints, information quality, and stochastic failures. The 75% OR target, established in 1985 without documented analytical foundation, does not govern system outcomes as policy intended. The practical consequence is that readiness assessment should shift from threshold compliance to diagnosing how units navigate the coupled system.

This chapter situates the argument historically, presents the OR-Usage framework that integrates the dissertation's findings, and states implications for commanders, the aviation enterprise, and prognostic investment decisions.

5.2 The Eighty-Year Conversation

This work does not introduce a new problem. It continues a conversation that began in 1943 and remains unresolved at the policy level today.

5.2.1 Waddington's Insight and the Coupling Problem

Chapter 1 traced the coupling problem to Waddington's 1943 analysis of RAF Coastal Command, where treating the 75% serviceability target as a measure of squadron efficiency proved "thoroughly misleading" [3]. His recommendations increased effective flying hours by more than 60% without adding aircraft. The insight was operationally useful during wartime but did not persist into peacetime measurement doctrine. Waddington's work remained classified for thirty years; when finally published in 1973, it found limited audience in the defense OR community.

5.2.2 The Recurring Critique

The critique of scalar readiness metrics is not new. Raffensperger and Schrage [5] proposed a "new paradigm" for measuring readiness, arguing that aggregate OR rates fail to capture what a commander actually needs and that readiness should be quantified in terms like time or resources to mission-ready status. Ignizio [123] explicitly coupled maintenance with force readiness, describing the vital role that understanding metric interdependence played in the origination of operations research itself. Harrison [9] criticized input-focused readiness metrics for "shedding little light" on actual mission performance, calling for output-oriented metrics that consider what units do with their readiness. Junor [10] argued that unit-level readiness rates "do not provide enough information" for strategic management and must be interpreted alongside demand and utilization.

The clearest contemporary articulation came from General David Goldfein, then USAF Chief of Staff, who stated that "the MC rate is actually not a very good measure of aviation readiness" [11].

5.2.3 A Documented but Under-Theorized Gap

Despite this recurring critique, the coupling argument remains under-theorized in the academic literature. A systematic search of Google Scholar (1985–2025) across three phrase combinations related to military aviation maintenance and readiness identified 4,754 publications. The vast majority optimize within the existing paradigm, such as scheduling maintenance, maximizing availability, improving fleet readiness, without questioning whether OR alone is the right metric. Among papers on "aircraft availability" and "optimization," the proportion also mentioning "doctrine" or "policy" declined from 100% in the late 1980s to 8% by 2020–2025 (Figure 5.1). The field has grown increasingly technical while engagement with policy foundations has atrophied.

The Army's readiness reporting apparatus, formalized through AR 220-1 and AR 700-138 (Chapter 1), made OR reportable and comparable across units but divorced the metric from how readiness was actually generated. Because a scalar percentage is easy to report and compare, it became the management object even when it was no longer diagnostic.

The contribution of this dissertation is to restore the missing second dimension. The OR-Usage framework developed here is a modern restatement of Waddington's serviceability-intensity logic, enabled by contemporary data and decision-analytic tools. It treats readiness as navigation in OR-Usage space rather than compliance with a scalar threshold.

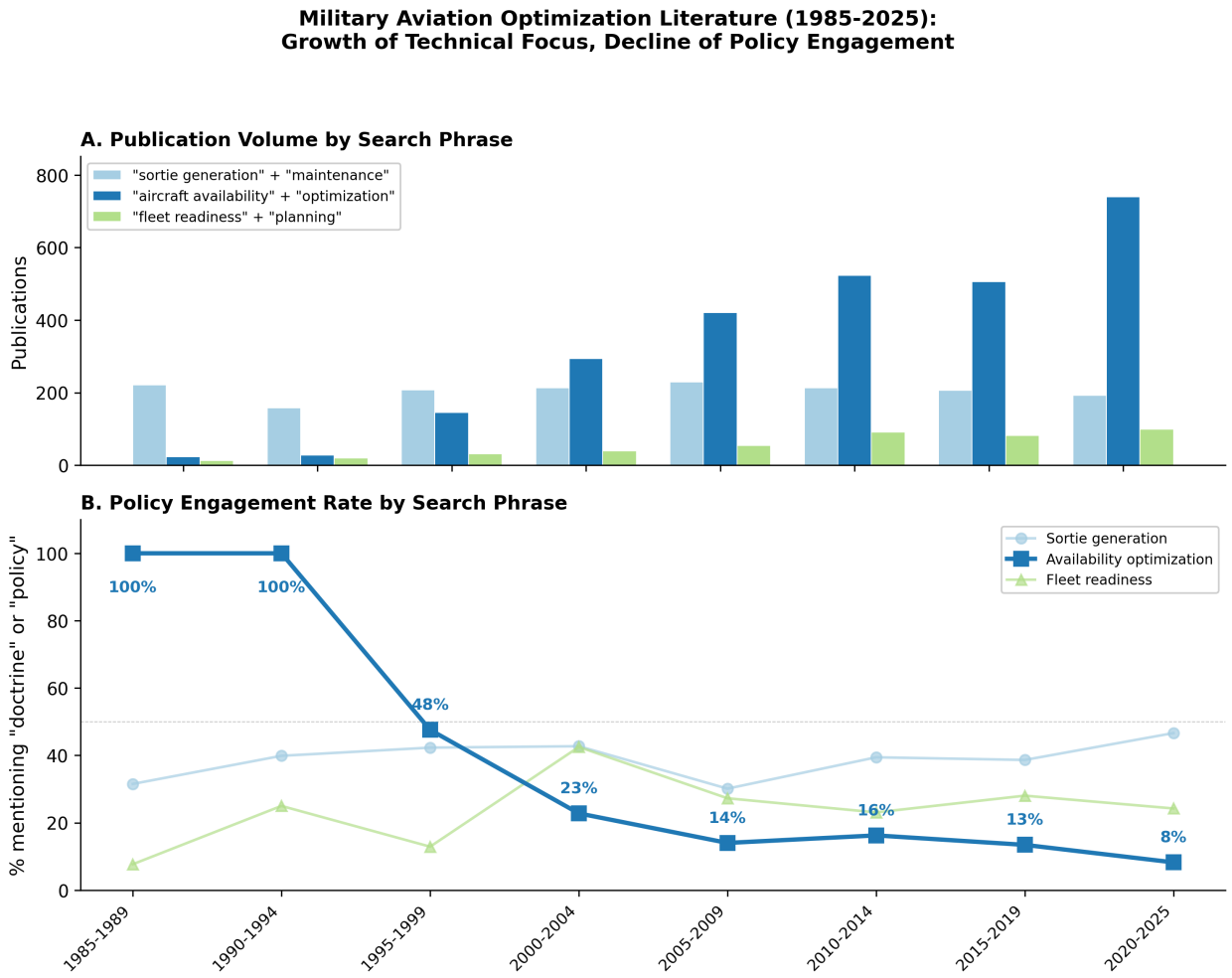

Figure 5.1: Military aviation optimization literature (1985–2025). Panel A shows publication volume by Google Scholar search phrase; Panel B shows the percentage of papers in each category also mentioning "doctrine" or "policy." Military aircraft-optimization research grew 32-fold while policy engagement collapsed from 100% to 8%.

5.3 The OR-Usage Framework

5.3.1 The Claim

OR and utilization are jointly generated by a coupled maintenance-usage system. Usage decisions load maintenance queues. Maintenance capacity determines how quickly aircraft return to service. Stochastic failures create unpredictable demand. Information quality shapes how well units can anticipate and respond. Evaluating OR in isolation systematically misrepresents unit efficiency. The framework developed across these chapters treats readiness and utilization as joint outputs to be interpreted together, not separate objectives to be

traded off.

5.3.2 Evidence

Chapter 2 (Behavioral Coupling) Analysis of 265,472 aircraft-day observations reveals that OR and usage do not behave as independent variables. Units fly more, not less, when OR falls below the 75% threshold. This is the opposite of what a binding constraint would produce. The strongest predictor of flight dispatch by magnitude is not OR but the day of the week: flight probability drops approximately 80% on weekends and 63% on Fridays relative to midweek. These effects dwarf the influence of readiness status. A post-hoc interaction analysis further reveals that units at low OR disproportionately ground aircraft approaching phase maintenance, consistent with risk-averse behavior that prioritizes metric preservation over efficient phase flow. OR does not function as a constraint in the decision process. Calendar structure, phase proximity, and their interaction with readiness status explain the variation that OR is assumed to govern.

Chapter 3 (Behavioral Heterogeneity) Among units operating under a common doctrinal and reporting environment, decision behavior varies in systematic, measurable ways. Behavioral profiles cluster into distinct types, and units exhibiting metric-preserving behaviors occupy Pareto-dominated positions in OR-Usage space. This finding bridges Chapters 2 and 4: it shows that efficiency differences are associated with behavioral variation (not fully explained by unobserved constraints), and it establishes that units are comparable in OR-Usage space in a way OR-alone cannot support.

Chapter 4 (Value of Information) The value-of-accuracy relationship is strongly concave: CV=25% captured 79% of achievable improvement, CV=10% captured 98%. These gains arise primarily through reduction in disruptive reactive maintenance events. The factorial analysis confirms that prediction accuracy, not policy sophistication, is the dominant driver of performance. Critically, information creates value only when doctrine adapts to use it. Layering sensors onto fixed usage-based schedules captures little benefit. Once prediction accuracy reaches moderate levels, the binding constraint shifts from information quality to maintenance capacity. Beyond that threshold, the binding factor becomes the doctrinal structures that determine whether prognostic information can influence decisions at all.

5.3.3 What the Evidence Reveals

Taken together, the three chapters converge on a specific diagnosis of the 75% OR threshold as a policy instrument.

First, the threshold does not constrain behavior as designed. Chapter 2 shows that units

fly *more* when OR falls below 75%, and that environmental factors such as day of the week and phase proximity, not readiness status, are the dominant determinants of whether a mission-capable aircraft flies on a given day. A metric that does not influence the decisions it is intended to govern has lost its regulatory function.

Second, where the threshold does appear to influence behavior, the effect is counterproductive. The OR-by-phase interaction in Chapter 2 reveals that units at low OR disproportionately ground aircraft near phase maintenance, consistent with risk-averse behavior aimed at preserving the metric rather than managing phase flow efficiently. Chapter 3 confirms this at the unit level: clusters exhibiting metric-preserving decision profiles occupy Pareto-dominated positions in OR-Usage space. The units most attentive to the threshold are, on average, the least efficient. This pattern is a textbook instance of Goodhart's Law: when a measure becomes a target, it ceases to be a good measure [124]. The fixed OR standard creates exactly this dynamic. Maintenance managers face incentives to prioritize serviceability benchmarks over combat proficiency. The organizational structure reinforces the problem. As described in Chapter 1, flight decisions are made at company and battalion level while maintenance capacity is managed at battalion level. This separation creates competing objectives between operations and maintenance that a single scalar metric cannot reconcile.

Third, the threshold becomes irrelevant once the binding constraint shifts. Chapter 4 demonstrates that as prognostic accuracy improves, the system transitions from information-constrained to capacity-constrained. At that point, further improvements in prediction precision cannot relax the physical maintenance slot constraint, and the scalar threshold adds no diagnostic value. The constraint that ultimately bounds fleet performance is not a reporting target but the maintenance capacity and doctrinal flexibility available to the unit.

These three findings do not merely suggest that OR is an incomplete metric. They indicate that the 75% threshold, as currently implemented, fails in each of the roles it might serve: it does not constrain utilization decisions, it may induce inefficient conservatism in units that do attend to it, and it provides no useful signal once the fleet is managed with even moderately accurate health information.

5.3.4 The Framework

Figure 5.2 shows the coupled system. Mission demand arrives exogenously. Usage decisions load maintenance queues, expose aircraft to failure, and directly affect fleet availability. Maintenance capacity determines return-to-service rates. Stochastic failures create unpredictable maintenance demand and can cause mission failure directly. Information quality

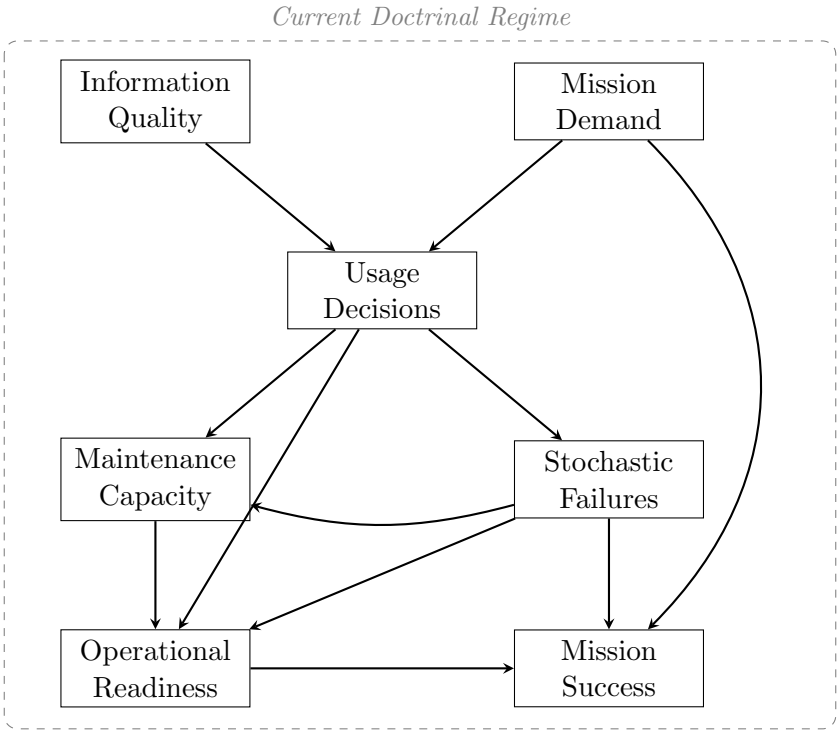

Figure 5.2: The OR-usage framework: readiness and mission success as emergent properties of a coupled system. The dashed boundary indicates that the entire framework operates within the current doctrinal regime.

shapes how well units anticipate and respond. OR and mission success both emerge from this system. Mission success depends on whether fleet availability meets demand. OR alone cannot distinguish whether low availability reflects a capacity shortfall or intensive utilization with capacity keeping pace. The entire framework operates within the current doctrinal regime, and changes to that regime can reshape all relationships simultaneously. Interpreting readiness in OR-Usage space converts reporting into diagnosis and clarifies which lever is binding: demand intensity, maintenance resources, information quality, or doctrinal flexibility.

5.3.5 Implications

For Commanders

The most immediate lesson is that calendar-driven volatility, not OR status, shapes fleet utilization patterns. Units that treat the training week as a fixed constraint concentrate flights midweek and stand down on Fridays, weekends, and Mondays. This pattern introduces predictable maintenance demand spikes that complicate phase flow management. Commanders

should examine their operating rhythm for these patterns and consider whether redistributing flights more evenly across the week would smooth maintenance loading. Phase flow itself should be managed as an explicit capacity-planning problem: the evidence from Chapter 2 shows that units tend to adjust flight patterns only for aircraft very near to or just out of phase, rather than managing the full fleet's phase spacing proactively. Most fundamentally, OR should be interpreted jointly with utilization, not treated as a standalone target. A unit at 72% OR with high flying hours and well-managed phase flow is in a fundamentally different operational posture than a unit at 72% OR with low flying hours and bunched phase entries, yet current reporting treats them identically.

For the Enterprise

The current reporting structure evaluates units on OR alone. This creates a measurement environment in which a unit can improve its reported readiness by flying less. This is exactly the behavior Waddington identified in 1943 and General Goldfein described in 2019. Chapter 3 provides empirical evidence that this incentive operates in practice: units whose decision profiles reflect metric-preserving tendencies occupy Pareto-dominated positions relative to peers that fly more aggressively while sustaining comparable readiness. Evaluating readiness without accounting for utilization risks rewarding behavior that preserves metrics rather than combat capability. The enterprise should evaluate units in OR-Usage space, reconsider the 75% threshold as the dominant management device, and develop reporting frameworks that jointly consider availability and usage intensity. FHPA alongside OR is a natural extension of existing reporting that requires no doctrinal revision, only the recognition that a single scalar percentage cannot capture what a two-dimensional system produces. The OR-Usage framework developed here provides the analytical foundation for a transition that the coupling literature has long implied: replacing the fixed 75% threshold with a variable serviceability standard indexed to operational intensity. Gordon's 1943 analysis for RAF Coastal Command derived exactly this relationship, a serviceability rate that varies as a function of flying intensity [3], yet the construct was never adopted into peacetime doctrine. The Pareto frontier in Chapter 3 operationalizes this logic with contemporary data: it evaluates whether a unit's readiness is commensurate with its utilization, rather than whether it exceeds an arbitrary fixed threshold.

For Prognostic Investment

The Army is investing substantially in predictive maintenance sensors and platforms. Chapter 4 provides a framework for evaluating where those investments yield operational returns

and where they encounter diminishing marginal value. The central finding is that moderate prediction accuracy (CV≈25%) captures approximately 79% of achievable improvement, and further refinement beyond CV≈10% yields negligible operational gains. Organizations should assess where the force currently operates on the accuracy-value frontier and direct investment toward crossing the threshold at which information quality ceases to be the binding constraint. Equally important, the value of prognostic information is bounded by the doctrinal structures that govern how it can be used. Fleets managed by fixed usage-based inspection intervals cannot benefit from prognostic signals regardless of prediction quality. The implication is that investment in sensors must be accompanied by investment in policies that can act on the information those sensors provide. Without doctrinal adaptation, prognostic technology becomes an overlay on a system that cannot use it. This is consistent with the GAO's finding that Army predictive maintenance efforts have yielded limited results in part because no standardized doctrine exists to translate condition monitoring into maintenance decisions [20]. Chapter 4 provides the formal demonstration: the value of prognostic information is bounded not by sensor technology but by the doctrinal structures that determine whether better information can change decisions.

5.4 Limitations and Future Work

5.4.1 Limitations

All empirical analysis focuses on AH-64 Apache operations in Active Component units. Specific findings may differ for other platforms, Reserve/Guard units, or operational contexts. Cost modeling is beyond this scope. The simulation models one RUL-tracked component per aircraft with stationary demand and an eight-aircraft fleet. Real aircraft have multiple subsystems with interacting failures, demand varies with deployment cycles, and larger fleets may exhibit different dynamics. The data do not indicate deployment status, mission requirements, or weather conditions; do not distinguish maintenance-related from supply-related non-mission capability; and span 2019–2022, a period affected by the COVID-19 pandemic. Observed patterns may partially reflect these unmeasured factors.

5.4.2 Future Work

Near-term extensions include multi-component RUL models with associated failure modes, supply chain integration to analyze how parts availability mediates prognostic value, longitudinal validation of behavioral recommendations, and cross-platform replication (UH-60,

CH-47, ground vehicles) to test generalizability. Strategic directions include cross-service analysis to determine whether patterns generalize beyond Army aviation, crew scheduling integration for combined equipment-personnel readiness models, development of SOM-based decision support tools for commanders, and formal testing of metrics that jointly incorporate OR and usage intensity.

5.5 Concluding Remarks

Readiness is produced, not reported. The 75% OR threshold, established in 1985 without documented analytical foundation, was designed to function as a regulatory signal, a constraint that would moderate utilization when availability declined. This dissertation demonstrates that it fails in that role. The threshold does not constrain the behavior it was intended to constrain: units fly more, not less, when OR falls below 75%, and the strongest determinants of flight dispatch are environmental factors such as day of the week and phase proximity, not readiness status. Where the threshold does influence behavior, the effect is counterproductive. Units that attend most closely to the metric exhibit risk-averse decision patterns, particularly the tendency to ground phase-proximate aircraft during low-OR periods. These units occupy Pareto-dominated positions in OR-Usage space. And the threshold becomes irrelevant as the system matures: once prognostic information reaches moderate accuracy, the binding constraint shifts from information to maintenance capacity, and from capacity to the doctrinal structures that determine whether better information can influence decisions at all.

The OR-Usage framework developed here does not require the Army to abandon readiness reporting. It requires the Army to stop treating a scalar percentage as the dominant management device for a system that produces two-dimensional outcomes. The framework restores the diagnostic dimension that Waddington identified in 1943 and that eighty years of peacetime reporting gradually stripped away. It enables incremental improvement through better phase flow management, more even operating rhythms, and joint evaluation of availability and utilization. These steps do not require perfect sensors or wholesale doctrinal revision. Readiness is not a number to be maximized. It is a system outcome that must be interpreted jointly with how the force is employed, and the tools to do so are now available.

REFERENCES

[1] HQDA. Army Techniques Publication 3-04.7: Army Aviation Maintenance, 10 2020. Headquarters, Department of the Army.

[2] Austin D. Semmel, Hans Sebastian Heese, and Brandon M. McConnell. Evaluating the implementation of operational readiness and maintenance policies in US Army aviation. *Journal of Defense Modeling and Simulation*, 2025. doi: 10.1177/15485129251328044.

[3] C. H. Waddington. *O.R. in World War 2: Operational Research Against the U-Boat*. Elek Science, London, 1973. Originally classified material from the RAF Coastal Command Operational Research Section, 1942-1945.

[4] Mike Busch. The waddington effect. *Sport Aviation*, 60(3):98–101, March 2011. https://www.savvyaviation.com/wp-content/uploads/articles_eaa/EAA_2011-03_the-waddington-effect.pdf.

[5] John F. Raffensperger and Linus E. Schrage. A new paradigm for measuring military readiness. *Military Operations Research*, 3(5):21–34, 1997. doi: 10.5711/morj.3.5.21. Proposes time- and resource-based readiness measures as alternatives to aggregate OR percentages.

[6] James L. Kays, William B. Carlton, Mark M. Lee, and Jr. Ratliff, William L. Analysis of operational readiness rates. Technical report, United States Military Academy, Department of Systems Engineering and Operations Research Center, West Point, New York, July 1998.

[7] Bradley W. Pippin. Allocating flight hours to Army helicopters. Master's thesis, Naval Postgraduate School, Monterey, California, 6 1998. https://apps.dtic.mil/sti/citations/ADA350138.

[8] B. Charles Tatum. Structural equation modeling applied to the analysis of readiness: Development of the naval readiness conceptual model. *Military Operations Research*, 7(3):5–15, 2002. Critiques SORTS reporting as measuring resource status rather than unit performance.

[9] Todd Harrison. Rethinking readiness. *Strategic Studies Quarterly*, 8(3):38–68, 2014. Critiques input-focused readiness metrics; calls for output-oriented measurement.

[10] Laura J. Junor. Managing military readiness. INSS Strategic Perspectives 20, National Defense University Press, 2017. https://digitalcommons.ndu.edu/inss-strategic-perspectives/20/.

[11] David L. Goldfein. Interview on air force readiness metrics. Air & Space Forces Magazine, September 2019. USAF Chief of Staff statement that "the MC rate is actually not a very good measure of aviation readiness".

[12] HQDA. Army Regulation 220 – 1: Army Unit Status Reporting and Force Registration–Consolidated Policies, 8 2022. Headquarters, Department of the Army.

[13] HQDA. Army Regulation 700-138: Army Logistics Readiness and Sustainability, 4 2018. Headquarters, Department of the Army.

[14] U.S. Department of Defense. AH-64E Apache Remanufacture: Modernized Selected Acquisition Report (MSAR). Technical report, U.S. Department of the Army, Washington, D.C., December 2023. https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/AH-64E_Remanufacture_MSAR_Dec_2023.pdf.

[15] HQDA. Army Regulation 95–1: Aviation Flight Regulations, 3 2018. Headquarters, Department of the Army.

[16] Jovani Dalzochio, Rafael Kunst, Jorge Luis Victória Barbosa, Pedro Clarindo da Silva Neto, Edison Pignaton, Carla Schwengber Ten Caten, and Alex de Lima Teodoro da Penha. Predictive maintenance in the military domain: A systematic review of the literature. *ACM Computing Surveys*, 55(13s):1–30, July 2023. doi: 10.1145/3586100.

[17] GAO. National Guard Helicopters: Additional Actions Needed to Prevent Accidents and Improve Safety. Report to the Committee on Armed Services, House of Representatives GAO-23-105219, United States Government Accountability Office, 3 2023. https://www.gao.gov/assets/gao-23-105219.pdf.

[18] GAO. MILITARY READINESS: Department of defense domain readiness varied from fiscal year 2017 through fiscal year 2019. Report to Congressional Committees GAO-21-279, United States Government Accountability Office, 441 G St. N.W., Washington, DC 20548, 4 2021.

[19] Kyle McDermott. Conversation on AH-64 Apache company command experiences. Personal communication, 2024. Major Kyle McDermott, U.S. Army, conversation on 21 February 2024.

[20] GAO. MILITARY READINESS: Actions needed to further implement predictive maintenance on weapon systems. Report to the Committee on Armed Services, House of Representatives GAO-23-105556, United States Government Accountability Office, 441 G St. N.W., Washington, DC 20548, 12 2022.

[21] GAO. Weapon System Sustainment: Aircraft Mission Capable Goals Were Generally Not Met and Sustainment Costs Varied by Aircraft. Report to Congressional Committees GAO-23-106217, United States Government Accountability Office, November 2022. https://www.gao.gov/assets/gao-23-106217.pdf.

[22] Jonathan Barbee. Griffin Program is a "Game Changer" for Maintenance. *Army Aviation Magazine*, 73(1):22, January 2024. Army Aviation Association of America (AAAA).

[23] A.J. Lipina. Identifying critical factors affecting combat mission ready status among USAF europe's aircrew. Graduate Research Project, School of Engineering and Management, Air Force Institute of Technology, 6 2009.

[24] Adam MacKenzie, J.O. Miller, Raymond R. Hill, and Stephen P. Chambal. Application of agent-based modelling to aircraft maintenance manning and sortie generation. *Simulation Modelling Practice and Theory*, 20:89–98, 2012. doi: 10.1016/j.simpat.2011.09.001.

[25] Nathaniel Choo, Darryl Ahner, and Lance Champagne. The effects of aircraft use and available repair spares on aircraft sortie generation: a long-duration logistical wargaming simulation tool. *Journal of Defense Modeling and Simulation*, 20:111–130, 09 2021. doi: 10.1177/15485129211040782.

[26] J. Ritschel, T. Ritschel, and N. York. Providing a piece of the puzzle: Insights into the aircraft availability conundrum. *Journal of Defense Analytics and Logistics*, 3(1): 29–40, 2019. doi: 10.1108/JDAL-09-2018-0015.

[27] V. McLean and A. D. Reiman. Transportation service level impact on aircraft availability. *Journal of Defense Analytics and Logistics*, 6(1):46–58, 2022. doi: 10.1108/JDAL-10-2021-0010.

[28] Andreas Gavranis and George Kozanidis. An exact solution algorithm for maximizing the fleet availability of an aircraft unit subject to flight and maintenance requirements. *European Journal of Operational Research*, 242:631–643, 04 2015. doi: 10.1016/j.ejor.2014.10.016.

[29] Stephane Barde, Soumaya Yacout, and Hayong Shin. Optimal preventive maintenance policy based on reinforcement learning of a fleet of military trucks. *Journal of Intelligent Manufacturing*, 30:147–161, 01 2019. doi: 10.1007/s10845-016-1237-7.

[30] Mikael Öhman, Markus Hiltunen, Kai Virtanen, and Jan Holmström. Frontlog scheduling in aircraft line maintenance: From explorative solution design to theoretical insight into buffer management. *Journal of Operations Management*, 67(2):120–151, 2021. doi: 10.1002/joom.1108.

[31] Simon N. Wood. Inference and computation with generalized additive models and their extensions. *TEST*, 29:307–339, April 2020. doi: 10.1007/s11749-020-00711-5.

[32] Julian J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2 edition, 2016. ISBN 9781498720960.

[33] Hercules Dalianis. *Clinical Text Mining*, chapter 6. Springer, 2018. ISBN 978-3-319-78502-8.

[34] Gavin L. Simpson. Modelling Palaeoecological Time Series Using Generalised Additive Models. *Frontiers in Ecology and Evolution*, 6, 2018. doi: 10.3389/fevo.2018.00149. `https://www.frontiersin.org/articles/10.3389/fevo.2018.00149`.

[35] Trevor Hastie and Robert Tibshirani. Generalized Additive Models: Some Applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987. ISSN 01621459. doi: 10.2307/2289439. http://www.jstor.org/stable/2289439.

[36] Simon N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman and Hall/CRC, 2 edition, 2017. doi: 10.1201/9781315370279.

[37] Giampiero Marra and Simon N. Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1): 53–74, 2012. doi: 10.1111/j.1467-9469.2011.00760.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2011.00760.x.

[38] Simon N. Wood. On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228, March 2013. ISSN 0006-3444. doi: 10.1093/biomet/ass048.

[39] D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. ISBN 9780521785167. https://books.google.com/books?id=Y4uEvXFP2voC.

[40] Blake E. Schwartz. *Performance Evaluation and Improvement of Expeditionary Military Sustainment Network Models*. PhD thesis, NC State University, Raleigh, NC, 2020. http://www.lib.ncsu.edu/resolver/1840.20/37489.

[41] Brandon M. McConnell, T.J. Hodgson, M.G. Kay, R.E. King, Y. Liu, G.H. Parlier, K. Thoney-Barletta, and J.R. Wilson. Assessing Uncertainty and Risk in an Expeditionary Military Logistics Network. *Journal of Defense Modeling and Simulation*, 18 (2):135–156, 2019. doi: 10.1177/1548512919860595. Online July 2019; print April 2021.

[42] Matthew B. Rogers, Brandon M. McConnell, Thom J. Hodgson, Michael G. Kay, Russell E. King, Greg H. Parlier, and Kristin Thoney-Barletta. A Military Logistics Network Planning System. *Military Operations Research*, 23(4):5–24, 2018. https://repository.lib.ncsu.edu/items/76464071-5a3f-4bf5-aa18-123b72cd4a3a.

[43] Thomas R. O'Neal. *Sortie-based Aircraft Component Demand Rate to Predict Requirements*. PhD thesis, Air Force Institute of Technology, 2020. Theses and Dissertations; 3199, https://scholar.afit.edu/etd/3199.

[44] M. Verhoeff, W.J.C. Verhagen, and R. Curran. Maximizing Operational Readiness in Military Aviation by Optimizing Flight and Maintenance Planning. *Transportation Research Procedia*, 10:941–950, 2015. doi: 10.1016/j.trpro.2015.09.048. 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015.

[45] David W. Lehman, Jungpil Hahn, Rangaraj Ramanujam, and Bradley J. Alge. The dynamics of the performance-risk relationship within a performance period: The moderating role of deadline proximity. *Organization Science*, 22(6):1613–1630, 2011. ISSN 10477039, 15265455. doi: 10.1287/orsc.1100.0626. http://www.jstor.org/stable/41303146.

[46] U.S. Department of Defense. 2022 National Defense Strategy, 10 2022. https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF.

[47] Kenneth J. Arrow. An extension of the basic theorems of classical welfare economics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 2:507–532, 1951.

[48] Gerard Debreu. The Coefficient of Resource Utilization. *Econometrica*, 19(3):273–292, 1951. https://www.jstor.org/stable/1906814.

[49] Kazutoshi Ando, Atsuhiko Kai, Yasunobu Maeda, and Kanuyuki Sekitani. Least distance based inefficiency measures on the pareto-efficient frontier in DEA. *Journal of the Operations Research Society of Japan*, 55(1):73–91, 2012. doi: 10.15807/jorsj.55.73.

[50] Tjalling C. Koopmans. Efficient allocation of resources. *Econometrica*, 19(4):455–465, 1951. doi: 10.7249/p116. https://www.jstor.org/stable/1907467.

[51] M. J. Farrell. The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3):253–290, 1957. doi: 10.2307/2343100. https://doi.org/10.2307/2343100.

[52] Justin L. Anderson, Jessica M. Astudillo, Zachary E. Butcher, Matthew D. Cornman, Anthony J. Correale, James B. Crumpacker, Nathaniel C. Dennie, Alex R. Gaines, Mark A. Gallagher, John C. Goodwill, Emily S. Graves, Jr. Hale, Donald B., Kimberly G. Holland, Benjamin D. Huffman, Michelle McGee, Nicholas A. Pollack, Rachel C. Ramirez, Camero Song, Emmie K. Swize, Erick A. Tello, Jesse G. Wales, Julius C. Walker, Andrew B. Wilson, Jr. Wilson, William F., Kylie E. Wooten, and Marcelo Zawadzki. Stochastic preemptive goal programming of air force weapon systems mix. *Journal of Defense Modeling and Simulation*, 20(2):147–158, 2023. doi: 10.1177/15485129211051751. https://doi.org/10.1177/15485129211051751.

[53] Abraham Charnes, William W Cooper, and Edward Rhodes. Measuring the Efficiency of Decision-making Units. *European Journal of Operational Research*, 2(6):429–444, 1978. doi: 10.1016/0377-2217(78)90138-8.

[54] Thiago Gomes Leal Ganhadeiro, Eliane da Silva Christo, Lidia Angulo Meza, Kelly Alonso Costa, and Danilo Pinto Moreira de Souza. Evaluation of Energy Distribution Using Network Data Envelopment Analysis and Kohonen Self-Organizing Maps. *Energies*, 11(10):2677, 2018. doi: 10.3390/en11102677.

[55] Aiyshwariya Paulvannan Kanmani, Renee Obringer, Benjamin Rachunok, and Roshanak Nateghi. Assessing Global Environmental Sustainability Via an Unsupervised Clustering Framework. *Sustainability*, 12(2):563, 2020. doi: 10.3390/su12020563.

[56] George Kozanidis. A multiobjective model for maximizing fleet availability under the presence of flight and maintenance requirements. *Journal of Advanced Transportation*, 43(2):155–182, 2009. doi: 10.1002/atr.5670430205. Named the Flight and Maintenance Planning (FMP) problem; added explicit capacity constraints.

[57] George Kozanidis, Andreas Gavranis, and Eftychia Kostarelou. Mixed integer least squares optimization for flight and maintenance planning of mission aircraft. *Naval Research Logistics*, 59(3-4):212–229, 2012. doi: 10.1002/nav.21483.

[58] David O. Marlow and Robert F. Dell. Optimal Flight and Maintenance Planning of a Military Aircraft Fleet through to Life-of-Type. *Military Operations Research*, 30(1): 5–27, 2025. doi: 10.5711/1082598330105. Key contributions: Phase flow staircase constraints, hours-based maintenance triggers, multi-objective optimization.

[59] Franco Peschiera, Robert Dell, Johannes Royset, Alain Haït, Nicolas Dupin, and Olga Battaïa. A novel solution approach with ML-based pseudo-cuts for the flight and maintenance planning problem. *OR Spectrum*, 43:635–664, 2021. doi: 10.1007/s00291-020-00591-z.

[60] George Kozanidis, Andreas Gavranis, and George Liberopoulos. Heuristics for flight and maintenance planning of mission aircraft. *Annals of Operations Research*, 221: 211–238, 2014. doi: 10.1007/s10479-013-1376-6.

[61] Douglas S. Altner, Paul R. Bartholomew, Erin M. Bongo, Susan S. M. Hanson, Jessica M. L. Matthews, Anthony C. Rojas, and Rebecca K. Rousseau. Military Aircraft Flight and Maintenance Scheduling with Many Additional Considerations. *Journal of Defense Modeling and Simulation*, pages 1–24, 2025. doi: 10.1177/15485129251383566. 16-criterion objective function; tolerance bands for phase timing flexibility.

[62] Ville Mattila and Kai Virtanen. Maintenance scheduling of a fleet of fighter aircraft through multi-objective simulation-optimization. *Simulation: Transactions of the Society for Modeling and Simulation International*, 90(9):1023–1040, 2014. doi: 10.1177/0037549714540008.

[63] David Marlow, Susan M. Sanchez, and Paul J. Sanchez. Testing policies and key influences on long-term aircraft fleet management using designed simulation experiments. *Military Operations Research*, 24(3):5–26, 2019.

[64] Thomas O'Neal, Hokey Min, Daniel Cherobini, and Seong-Jong Joo. Benchmarking aircraft maintenance performances using data envelopment analysis. *International Journal of Quality & Reliability Management*, 38(6):1328–1341, 2021. doi: 10.1108/IJQRM-05-2020-0157.

[65] Mansik Hur, Seong-Jong Joo, and Jaeyoung Cho. Performance measure of maintenance practices for F-16 fighter jets by data envelopment analysis. *International Journal of Quality & Reliability Management*, 39(1):280–296, 2022. doi: 10.1108/IJQRM-08-2020-0272.

[66] Robert Meissner, Antonia Rahn, and Kai Wicke. Developing prescriptive maintenance strategies in the aviation industry based on a discrete-event simulation framework for post-prognostics decision making. *Reliability Engineering & System Safety*, 214: 107812, 2021. doi: 10.1016/j.ress.2021.107812.

[67] I. Tseremoglou and B. F. Santos. Condition-based maintenance scheduling of an aircraft fleet under partial observability: A deep reinforcement learning approach. *Reliability Engineering & System Safety*, 241:109582, 2024. doi: 10.1016/j.ress.2023.109582.

[68] Ron Wehrens and Johannes Kruisselbrink. Flexible Self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software*, 87(7), 2018. doi: 10.18637/jss.v087.i07. `https://www.jstatsoft.org/article/view/v087i07`.

[69] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, 3 edition, 2013. ISBN 9780429113079. doi: 10.1201/b16018.

[70] Stan Development Team. Stan modeling language users guide and reference manual, 2024. `https://mc-stan.org`.

[71] Kristin K. Nicodemus, James D. Malley, Carolin Strobl, and Andreas Ziegler. The Behaviour of Random Forest Permutation-based Variable Importance Measures under Predictor Correlation. *BMC Bioinformatics*, 11:110, 2010. doi: 10.1186/1471-2105-11-110. `http://www.biomedcentral.com/1471-2105/11/110`.

[72] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8(1):25, 2007. doi: 10.1186/1471-2105-8-25.

[73] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001. doi: 10.1111/1467-9868.00293.

[74] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, 2 edition, 1997. ISBN 978-3-540-62011-5. doi: 10.1007/978-3-642-56927-2.

[75] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234:34–35, 1971. doi: 10.1038/234034a0.

[76] Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2): 37–50, 1912. doi: 10.1111/j.1469-8137.1912.tb05611.x.

[77] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. doi: 10.1109/CVPR.2019.00075.

[78] Teuvo Kohonen. Essentials of the Self-organizing Map. *Neural Networks*, 37:52–65, 2013. doi: 10.1016/j.neunet.2012.09.018.

[79] Chotirat Ann Ratanamahatana and Eamonn Keogh. Making time-series classification more accurate using learned constraints. In *Proceedings of the SIAM International Conference on Data Mining*, pages 11–22, Philadelphia, PA, 2004. Society for Industrial

and Applied Mathematics. doi: 10.1137/1.9781611972740.2. https://proxying.lib.ncsu.edu/index.php?url=https://www.proquest.com/conference-papers-proceedings/making-time-series-classification-more-accurate/docview/940859561/se-2.

[80] Joe H. Ward, Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.

[81] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[82] A.D̃. Forbes. Classification-algorithm evaluation: five performance measures based on confusion matrices. *Journal of Clinical Monitoring*, 11(3):189–206, 1995. doi: 10.1007/BF01617722.

[83] Hans van der Hoef and Matthijs J. Warrens. Understanding information theoretic measures for comparing clusterings. *Behaviormetrika*, 46:353–370, 2019. doi: 10.1007/s41237-018-0075-7.

[84] Xin Liu, Hui-Min Cheng, and Zhong-Yuan Zhang. Evaluation of community detection methods. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1736–1746, 2020. doi: 10.1109/TKDE.2019.2911943.

[85] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

[86] Xin-Sheng Si, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. Remaining useful life estimation—a review on the statistical data-driven approaches. *European Journal of Operational Research*, 213(1):1–14, 2011. doi: 10.1016/j.ejor.2010.11.018.

[87] Joanna Z. Sikorska, Melinda Hodkiewicz, and Liang Ma. Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, 25(5):1803–1836, 2011. doi: 10.1016/j.ymssp.2010.11.018.

[88] Kamran Javed, Rafael Gouriveau, and Noureddine Zerhouni. State of the art and taxonomy of prognostics approaches, trends of prognostics applications and open issues towards maturity at different technology readiness levels. *Mechanical Systems and Signal Processing*, 94:214–236, 2017. doi: 10.1016/j.ymssp.2017.01.050.

[89] I. de Pater, A. Reijns, and M. Mitici. Alarm-based predictive maintenance scheduling for aircraft engines with imperfect remaining useful life prognostics. *Reliability Engineering & System Safety*, 221:108341, 2022. doi: 10.1016/j.ress.2022.108341.

[90] M. Memarzadeh and M. Pozzi. Value of information in sequential decision making: Component inspection, permanent monitoring, and system-level scheduling. *Reliability Engineering & System Safety*, 154:137–151, 2016. doi: 10.1016/j.ress.2016.05.014.

[91] W. Fauriat and E. Zio. Optimization of an aperiodic sequential inspection and condition-based maintenance policy driven by value of information. *Reliability Engineering & System Safety*, 204:107133, 2020. doi: 10.1016/j.ress.2020.107133.

[92] C. Song, C. Zhang, A. Shafieezadeh, and R. Xiao. Value of information analysis in non-stationary stochastic decision environments: A reliability-assisted POMDP approach. *Reliability Engineering & System Safety*, 217:108034, 2022. doi: 10.1016/j.ress.2021.108034.

[93] Nima Safaei and Andrew K. S. Jardine. Aircraft routing with generalized maintenance constraints. *Omega*, 80:111–122, 2018. ISSN 0305-0483. doi: 10.1016/j.omega.2017.08.013.

[94] Guesik Cha, Junseok Park, and Ilkyeong Moon. Military aircraft flight and maintenance planning model considering heterogeneous maintenance tasks. *Reliability Engineering & System Safety*, 239:109497, 2023. ISSN 0951-8320. doi: 10.1016/j.ress.2023.109497.

[95] Danny Parker and Andy Bellocchio. A modern decision-making framework for prognostic & predictive maintenance (PPMx). *Army Aviation Magazine*, January 2022.

[96] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988. doi: 10.2307/3214163.

[97] Donghyun Cho, Jooho Song, and Sungsoo Choi. Aircraft maintenance scheduling considering remaining useful life. *Computers & Industrial Engineering*, 109:179–191, 2017.

[98] Jonathan L. Paynter. *The Value of a Predicted Fault in Maintenance Planning*. PhD thesis, Massachusetts Institute of Technology, 2022.

[99] Antonios Kamariotis, Konstantinos Tatsis, Eleni Chatzi, Kai Goebel, and Daniel Straub. A metric for assessing and optimizing data-driven prognostic algorithms for predictive maintenance. *Reliability Engineering & System Safety*, 242:109723, 2024. doi: 10.1016/j.ress.2023.109723. Key contributions: Decision-oriented metric M comparing achieved cost to perfect-prognostics cost; explicit variation of RUL accuracy via virtual simulator; shows diminishing returns beyond moderate accuracy; C-MAPSS turbofan case study; single-component replacement context.

[100] J. Lee and M. Mitici. Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics. *Reliability Engineering & System Safety*, 230:108908, 2023. doi: 10.1016/j.ress.2022.108908.

[101] Yang Hu, Xuewen Miao, Jun Zhang, Jie Liu, and Ershun Pan. Reinforcement learning-driven maintenance strategy: A novel solution for long-term aircraft maintenance decision optimization. *Computers & Industrial Engineering*, 153:107056, 2021. doi: 10.1016/j.cie.2020.107056.

[102] Fuat Kosanoglu, Mahir Atmis, and Hasan Hüseyin Turan. A deep reinforcement learning assisted simulated annealing algorithm for a maintenance planning problem. *Annals of Operations Research*, 339:79–110, 2024. doi: 10.1007/s10479-022-04612-8.

[103] K. Vos, Z. Peng, E. Lee, and W. Wang. Aircraft fleet availability optimisation: A reinforcement learning approach. *The Aeronautical Journal*, 127(1318):2204–2218, 2023. doi: 10.1017/aer.2023.104.

[104] Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, 1961.

[105] Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

[106] Jan M. van Noortwijk. A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1):2–21, 2009. doi: 10.1016/j.ress.2007.03.019.

[107] Vytautas Blechertas, Abdel Bayoumi, Nicholas Goodman, Ronak Shah, and Yong-June Shin. CBM Fundamental Research at the University of South Carolina: A Systematic Approach to U.S. Army Rotorcraft CBM and the Resulting Tangible Benefits. In *Proceedings of the American Helicopter Society Technical Specialists' Meeting on Condition Based Maintenance*, Huntsville, AL, February 2009. American Helicopter Society International.

[108] Harold S. Balaban, Robert T. Brigantic, Samuel A. Wright, and Anthony F. Papatyi. A simulation approach to estimating aircraft mission capable rates for the United States Air Force. In J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, editors, *Proceedings of the 2000 Winter Simulation Conference*, pages 1035–1042, 2000. doi: 10.1109/wsc.2000.899908.

[109] J. R. Looker, V. Mak-Hau, and D. O. Marlow. Optimal policies for aircraft fleet management in the presence of unscheduled maintenance. In *Proceedings of the 22nd International Congress on Modelling and Simulation (MODSIM2017)*, pages 1392–1398, Hobart, Tasmania, Australia, 2017.

[110] Austin Semmel, Hans Sebastian Heese, Brandon McConnell, and Benjamin Rachunok. A framework for analyzing operational efficiency in U.S. Army aviation: Self-Organizing Map-based clustering of flight dispatch decisions. In *Proceedings of the Military Operations Research Society (MORS) Symposium*, Leesburg, VA, June 2025. Selected as Best in Working Group: Readiness.

[111] Rodrigo Coelho Barros, Márcio Porto Basgalupp, André C. P. L. F. de Carvalho, and Alex A. Freitas. A Survey of Evolutionary Algorithms for Decision-Tree Induction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(3):291–312, 2012. doi: 10.1109/tsmcc.2011.2157494.

[112] Francisco Herrera, Manuel Lozano, and José Luis Verdegay. Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis. *Artificial Intelligence Review*, 12:265–319, 1998. doi: 10.1023/a:1006504901164.

[113] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Natural Computing Series. Springer, Berlin, Heidelberg, 2 edition, 2015. ISBN 978-3-662-44873-1. doi: 10.1007/978-3-662-44874-8.

[114] Darrell Whitley, Soraya Rana, and Robert B. Heckendorn. The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. *Journal of Computing and Information Technology*, 7(1):33–47, 1999. `http://cit.fer.hr/index.php/CIT/article/view/2919`.

[115] James P Cohoon, Shailesh U Hegde, Worthy N Martin, and Dana Richards. Punctuated Equilibria: A Parallel Genetic Algorithm. In *Proceedings of the 2nd International Conference on Genetic Algorithms*, pages 148–154, 1987.

[116] Yiyuan Gong and Alex Fukunaga. Distributed Island-model Genetic Algorithms Using Heterogeneous Parameter Settings. In *IEEE Congress on Evolutionary Computation*, pages 820–827. IEEE, 2011. doi: 10.1109/cec.2011.5949703.

[117] Ágoston E Eiben, Robert Hinterding, and Zbigniew Michalewicz. Parameter Control in Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2): 124–141, 1999. doi: 10.1109/4235.771166.

[118] Erick Cantú-Paz. Migration Policies, Selection Pressure, and Parallel Evolutionary Algorithms. *Journal of Heuristics*, 7(4):311–334, 2001. doi: 10.1023/a:1011375326814.

[119] Larry J Eshelman and J David Schaffer. Real-coded Genetic Algorithms and Interval-schemata. In *Foundations of Genetic Algorithms*, volume 2, pages 187–202. Elsevier, 1993. doi: 10.1016/b978-0-08-094832-4.50018-0.

[120] Brianna L Greenstein, Danielle C Elsey, and Geoffrey R Hutchison. Determining Best Practices for Using Genetic Algorithms in Molecular Discovery. *The Journal of Chemical Physics*, 159(9):091501, 2023. doi: 10.1063/5.0158053.

[121] John W. Tukey. Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99–114, 1949. doi: 10.2307/3001913.

[122] Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, Hoboken, NJ, 9th edition, 2017. ISBN 978-1119113478.

[123] James P. Ignizio. The waddington effect, c4u-compliance, and subsequent impact on force readiness. *Phalanx*, 43(3):17–21, 2010. `http://www.jstor.org/stable/24910488`.

[124] Marilyn Strathern. "improving ratings": Audit in the British university system. *European Review*, 5(3):305–321, 1997. doi: 10.1002/(SICI)1234-981X(199707)5:3⟨305::AID-EURO184⟩3.0.CO;2-4.

[125] Eva Cantoni and Trevor Hastie. Degrees-of-Freedom Tests for Smoothing Splines. *Biometrika*, 89(2):251–263, 06 2002. doi: 10.1093/biomet/89.2.251.

[126] Simon N. Wood. mgcv: Mixed GAM computation vehicle with automatic smoothness estimation, 2023. R package version 1.8-40, `https://CRAN.R-project.org/package=mgcv`.

[127] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. `https://ggplot2.tidyverse.org`.

[128] Jacolien Van Rij, Martijn Wieling, R. Harald Baayen, and Hedderik Van Rijn. itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs, 2022. R package version 2.4.1.

[129] George E. P. Box and Norman R. Draper. *Empirical Model-Building and Response Surfaces*. John Wiley and Sons, New York, 1987. ISBN 9780471810339.

[130] Hans Mittelmann. Decision Tree for Optimization Software, 2024. `https://plato.asu.edu/sub/benchm.html`.

[131] Peter J. Huber. *Robust Statistics*. Wiley-Interscience, New York, 1981.

[132] Robert C. MacCallum, Shaobo Zhang, and Kristopher J. Preacher. On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods*, 7(1):19–40, 2002. doi: 10.1037/1082-989x.7.1.19.

[133] Rodney X. Sturdivant, David W. Hosmer, and Stanley Lemeshow. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, 3 edition, 2013.

[134] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297–318, 1986. doi: 10.1214/ss/1177013604.

[135] Teuvo Kohonen. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43:59–69, 1982. doi: 10.1007/bf00337288.

[136] Teuvo Kohonen. The Self-Organizing Map. *Proceedings of the IEEE*, 78(9):1464–1480, Sep 1990. doi: 10.1109/5.58325.

[137] Aiyshwariya Paulvannan Kanmani, Renee Obringer, Benjamin Rachunok, and Roshanak Nateghi. Assessing Global Environmental Sustainability Via an Unsupervised Clustering Framework. *Sustainability*, 12(2):563, 2020. doi: 10.3390/su12020563.

[138] Ron Wehrens and Lutgarde M.C. Buydens. Self- and Super-organizing Maps in R: The kohonen Package. *Journal of Statistical Software*, 21(5):1–19, 2007. doi: 10.18637/jss.v021.i05.

[139] Sheldon Lou, Jiong Jiang, and Kenneth Keng. Clustering Objects Generated by Linear Regression Models. *Journal of the American Statistical Association*, 88(424):1356–1362, 1993. doi: 10.2307/2291277.

[140] Geoffrey J. McLachlan and Kaye E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York, 1988. ISBN 0-8247-7691-7.

[141] Stuart P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/tit.1982.1056489.

[142] E. W. Forgy. Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications. *Biometrics*, 21:768–780, 1965.

[143] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. OPTICS: Ordering Points To Identify the Clustering Structure. *ACM SIGMOD Record*, 28: 49–60, 1999. doi: 10.1145/304181.304187.

[144] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231, Portland, OR, 1996. AAAI Press.

[145] Carl Edward Rasmussen. The Infinite Gaussian Mixture Model. In *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS)*, pages 554–560. MIT Press, 1999.

[146] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based Clustering for Social Networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007. doi: 10.1111/j.1467-985x.2007.00471.x.

[147] R. A. Hahn and Alexandra M. Newman. Scheduling United States Coast Guard helicopter deployment and maintenance at Clearwater Air Station, Florida. *Computers & Operations Research*, 35(6):1829–1843, 2008. doi: 10.1016/j.cor.2006.09.015.

[148] Heungseob Kim. Aircraft-to-mission assignment model for engine life management. *Military Operations Research*, 25(3):43–60, 2020. doi: 10.5711/1082598325343.

[149] Dong-Hoon Lee, Bong-Joo Jeong, Bong-Kyun Kim, Ju-Yong Lee, Sebong Jang, and Yeong-Dae Kim. Algorithms for aircraft flight and maintenance scheduling in an army aviation unit. *Military Operations Research*, 21(3):5–18, 2016.

[150] Mahrukh Iftikhar, Muhammad Shoaib, Ayesha Altaf, Faiza Iqbal, Santos Gracia Villar, and Imran Ashraf. A Deep Learning Approach to Optimize Remaining Useful Life Prediction for Li-ion Batteries. *Scientific Reports*, 14:25838, 2024. doi: 10.1038/s41598-024-77427-1.

[151] Guorong Ding, Wenbo Wang, and Jiaojiao Zhao. Prediction of Remaining Useful Life of Rolling Bearing Based on Fractal Dimension and Convolutional Neural Network. *Measurement and Control*, 55(1-2):79–93, 2022. doi: 10.1177/00202940211065674.

[152] Cheng Peng, Yufeng Chen, Weihua Gui, Zhaohui Tang, and Changyun Li. Remaining Useful Life Prognosis of Turbofan Engines Based on Deep Feature Extraction and Fusion. *Scientific Reports*, 12(1):6491, 2022. doi: 10.1038/s41598-022-10191-2.

[153] Dirk Thierens. Adaptive Mutation Rate Control Schemes in Genetic Algorithms. In *Proceedings of the 2002 Congress on Evolutionary Computation*, pages 980–985. IEEE, 2002. doi: 10.1109/cec.2002.1007058.

[154] Yaochu Jin and Jürgen Branke. Evolutionary Optimization in Uncertain Environments—A Survey. *IEEE Transactions on Evolutionary Computation*, 9(3): 303–317, 2005. doi: 10.1109/tevc.2005.846356.

APPENDICES

Appendix A

Acronyms, Systems, and Doctrine

This appendix consolidates the acronyms, systems of record, and doctrinal publications referenced throughout this dissertation.

A.1 Systems of Record

Table A.1: Relevant US Army systems of record

Acronym	Definition
ERDC	Engineer Research and Development Center
CAMMS	Consolidated Aviation Maintenance Management System
CAFR	Central Aviation Flight Records
LOGSA	Logistics Support Activity

A.2 Acronyms and Chapter References

Table A.2: Acronyms and chapter references

Acronym	Definition	Chapters
AH-64	Attack Helicopter (Apache)	1, 2
ANOVA	Analysis of Variance	4, D
ARB	Attack Reconnaissance Battalion	1
ASB	Aviation Support Battalion	1
BIC	Bayesian Information Criterion	4, D
CAB	Combat Aviation Brigade	1
CAFR	Central Aviation Flight Records	1
CAMMS	Consolidated Aviation Maintenance Management System	1
CBM	Condition-Based Maintenance	1, 4

Acronym	Definition	Chapters
CI	Confidence Interval	2
CTC	Combat Training Center	3, 4
CV	Coefficient of Variation	1, 2, 4, D
DADE	Department of the Army Directed Event	2
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	C
DEA	Data Envelopment Analysis	1, 3
DoW	Day of Week	3, 4
DTW	Dynamic Time Warping	1, 3
EDF	Effective Degrees of Freedom	2, B
ERDC	Engineer Research and Development Center	1, 2
FIELD	Field Maintenance	2
FHPA	Flying Hours per Aircraft	1–5
FI	Fixed-Interval (maintenance policy)	4, D
FMC	Fully Mission Capable	1–5
FMP	Flight and Maintenance Planning	3, 4
FY	Fiscal Year	1, 2
GA	Genetic Algorithm	4, D
GAM	Generalized Additive Model	1–3, B
GAO	Government Accountability Office	1–4
GINI	Gini Impurity (importance metric)	3
GLM	Generalized Linear Model	2
GMM	Gaussian Mixture Models	C
HDBSCAN	Hierarchical DBSCAN	C
HQDA	Headquarters, Department of the Army	1
HSD	Honest Significant Difference (Tukey's)	4, D
LDA	Latent Dirichlet Allocation	C
LOESS	Locally Estimated Scatterplot Smoothing	2, C
LOGSA	Logistics Support Activity	1
MAPE	Mean Absolute Percentage Error	4, D
MID	Minimum Improving Distance	3
MS	Mission Success	1, 4, 5
NMC	Not Mission Capable	1–5
NMCM	Non-Mission Capable: Maintenance	2, 4
NMCS	Non-Mission Capable: Supply	2, 4
NMI	Normalized Mutual Information	1, 3

Acronym	Definition	Chapters
OPTEMPO	Operational Tempo	1
OPTICS	Ordering Points To Identify Clustering Structure	C
OR	Operational Readiness	1–5, B
PCA	Principal Component Analysis	C
PdM	Predictive Maintenance	1, 2
PMC	Partially Mission Capable	2, 4
PMCM	Partially Mission Capable: Maintenance	2
PMCS	Partially Mission Capable: Supply	2
POMDP	Partially Observable Markov Decision Process	3, 4
PPMx	Prognostic and Predictive Maintenance	1, 4
RAF	Royal Air Force	1
RTL	Ready to Launch	1
RUL	Remaining Useful Life	1, 4, D
SME	Subject Matter Expert	4, D
SOM	Self-Organizing Map	1, 3, C
SUST	Sustainment/Depot-level Maintenance	2
TAP	Time Above Pareto	1, 3
USAF	United States Air Force	1
VOI	Value of Information	1, 4

A.3 Doctrine References

Table A.3: Doctrine references

Document	Purpose
AR 95-1	Flight Regulations
AR 220-1	Army Unit Status Reporting
AR 700-138	Army Logistics Readiness and Sustainability
ATP 3-04.7	Army Aviation Maintenance

Appendix B

Chapter 2 Supporting Tables and Figures

B.1 Background Data & Model Results

Table B.1: OR and Hours until Phase (Model Data)

	OR			
Hours until Phase	**High**	**Medium**	**Low**	**Total**
High	25,522	25,953	10,527	62,002
	(9.6%)	(9.8%)	(4.0%)	(23.4%)
Medium	65,069	55,881	22,286	143,236
	(24.5%)	(21.0%)	(8.4%)	(54.0%)
Low	29,698	22,407	8,129	60,234
	(11.2%)	(8.4%)	(3.1%)	(22.7%)
Total	120,289	104,241	40,942	265,472
	(45.3%)	(39.3%)	(15.4%)	(100%)

Table B.2: OR and Bank Hour Percentage (Battalion Aggregated Data)

	OR			
Bank Hour Category	**High**	**Medium**	**Low**	**Total**
High (> 65%)	964	1,577	1,129	3,670
	(3.2%)	(8.5%)	(6.1%)	(19.8%)
Medium (45% $\leq$ x $\leq$ 65%)	4,262	4,295	2,583	11,140
	(23.1%)	(23.2%)	(14.0%)	(60.3%)
Low (< 45%)	2,424	1,033	220	3,677
	(13.1%)	(5.6%)	(1.2%)	(19.9%)
Total	7,650	6,905	3,932	18,487
	(41.4%)	(37.3%)	(21.3%)	(100%)

Table B.3: Outlier Analysis by Battalion

Battalion	Outliers	Total Obs	Outlier Percentage
A	441	17,123	2.58%
B	363	13,002	2.79%
C	250	12,483	2.00%
D	364	14,494	2.51%
E	389	14,391	2.70%
F	6	11,930	0.05%
G	418	6,341	6.59%
H	257	16,851	1.53%
I	226	7,502	3.01%
J	199	17,079	1.17%
K	346	10,450	3.31%
L	306	10,768	2.84%
M	380	13,561	2.80%
N	401	14,984	2.68%
O	352	14,669	2.40%
P	317	24,648	1.29%
Q	333	16,680	2.00%
R	417	13,924	2.99%
S	366	12,392	2.95%

Outliers found via Cook's Distance

Table B.4: Spline Contribution and Odds Ratio Comparisons (Full Model D)

Measure	Value	Contribution	Odds Ratio	Difference from Max
OR				
OR	0.900	-0.651	0.343	-0.220
	0.750	-0.089	0.492	-0.071
	0.600	0.252	0.563	–
Hours until Phase				
Hours until Phase	10.000	-0.233	0.442	-0.088
	250.000	-0.022	0.494	-0.036
	400.000	0.119	0.530	–
Days to Report				
Days to Report	1.000	-0.144	0.465	-0.058
	15.000	-0.085	0.478	0.045
	25.000	0.091	0.523	–

Table B.5: GAM with Tensor Product Spline Model Output Summary

Term	**Estimate**	**Std. Error**	**p-Value**
(Intercept)	0.529	0.047	<0.001
Year 2020	-0.324	0.015	<0.001
Year 2021	-0.311	0.015	<0.001
Year 2022	-0.241	0.018	<0.001
February	0.106	0.016	<0.001
March	0.110	0.016	<0.001
April	0.098	0.016	<0.001
May	0.039	0.016	0.013
June	-0.028	0.018	0.113
July	-0.037	0.018	0.037
August	0.142	0.018	<0.001
September	0.135	0.018	<0.001
October	-0.023	0.017	0.173
November	-0.088	0.017	<0.001
December	-0.498	0.018	<0.001
Sunday	-1.614	0.014	<0.001
Monday	-0.243	0.012	<0.001
Tuesday	0.014	0.011	0.228
Thursday	-0.213	0.011	<0.001
Friday	-0.993	0.013	<0.001
Saturday	-1.628	0.014	<0.001
Approx. Sig. of Smooth Terms	**EDF**	**Ref. df**	**p-Value**
OR	3.688	3.750	<0.001
Hours until Phase	3.994	4.000	<0.001
Days Until Report	1.540	1.788	<0.001
Hours until Phase $\otimes$ OR	18.851	19.935	<0.001
Battalion	17.873	18.000	<0.001

Adjusted $R^2 = 0.113$; Explained Deviance = 9.81%; n = 265,472

B.2 Supplementary Material

Algorithm 3 Impute Aircraft Hours Until Phase Maintenance

1: **for** each aircraft j **do**
2: Calculate total observed flying hours over dataset duration
3: Estimate number of 500-hour phase maintenance cycles
4: Identify top five longest consecutive NMC days
5: **for** each NMC period **do**
6: Calculate flying hours leading up to the NMC period
7: Determine the probability of the NMC being a phase maintenance cycle, considering both flying hours and duration of downtime
8: **if** probability indicates likely phase maintenance **then**
9: Allocate as a phase maintenance cycle
10: Reset phase maintenance cycle to 500 hours
11: **end if**
12: **end for**
13: **for** each subsequent flight of aircraft j **do**
14: Decrement hours remaining in phase maintenance cycle based on flight duration
15: **end for**
16: **end for**

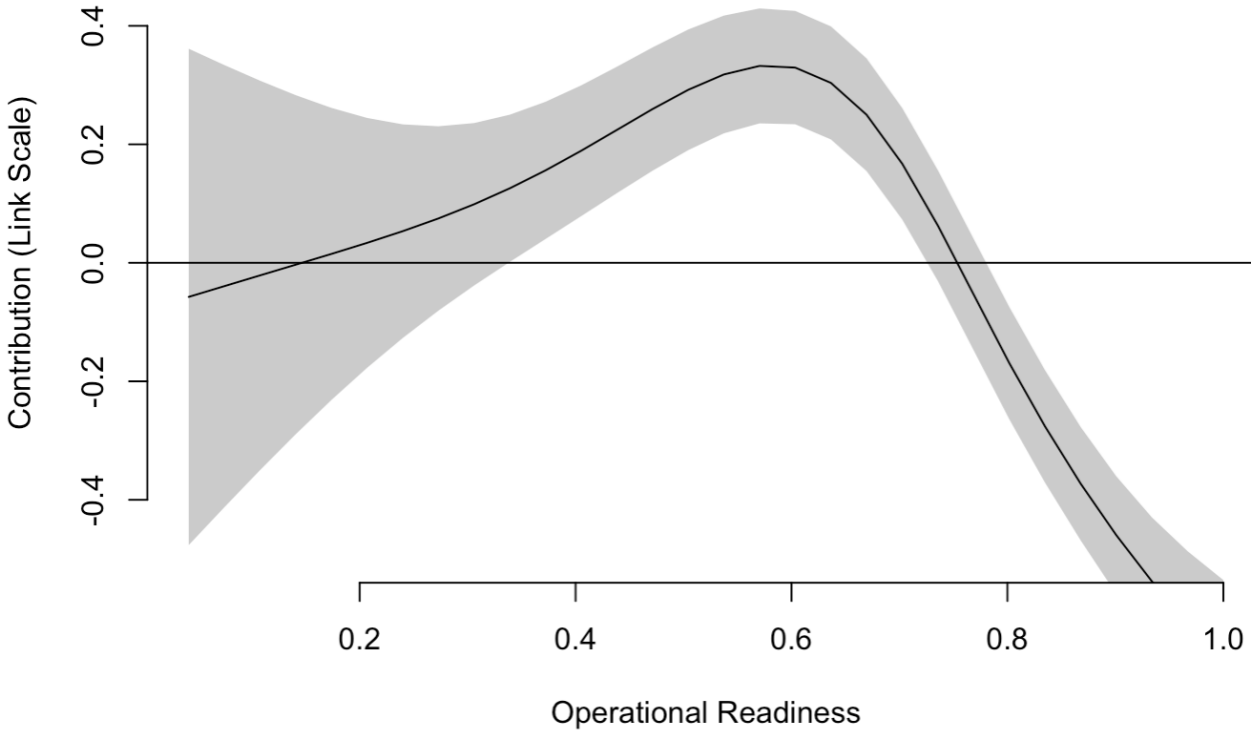

Figure B.1: Spline for OR (Tensor Model)

95% Credible Interval

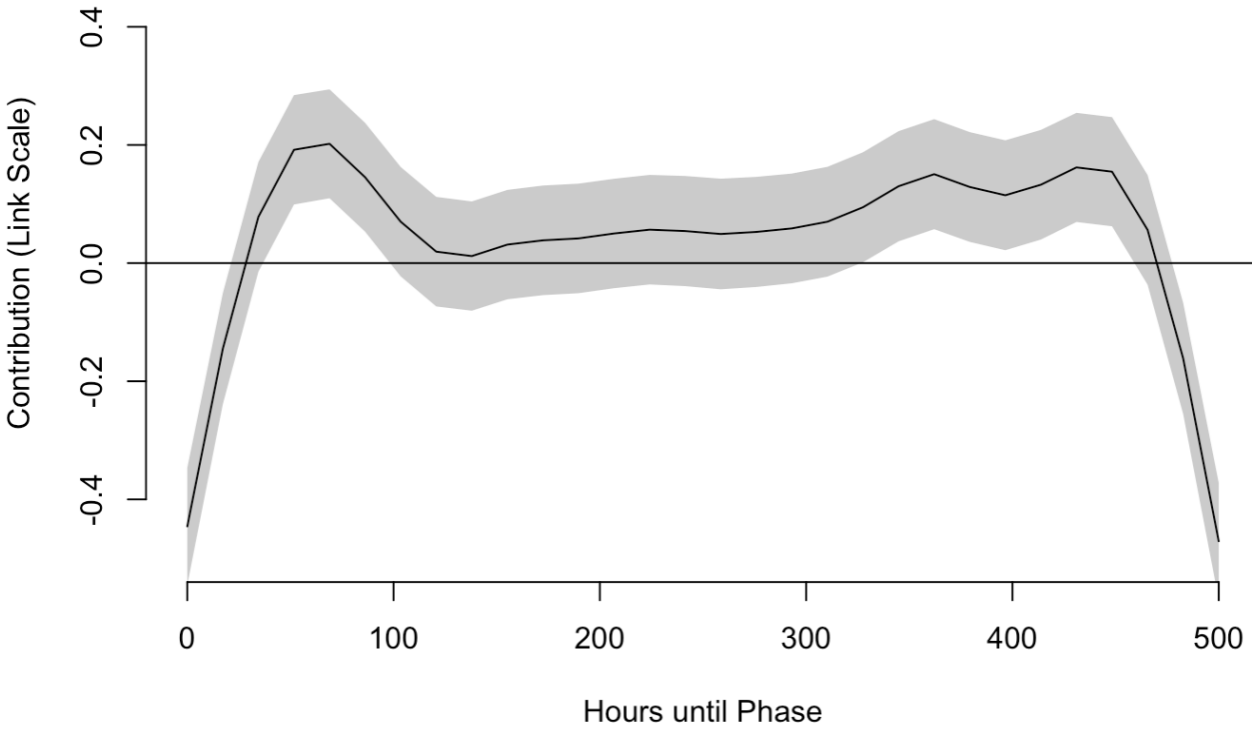

Figure B.2: Spline for Hours until Phase (Tensor Model)

95% Credible Interval

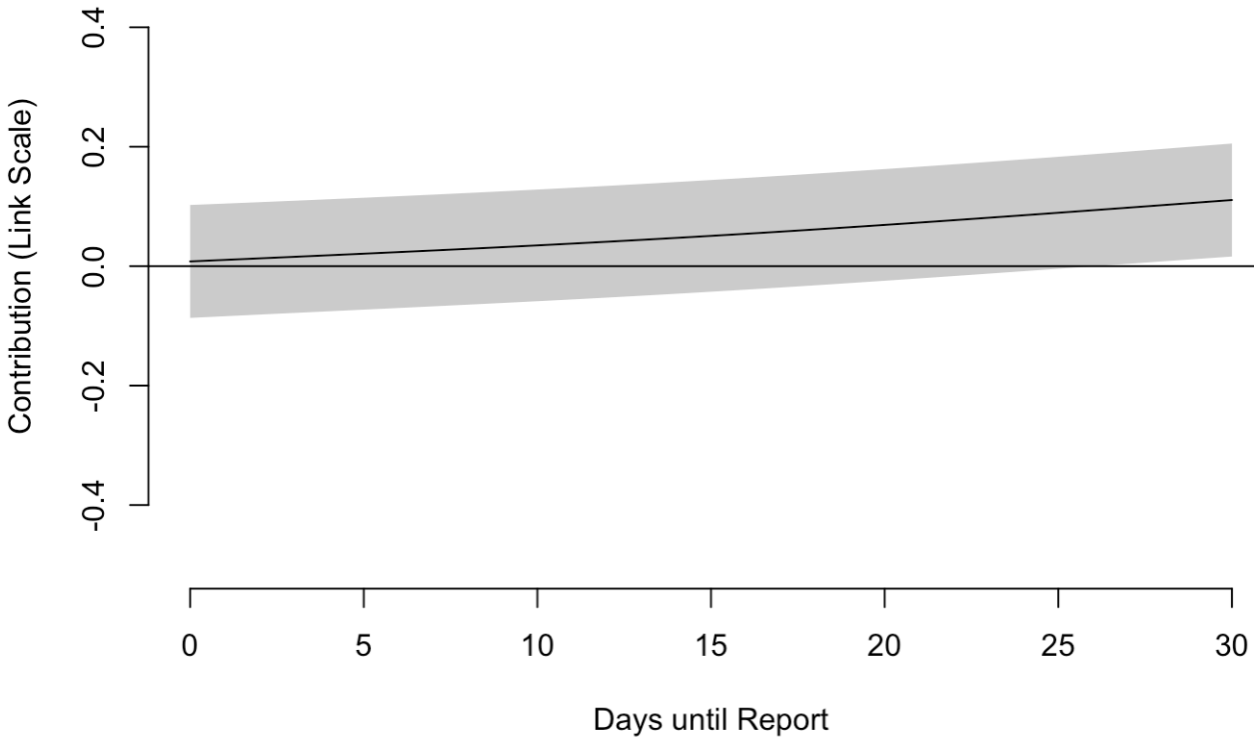

Figure B.3: Spline for Days until Report (Tensor Model)

95% Credible Interval

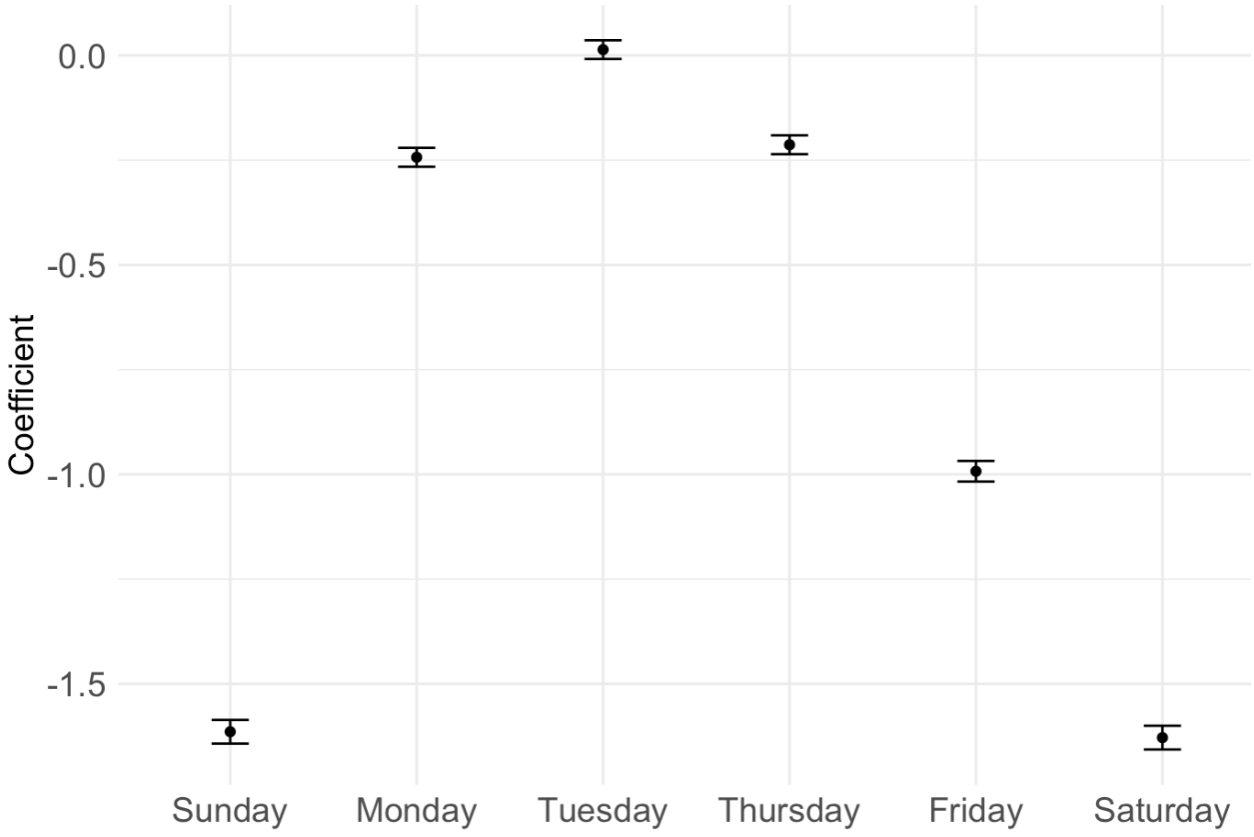

Figure B.4: Estimated Fixed Effect of Day of the Week (Tensor Model)

95% Confidence Interval

B.3 GAM Assumptions and Diagnostics

In modeling both GAMs, we make necessary model assumptions and conduct diagnostics to ensure their validity. Tests for multicollinearity among covariates indicated no significant issues. Table B.6 shows the Spearman and Pearson correlation coefficients for each two-way combination of covariates. Numbers approaching ± 1 indicate near-perfect collinearity in the data; we observe no values greater than 0.05, indicating a lack of linear relationship between the variables. *Day of the week* versus *Days until report* is intentionally omitted as it is completely determined by a calendar alone and provides no meaningful interpretation. The linearity of covariates and the impact of outliers are verified through diagnostic plots and Cook's distance, respectively, detailed in Table B.3 in Appendix B.1. We do not omit outliers at any point during the modeling stage.

Table B.6: Correlation Coefficients between Independent Variables

Variable 1	**Variable 2**	**Pearson**	**Spearman**
OR	Day of the Week	0.0004	0.0018
OR	Hours until Phase Maint.	-0.0484	-0.0477
OR	Days until Report	0.0119	0.0116
Hours until Phase Maint.	Day of the Week	0.004	0.0041
Hours until Phase Maint.	Days until Report	0.0041	0.0041

An additional test for concurvity, the non-linear analog to multicollinearity, indicates a lack of correlation between the smoothing spline terms. Table B.7 below illustrates how removing the Battalion random effect term eliminates concurvity concerns, as many units tend to exhibit distinct flying personalities (never allowing OR to drop below 75%).

Low values indicate a lack of non-linear relationships among the covariates. In certain units, low OR rarely (or never) occurs and can exhibit increased uncertainty in these outlier scenarios.

We select the optimal number of knots for each smoothing term by examining the BIC of each model over different k values according to standard methodologies from Cantoni and Hastie [125] and implemented in Wood [126]. We do so for each smoothing term individually and select $k = 4$ knots for OR; $k = 5$ knots for hours until phase; and $k = 5$ knots for day of the week. Next, we perform a k-Index test to verify if additional complexity is required. Results for this process are collectively found in Table B.8 and Figure B.5.

High concurvity here is to be expected, given that OR and hours until phase main-effect

Table B.7: Concurvity Measures for GAM Models with and without Battalion Random Effect (Equation 2.1)

Term	With Random Effect	Without Random Effect
Worst		
Hours until Phase	0.10	0.09
OR	0.87	0.09
Days until Report	0.02	0.01
Battalion	1.00	–
Observed		
Hours until Phase	0.03	0.01
OR	0.19	0.08
Days until Report	0.00	0.00
Battalion	0.02	–
Estimate		
Hours until Phase	0.08	0.06
OR	0.68	0.06
Days until Report	0.00	0.00
Battalion	0.07	–

Table B.8: k-Index Test for Knot Complexity in GAM Spline Terms (Equation 2.1)

Term	k'	edf	k-index	p-value
OR	3.0	3.0	0.89	0.55
Hours Until Phase	4.0	4.0	0.88	0.26
Days Until Report	4.0	4.0	0.88	0.17

splines are still included. Model interpretability and generalization suffer if they are removed, however. Further, as seen in Model 1, the presence of the random effect induces a high degree of concurvity.

All analyses were performed using `R` Statistical Software (v4.3.1; R Core Team 2023). Data visualization uses the `ggplot2` and `itsadug` `R` packages [127, 128], while statistical modeling and cross-validation were conducted with the `mgcv` `R` package [126].

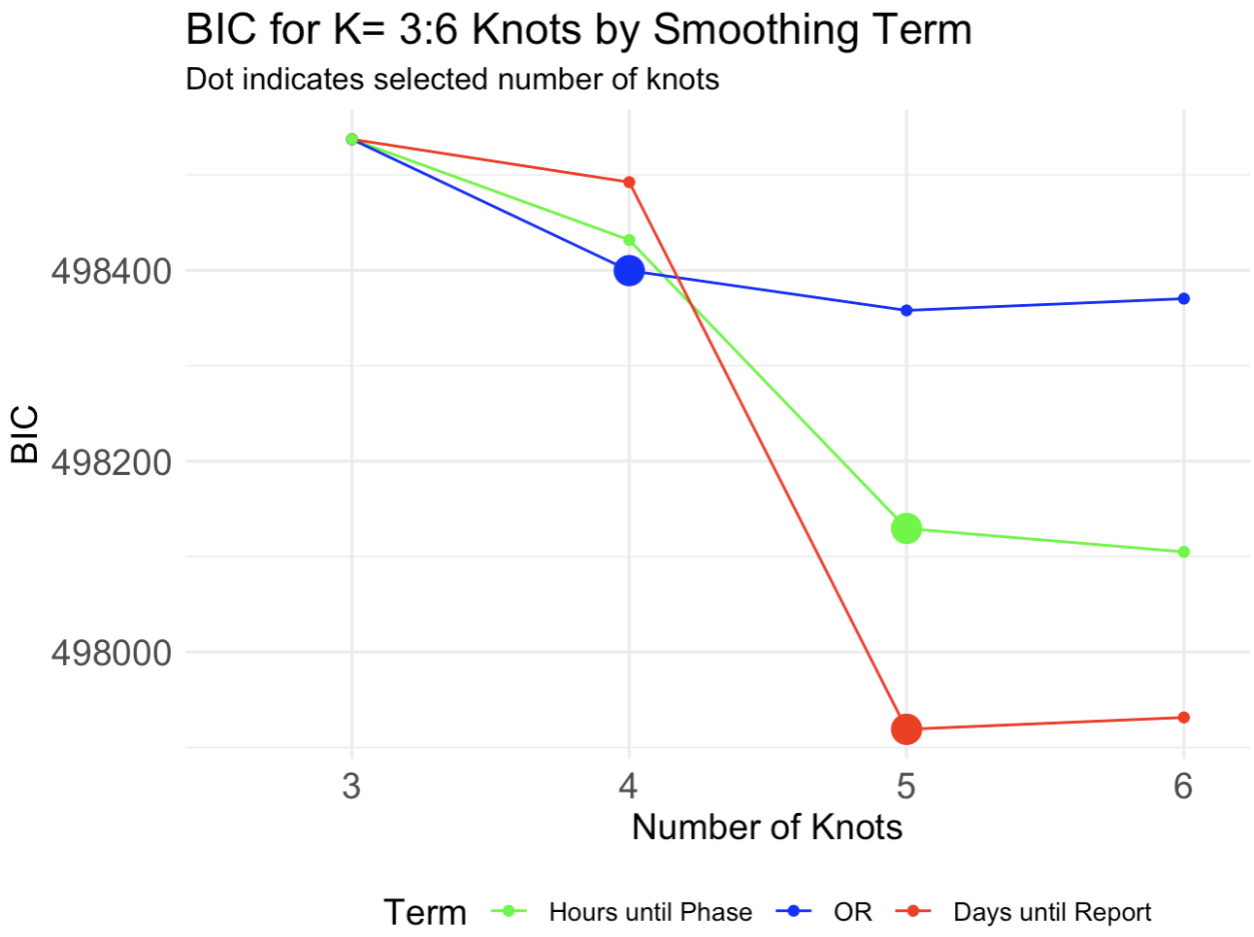

Figure B.5: Optimal number of knots chosen via elbow method using model Bayesian Information Criterion (Equation 2.1)

Table B.9: Concurvity Measures for Tensor Product GAM

Term	**Worst**	**Observed**	**Estimate**
Days Until Report	1.000	1.000	1.000
Hours Until Phase	0.005	0.004	0.004
OR	0.999	0.999	0.999
Hours Until Phase $\otimes$ OR	0.999	0.999	0.998
Battalion (Random Effect)	1.000	0.029	0.081

Appendix C

Chapter 3 Supporting Tables and Figures

C.1 Interaction Layer Constraint Formulation

This appendix provides the constrained optimization formulation for the interaction layer (Section 3.4.4). The interaction layer ($\ell = 5$) contains four coefficients representing products of two main-effect layers: OR ($\ell = 1$) and Hours until Phase ($\ell = 2$). Unlike independent layers where perturbations are unconstrained, the interaction layer requires that each coefficient equal the product of its corresponding main effects. This constraint ensures that recommended adjustments remain internally consistent. More specifically, we prevent contrary recommendations like increasing sensitivity to high-OR conditions while simultaneously decreasing the interaction between high OR and flying hours.

Main Effects and Interaction Structure. The two main-effect layers are:

$$\begin{aligned} \text{OR layer } (\ell = 1): &\quad \beta_{(1_m)} \text{ for } m \in \{1, 2\} \text{ (OR High, OR Low)} \\ \text{Hours layer } (\ell = 2): &\quad \beta_{(2_k)} \text{ for } k \in \{1, 2\} \text{ (Hours High, Hours Low)} \end{aligned}$$

The interaction layer ($\ell = 5$) has four coefficients, one for each combination of m and k. We denote these as $\beta_{(5_{m,k})}$, constrained to equal the product of the corresponding main effects:

$$\begin{aligned} \beta_{(5_{1,1})} &= \beta_{(1_1)} \cdot \beta_{(2_1)} \quad \text{(OR High} \times \text{Hours High)} \\ \beta_{(5_{1,2})} &= \beta_{(1_1)} \cdot \beta_{(2_2)} \quad \text{(OR High} \times \text{Hours Low)} \\ \beta_{(5_{2,1})} &= \beta_{(1_2)} \cdot \beta_{(2_1)} \quad \text{(OR Low} \times \text{Hours High)} \\ \beta_{(5_{2,2})} &= \beta_{(1_2)} \cdot \beta_{(2_2)} \quad \text{(OR Low} \times \text{Hours Low)} \end{aligned} \tag{C.1}$$

This multiplicative constraint is a modeling choice that ensures directional consistency. If a unit increases its responsiveness to both high OR and high hours-to-phase, the interaction term adjusts proportionally to reflect this combined change. While interaction coefficients in logistic regression are not inherently products of main effects, this functional form prevents

the optimization from recommending interaction adjustments that contradict the underlying main-effect perturbations.

SOM Prototype Notation. Let $\mathbf{p}^c_{(\ell)}$ and $\mathbf{p}^n_{(\ell)}$ denote the SOM prototype vectors for layer ℓ in clusters c (current) and n (neighbor), respectively. Clusters are indexed by efficiency ordering: Cluster 1 lies on the Pareto frontier and Cluster 6 is farthest from it. The decision boundary normal vector is $\mathbf{u}_{(\ell)} = \mathbf{p}^n_{(\ell)} - \mathbf{p}^c_{(\ell)}$.

Optimization Problem. The optimization finds the minimum perturbations $\Delta\beta_{(1_m)}$ and $\Delta\beta_{(2_k)}$ to the main-effect coefficients such that the perturbed interaction coefficients cross into a more efficient cluster. We introduce auxiliary variables to ensure no coefficient is simultaneously increased and decreased:

$$\delta^+_{(1_m)}, \delta^-_{(1_m)} \geq 0 \quad \text{for } m \in \{1, 2\} \tag{C.2}$$

$$\delta^+_{(2_k)}, \delta^-_{(2_k)} \geq 0 \quad \text{for } k \in \{1, 2\} \tag{C.3}$$

The actual adjustment is $\Delta\beta_{(1_m)} = \delta^+_{(1_m)} - \delta^-_{(1_m)}$ (and similarly for layer 2). Here δ^+ and δ^- represent the magnitudes of any increase or decrease, respectively, where their difference yields the signed perturbation:

Table C.1: Example perturbation outcomes from auxiliary variable decomposition.

δ^+	δ^-	$\Delta\beta = \delta^+ - \delta^-$	Effect
0.3	0	$+0.3$	Coefficient increases
0	0.5	-0.5	Coefficient decreases
0	0	0	No change

At most one of δ^+ or δ^- can be non-zero, which we enforce using a near-zero tolerance of $\epsilon = 10^{-6}$. This ensures commanders receive clear, unambiguous guidance. The complete

optimization problem minimizes the squared MID from Equation 3.4:

$$\begin{aligned}\min_{\boldsymbol{\delta}^+,\boldsymbol{\delta}^-}\ & \alpha_{(5)} \cdot \|\Delta\boldsymbol{\beta}_{(5)}\|^2 \\ &= \alpha_{(5)} \cdot \sum_{m=1}^{2}\sum_{k=1}^{2}\big(\Delta\beta_{(5_{m,k})}\big)^2 \end{aligned} \tag{C.4}$$

subject to:

$$\mathbf{u}_{(5)}^{\top}\big(\boldsymbol{\beta}_{(5)} + \Delta\boldsymbol{\beta}_{(5)}\big) - \tfrac{1}{2}\big(\|\mathbf{p}_{(5)}^{n}\|^2 - \|\mathbf{p}_{(5)}^{c}\|^2\big) \leq 0 \tag{C.5}$$

$$\beta_{(5_{m,k})} + \Delta\beta_{(5_{m,k})} = \big(\beta_{(1_m)} + \Delta\beta_{(1_m)}\big) \cdot \big(\beta_{(2_k)} + \Delta\beta_{(2_k)}\big) \quad \forall\, m,k \in \{1,2\} \tag{C.6}$$

$$\Delta\beta_{(1_m)} = \delta^{+}_{(1_m)} - \delta^{-}_{(1_m)} \quad m \in \{1,2\} \tag{C.7}$$

$$\Delta\beta_{(2_k)} = \delta^{+}_{(2_k)} - \delta^{-}_{(2_k)} \quad k \in \{1,2\} \tag{C.8}$$

$$\delta^{+}_{(1_m)} \cdot \delta^{-}_{(1_m)} \leq \epsilon \quad m \in \{1,2\} \tag{C.9}$$

$$\delta^{+}_{(2_k)} \cdot \delta^{-}_{(2_k)} \leq \epsilon \quad k \in \{1,2\} \tag{C.10}$$

$$\delta^{+}_{(1_m)}, \delta^{-}_{(1_m)}, \delta^{+}_{(2_k)}, \delta^{-}_{(2_k)} \geq 0 \tag{C.11}$$

Constraint C.5 ensures the perturbed coefficients cross the decision boundary into the more efficient cluster n. Constraint C.6 enforces that interaction coefficients equal the product of their corresponding main effects. Constraints C.7–C.11 implement mutual exclusivity. They prevent contradictory recommendations by ensuring each main-effect coefficient can only be adjusted in one direction.

C.2 Background Data & Model Results

SOM Codebook Values

Table C.2: Codebook Values of Days until Report Layer

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Days until Report	-1.083	0.015	1.603	-1.357	-0.619	0.449

Table C.3: Codebook Values of Hours until Phase Layer

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Hours until Phase (High)	-0.156	-1.614	0.244	1.860	-0.752	0.573
Hours until Phase (Low)	0.154	-1.065	-0.286	2.082	-0.843	0.667

Table C.4: Codebook Values of OR Layer

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
OR (High)	0.187	3.447	-0.660	-0.788	-0.467	0.225
OR (Low)	-0.026	-1.195	0.143	1.544	-0.840	0.599

Table C.5: Codebook Values of Hours until Phase * OR Interaction Layer

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Interaction: Hours until Phase (High) * OR (High)	0.445	-2.149	0.555	-1.123	0.363	-0.278
Interaction: Hours until Phase (Low) * OR (High)	-0.734	-1.059	0.718	-0.449	0.482	-0.215
Interaction: Hours until Phase (High) * OR (Low)	-0.685	0.891	0.477	-1.200	0.583	-0.391
Interaction: Hours until Phase (Low) * OR (Low)	0.163	0.320	0.258	-0.102	-0.464	0.182

Table C.6: Codebook Values of Day of the Week Layer

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Sunday	-0.996	2.721	0.576	-0.631	0.149	-0.299
Monday	-1.424	0.971	0.155	1.889	-0.178	0.268
Tuesday	-1.328	0.195	0.366	0.377	-0.188	0.535
Thursday	-0.565	1.912	0.120	-1.657	-0.520	0.614
Friday	-0.460	3.174	0.288	-0.249	-0.243	-0.183
Saturday	-0.941	2.657	0.628	0.057	-0.110	-0.244

Table C.7: Summary statistics for the 500 posterior samples of the beta coefficients.

Variable	Mean	Standard Error
Intercept	-0.114	0.030
Fiscal Year: 2020	0.211	0.003
2021	0.260	0.005
2022	0.373	0.004
February	0.079	0.004
March	0.107	0.004
April	0.060	0.004
May	0.026	0.004
June	0.009	0.004
July	-0.013	0.005
August	0.165	0.006
September	0.137	0.006
October	-0.022	0.004
November	-0.100	0.003
December	-0.478	0.003
Hours until Phase (High)	0.074	0.002
Hours until Phase (Low)	0.009	0.003
Days until Report	0.003	0.000
OR (High)	-0.327	0.004
OR (Low)	0.231	0.003
Sunday	-1.785	0.005
Monday	-0.272	0.001
Tuesday	0.004	0.001
Thursday	-0.239	0.002
Friday	-1.044	0.002
Saturday	-1.800	0.006
Interaction: Hours until Phase (High) OR (High)	-0.024	0.004
Interaction: Hours until Phase (Low) OR (High)	-0.053	0.004
Interaction: Hours until Phase (High) OR (Low)	0.019	0.004
Interaction: Hours until Phase (Low) OR (Low)	-0.585	0.027

Table C.8: Distances Between Clusters Based on SOM Mapping

From Cluster	To Cluster	Distance
1	2	7.65
1	3	7.70
1	4	8.96
1	5	7.54
1	6	9.93
2	3	3.50
2	4	3.63
2	5	2.77
2	6	4.29
3	4	3.32
3	5	3.12
3	6	5.88
4	5	4.81
4	6	5.51
5	6	5.86

C.3 Supplementary Figures

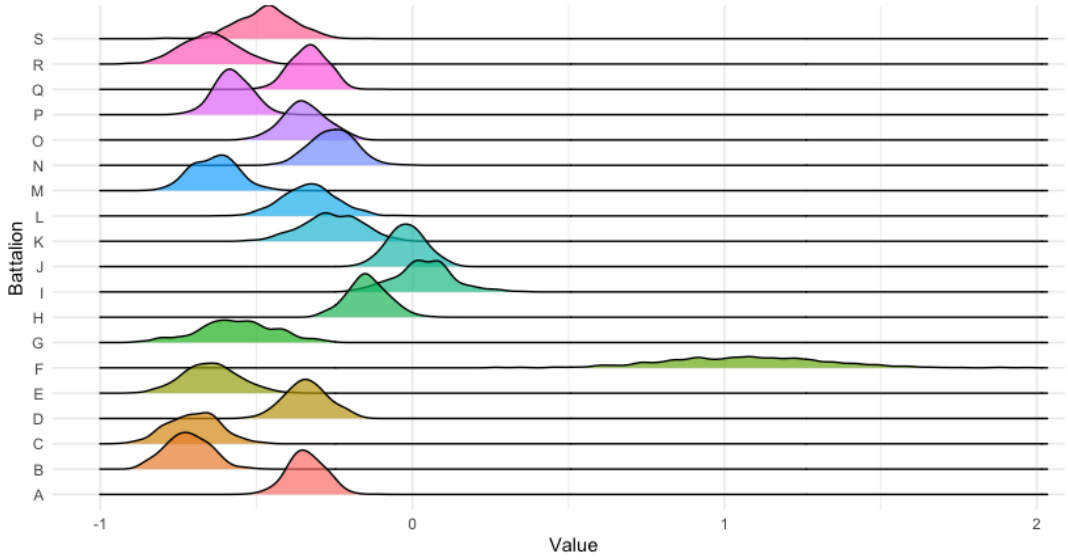

Figure C.1: Density Ridge Plot: OR (High)

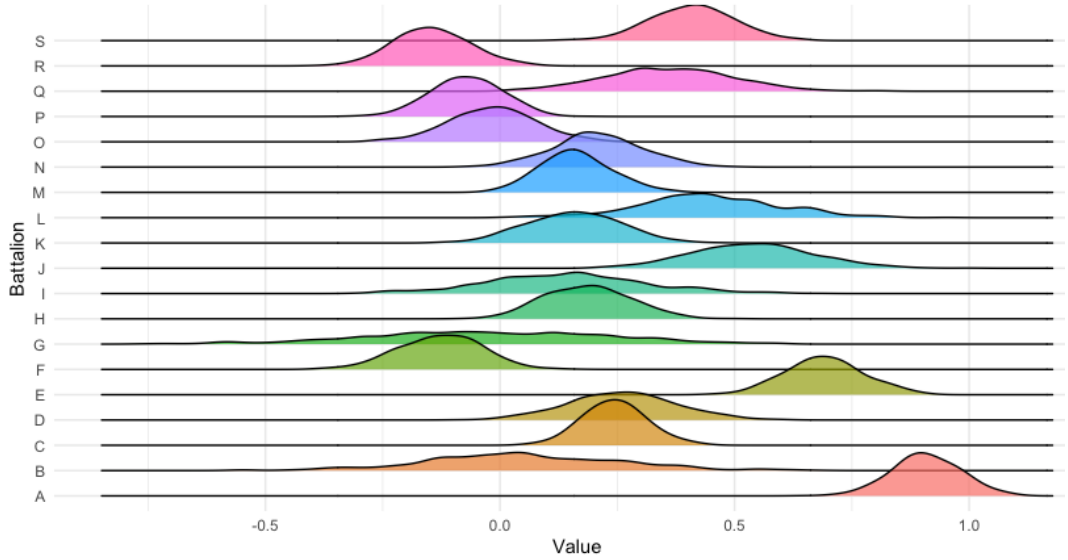

Figure C.2: Density Ridge Plot: OR (Low)

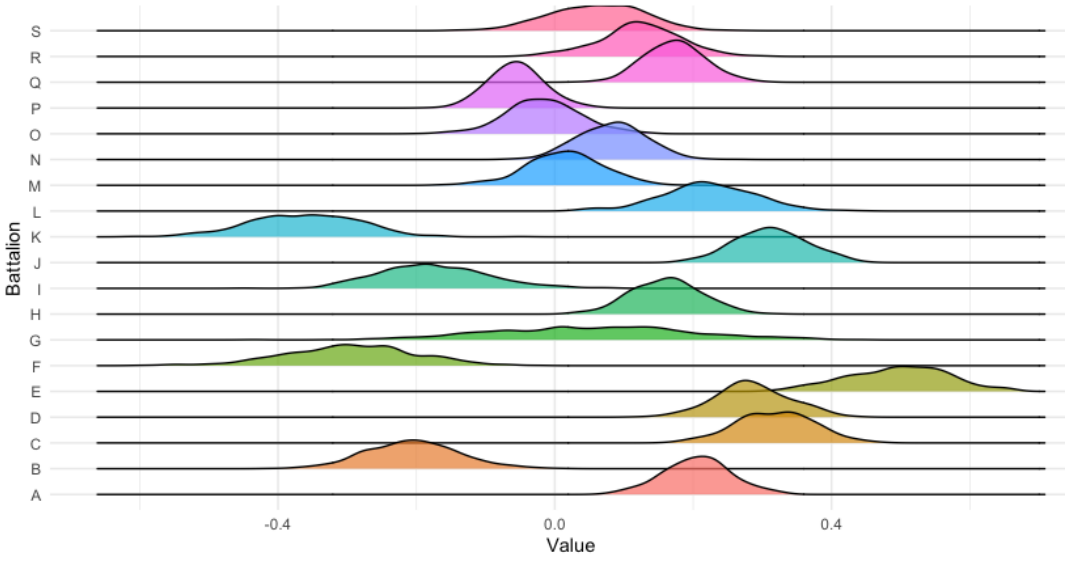

Figure C.3: Density Ridge Plot: Hours (High)

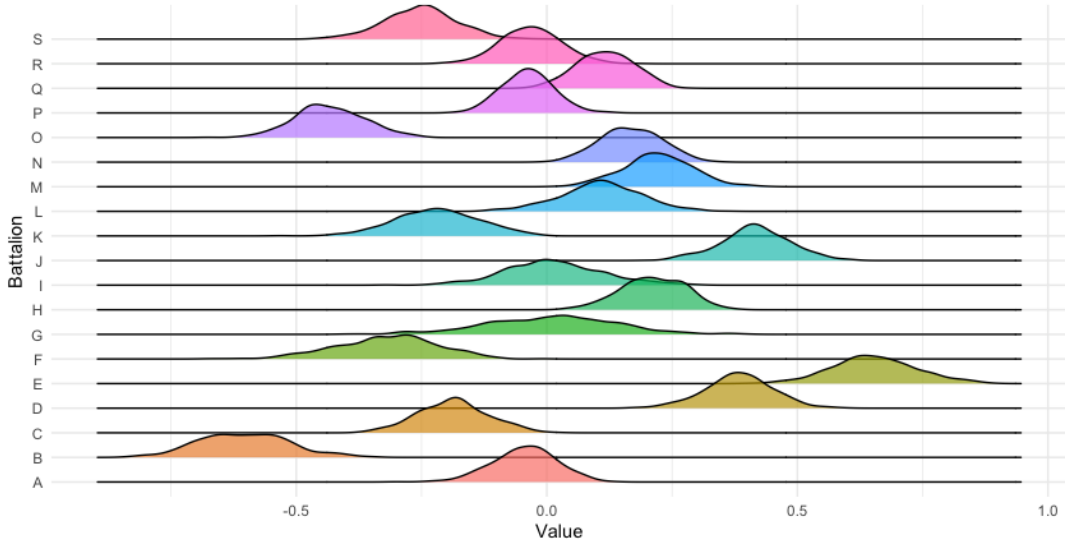

Figure C.4: Density Ridge Plot: Hours (Low)

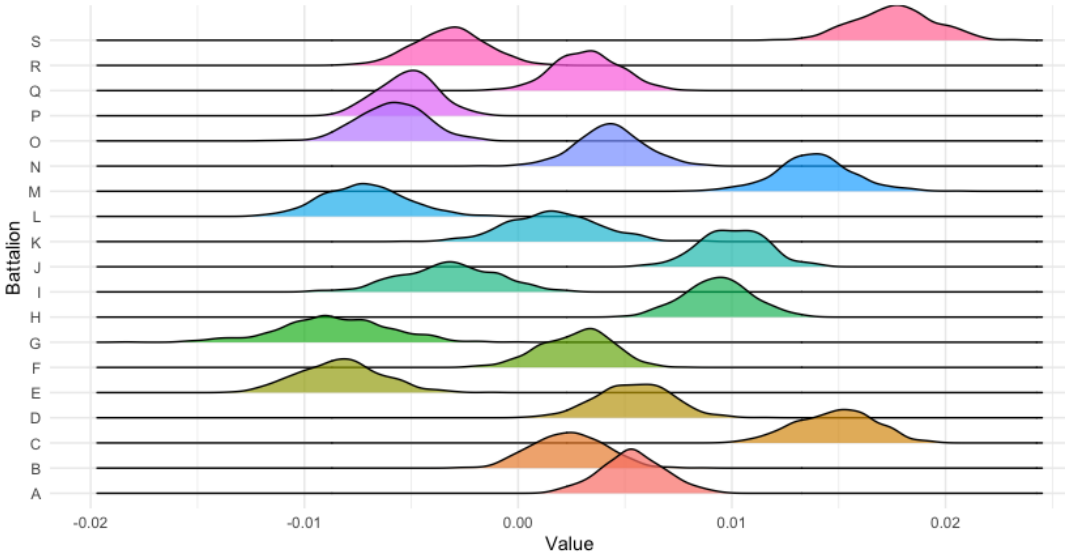

Figure C.5: Density Ridge Plot: Days Remaining in Reporting Period

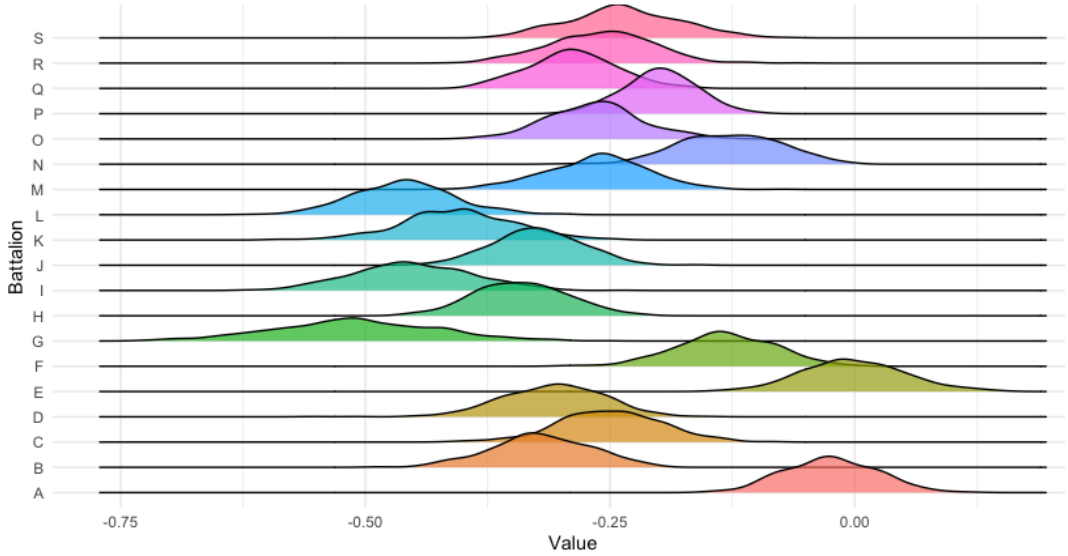

Figure C.6: Density Ridge Plot: Monday

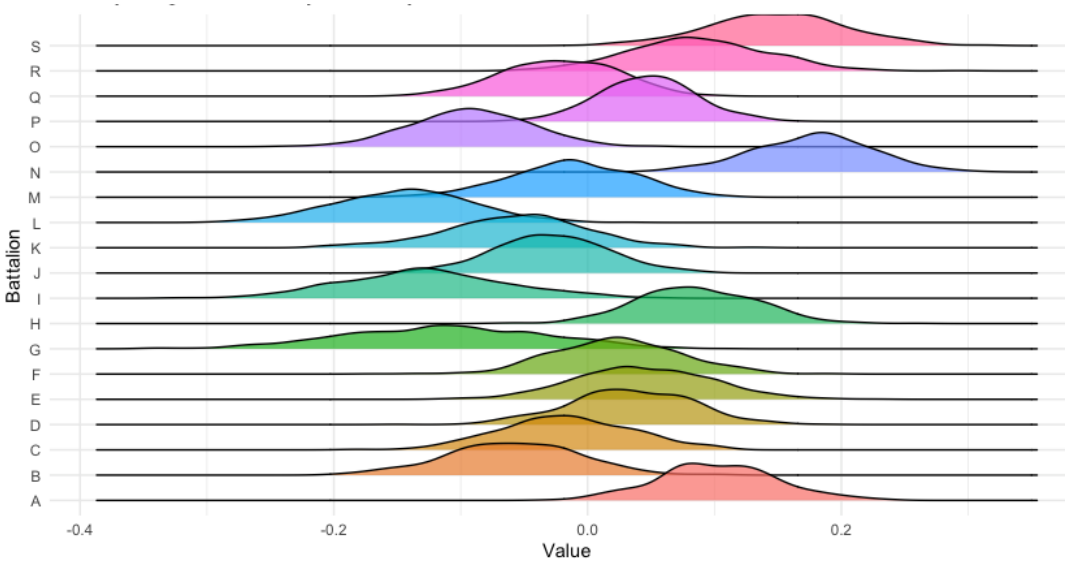

Figure C.7: Density Ridge Plot: Tuesday

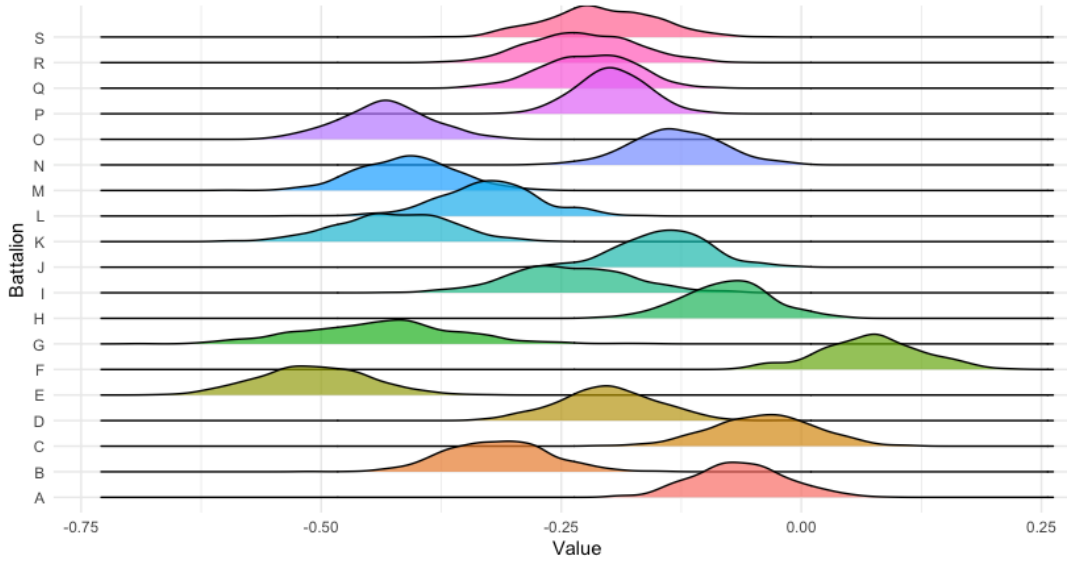

Figure C.8: Density Ridge Plot: Thursday

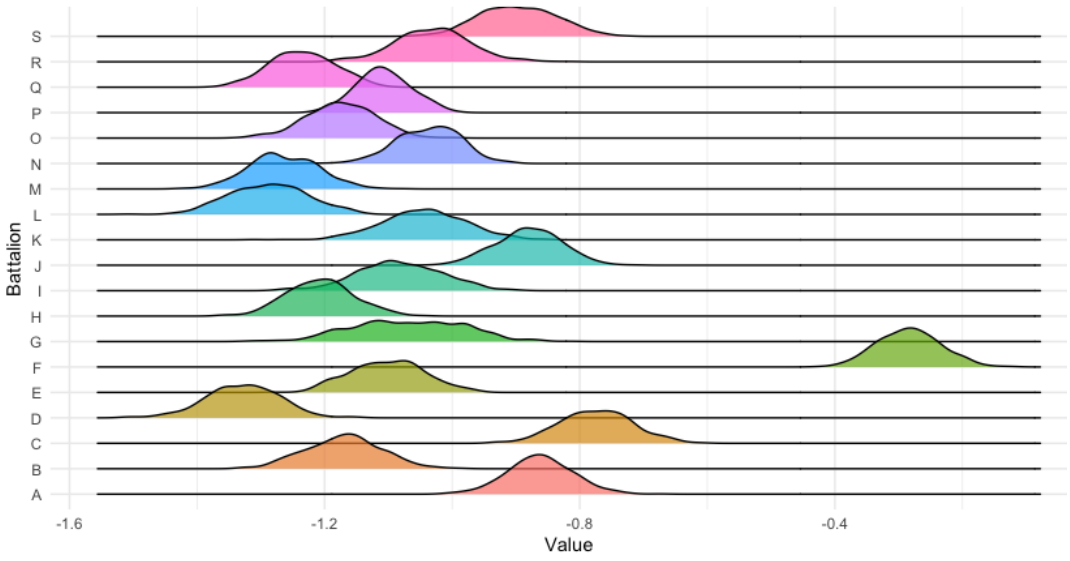

Figure C.9: Density Ridge Plot: Friday

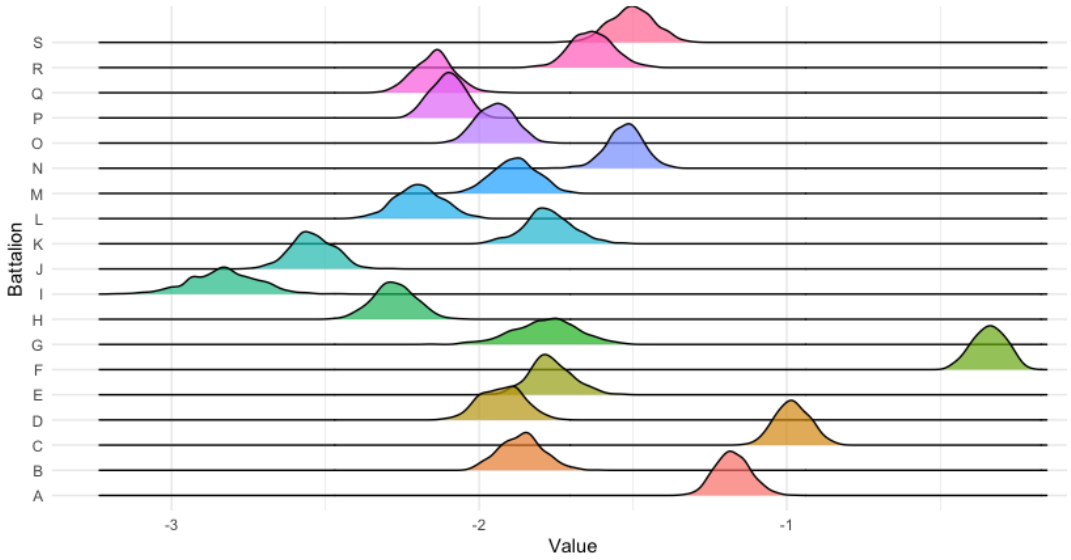

Figure C.10: Density Ridge Plot: Saturday

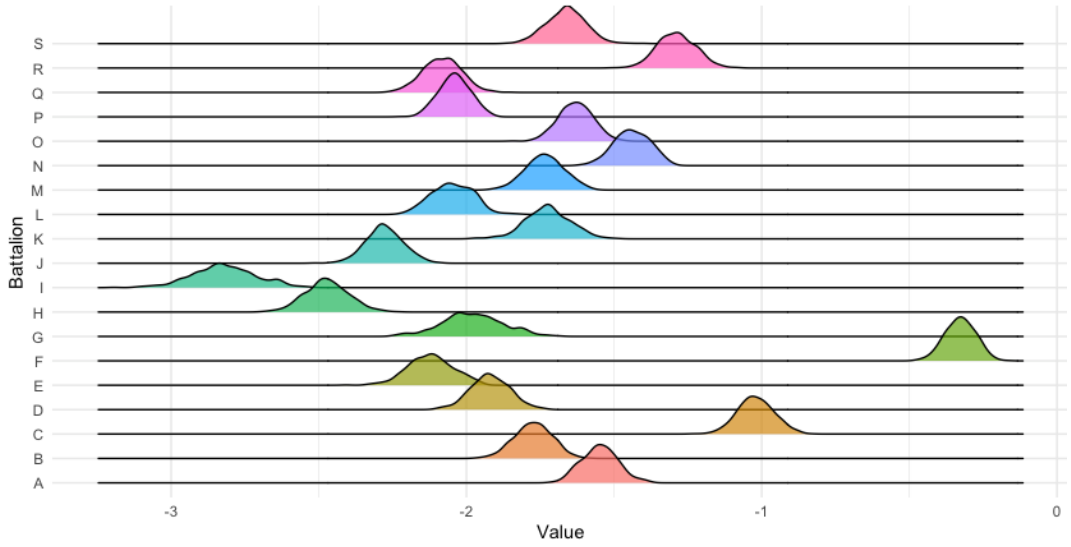

Figure C.11: Density Ridge Plot: Sunday

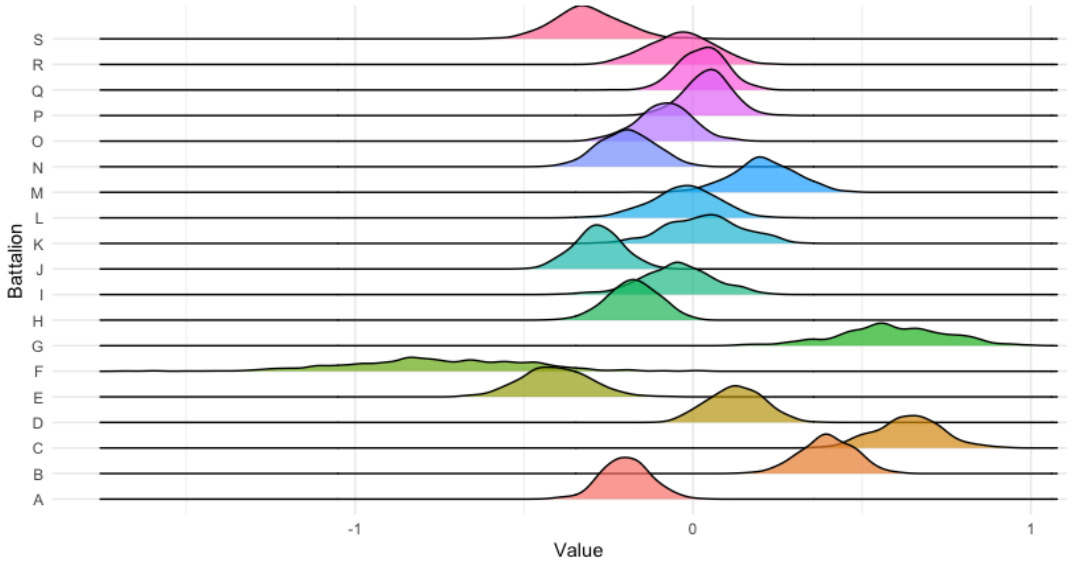

Figure C.12: Density Ridge Plot: OR (High) * Hours (High)

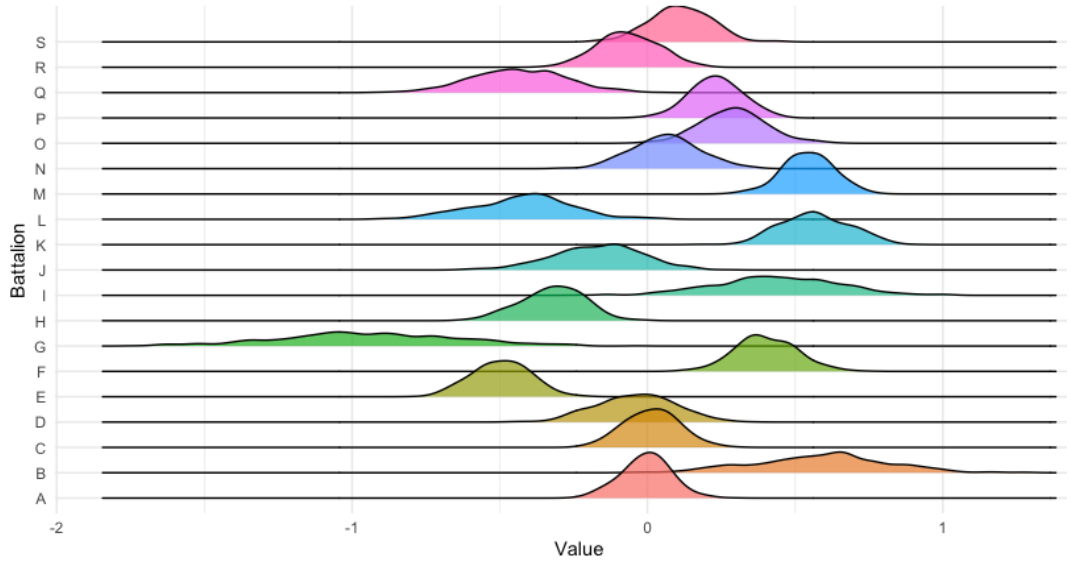

Figure C.13: Density Ridge Plot: OR (High) * Hours (Low)

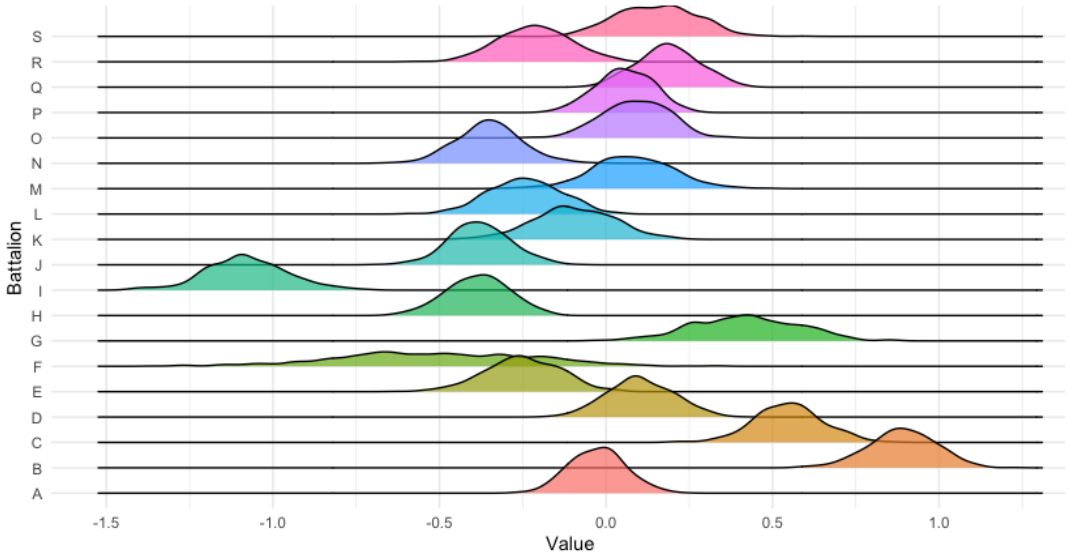

Figure C.14: Density Ridge Plot: OR (Low) * Hours (High)

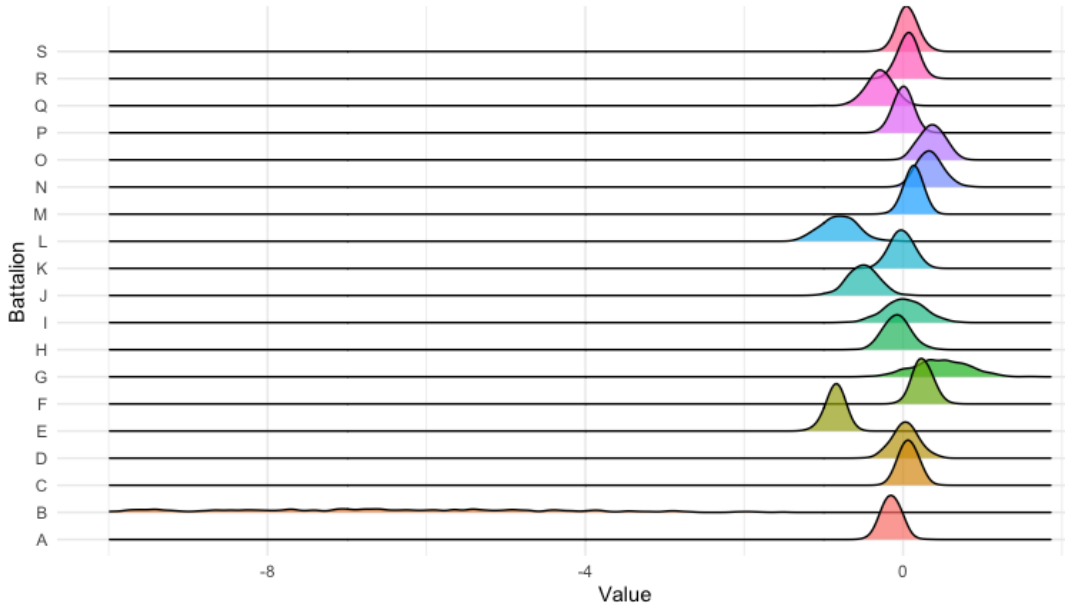

Figure C.15: Density Ridge Plot: OR (Low) * Hours (Low)

C.4 A Note on Assessing Robustness to Discretization Thresholds in US Army Apache Helicopter Flight Data through a Generalized Additive Model

"All models are wrong, but some are useful" –George E.P. Box [129]

The adage from George E.P. Box continues to resonate decades later and offers a cautionary tale that is increasingly relevant in today's high-performance computing environment. With continued advancements in computing power, statistical modeling has become more accessible and allows resolution of large-scale optimization problems in mere seconds [130]. However, this ease of computation introduces new challenges, particularly in the interpretation of model parameters, often one of the most overlooked aspects of statistical analysis. Our goal to construct robust models requires a deep understanding of their sensitivity to slight variations in the data and its parameterization. The principles of robust statistics capture the objectives we aim for in model development, as outlined explicitly by Huber:

1. It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.

2. It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly, that is, the latter (described, say, in terms of the asymptotic variance of an estimate, or of the level and power of a test) should be close to the nominal value calculated at the model.

3. Somewhat larger deviations from the model should not cause a catastrophe.[131]

When performing classification on operational data, we may be interested in a range of values of a predictor variable for which the response remains classified in the same manner. For example, if it is 90 degrees Fahrenheit outside, one might predict that a local pool will be open, and the prediction is *robust* to a change of $\pm$ 10 degrees. In the context of US Army Apache helicopter data, we desire a model to predict a single aircraft's likelihood of usage on a given day as a function of various relevant operational data points. Definitions and the support of the model's predictors are quickly summarized:

- **Operational Readiness:** 0 – 100%; A measure of a unit's aircraft availability, reported monthly to a unit's higher headquarters. The standard for Army aviation equipment is 75% [12].

- **Hours until Phase Maintenance:** 0 – 500 hours; AH-64 Apache helicopters are required to undergo major phase maintenance following every 500^{th} flight hour. These maintenance periods span days to weeks and, by doctrine, should be completed in no more than 44 days.
- **Days Remaining in the Reporting Period:** 0 – 31 days; the reporting period closes on the 15^{th} of each month.
- **Day of the Week:** Sunday – Saturday; a factor variable treated with Wednesday as the reference value.

The response is binary–whether a specific aircraft flew on that day or not. N = 265,472 observations of capable aircraft that fly on 16.7% of days. The goal of this paper is to determine the cutoff points for the discretization of the predictor *hours until phase maintenance* at which a logistic regression model remains robust.

Discretization and Robustness Indeed, MacCallum et al. argue that the discretization (what they refer to as *dichotomization*) of random variables into bins must be avoided at all costs unless justified by additional context [132]. The operation has the potential to skew results by over-generalizing a situation. However, if a model's assumptions are violated, discretization may be warranted. For example, in logistic regression, we assume that the relationship between the log-odds and each covariate is linear in parameters [133]. A quick inspection of the bivariate relationship between the hours until phase maintenance and flying in Figure C.16 shows this assumption is indeed violated. The log-odds of flight increase between 500–405 hours; decrease slightly between 405 and 90; and decrease at a more rapid rate when under 90 hours. Discretizing hours until phase maintenance into high / medium / and low categories may provide more interpretable and actionable results. This naturally begs the question: *how do we choose the cutoff points for the discretization of hours until phase maintenance?* We could select based on observed quantiles, the inflection points of 405 and 90 hours, ± 1 standard deviation from the mean (251.4 ± 154.4), or another heuristic of choice. Whichever choice is made, the model should be robust to some given range of cutoff values. More simply: we want a robust model that produces similar results regardless of our choice of truncation threshold.

Generalized Additive Model A widely accepted method for creating a model that includes a continuous predictor that exhibits a non-linear relationship with the response is to fit a Generalized Additive Model (GAM) [134]. GAMs assume that the relationship between the covariates and the response can be captured by additive smooth functions (splines), and

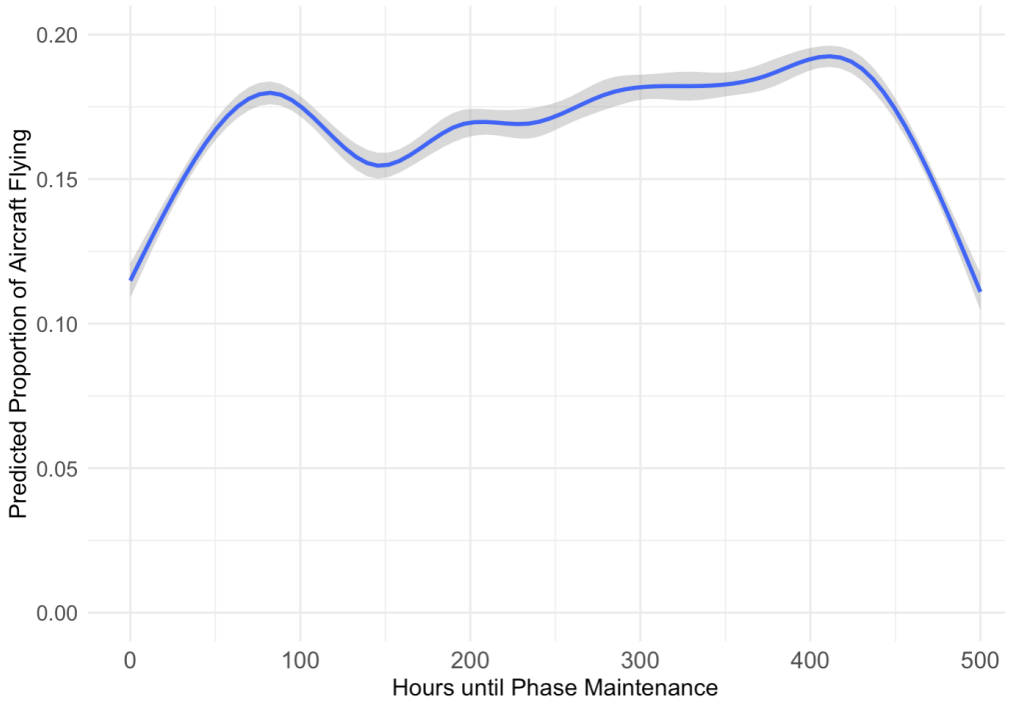

Figure C.16: Flying vs OR: fitted GLM via LOESS (95% CI)

that the residuals are independently and identically distributed with a mean of zero. Rather than a linear relationship with the response, covariates in a GAM are fit with a regression spline over various intervals of the feature space where the number of inflection points, or knots, is chosen via cross-validation. We fit a cubic spline according to Equation C.12 using the **mgcv** package in R and perform cross-validation to obtain the optimal number of knots, $k = 3$, via AUC–ROC [126].

$$\begin{aligned} \log\left(\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right) &= \beta_0 + \beta_1 \times \text{Year}_i + \beta_2 \times \text{Month}_i \\ &+ f_1 \times \text{Operational Readiness}_i \\ &+ f_2 \times \text{Hours until phase maintenance}_{ij} \\ &+ f_3 \times \text{Days remaining in reporting period}_i \\ &+ \beta_3 \times \text{Day of the week}_i + \varepsilon_{ij} \end{aligned} \tag{C.12}$$

The smooth functions f_1, f_2, and f_3 for continuous predictors are estimated non-parametrically and each can be expressed as a sum of basis functions:

$$f_i(x) = \sum_{k=1}^{K_i} \alpha_{ik} b_{ik}(x) \tag{C.13}$$

where $b_{ik}(x)$ are the basis functions, α_{ik} are the estimated coefficients for the i-th smooth function, and K_i is the number of basis functions used for the i-th smooth term. The β coefficients represent the fixed effects in the logistic regression model; the α coefficients are associated with the non-linear effects captured by the smooth functions; and ε_{ij} represents the random error or unexplained variation in the data for aircraft j on day i. The splines are fit by minimizing a penalized likelihood criterion where the penalty is determined by the smoothing parameter, optimized via generalized cross-validation. Credible intervals for the fitted splines are derived from the estimated covariance matrix of the coefficients. These intervals reflect the uncertainty around the spline estimates under a Gaussian distribution for the coefficients [31].

We test whether the effective degrees of freedom (approximately k-1 for each predictor) allow for adequate flexibility in each spline using standard diagnostic tests for the number of basis functions. We find that $k = 3$ is adequate for *hours until phase maintenance* and *days remaining in the reporting period* but that *operational readiness* would benefit from the addition of another knot. Allowing $k = 4$ for the operational readiness spline yields an improved fit according to Table C.9.

Table C.9: GAM fit for $K = 3$ vs $K = 4$ knots for operational readiness

Term	**K**	**K-Index**	**p-Value**
Hours until Phase Maint.	3	0.97	0.69
Operational Readiness	3	0.93	**0.015**
Days until Report	3	0.97	0.755

Term	**K**	**k-index**	**p-value**
Hours until Phase Maint.	3	0.96	0.23
Operational Readiness	4	0.95	0.18
Days until Report	3	0.97	0.52

Note: **Bold** font indicates statistical significance at the $\alpha = 0.05$ level.

Model Results Overall, we achieve a maximum F1 score of 0.38 after performing 100 iterations of cross-validation using an 80/20 training and testing splits on each prediction threshold from 0.05 to 0.5 according to Figure C.18. This is analogous to weighting flying days at 80%. The high count of non-flying days compels model weighting to achieve any predicted flying days. All regression splines are statistically significant at the $\alpha = 0.05$ level. A likelihood ratio test comparing the full model against a null model indicates a significant improvement in model fit $\chi^2 = 17,779$, $p < 2.2 \times 10^{-16}$. This result provides statistical evidence that the included predictors and smooth terms significantly enhance the model's ability to explain the variability in daily aircraft usage.

Table C.10: Summary of Model Fit

Parametric coefficients				
Term	Estimate	Std. Error	Z-Value	p-Value
(Intercept)	-0.906	0.031	-29.24	<0.001
Year: 2020	-0.313	0.023	-13.55	<0.001
Year: 2021	-0.269	0.023	-11.47	<0.001
Year: 2022	-0.233	0.028	-8.23	<0.001
February	0.086	0.025	3.41	0.001
March	0.113	0.024	4.64	<0.001
April	0.113	0.025	4.57	<0.001
May	0.057	0.025	2.29	0.022
June	-0.013	0.028	-0.44	0.659
July	-0.016	0.029	-0.57	0.568
August	0.152	0.028	5.51	<0.001
September	0.137	0.028	4.93	<0.001
October	-0.009	0.027	-0.34	0.733
November	-0.078	0.027	-2.90	0.004
December	-0.491	0.028	-17.22	<0.001
Sunday	-1.569	0.025	-63.57	0.001
Monday	-0.240	0.018	-13.66	<0.001
Tuesday	0.013	0.017	0.75	0.453
Thursday	-0.204	0.017	-11.78	<0.001
Friday	-0.968	0.020	-47.73	<0.001
Saturday	-1.584	0.025	-63.95	<0.001
Approximate significance of smooth terms				
Term	Effective degrees of freedom	Ref.df	Chi.sq	p-value
Days until Report Spline	1	1	27.87	<0.001
Hours until Phase Spline	1.998	2	254.68	<0.001
Operational Readiness Spline	2.993	3	2398.74	<0.001

Adjusted $R^2 = 0.0612$. Deviance explained = 7.43%.

UBRE = -0.0128. Scale est. = 1. n = 265,472.

Hours until Phase Maintenance Cutoff Thresholds Given the adequate fit of the model, statistical significance of each of the main covariates, and ultimate goal of finding discretization points for the *hours until phase maintenance*, we now examine the regression splines individually. We are able to isolate the effect of the hours until phase maintenance on the decision to fly a capable aircraft in Figure C.17. As expected, we observe a quadratic relationship between hours until phase maintenance and the likelihood of flying a given aircraft. Aircraft are less likely to fly in the early stages after leaving phase maintenance or when approaching phase maintenance. We find the empirical roots of the quadratic at 404.5 and 104.7 hours with associated empirical roots of the credible intervals of (396.7 – 412.9) and (96.6 – 112.3) hours, respectively. As such, choices within the given intervals will reasonably provide a robust estimate of the actual manner in which practitioners manage their aircraft in terms of hours until phase maintenance. Therefore, we propose establishing generalized thresholds at the economically interpretable values of 400 and 100 hours until phase maintenance. These thresholds delineate high, medium, and low stages within each maintenance cycle.

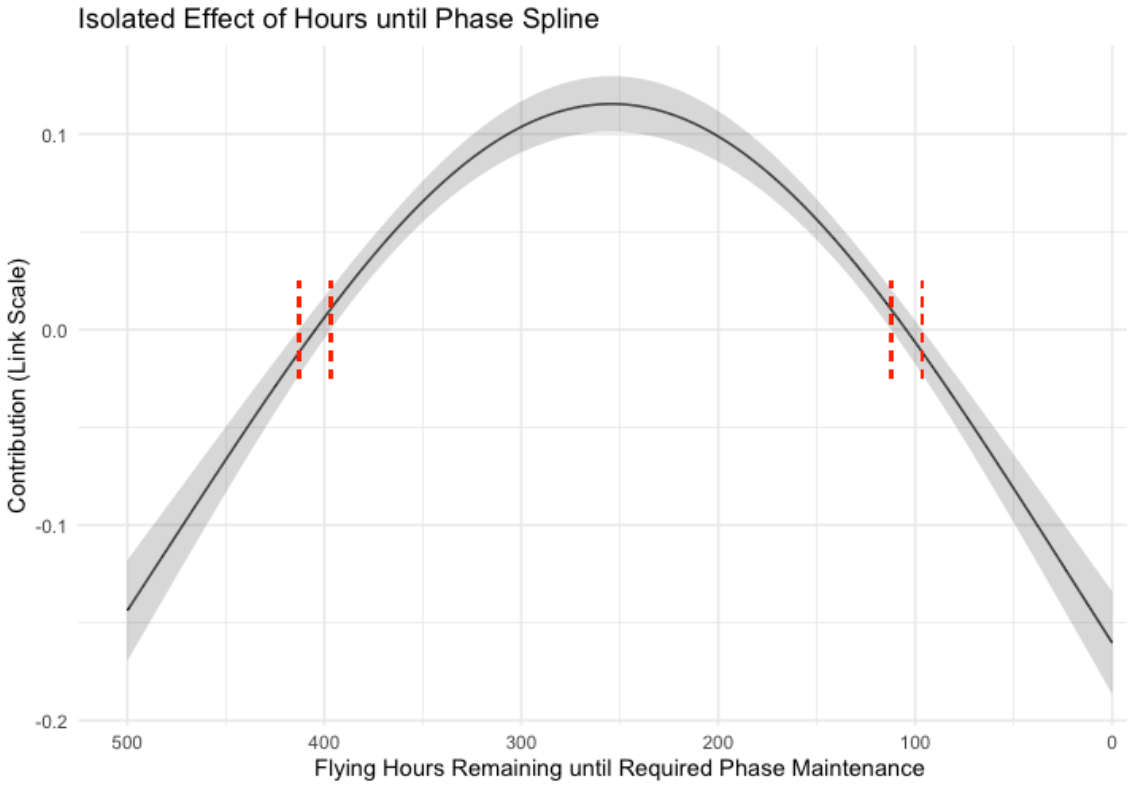

Figure C.17: Regression Spline for Hours until Phase (95% Credible Interval)
Empirical roots are found at 404.5 (396.7 – 412.9) and 104.7 (96.6 – 112.3) hours

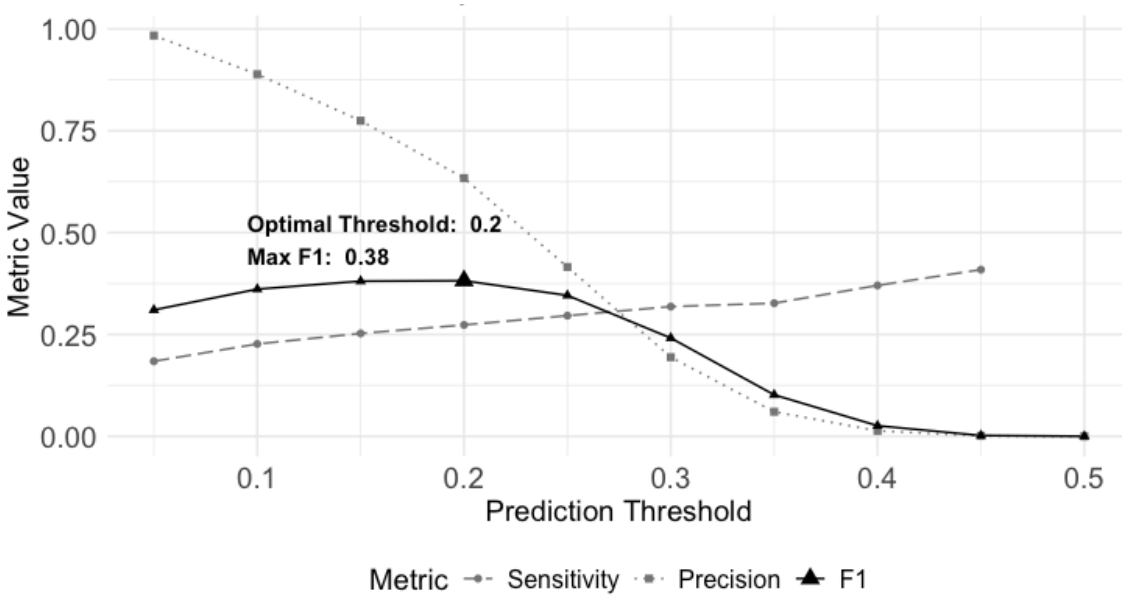

Figure C.18: Model performance metrics across prediction acceptance thresholds using 100 iterations of 5-fold cross-validation.

Conclusion In this section, we have determined robust cutoff thresholds for the discretization of hours until phase maintenance to refine the predictive accuracy of aircraft maintenance schedules. Establishing proposed thresholds at 400 and 100 hours, this model offers practitioners and decision-makers on the ground a meaningful manner in which to frame decisions surrounding their maintenance cycles and their associated impact on aircraft usage.

C.5 Background on Latent Pattern Detection Models

The primary objective of this section is to provide a concise background on models used for model-based latent pattern detection, including our methodology of choice—the self-organizing map—and to demonstrate their application in detecting latent patterns within an unsupervised learning context. SOMs, developed by Teuvo Kohonen, are a type of artificial neural network that perform a discrete dimensionality reduction (typically, to nodes on a 2-D map) while preserving the topological properties of the identified clusters [135]. This preservation means that the neighboring clusters on the map are more similar to each other than those farther apart, providing a significant advantage over other unsupervised clustering techniques for certain use cases [68]. Specifically, for latent pattern detection, this property allows for more meaningful interpretations and facilitates sensitivity analysis [136]. By maintaining an ordered structure of the data, SOMs enable comparisons between clusters and may offer actionable "pathways" for commanders seeking to improve their unit's positioning on the Pareto frontier [137, p. 1].

In SOMs, clustering is performed by mapping rows of input data to a "Best Matching Unit" (neuron) that minimizes the distance to a codebook vector, which functions as a prototypical observation seen in that node [68, p. 2]. This distance is often referred to as the *quantization error* [78, p. 53]. The algorithm uses either online or batch updating to change all nodes' codebook vectors within some distance as defined by a neighborhood function to become more similar to the new data [68]. This neighborhood function converges over time and performs exactly like k-means clustering when the number of affected nodes by an update becomes one [138, p. 3]. More information on Self-Organizing Maps can be found in Kohonen's works (Kohonen [78, 135, 136]) and textbook [74] along with a practical walkthrough of the R package *kohonen* written by Wehrens and Kruisselbrink [68].

We seek to perform unsupervised clustering rather than induce a label to different units as defined by their position or quadrant on the Pareto frontier and use a supervised method such as K-Nearest Neighbors to perform classification, which can fail when presented with multiple overlapping distributions. This technique aims to recognize latent flight decision-making profiles as characterized by a linearized model that predicts flight decisions [74, p. 26]. Specifically, we wish to determine if flight decision-making patterns correspond to different regions on the Pareto frontier itself. This form of clustering on regression output falls under the more general category of model-based clustering, which Lou et al. popularized in 1993, as a problem of clustering data based on its underlying data generation process. Model-based latent pattern detection and dimensionality reduction are well-studied concepts. However, many dimensionality reduction techniques, such as principal component analysis (PCA) and

Latent Dirichlet Allocation (LDA) fail to "incorporate information on how objects should be compared," [138, p. 1].

Common unsupervised clustering methods include mixture models, k-means, and density-based approaches like DBSCAN, HDBSCAN, and OPTICS [140–144]. While highly accurate, these methods prioritize clustering performance over interpretability of distributions within clusters. In contrast, SOMs preserve topology by keeping similar high-dimensional objects proximate in the two-dimensional plane, which enables more interpretable post hoc analyses [138]. Additionally, infinite GMMs and OPTICS produce inconsistent numbers of clusters across trials, complicating consensus groupings and stability analysis [143, 145].

Although the OPTICS algorithm achieved remarkable performance in our context (clustering 19 battalions into 19 distinct clusters with 99.8% accuracy when tuned by silhouette score), this precision suggested overfitting rather than the identification of broader similarity patterns. HDBSCAN offered better control over cluster numbers, but even promising configurations (e.g., six clusters with 300-point minimum) produced unproductive divisions—four clusters containing only outliers with remaining observations in one large cluster. Given these limitations, we ultimately selected SOMs for our clustering approach.

Handcock et al. [146] introduced a two-stage Bayesian model for unsupervised clustering of social media usage, while Kanmani et al. [137] clustered countries' sustainability profiles to recommend practical improvements for environmental performance. Both frameworks inspired our two-stage approach combining Bayesian logistic regression with SOM clustering. Applied to military aviation data, this methodology offers units actionable pathways to improve operational efficiency by utilizing the SOM's inherent ability to maintain meaningful relationships between clusters.

C.6 Supplemental Results

Table C.11: Minimum perturbations by unit to induce an improving change in clustering (only the layer associated with the minimum change is shown. Units in Clusters 1 or 2 not shown.). Values are presented on the original beta coefficient scale after rescaling from the normalized scale used during clustering.

Unit	Original Cluster	Improved Cluster	Focus Layer	Adjustment
B	3	2	OR	High OR (+0.0434) Low OR (+0.0906)
C	5	1	Hours until Phase	High Hours (-1.0275) Low Hours (-0.4311)
E	6	3	Hours until Phase	High Hours (+0.0132) Low Hours (+0.0147)
G	4	1	Hours until Phase	High Hours (-0.7836) Low Hours (-0.6551)
I	4	1	Hours until Phase	High Hours (-0.6509) Low Hours (-0.5442)
K	3	2	Day of Week	Sun (-0.0006) Mon (+0.0007) Tue (+0.0011) Thu (+0.0016) Fri (<0.0001) Sat (-0.0002)
L	4	1	Hours until Phase	High Hours (-0.9309) Low Hours (-0.7782)
M	5	1	Hours until Phase	High Hours (-0.9143) Low Hours (-0.3836)
O	3	2	OR	High OR (<0.0001) Low OR (<0.0001)
P	3	2	Hours until Phase	High Hours (-0.0371) Low Hours (-0.0422)
R	3	2	Interaction	High Hours, High OR (-0.0450) Low Hours, High OR (+0.2357) High Hours, Low OR (+0.0489) Low Hours, Low OR (-0.0588)
S	5	1	Hours until Phase	High Hours (-0.7852) Low Hours (-0.3294)

The clusters identify operational patterns among units, with each cluster representing a unique approach to navigating the OR-FHPA tradeoff. To translate our decision support framework into practical recommendations, we present a case study examining three units (B, E, and R) across all stages of our analysis. Table 3.4 provides the operational context for these units, while Figure 3.5 illustrates the minimum adjustments needed at both unit and variable levels. We leave it to the commander to determine which pathway best suits their

Table C.12: Minimum Cumulative Perturbation by Layer for Units Improving from **Cluster 3 to 2**. Values denote the L2 norm of each layer's perturbation vector on the original coefficient scale with layer-specific optimization weights applied.

Unit	OR	Day of Week	Days Until Report	OR × Hours Interaction	Hours Until Phase
B	0.007	0.016	<0.001	0.546	0.014
K	0.014	<0.001	<0.001	0.028	0.008
O	<0.001	<0.001	<0.001	0.020	0.006
P	0.011	0.052	<0.001	0.010	0.002
R	0.018	0.015	<0.001	0.011	0.005

Table C.13: Minimum Cumulative Perturbation by Layer for Units Improving from **Cluster 4 to 1**. Values denote the L2 norm of each layer's perturbation vector on the original coefficient scale with layer-specific optimization weights applied.

Unit	OR	Day of Week	Days Until Report	OR × Hours Interaction	Hours Until Phase
G	0.181	0.557	0.001	1.368	0.030
I	0.143	0.669	0.001	1.871	0.025
L	0.176	0.595	0.001	1.198	0.035

operational context. We fully acknowledge that environmental considerations may exist that our framework cannot fully capture. Table C.16 in Appendix C.6 provides detailed coefficient changes for these three units, while Tables C.12 through C.15 show the perturbation magnitudes by layer for all dominated units.

Table C.14: Minimum Cumulative Perturbation by Layer for Units Improving from **Cluster 5 to 1**. Values denote the L2 norm of each layer's perturbation vector on the original coefficient scale with layer-specific optimization weights applied.

Unit	OR	Day of Week	Days Until Report	OR × Hours Interaction	Hours Until Phase
C	0.165	0.589	0.002	0.959	0.032
M	0.159	0.744	0.002	1.063	0.029
S	0.152	0.678	0.002	1.438	0.025

Table C.15: Minimum Cumulative Perturbation by Layer for Units Improving from **Cluster 6 to 3**. Values denote the L2 norm of each layer's perturbation vector on the original coefficient scale with layer-specific optimization weights applied.

Unit	OR	Day of Week	Days Until Report	OR × Hours Interaction	Hours Until Phase
E	0.025	0.065	0.002	0.070	<0.001

Table C.16: Minimum Change by Layer for Selected Units **B**, **E**, and **R**

Layer Name	**Variable Label**	**B**	**E**	**R**
OR	High OR	0.0176	0.0183	0.0436
		(0.0434)	(0.0453)	(0.1079)
OR	Low OR	0.0271	-0.1013	0.0675
		(0.0906)	(-0.3381)	(0.2252)
Hours until Phase	High Hours to Phase	0.0703	0.0030	-0.0270
		(0.3085)	(0.0132)	(-0.1183)
Hours until Phase	Low Hours to Phase	0.1088	0.0046	-0.0417
		(0.3514)	(0.0147)	(-0.1347)
Day of the Week	Friday	-0.0000	-0.0000	-0.0000
		(-0.0047)	(-0.0015)	(-0.0047)
Day of the Week	Monday	-0.0000	0.0004	-0.0000
		(-0.0365)	(0.4055)	(-0.0359)
Day of the Week	Saturday	0.0000	0.0000	0.0000
		(0.0107)	(0.0325)	(0.0105)
Day of the Week	Sunday	0.0000	-0.0003	0.0000
		(0.0357)	(-0.1519)	(0.0351)
Day of the Week	Thursday	-0.0003	-0.0007	-0.0003
		(-0.0913)	(-0.2215)	(-0.0897)
Day of the Week	Tuesday	-0.0000	0.0001	-0.0000
		(-0.0584)	(0.1111)	(-0.0574)
# of Days until Report Period	Days Until Report	-0.0005	-0.0067	-0.0005
		(-0.0892)	(-0.9813)	(-0.0837)
Interaction: OR & Hours to Phase	High Hours to Phase & High OR	-0.0814	0.0330	-0.0154
		(-0.3098)	(0.0872)	(-0.0450)
Interaction: OR & Hours to Phase	High Hours to Phase & Low OR	-0.2734	0.3591	0.0210
		(-0.5895)	(0.8371)	(0.0489)
Interaction: OR & Hours to Phase	Low Hours to Phase & High OR	-0.2717	-0.0693	0.1028
		(-0.5538)	(-0.1715)	(0.2357)
Interaction: OR & Hours to Phase	Low Hours to Phase & Low OR	2.6235	0.3365	-0.1531
		(10.1826)	(1.3027)	(-0.0588)

Note: Values shown are the beta coefficients on original scale with normalized values used in the SOM clustering and optimization stages shown in parentheses underneath.

C.7 Unit Trace Plots by Cluster

Figure C.19: Cluster 1 Trace: Pareto frontier of Monthly OR vs Average FHPA

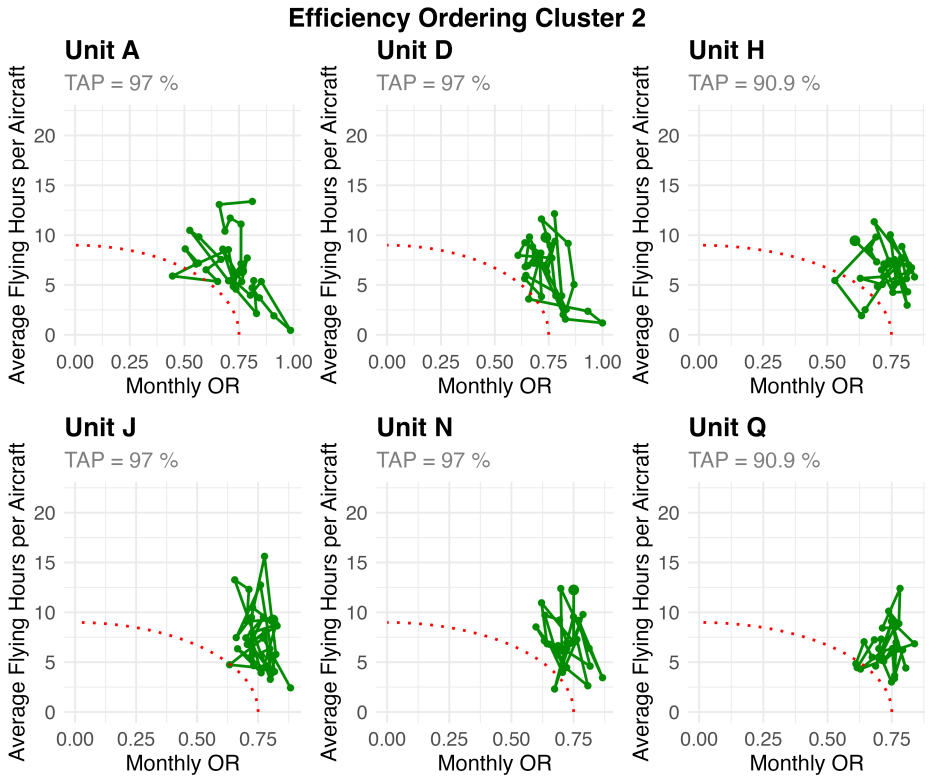

Figure C.20: Cluster 2 Traces: Pareto frontier of Monthly OR vs Average FHPA

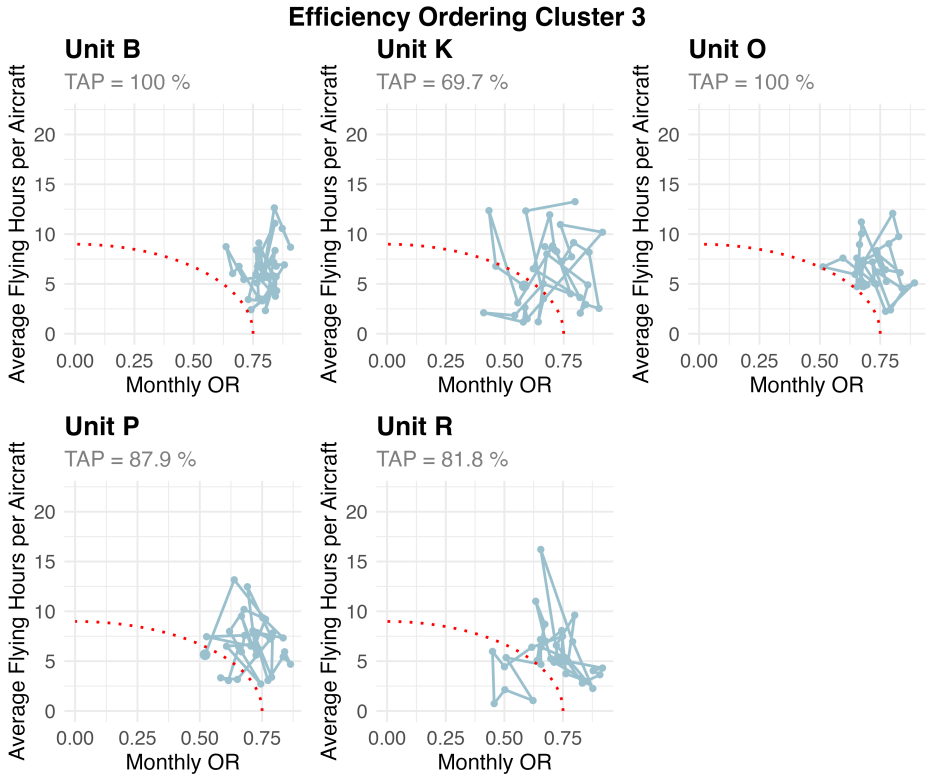

Figure C.21: Cluster 3 Traces: Pareto frontier of Monthly OR vs Average FHPA

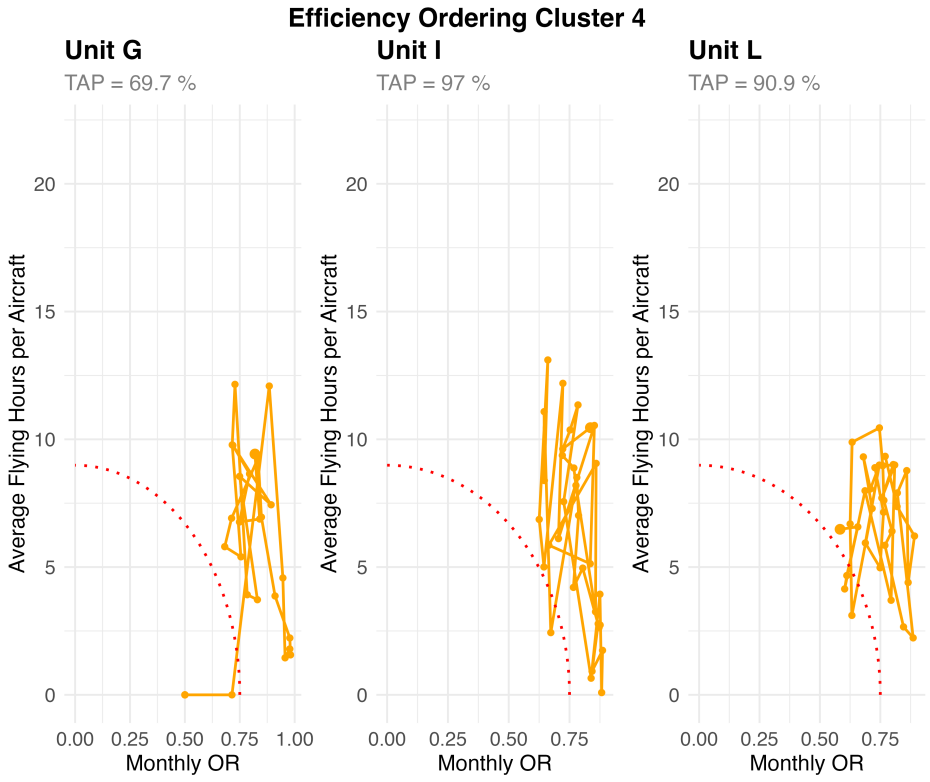

Figure C.22: Cluster 4 Traces: Pareto frontier of Monthly OR vs Average FHPA

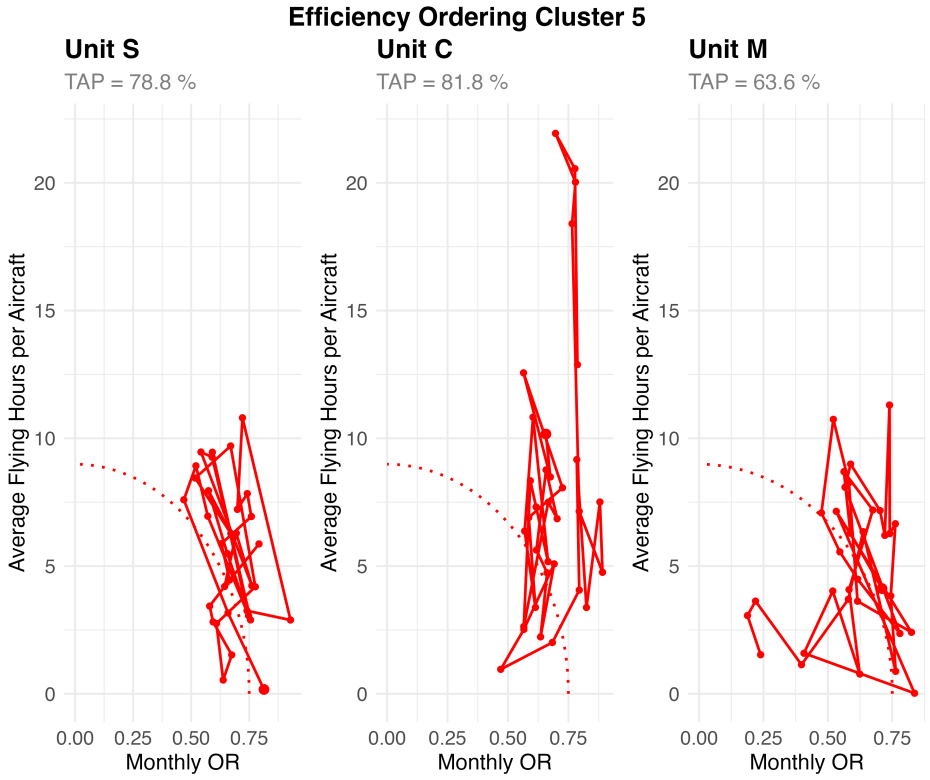

Figure C.23: Cluster 5 Traces: Pareto frontier of Monthly OR vs Average FHPA

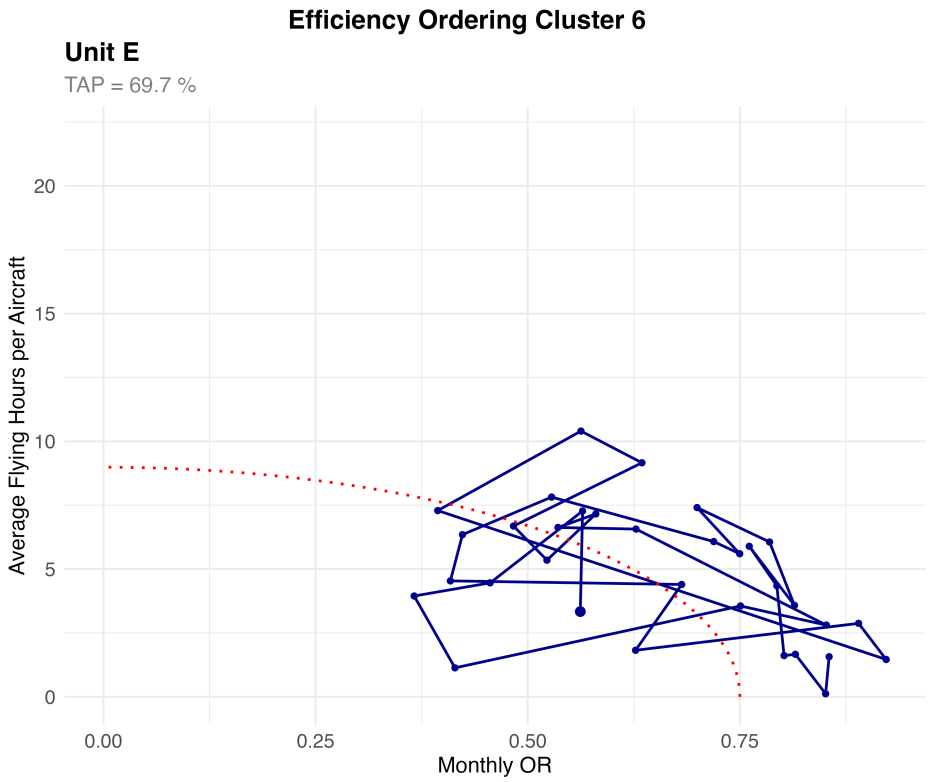

Figure C.24: Cluster 6 Trace: Pareto frontier of Monthly OR vs Average FHPA

C.8 Profile Concordance: Formal Definitions

This appendix provides formal definitions for the profile concordance metrics introduced in Section 3.4.4. All metrics compare a unit's behavioral profile against each candidate cluster prototype. The comparison is per unit, not per cluster. Two units in the same cluster produce different metric values because their individual coefficient profiles differ. Each metric independently selects the candidate target that maximizes similarity (GJ, IoU) or minimizes distance (QE).

Generalized Jaccard Index (GJ). The Generalized Jaccard Index [75] measures the overlap between a unit's behavioral profile and a cluster prototype. Each SOM layer contributes one element to the profile vector: the mean absolute coefficient magnitude for that layer. Given a unit's profile vector $\mathbf{a}$ and a target prototype's vector $\mathbf{b}$, GJ is defined as

$$J(\mathbf{a}, \mathbf{b}) = \frac{\sum_i \min(a_i, b_i)}{\sum_i \max(a_i, b_i)}.$$

Values range from 0 (completely disjoint profiles) to 1 (identical profiles). Higher values indicate greater similarity. The metric is order-invariant: permuting the SOM layers does not change the result.

The SOM groups units by relative behavioral profile, not by identical coefficient vectors. No unit perfectly replicates any cluster prototype. GJ values at the MID-selected target range from 0.215 to 0.549 in this dataset. These values reflect the inherent heterogeneity within SOM clusters rather than poor target selection. The relevant comparison is relative, not absolute: among candidate targets, which one maximizes overlap with the unit's existing profile?

Intersection over Union (IoU). IoU, adapted from the Jaccard Index [76, 77], maps GJ's overlap concept onto the radar chart visualization. Each SOM layer defines one axis of the radar plot, with the radius equal to the mean absolute coefficient magnitude. The intersection polygon uses radius $\min(a_i, b_i)$ at each axis and the union polygon uses $\max(a_i, b_i)$. IoU equals the ratio of intersection area to union area. Values range from 0 to 1 with the same interpretation as GJ. IoU values are consistently lower than GJ for the same comparison because polygon area involves products of adjacent radii rather than direct sums. This adjacency dependence means that IoU is sensitive to the ordering of axes on the radar chart. GJ is not. In this dataset, IoU selects the same target as GJ for all 18 non-frontier units, which suggests that the radar chart comparisons are not artifacts of axis ordering.

Quantization Error (QE). The same quantization error that guided grid selection (Table 3.2) also measures how well an individual unit fits a specific cluster prototype. In grid selection, we minimized aggregate QE across all units to choose the 3×2 lattice. Here, we compute QE between each unit and each candidate target prototype:

$$\text{QE} = \sqrt{\sum_{\ell} \alpha_\ell \, \|\mathbf{x}_\ell - \mathbf{p}_\ell\|^2},$$

where $\mathbf{x}_\ell$ is the unit's coefficient vector for layer ℓ, $\mathbf{p}_\ell$ is the target prototype, and α_ℓ is the SOM's learned layer importance weight [78, p. 53]. Lower QE indicates greater similarity. Unlike GJ and IoU, which summarize each layer as a single scalar, QE preserves within-layer detail across all 15 coefficients.

QE provides a natural baseline: each unit's QE to its own cluster prototype measures current fit. When a unit's QE to a candidate target is comparable to or lower than its own-cluster QE, the transition is natural rather than a forced behavioral shift. Among the 12 non-frontier units in Clusters 3–6, 7 have lower QE to the MID-recommended target than to their own cluster prototype. The three Cluster 5 units (C, M, S) are substantially closer to the Cluster 2 prototype than to their own (QE ratios of 0.64, 0.71, and 0.66 respectively). These units already resemble the recommended destination more than the cluster they currently occupy.

C.9 Literature Positioning in Flight and Maintenance Planning

Table C.17 positions this paper relative to the flight and maintenance planning (FMP) literature. The main distinction is scope and orientation: the prescriptive optimization literature focuses on a single unit and asks what schedule is optimal; purely diagnostic approaches benchmark outcomes without modeling decisions; this paper analyzes all units simultaneously, identifies behavioral patterns that distinguish efficient from inefficient units, and quantifies how dominated units differ from their more efficient peers in terms of specific decision dimensions.

Table C.17: Literature Positioning in Flight and Maintenance Planning

Citation	**Scope of Analysis**	**Role of Units**	**Phase Maintenance**	**Primary Question**	**Method**
Altner et al. (2025)	Single unit (wing-level)	Squadrons as constraints	Scheduling constraint (phase staircase)	What is the optimal FMP schedule?	MILP
Marlow and Dell (2025)	Single unit	Squadrons as assignment locations	Scheduling constraint with induction windows	What is the optimal life-of-type FMP plan?	MIP
Safaei and Jardine (2018)*	Single unit	Aircraft as decision variables	Maintenance capacity constraint	What routing minimizes maintenance misalignment?	MIP + heuristic
Peschiera et al. (2021)	Single unit	Predefined aircraft types/functional clusters	Remaining flight time constraints	What assignment maximizes long-horizon objectives?	MIP + ML cuts
Gavranis and Kozanidis (2015)	Single unit	Individual aircraft; multi-squadron extension	Residual flight time constraints	What schedule maximizes fleet availability?	MILP + cuts
Kozanidis et al. (2012)	Single unit	Aircraft as decision variables	Aircraft flowchart diagonal	What schedule minimizes deviation from targets?	MINLP

Continued on next page

Table C.17 – *Continued from previous page*

Citation	**Scope of Analysis**	**Role of Units**	**Phase Maintenance**	**Primary Question**	**Method**
Kozanidis et al. (2014)	Single unit	Squadrons as subunits	Residual flight time constraints	How can large FMP instances be solved efficiently?	Heuristics
Kozanidis (2009)	Single unit	Squadrons as subunits	Residual flight time constraints	What schedule maximizes availability?	MILP + heuristics
Verhoeff et al. (2015)	Single unit	Aircraft as decision variables	Phase flow chart constraints	What schedule maximizes OR components?	MILP
Hahn and Newman (2008)	Single unit	Helicopters as decision variables	Flight-hour inspection constraints	What schedule ensures operational availability?	MILP
Kim (2020)	Single unit	Aircraft/engines as decision variables	Engine residual life distribution	What assignment maintains uniform depot timing?	MIP (OLS/LAD)
Lee et al. (2016)	Single unit	Aircraft as decision variables	Flight-hour-based (STMA/LTMA)	What schedule maximizes mission completion?	MIP + heuristics
Mattila and Virtanen (2014)	Single unit	Aircraft as decision variables	Flight-hour-based with simulation	What schedule balances availability and adherence?	Simulation-optimization
Marlow et al. (2019)	Single unit	Squadrons as subunits	Multiple service levels	Which policies most influence fleet performance?	DES + DOE
Tseremoglou and Santos (2024)*	Single unit	Components as decision variables	Component RUL-based (POMDP)	What CBM schedule is optimal under uncertainty?	POMDP + DRL
O'Neal et al. (2021)	Cross-period comparison	Time periods as DMUs	Not modeled	How efficient is maintenance over time?	DEA + regression
Hur et al. (2022)	Cross-unit comparison	Units as efficiency benchmarks	Not modeled	How efficient is maintenance across units?	DEA + regression

Continued on next page

Table C.17 – *Continued from previous page*

Citation	Scope of Analysis	Role of Units	Phase Maintenance	Primary Question	Method
Semmel et al. (2025)	Fleet-wide analysis	Units as explanatory factors (random effects)	Observed state (hours until phase)	Does OR regulate flying behavior as doctrine intends?	GAM
This paper	**All units simultaneously**	**Units as behavioral entities**	**Observed state influencing behavior**	**What patterns distinguish efficient from inefficient units?**	**Bayesian logit + SOM**

*Commercial airline context (not military aviation).

Key: MILP = Mixed-Integer Linear Program; MIP = Mixed-Integer Program; MINLP = Mixed-Integer Nonlinear Program; DES = Discrete-Event Simulation; DOE = Design of Experiments; DEA = Data Envelopment Analysis; POMDP = Partially Observable Markov Decision Process; DRL = Deep Reinforcement Learning; GAM = Generalized Additive Model; SOM = Self-Organizing Map.

Appendix D

Chapter 4 Supporting Tables and Figures

D.1 Simulation Parameter Selection and Justification

This appendix documents the rationale for simulation parameter choices. Model parameters follow a precedence hierarchy: (1) operational data are used when available, (2) Army doctrine is applied when data are unavailable or inappropriate, and (3) subject matter expert (SME) judgment is used only for parameters not specified by either source. This hierarchy ensures operational realism while maintaining transparency about abstraction choices. Where neither data nor doctrine provided guidance, we made design choices to produce a realistic and functional constraint regime. Table D.1 summarizes key parameters and their justification sources.

Table D.1: Simulation parameter summary with justification sources.

Parameter	Value	Source	Justification
Fleet size (N)	8 aircraft	Doctrine	Typical company size
Simulation horizon	365 days	Doctrine	Fiscal year alignment
Phase durations	11 / 44 days	Doctrine	ATP 3-04.7 planning factors
Expected demand	2.0 ac/day	Data	25% of fleet; above 16.7% Army average
Token budget (K)	120 tokens	SME	≈5% exhaustion probability
Slot capacity	2 routine, 1 phase	SME	Realistic unit throughput
Token costs	1, 1, 3, 10	SME	Relative effort scaling
CV levels	0–50%	Literature	Spans state-of-art to noisy sensors

D.1.1 Fleet Structure

The fleet consists of $N = 8$ aircraft operating over a 365-day horizon. This fleet size reflects a typical Army aviation company, which maintains approximately 8–12 aircraft depending on airframe type [1]. The 365-day horizon aligns with Army fiscal year planning cycles

and allows sufficient time for multiple phase maintenance cycles and policy performance assessment.

D.1.2 Maintenance Capacity and Token Budget

Slot constraints The simulation provides two routine maintenance slots and one phase maintenance slot. This structure reflects realistic unit throughput: a company-level maintenance section can typically support one to two short-duration events (preventive or reactive) concurrently, while major phase maintenance requires dedicated resources that preclude parallel phase events. Under this capacity structure, the fleet cannot have two aircraft in phase maintenance simultaneously—a constraint consistent with organic maintenance capability.

Token budget The annual token budget ($K = 120$) was selected so that capacity is tight enough to differentiate policies but not so restrictive that all policies fail. Figure D.1 shows the relationship between token budget and annual exhaustion probability under baseline operating conditions. At $K = 120$, approximately 5% of simulation runs exhaust the maintenance budget before year-end, representing a moderately constrained but feasible operating environment. The resource-constrained sensitivity scenario uses $K = 100$, which increases exhaustion probability to approximately 25%.

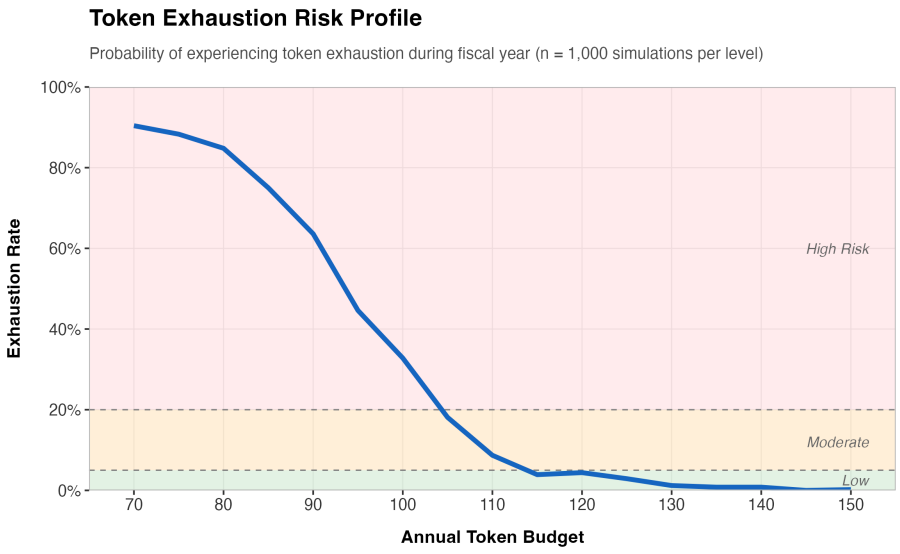

Figure D.1: Token exhaustion probability as a function of annual budget K. The baseline budget ($K = 120$) yields approximately 5% exhaustion probability, creating meaningful resource constraints without forcing policy failure. The resource-constrained scenario ($K = 100$) increases exhaustion risk to approximately 25%.

Token costs Token costs (1 for preventive/reactive, 3 for minor phase, 10 for major phase) represent relative effort scaling rather than precise labor-hour equivalents. The 3:1 and 10:1 ratios approximate the relative resource intensity of scheduled phase events versus routine maintenance actions. These values were selected via SME consultation to produce realistic maintenance workload distributions.

D.1.3 Maintenance Duration Distributions

Maintenance durations are stochastic and differ by event type. Figure D.2 illustrates the duration distributions used in the simulation. Reactive (unscheduled) maintenance follows a lognormal distribution with mean approximately 11 days and a right-skewed tail, reflecting the unpredictability and potential complexity of unplanned repairs. Preventive maintenance follows a bounded uniform distribution (1–4 days), representing short, predictable interventions. Phase maintenance durations (minor: 8–14 days; major: 38–50 days) are calibrated to doctrinal planning factors [1].

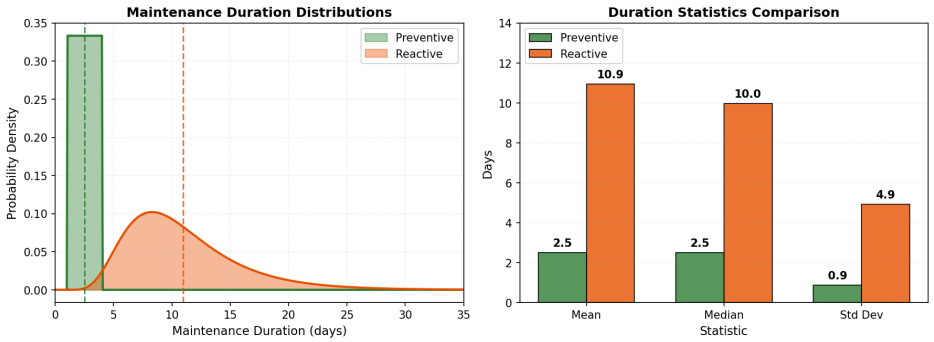

Figure D.2: Maintenance duration distributions by event type. Reactive maintenance exhibits a right-skewed lognormal distribution with long tail, while scheduled events (preventive, minor phase, major phase) follow bounded distributions. This asymmetry (reactive events are both longer on average and more variable) drives the operational cost of unplanned failures.

The key asymmetry is that reactive maintenance is both longer on average (mean ≈11 days vs. 2.5 days for preventive) and more variable (right-skewed tail extending beyond 20 days). This asymmetry underlies a central finding of the paper: policies that reduce reactive failures improve fleet outcomes not just by avoiding downtime, but by reducing downtime *uncertainty*.

D.1.4 Prognostic Accuracy Levels

The coefficient of variation (CV) levels used in the experimental design span the range of prognostic accuracy reported in recent literature. Table D.2 maps each CV level to representative prediction error metrics from published RUL prediction studies.

Table D.2: CV-to-literature mapping. Each experimental CV level corresponds to a range of prediction accuracies observed in published RUL studies. MAPE = Mean Absolute Percentage Error.

CV	Interpretation	Literature MAPE	Representative Source
0%	Perfect knowledge	—	Theoretical baseline
5%	State-of-the-art	2–5%	Battery degradation [150]
10%	Good deployment	10–15%	Bearing fault prognosis [151]
25%	Typical	15–20%	Turbofan engine RUL [152]
50%	Noisy / worst-case	20–30%+	Challenging field conditions

The 0% CV baseline represents an unattainable ideal that isolates the upper bound on policy performance. The 5% and 10% levels represent current best-in-class laboratory and deployment accuracy, respectively. The 25% level reflects typical field performance with sensor degradation and environmental variability. The 50% level represents a pessimistic scenario with substantial sensor noise or model mismatch. This range ensures that results characterize the full spectrum of practically relevant prognostic quality.

D.2 Sensitivity Scenario Transition Matrices

This appendix presents the transition matrices used in the sensitivity analyses. Each matrix is a self-exciting Markov chain governing daily mission demand, with states 0–7 representing the number of aircraft required per day. The self-exciting structure mirrors the cyclical nature of military operations: high-tempo periods tend to persist due to sustained training exercises or deployments, while low-tempo periods reflect recovery and maintenance windows. The stationary distribution $\boldsymbol{\pi}$ is shown in the final row of each table.

D.2.1 High Optempo Scenario

The high optempo transition matrix shifts probability mass toward higher demand states while maintaining similar entropy (predictability) to the baseline. Expected demand in-

creases from 2.0 to 3.0 aircraft per day, representing a 50% increase in operational tempo. This scenario tests policy robustness under sustained high mission requirements.

Table D.3: High optempo transition matrix $\mathbf{P}$ with stationary distribution $\boldsymbol{\pi}$. Expected demand $\mathbb{E}[D] \approx 3.0$ aircraft/day.

	To State (Aircraft Required)							
From	0	1	2	3	4	5	6	7
0	.650	.200	.100	.050	.000	.000	.000	.000
1	.200	.425	.250	.075	.050	.000	.000	.000
2	.100	.225	.350	.200	.075	.050	.000	.000
3	.050	.100	.200	.300	.200	.100	.050	.000
4	.000	.050	.100	.175	.350	.175	.100	.050
5	.000	.025	.050	.100	.200	.350	.200	.075
6	.000	.000	.025	.050	.100	.250	.425	.150
7	.000	.000	.000	.025	.050	.150	.400	.375
π	.150	.152	.155	.130	.128	.121	.112	.052

D.2.2 High Variance Scenario

The high variance transition matrix maintains the same expected demand as baseline ($\mathbb{E}[D] \approx 2.0$) but increases entropy by 32% (achieving 91% of maximum entropy). All 64 state transitions have non-zero probability, allowing the system to jump from state 0 (no demand) to state 7 (maximum demand) in a single day. This scenario represents unpredictable operational environments where demand fluctuates dramatically and tests whether policies can adapt to high uncertainty.

D.2.3 Scenario Comparison Summary

Table D.5 summarizes the key differences between the three mission demand scenarios used in the experimental design.

D.3 CV-to-Decision Behavior Mapping

This section provides operational interpretation of the CV tiers used in the experimental design. Table D.6 translates prognostic accuracy levels into decision-relevant probabilities: how often does a given CV level produce early triggers (preventive maintenance when failure was distant) or late triggers (missed intervention windows)?

Table D.4: High variance transition matrix $\mathbf{P}$ with stationary distribution $\boldsymbol{\pi}$. Expected demand $\mathbb{E}[D] \approx 2.0$ aircraft/day; entropy 1.32× baseline.

	To State (Aircraft Required)							
From	0	1	2	3	4	5	6	7
0	.375	.250	.125	.075	.050	.050	.050	.025
1	.350	.275	.125	.075	.050	.050	.050	.025
2	.325	.250	.125	.075	.075	.050	.050	.050
3	.300	.200	.125	.125	.075	.075	.050	.050
4	.225	.200	.125	.125	.125	.075	.075	.050
5	.225	.175	.125	.100	.075	.125	.075	.100
6	.225	.150	.125	.100	.075	.100	.100	.125
7	.200	.150	.100	.100	.100	.100	.100	.150
π	.320	.233	.124	.087	.066	.064	.059	.049

Table D.5: Mission demand scenario comparison.

Property	**Baseline**	**High Optempo**	**High Variance**
Expected demand (aircraft/day)	2.0	3.0	2.0
Entropy (relative to baseline)	1.00	≈1.00	1.32
Zero-probability transitions	22	22	0
Max single-day jump	±3 states	±3 states	±7 states

Interpretation At CV = 5% (upper-bound laboratory accuracy), an aircraft with true RUL of 50h has only a 2% chance of being observed below threshold when it should be above, or vice versa. At CV = 50% (noisy sensors), the same aircraft has a 30% chance of triggering early and 25% chance of being missed entirely. The asymmetry at higher CV arises from the Gamma distribution's positive skew: high-CV observations more frequently underestimate than overestimate true RUL.

Policy implications The diminishing returns result can be understood through this lens: moving from CV = 50% to 25% reduces false-alarm and missed-detection rates substantially (30%→15%, 25%→12%). Further improvement from CV = 10% to 5% provides smaller absolute reductions (5%→2%), explaining why operational gains plateau in the high-accuracy regime.

D.4 Genetic Algorithm Hyperparameters

This section documents the genetic algorithm configuration used for policy optimization. The heterogeneous island model employs three subpopulations with differentiated exploration-exploitation tradeoffs.

Table D.6: CV-to-decision behavior mapping for alarm threshold $\tau = 50$h. Probabilities computed analytically assuming Gamma-distributed observation errors with the indicated CV, evaluated at true RUL = 50h (near threshold).

CV (%)	P(Early Trigger)	P(Late Trigger)	Operational Meaning
5	∼2%	∼2%	Tight bounds; rarely wrong
10	∼5%	∼5%	Occasional early/late triggers
25	∼15%	∼12%	Frequent early; some late
50	∼30%	∼25%	High uncertainty both ways

D.4.1 Island Configuration

Table D.7: Island model configuration. Each island operates with different selection pressure to balance exploration and exploitation.

Parameter	Explorer	Refiner	Validator
Tournament size (k)	2	3	4
Elitism rate	10%	15%	16%
Role	Diversity generation	Intensification	Final refinement

Migration occurs every 40 generations in a feed-forward ring topology (Explorer → Refiner → Validator), with each island exporting its top performers to the downstream population [114, 118].

D.4.2 Genetic Operators

Crossover Continuous genes (threshold values) use BLX-α crossover with $\alpha = 0.5$, which samples offspring uniformly from an interval extending beyond the parents' values [112, 119]. Discrete genes (feature selections) use uniform crossover with 50% swap probability per gene.

Mutation Mutation is adaptive to transition from exploration to refinement as evolution progresses [117, 153]. Continuous genes receive Gaussian perturbations with standard deviation decreasing exponentially from $\sigma = 0.15$ (early generations) to $\sigma = 0.02$ (late generations). Discrete genes mutate with 10% probability per gene throughout training.

D.4.3 Termination and Evaluation

During GA search, each chromosome is evaluated over 50 Monte Carlo episodes to balance fitness estimation reliability against computational cost [154]. Training terminates after 121 generations without fitness improvement (corresponding to three complete migration cycles). The best policy is extracted from the validator island. For final performance reporting and statistical inference, selected policies are re-evaluated on 10,000 independent replications.

D.5 GA Results

Table D.8: Performance comparison: benchmarks vs. GA-optimized policies across RUL accuracy levels (pooled across alarm thresholds τ). MS = Mission Success (%), OR = Operational Readiness (%), Reactive = mean annual reactive failures per aircraft. Bold indicates improvement over all benchmarks.

	Benchmarks			**GA by CV**				
Metric	**FI-25h**	**FI-50h**	**Heuristic**	**50%**	**25%**	**10%**	**5%**	**0%**
Mission-Focused ($w_{\text{ms}} = 0.7$)								
Mission Success (%)	41.1	56.2	62.1	**64.5**	**68.5**	**70.1**	**70.1**	**70.3**
Operational Readiness (%)	74.1	63.0	55.7	56.8	60.2	61.3	61.5	61.6
Reactive Failures	1.9	6.8	15.8	4.8	2.9	3.9	3.2	3.4
Readiness-Focused ($w_{\text{ms}} = 0.3$)								
Mission Success (%)	41.1	56.2	62.1	**65.3**	**69.5**	**71.6**	**70.5**	**70.3**
Operational Readiness (%)	74.1	63.0	55.7	58.8	61.3	**63.6**	61.8	62.2
Reactive Failures	1.9	6.8	15.8	4.9	3.6	4.1	3.8	3.1

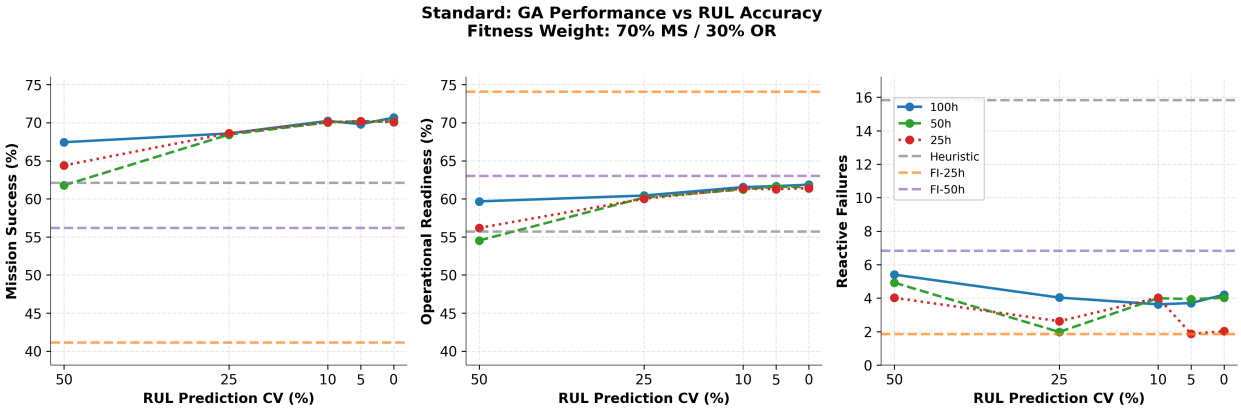

Figure D.3: Effect of prediction accuracy on MS, OR, and reactive failures under standard conditions for the mission-focused preference profile ($w_{\text{ms}} = 0.7$). GA-optimized policies outperform all benchmarks across the full CV range. Diminishing returns are evident: most improvement is captured by CV=25%.

D.6 Maintenance Event Breakdown and Slot Utilization

This section provides a complete breakdown of maintenance events and maintenance capacity utilization across all policies under baseline operating conditions. The fleet operates with two preventive/reactive maintenance slots (730 slot-days per year) and one phase maintenance slot (365 slot-days per year). Slot utilization is calculated as: (events $\times$ mean duration) / available slot-days. Mean durations are: preventive = 2.5 days, reactive = 10.9 days, minor phase = 11 days, major phase = 44 days.

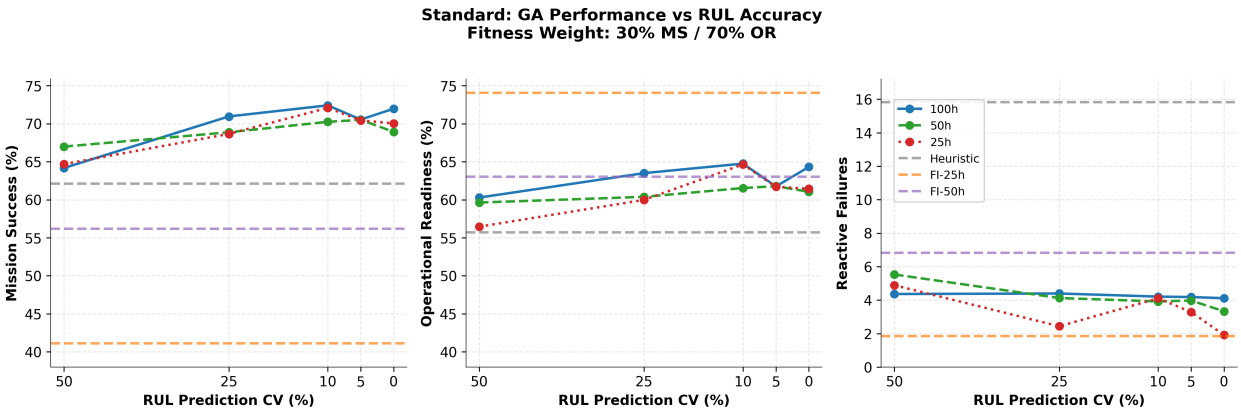

Figure D.4: Effect of prediction accuracy on MS, OR, and reactive failures under standard conditions for the readiness-focused preference profile ($w_{\text{ms}} = 0.3$). GA-optimized policies dominate all benchmarks across the full CV range.

Table D.9: Maintenance event breakdown and slot utilization under baseline conditions (n = 10,000 replications per policy). P/R Util = preventive/reactive slot utilization; Phase Util = phase slot utilization; Flt Hrs = annual flight hours; Hrs/Reset = flight hours ÷ total maintenance events.

Policy	CV / T	MS	Prev	React	Minor	Major	Total	P/R Util	Phase Util	Flt Hrs	Hrs/Reset
Benchmarks											
Heuristic	—	62.1	0.0	15.8	7.4	5.8	29.0	21.6%	92.2%	—	—
FI-25h	25h	41.1	73.1	1.9	5.5	2.9	83.4	27.6%	51.5%	2,436	29
FI-50h	50h	56.2	45.1	6.8	—	—	51.9	24.8%	—	∼3,100	∼60
GA: CV = 0% (Perfect)											
GA	0% / 100h	70.7	39.9	4.2	8.0	4.6	56.7	19.4%	79.6%	3,663	65
GA	0% / 50h	70.2	40.4	4.0	7.9	4.6	56.9	19.3%	79.3%	3,660	64
GA	0% / 25h	70.1	44.0	2.0	7.9	4.5	58.4	17.8%	78.1%	3,644	62
GA: CV = 10%											
GA	10% / 100h	70.2	41.7	3.6	7.9	4.6	57.8	19.2%	79.3%	3,651	63
GA	10% / 50h	70.0	41.2	4.0	7.9	4.6	57.7	19.6%	79.3%	3,649	63
GA	10% / 25h	70.1	41.3	4.0	7.9	4.6	57.8	19.6%	79.3%	3,652	63
GA: CV = 25%											
GA	25% / 100h	68.6	44.5	4.0	7.7	4.4	60.6	20.7%	76.2%	3,591	59
GA	25% / 50h	68.4	48.4	2.0	7.7	4.3	62.4	19.3%	75.0%	3,573	57
GA	25% / 25h	68.6	47.1	2.6	7.7	4.3	61.7	19.7%	75.0%	3,595	58
GA: CV = 50%											
GA	50% / 100h	67.4	51.3	5.4	7.3	3.8	67.8	25.0%	67.8%	3,459	51
GA	50% / 50h	61.8	52.9	4.9	7.2	3.8	68.8	24.8%	67.5%	3,386	49
GA	50% / 25h	64.4	53.2	4.0	7.4	3.8	68.4	23.7%	68.1%	3,452	50

Notes: All values are annual means per 8-aircraft fleet. The heuristic's 92.2% phase utilization reflects reactive cascade into phase maintenance, not planned entries. FI-50h minor/major phase data unavailable. CV = coefficient of variation; T = alarm threshold. MS70 fitness weight shown.

D.7 Tukey HSD Block Analysis

The significant 4-way interaction (rejected in ANOVA) indicates that the A×B interaction magnitude varies across blocks. To determine which blocks differ significantly, we apply Tukey's Honest Significant Difference (HSD) procedure [121], which controls the family-wise error rate across all $\binom{6}{2} = 15$ pairwise block comparisons.

Table D.10 reports the A×B interaction estimates by block. The ms30/100h block (readiness-focused objective with conservative alarm threshold) shows a +3.2 pp synergy effect—a statistically significant and practically meaningful interaction. In this configuration, policies trained under accurate conditions perform disproportionately well when tested with accurate sensors. However, this exception does not generalize: in 5 of 6 blocks, the A×B interaction is <1 pp and not practically significant.

Table D.10: A×B interaction by block. Only the ms30/100h block shows a practically significant synergy effect.

Block	**A×B (MS, pp)**	**Practical Significance**
ms30/100h	+3.2	Yes
ms30/50h	+0.8	No
ms30/25h	+0.5	No
ms70/100h	+0.8	No
ms70/50h	−0.9	No
ms70/25h	+0.4	No

D.8 GA Training Variability at High Accuracy

Table D.11 reports the best training fitness achieved by the genetic algorithm at each combination of prediction accuracy (CV) and alarm threshold (τ) for the mission-focused objective ($w_{\text{ms}} = 0.7$). At CV=5% and CV=10%, best training fitness differs by less than 0.5 percentage points at each threshold. At $\tau = 25$h the two levels are effectively identical. These differences are small relative to the variation across alarm thresholds within a single CV level (2–3 percentage points), which confirms that the reversals observed in Section 4.6.2 reflect optimization noise rather than a true performance difference between CV=5% and CV=10%.

Table D.11: Best GA training fitness by prediction accuracy and alarm threshold ($w_{\text{ms}} = 0.7$). The CV=5% versus CV=10% gap is less than 0.5 percentage points at every threshold.

CV	$\tau = 100$h	$\tau = 50$h	$\tau = 25$h
0% (perfect)	0.837	0.857	0.860
5%	0.832	0.850	0.856
10%	0.835	0.854	0.856
5% vs. 10% gap	0.003	0.004	$<$0.001

D.9 Gamma Observation Noise

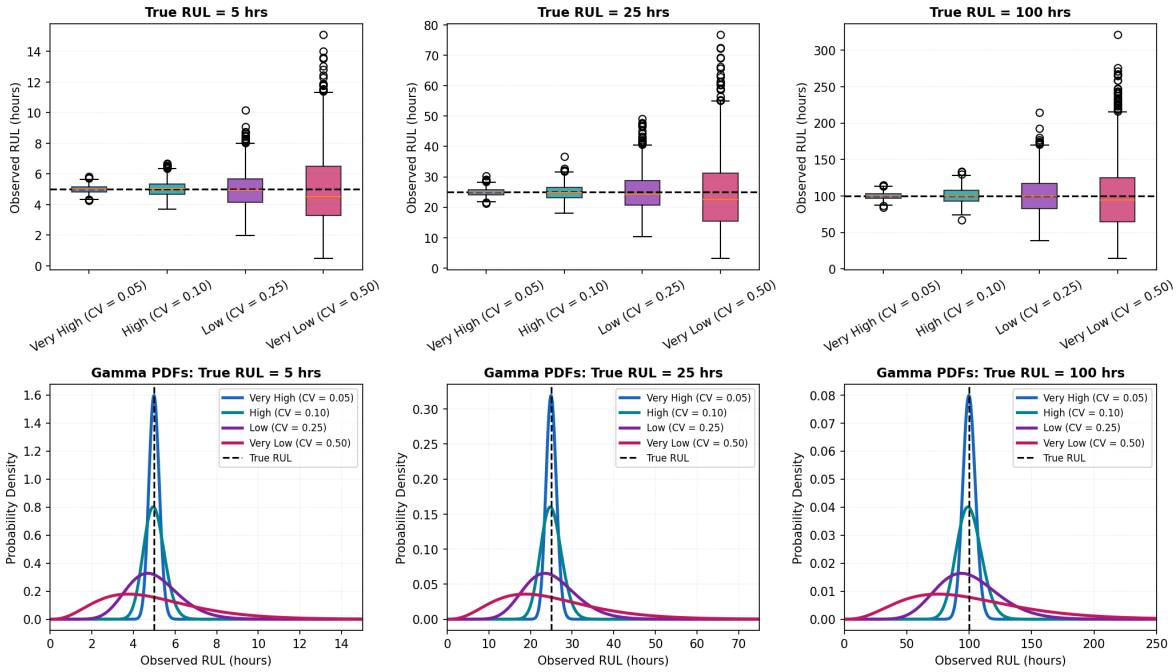

Figure D.5: Gamma observation noise for RUL at selected true values and coefficients of variation. Observations are unbiased with standard deviation proportional to true RUL, increasing with CV.

D.10 Post-Hoc Interaction Plots

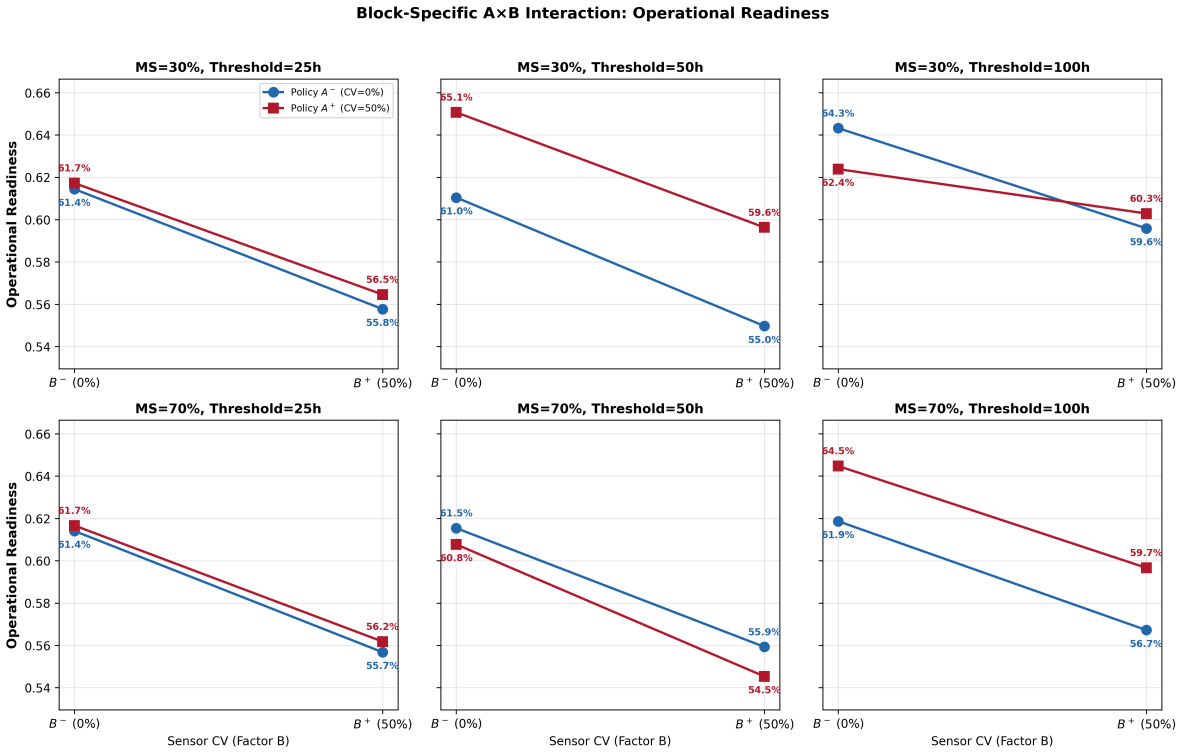

Figure D.6: Factorial interaction plots for Operational Readiness. Near-parallel lines across most blocks indicate that the sensor effect dominates, with minimal A×B interaction.

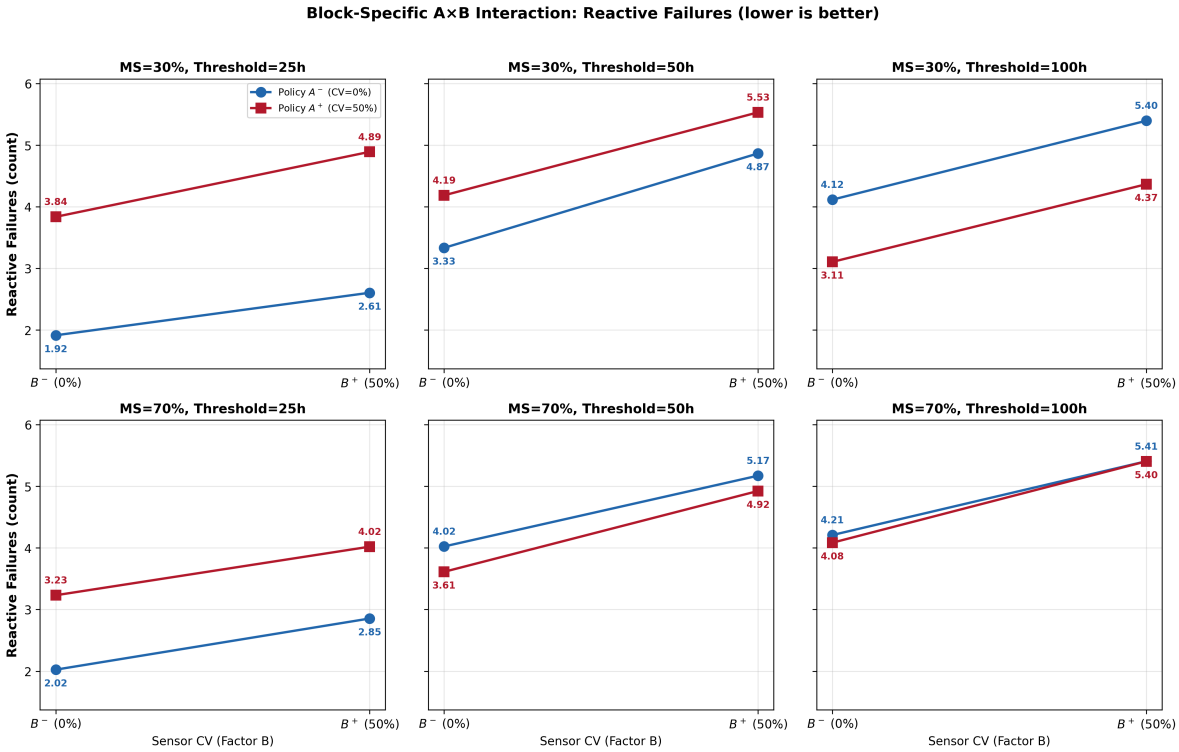

Figure D.7: Factorial interaction plots for Reactive Failures. Near-parallel lines across most blocks indicate that the sensor effect dominates, with minimal A×B interaction.

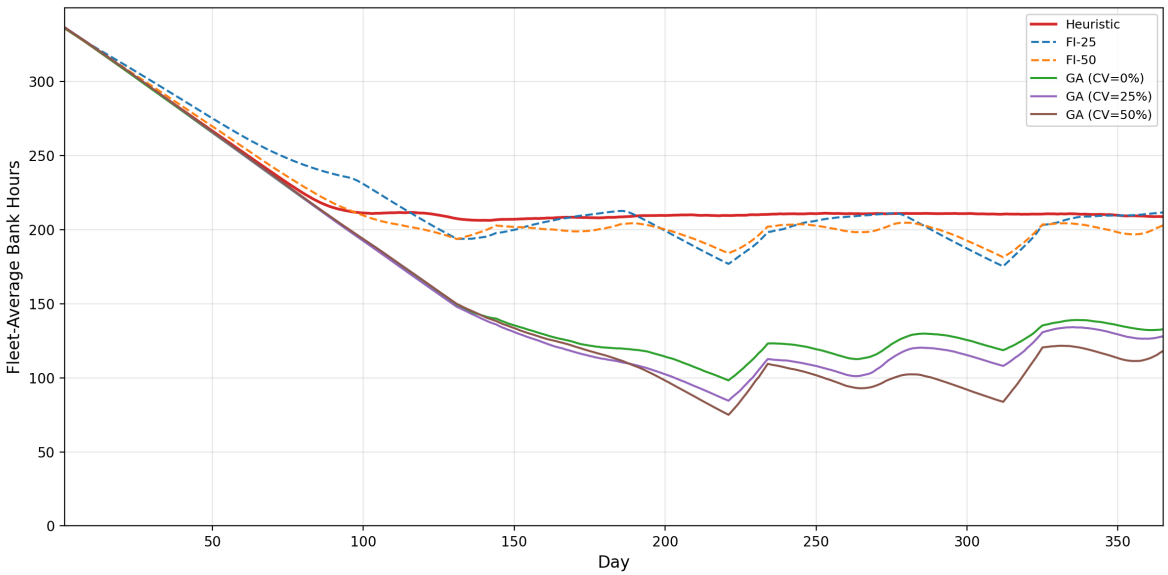

Figure D.8: Fleet-average bank hours (hours until major phase maintenance) over the simulation horizon by policy, averaged across 10,000 replications under the standard scenario ($w_{\text{ms}} = 0.7$, $\tau = 100$h). Benchmark policies (Heuristic, FI-25, FI-50) stabilize near 200 hours, conserving bank hours and cycling through phase infrequently. GA-optimized policies drive utilization substantially harder, stabilizing near 100–130 hours. Within the GA group, higher CV produces lower fleet-average bank hours, consistent with noisy signals causing misallocation of utilization across aircraft.