A FRAMEWORK FOR DESIGNING TECHNOLOGY DEVELOPMENT ACTIVITIES

A Thesis
Presented to
The Academic Faculty

by

Ryan B. Jacobs

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Aerospace Engineering

Georgia Institute of Technology
August 2016

A FRAMEWORK FOR DESIGNING TECHNOLOGY DEVELOPMENT ACTIVITIES

Approved by:

Professor Dimitri N. Mavris, Advisor
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Kelly Griendling
School of Aerospace Engineering
Georgia Institute of Technology

Professor Daniel P. Schrage
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Jeff Schutte
General Electric Aviation

Professor Brani Vidakovic
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: July 6, 2016

ACKNOWLEDGMENTS

During my time at Georgia Tech I have received encouragement and support from many individuals. This dissertation would not have been possible without my advisor and mentor, Professor Dimitri Mavris. I am grateful for the many learning opportunities he has provided me with. I would like to thank the members of my committee: Dr. Griendling, Professor Schrage, Dr. Schutte, and Professor Vidakovic. Their guidance is greatly appreciated, and I feel privileged to have had the opportunity to learn from them.

I would like to thank Dr. Fayette Collier from NASA Langley Research Center for funding my research. Working with you on a real technology development program has provided me with invaluable experience and insights that helped to form a practical basis for the ideas in this dissertation.

I would like to express my gratitude to my family. Without the love and support of my parents, none of my achievements would have been possible. To my wife, Allison, I cannot thank you enough for your support throughout graduate school. I look forward to our future in Boston together.

To all of my friends at the Aerospace Systems Design Laboratory, thank you for making my time at Georgia Tech enjoyable and enlightening. I would like to extend a special thanks to Justin Kizer and Michael Steffens, who provided constructive feedback during the formulation of the ideas in this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

To fulfill the future aviation needs of the public and military, there are efforts in industry and government to integrate aircraft with enabling technologies to achieve aggressive goals and requirements for performance and capabilities. However, many enabling technologies are immature, and system integrators incur the associated risk when they integrate these technologies. This risk can be reduced through technology development programs, but these programs often require over ten years and significant resources before the technology can be transitioned to the vehicle. Ideally, the process could be accelerated and the required resources reduced by creating the development activities, such as physical experiments and tests, such that they maximize performance improvement, maturation, and risk reduction during the development program. The motivating question is *How should technology development activities be designed?* The research in this dissertation comprises contributions toward a solution this problem.

A review of the literature pertaining to the design of technology development activities revealed that current practices are driven by a qualitative criterion called Technology Readiness Level that does not provide a clear picture of the state of knowledge about technology impacts. The immediate consequence of using this criterion for decision making is that it does not capture all of the critical dimensions of the consequence space for evaluating alternative activity designs and may result in misinformed decisions. Existing technology development activity design methodologies were identified that improve upon current practices, but they fall short of providing a complete path to designing a portfolio of technology development activities. To address the gaps from the literature, a novel framework was proposed that comprises

three phases: (1) thought experimentation, (2) detailed definition of the activities, and (3) statistical design of experiments. Although the proposed framework can be implemented as is for a given technology development program, opportunities were identified to enhance the framework by adding rigor to the decision making processes.

Three enhancements to the proposed solution framework are presented in this dissertation. Each enhancement improves upon methods from the literature by addressing research gaps. First, existing methodologies for planning and managing technology development leverage sensitivity analyses to inform decisions regarding which classes of development activities to pursue. It was argued that this approach does not explicitly evaluate alternatives, but rather provides measures of the potential of *any* development activities to affect system-level uncertainty and performance. Thus, a need was identified for an appropriate way for decision makers to evaluate the alternatives for downselection. Second, existing quantitative methodologies make the assumption that the combined epistemic and aleatory uncertainty surrounding technology integration impacts can be quantified from a combination of data and expert elicitation. Bayesian inference has been proposed for sequentially updating initial probability distributions with data from technology development activities, but misleading inferences can arise when the data sources are heterogeneous. To overcome this issue, there is a need for an appropriate way to quantify technology integration impact uncertainty in light of data from multiple, heterogeneous experiments. Finally, as part of any decision process for the detailed design of the development activities, there are multiple criteria that are important to include when evaluating the alternatives. One of the most prominent criteria that is mentioned in the literature is uncertainty reduction. To enable the evaluation of alternatives, a need was identified for an appropriate way to quantitatively estimate expected uncertainty reduction for planned technology development activities.

The first research gap was addressed with a normative decision support methodology that incorporates techniques from multiattribute utility theory. The methodology entails establishing objectives and attributes, constructing a utility model to represent decision makers' values, modeling the impacts of the alternatives, and evaluating the alternatives with expected utility. The product of the methodology is not simply a single expected utility for each alternative but rather a capability that enables quantitative tradeoffs and sensitivity analyses to provide insights and stimulate deeper thinking about the problem on the part of the decision makers. Compared with the state of the art, the proposed methodology is an improvement because it was shown to enable explicit evaluation of alternatives rather than only providing measures of potential for each technology.

The second and third research gaps were addressed for two types of technology development activities: computer experiments and physical experiments. Although there are many types of technology development activities, these were the focus because they are crucial to development; technologies cannot be matured without them. The ingredients for a solution were identified in the statistics and machine learning literature. These ingredients were synthesized and adapted for the technology development context to formulate a methodology that addresses the research gaps. The first three steps of the methodology were borrowed from the data analysis literature. These steps comprise the traditional pipeline of cleaning a data set, identifying a set of predictive models, and evaluating and selecting from the set of models. The fourth step is a novel contribution because it provides an approach for incorporating epistemic technology maturity uncertainty in Gaussian process model predictions. The fifth step is also a novel contribution because it fuses a rigorous information theoretic framework for quantifying uncertainty reduction with predictive models that incorporate the additional layer of epistemic uncertainty associated with technology maturity.

The second gap was also investigated for success/failure reliability tests. An adaptation of the traditional Bayesian beta-binomial probability model was formulated to address the research gap. The novel Bayesian reliability analysis methodology begins with traditional Bayesian data analysis steps. Then, a maturity weight is introduced in the posterior beta distribution to enable discounting of the reliability data at a given point in the development process. The flexibility provided by the infusion of a maturity weight was shown to enable an analyst to inject additional subjective uncertainty into the inference process, thereby enabling estimates of failure probabilities that reflect this maturity uncertainty.

The objective of this research was to establish a framework for designing technology development activities that improves the state of decision support capabilities. Although the framework has been established so that it can be populated with additional improvements in the future, the research objective was achieved because all of the contributions presented in this dissertation have been shown to improve upon existing methods and current practices.

CHAPTER I

INTRODUCTION

To fulfill the future aviation needs of the public and military, systems integrators and government organizations are striving to determine how to achieve aggressive goals and requirements for performance and capabilities. One approach being pursued is to infuse enabling technologies into aircraft. In the commercial aviation sector, current goals and requirements are primarily motivated by concerns about the environmental impacts of aviation and the cost of fuels. For instance, Table 1 lists noise, NOx emissions, and fuel burn goals for subsonic transport aircraft that are being targeted by NASA under the Environmentally Responsible Aviation (ERA) project. In an attempt to simultaneously meet these goals, airframe and propulsion technologies are being pursued, in addition to advanced vehicle concepts and improved airspace operations. On the military side, capability goals and requirements are driving the development of aircraft concepts that introduce new technical challenges. For example, the U.S. Navy's interest in unmanned, autonomous aircraft systems for the suppression of enemy air defenses led to the design and test of the Northrop Grumman X-47B [1]. Due to the tailless design for low observability of the X-47B, achieving satisfactory high lift and control during low-speed carrier operations is difficult with conventional control methods. As a result, innovative control effectors, such as active flow control (AFC) actuators, have been investigated to overcome stability and control issues for tailless configurations.

AFC is an enabling technology that is used as one of the examples in this dissertation. Here, the history of flow control serves as an example of the challenges involved in transitioning enabling technologies to aircraft. Flow control involves the

Table 1: NASA subsonic transport system-level goals (data from Ref. [2])

Technology benefits[a]	N+1 (2015)	N+2 (2020[b])	N+3 (2025)
Noise (cumulative below stage 4)	-32 dB	-42 dB	-52 dB
LTO NOx (below CAEP6)	-60%	-75%	-80%
Cruise NOx emissions (relative to 2005 best in class)	-55%	-70%	-80%
Aircraft fuel/energy consumption (relative to 2005 best in class)[c]	-33%	-50%	-60%

[a] Projected benefits once technologies are matured and implemented by industry. Benefits vary by vehicle size and mission. N+1 and N+3 values are referenced to a 737-800 with CFM56-7B engines, N+2 values are referenced to a 777-200 with GE90 engines.

[b] ERA's time-phase approach includes advancing "long-pole" technologies to TRL 6 by 2015.

[c] CO_2 emission benefits dependent on life-cycle CO_2e per megajoule for fuel and/or energy source used.

use of active or passive devices to achieve a desired change in wall-bounded or free-shear flows [3]. Passive flow control does not require auxiliary power or a control loop, whereas AFC entails energy addition to the flow with devices called actuators. Although scientific AFC research has been ongoing since Prandtl's suction flow control experiments over 100 years ago [4], few production vehicles currently operate with applied AFC techniques. Most of these aircraft employ boundary layer control (BLC) for lift augmentation. Examples include the Mikoyan-Gurevich MiG-21, which has an internally-driven BLC system, and the Boeing C-17 Globemaster III, which uses externally-blown flaps. Also, some helicopters, such as the MD Helicopters MD 600N, use BLC for anti-torque control in lieu of a tail rotor. Internally-driven BLC systems became unpopular for aircraft applications by the late 1960s, primarily because of integration issues. The ducting required for a BLC system introduces additional weight and complexity to the vehicle. Efficiency is also a concern because of the amount of compressed flow needed for effectiveness. According to Williams and MacMynowski [5], in the 1970s the application of BLC shifted toward externally-blown flaps for high lift during takeoff and landing, but there was still a demand for more practical and reliable flow control techniques. In the 1980s, upon the acceptance of the notion that organized flow structures are abundant in turbulent shear flows, the

AFC paradigm changed from using BLC to modify the mean boundary-layer flow behavior to using "modern" AFC devices to operate on flow instabilities. Modern AFC, hereafter referred to with the initialism AFC, is often proclaimed in the literature as being superior to BLC and offering significant performance improvement for aircraft. However, there are few examples of successful transition of AFC from a laboratory setting to practical applications. A question naturally follows from this observation: *Why is the application of an enabling technology, such as AFC, to current and future flight vehicles challenging?*

Potential answers to this question are related to the uncertainty surrounding immature technologies and the consequences of integration with aircraft. For instance, efforts to transition AFC devices to full-scale applications began in the early 2000s, and the system-level integration effects are still not well understood. Also contributing to the uncertainty is a lack of understanding of the governing physics of many technologies. In addition to uncertainty, vehicle integrators may be wary of immature technologies because of the possible business repercussions. The additional complexity that enabling technologies add to aircraft can increase costs incurred by the manufacturer and operator. For example, Liddle et al. [6] argued that integration of AFC devices will result in increased cost associated with meeting safety standards, particularly for application scenarios in which failure of the AFC system would be catastrophic. They also claimed that unsuccessful AFC implementation could put an aircraft manufacturer out of business. The perennial problem for the integration of any immature technology can be summarized with one concept: risk.

1.1 Risk and Uncertainty in the Technology Development Context

Sources of uncertainty combined with unwanted consequences hinder the application of promising but immature technologies. Uncertainty and consequences are common

components of definitions for risk. Risks associated with adopting an immature technology must be managed to ensure successful application. In order to understand how risk can be managed, precise definitions of risk and uncertainty are required.

Unfortunately, there is not one general definition of risk that is widely accepted across all professional fields. In medicine, risk is the probability of an undesirable event. For example, doctors report cancer risk to their patients, indicating the probability that the patients will develop cancer. Some economists view risk as uncertainty, with variance as a measure of the uncertainty. Consequences are implicit in both of these definitions. In engineering, risk definitions are typically based on events that result in unwanted consequences and the probability of those events. The same can be said of technology development, as seen in the following examples. Moorhouse [7] defined risk as "the judgment of probability and consequence to the system application of failure of that technology to match predictions adequately." Smaling and de Weck [8] defined risk as "the likelihood that a system design or architecture will not satisfy the performance objectives and the negative consequences thereof." Many organizations and researchers, such as Mankins [9], promote the use of a risk matrix for assessing risk in a technology development program. The risk matrix captures the interaction of probability of technical failure and the consequences of failure. Although there are differences in the risk definitions from these technology development examples, all of them contain probability (or likelihood) and consequences. A general definition of risk that encapsulates these elements states that risk is the combination of possible consequences and associated uncertainties [10]. This can be written more compactly as (A, C, U), where A represents possible events, C represents consequences of A occurring, and U is the uncertainty surrounding A and C.

In the context of interest, risk is a burden of decision makers who must determine whether to transition a technology from development to vehicle application. If decision makers could attain a state in which there is no uncertainty and thus no risk,

a decision could be made that would guarantee desired outcomes. A definition of uncertainty that aligns with this notion is provided by Nikolaidis [11], who defined uncertainty indirectly from certainty. Nikolaidis defined *certainty* as the condition of possessing all knowledge that is required to choose the action with the most desirable consequences. *Uncertainty* is the gap between certainty and a decision maker's present state of knowledge, as shown by the top bar in Fig. 1. A decision maker's uncertainty can be decomposed using a taxonomy that the risk assessment community has developed over the past couple of decades [12]:

- *Aleatory uncertainty*: uncertainty due to inherent randomness
- *Epistemic uncertainty*: uncertainty due to lack of knowledge

These are the two kinds of uncertainty that surround the integration impacts of a technology at any point in time. Aleatory uncertainty is a property of the system being observed, and variability exhibited by the system cannot be reduced unless the system itself is modified. No matter how much information is attainable regarding an observable system (e.g., a die rolled by a human), sources of aleatory uncertainty are unpredictable. Aleatory uncertainty is often treated as irreducible because of this perception. The term "aleatory" was derived from the Latin *alea*, which translates to English as "die" (i.e., the singular form of "dice"). An engineering example of a source of aleatory uncertainty is Young's modulus of a material. Although Young's modulus is reported as a constant, there is variability between material samples due to the manufacturing process. Variability in Young's modulus can be reduced only by improving the manufacturing process, not by simply observing measurements of the material samples. The term "epistemic" comes from the Greek "episteme", meaning knowledge; hence, epistemic uncertainty can be reduced by acquiring additional knowledge. An example of an epistemic uncertainty source is a calibration parameter in a deterministic computer model of a physical system. The value of the calibration parameter, such that the model predictions will match reality, is uncertain. After

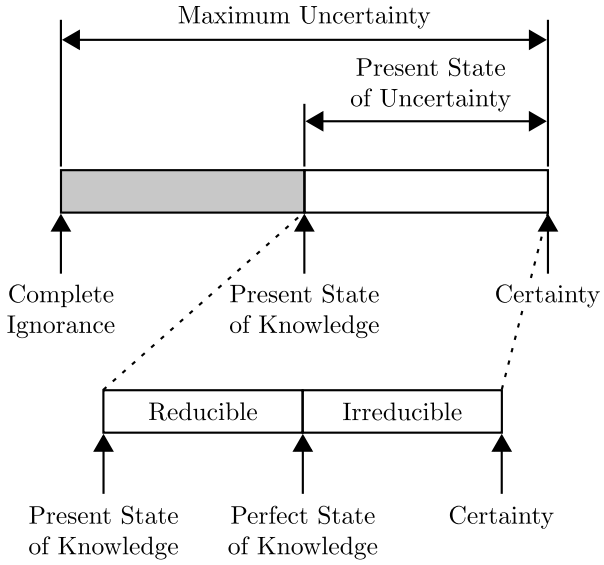


Figure 1: Depiction of the definition of uncertainty (adapted from Ref. [11]).

obtaining data from physical experiments, discrepancies between the model predictions and the system behavior can be minimized using a calibration process. Another example of an epistemic uncertainty is a person's knowledge of the current population of Atlanta, which would vary from one person to another. One could reduce this uncertainty by observing the latest U.S. census results. These examples help to illustrate that epistemic uncertainty is a property of an observer and not the observable. Aleatory (irreducible) and epistemic (reducible) uncertainties are the components of a decision maker's total uncertainty, as illustrated by the bottom bar of Fig. 1.

The research presented in this dissertation requires quantitative representation of aleatory and epistemic uncertainties. Although it is generally accepted that probability theory is appropriate for representing aleatory uncertainties, epistemic uncertainty representation is a controversial topic. Multiple frameworks for epistemic uncertainty representation have been proposed, such as interval analysis, evidence theory, and possibility theory. However, there is not a unified, authoritative position regarding which technique is the most appropriate. Probability theory is used to represent aleatory and epistemic uncertainties in this dissertation because it provides a well-established

mathematical framework for quantifying uncertainty, and, as argued by Zang [13], it seems that many engineers and scientists (decision makers) desire answers to their questions in terms of subjective, probabilistic interpretations of uncertainty.

1.1.1 The Nature of Technology Integration Impact Uncertainty

The integration impact of a technology must be observed using a scientific approach to reduce uncertainty. When the integration impacts are measured in a physical experiment, measurement errors are present due to bias (epistemic) and precision (aleatory) sources [14]. If a deterministic computer model is used to estimate technology performance, then uncertainty surrounding the output can be due to model input variables (epistemic/aleatory), model form (epistemic), and numerical approximations (epistemic) [15]. With any kind of activity designed to characterize technology integration impacts, there is an additional layer of epistemic uncertainty associated with forecasting the future.

Since the future performance of a technology is of particular interest for stakeholders, forecasting is a necessity. The result of any technology forecasting exercise is laden with epistemic uncertainty because the impacts of the technology on the system cannot be known with a high degree of certainty until the technology has been fully integrated with the system and demonstrated in real operations. Prior to this, knowledge of things such as technology design variable settings, system design variable settings, and technology scalability is limited.

Kirby and Mavris [16] proposed a model that helps to illustrate how technology integration impact uncertainty changes with technology maturation. They proposed the characterization of technology impact uncertainty with Weibull distributions that are a function of technology readiness levels (TRL)s. TRLs are a commonly used maturity metric for technologies, and they are described further in Chapter 2. A notional example of their approach for wing weight reduction, due to the use of a composite

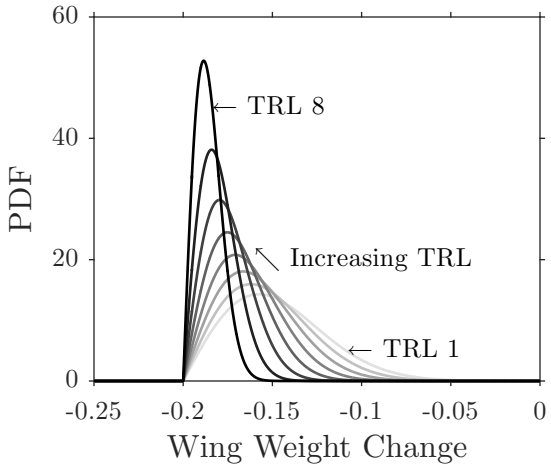


Figure 2: Notional technology impact distributions for a range of TRLs (adapted from Ref. [16]).

structure on a supersonic transport aircraft, is shown in Fig. 2. As illustrated by the probability density function (PDF) shapes in the figure, they argued that relatively large uncertainty is present at lower TRLs, when there is significant lack of knowledge about the technology. As TRL increases with knowledge acquisition, the variability of the distributions reduces, and the mode value shifts toward an expert-defined value at the highest maturity level of TRL 9. The expert-defined value, 20% wing weight reduction, represents the desired level of performance improvement for the technology.

1.1.2 Risk Management

For a system development program, risks are commonly grouped into three categories: schedule, budget, and performance. In addition to demonstrating improved performance and maturation, one of the goals of technology development is to reduce these risks for subsequent research [9]. If technology development is not executed properly, the product development program that incorporates the immature technology can encounter schedule delays, cost overruns, and performance shortfalls. For example, the F-22 program was justified as being a more capable aircraft than other fighters of the time. This motivated the inclusion of immature technologies that are critical

to the F-22's performance and distinguish it from other fighters. According to a 2006 U.S. Government Accountability Office (GAO) analysis of the F-22 aircraft program, technology maturation issues substantially contributed to the 189% cost growth per aircraft that the U.S. Air Force incurred [17]. The Air Force responded by reducing the number of procured aircraft by over 70%.

Successful organizations avoid undesirable consequences by managing technology risk before transitioning a technology to a product development program. Generic options for risk management include risk reduction, risk transfer, self-retention, and risk avoidance. Technology risk reduction is often implemented by laboratories within the same company as the product developers or in external organizations such as NASA. Common technology risk reduction phases are exploration, development, and transition of technologies. Figure 3 illustrates the relationship between technology development and product development activities. The exploration phase is where technology application ideas are proposed and evaluated based on factors such as relevance to future products, competitiveness in the market, cost, manufacturing issues, and life cycle management issues. The development phase entails improving understanding of the technology, maturing the technology (typically to TRL 6 or 7), and refining the solution with a particular product line in mind. Once the technology has been demonstrated in an operational environment, decision makers must determine whether their confidence in the success of the technology for a given product line is enough to warrant transition.

In order to build confidence in the success of a technology, the uncertainty component of risk is reduced. This requires costly experimentation and research. For example, over 17% of the $69.7 billion (2013 dollars) U.S. Department of Defense (DoD) budget for research, development, test, and evaluation was allocated for technology development in fiscal year 2013 [18]. The time required to reduce technology risk to an acceptable level for transition is also substantial; the time from start of

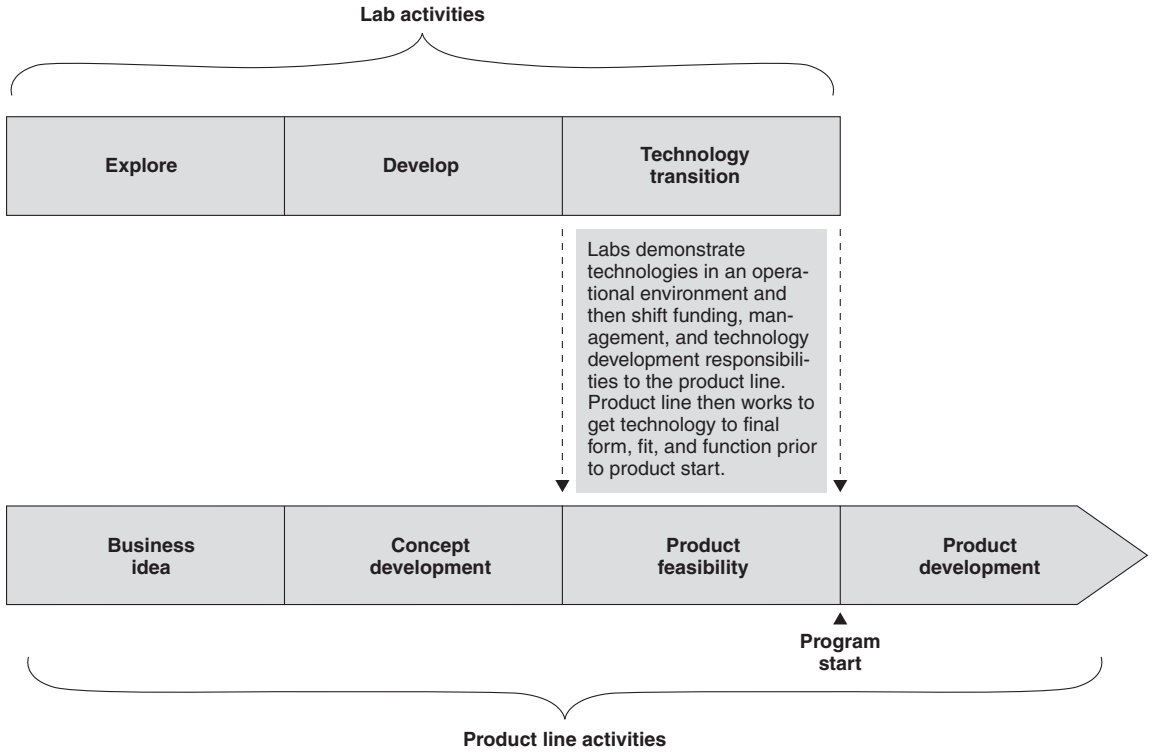


Figure 3: The relationship between technology development (top) and product development (bottom) (from Ref. [17]).

technology development to transition is frequently more than ten years. Naturally, accelerating the technology development process is a goal in both commercial and military organizations. Successful implementation of novel technologies by a company before others can result in valuable gains in market share. In the military, accelerating technology development can help countries keep pace with or exceed the technological progress of adversaries.

1.2 Technology Development Activities

The development progress of advanced technologies viewed over time or effort has been shown to follow an S-shaped evolutionary path [19], as notionally depicted in Fig. 4. Performance of technologies is relatively poor at the initial stage, then it improves rapidly as research and development is conducted. The growth becomes approximately linear once a threshold of knowledge and understanding is reached. At

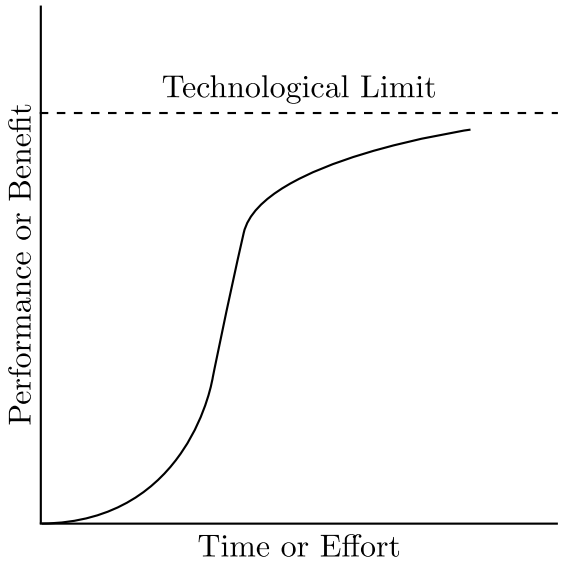


Figure 4: Notional technology S-curve evolution.

this point, the technology is commercially exploited. With additional time or effort, improvements become more difficult to achieve, and the technology asymptotically approaches a technological limit. This limit can be set by social considerations, such as safety regulations, or it can be due to constraints imposed by nature.

During the research and development stages of technology evolution, reduction of technology integration impact uncertainty is effected through the acquisition of knowledge and understanding of the technology by conducting development activities, such as computer experiments, physical experiments, tests, and system studies. Experiments and tests can be differentiated by their objectives. Experiments are often performed to improve the understanding of a physical process, improve mathematical models of well-understood physical processes, or to validate mathematical models, whereas tests are usually conducted to measure the goodness of a particular component, subsystem, or system in terms of reliability, performance, or safety [12]. The relationship between data that are generated by development activities and the knowledge built from them provides the connection between technology risk reduction and development activities.

An example of real technology development activities that were conducted by

NASA and Boeing for an AFC technology is depicted in Fig. 5. The integration approach for this technology is to install an AFC architecture on board an aircraft to control flow separation over the vertical tail, thereby substantially increasing directional control authority. The vertical tail of a commercial transport airplane is sized to counteract asymmetric thrust during rare engine failure scenarios at low speeds, and it is oversized for the majority of flight conditions during a typical mission, such as cruise. Also, the vertical tails of many transport aircraft families are the same size within a given model family. The vertical tail of a model family is sized for the member with the shortest fuselage length. Thus, the longer family members carry an oversized vertical tail due to the longer moment arms of each. Integration of an AFC architecture with sweeping jet actuators has the potential to enable vertical tail area reduction for the entire model family while still meeting the constraints of emergency scenarios. The primary benefit of this integration approach is that the drag of the vehicles will be reduced throughout the flight envelope, resulting in fuel cost savings for airlines. Sub-scale wind tunnel experiments were completed to graduate the technology to a maturity of TRL 4, a full-scale wind tunnel experiment was conducted to mature the technology to TRL 5, and the technology was demonstrated in flight to achieve TRL 6. Depending on the risk tolerance of a system integrator, technologies are typically transitioned from the technology development program to the system development program at TRL 6 or 7. As the development activities were completed, greater confidence was gained in the side force enhancement achieved by the AFC technology, and this is depicted in the top of the figure. As epistemic uncertainty surrounding this integration impact was reduced, the epistemic uncertainty surrounding the cruise drag reduction for a vision large twin aisle (LTA) airplane integrated with the AFC technology was also reduced.

As a technology development program progresses and epistemic uncertainty is reduced, the technology should mature through demonstration in increasingly complex

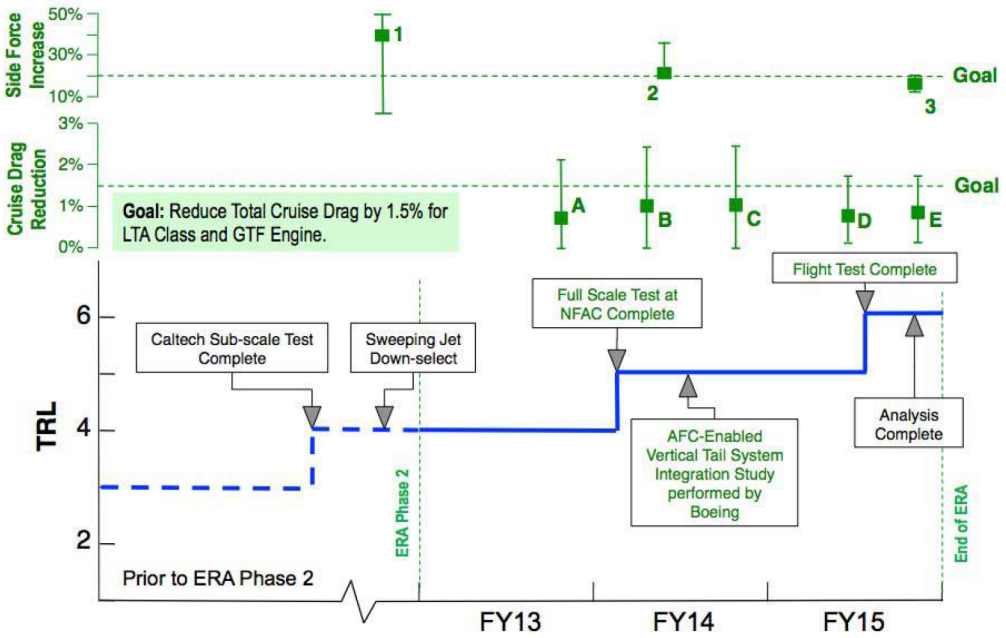


Figure 5: AFC technology development activities conducted during the ERA project and the associated technology integration impact uncertainty (from Ref. [20]).

integration scenarios. Also, the S-curve model suggests that performance of a technology should be improved through development activities. If development activities do not help technologists achieve these objectives, then resources will be squandered, and a critical technology may not be transitioned to a system. Or, a poorly developed technology may be transitioned and negatively impact the system development program. This result highlights the importance of intelligently designing the development activities, which is a perennial problem in technology development. This problem is summarized with the following question that motivated the work in this dissertation.

Motivating Question: How should technology development activities be designed?

This problem is difficult for a number of reasons. One of the key reasons is that there exists a wide variety of alternatives. Largent [21] enumerated some possible

classes of technology development activities. These activities varied greatly in complexity and cost. For example, at the early stages of development a simple paper study may be appropriate to define the technology and estimate performance using low-order methods. As the technology matures, expensive physical experiments and demonstrations may be of more value. There are some activities that Largent claimed would be appropriate for any maturity level, including design space exploration using physics-based models and creation of analysis capabilities for predicting technology performance. Even when a class of activity has been chosen, there can remain many degrees of freedom for fully specifying the activity to be conducted. For example, defining a physical experiment requires selecting the facility, the scale, the measurement equipment, the experimental apparatus, the independent and dependent variables, etc. Given the large variety of development activities, it is important to have criteria to compare alternatives. However, some or all of the criteria that characterize value of the activities are not easily quantified before the activities are conducted, and some of the criteria may be conflicting with others. For example, how can one estimate the performance improvement and uncertainty reduction obtained from conducting a given set of development activities, and how will these criteria conflict with cost and time? There is an additional layer of difficulty associated with the fact that there is uncertainty surrounding the attributes of each alternative. For instance, the cost and time required to complete a specific type of technology development activity may be known with low precision before details of the activity are determined. And finally, the number of alternatives available to decision makers grows with the number of technologies that are being developed within a program. Nevertheless, decision maker must somehow prioritize technologies since not all of the technologies in a given program will be affected by the selected development activities.

1.3 Research Objective

The objective of the research presented in this dissertation is to establish a framework for designing technology development activities that improves the state of decision support capabilities. As a first step toward meeting this objective, research gaps in the literature are identified in Chapter 2. Then, a novel framework for addressing these gaps is presented, and the need for multiple quantitative capabilities to add rigor to the framework is defined in Chapter 3. Next, three contributions toward improving the framework are presented in Chapters 4, 5, and 6. Arguments are established in the beginning of each of these three chapters. The common theme of the claims is that these contributions fill previous capability gaps in an appropriate manner. Finally, the contributions of this dissertation are summarized, limitations are discussed, future research opportunities are described, and the overarching thesis statement is presented in Chapter 7.

CHAPTER II

LITERATURE REVIEW

In the first part of this chapter, the existing body of research that is pertinent to the design of technology development activities is reviewed. Then, current practices are discussed. Finally, research gaps are identified with regard to what is needed to address the motivating question from the introduction chapter.

The literature that is germane to the problem area of interest can be partitioned into two categories: writings that provide guidance for designing technology development activities and those that present methods for designing a portfolio of tests and experiments in a system development program. The latter body of literature is considered relevant because although the focus is on integrated system development, one of the cardinal objectives of executing both tests and experiments in this context is to reduce uncertainty surrounding system performance.

2.1 Overview of the Literature: Guidance for Designing Technology Development Activities

Existing guidance for the type of technology development activities that should be implemented as a technology matures is based on technology readiness level (TRL) scales. Introduced by NASA in the mid-1970s, TRL is a discipline-independent, programmatic figure of merit that is used to assess and communicate the maturity of technologies [22]. The first TRL definitions were later established in a white paper by Mankins [23]. Since then, TRLs have been widely used by government and international organizations such as the U.S. DoD, the U.S. Department of Energy, and the North Atlantic Treaty Organization. TRL scales are ordinal, with most containing the integers 1 through 9. A TRL of 1 represents the lowest maturity level and a

Table 2: NASA technology readiness levels (definitions from Ref. [24])

TRL	Definition
1	Basic principles observed and reported
2	Technology concept and/or application formulated
3	Analytical and experimental critical function and/or characteristic proof-of-concept
4	Component and/or breadboard validation in laboratory environment
5	Component and/or breadboard validation in relevant environment
6	System/subsystem model or prototype demonstration in a relevant environment (ground or space)
7	System prototype demonstration in a target/space environment
8	Actual system completed and "flight qualified" through test and demonstration (ground or flight)
9	Actual system "flight proven" through successful mission operations

TRL of 9 represents a technology that has been proven through successful mission operations. Each readiness level is accompanied by a definition that indicates the development status of a technology. An example TRL scale from NASA is shown in Table 2.

There is not an explicit indication of the uncertainty or risk associated with each maturity level in most TRL definitions. The NASA TRL definitions in Table 2 serve as an example of this fact. Additionally, TRL scales do not capture how difficult or costly the maturation process is. Mankins [25] introduced a measure called the "Research and Development Degree of Difficulty" (R&D^3) to help decision makers understand and communicate the difficulty and likelihood of achieving technology research and development (R&D) objectives. The R&D^3 scale is also ordinal, with integers from 1 to 5. A level of 1 means that a very low degree of difficulty is anticipated and that the probability of success with "normal" R&D effort is 99%. On the opposite end of R&D^3, level 5 means that the anticipated difficulty is high enough that a fundamental breakthrough is required and that the probability of success with "normal" R&D effort is 20%. Mankins [9] later developed an approach for integrated technology readiness and risk assessment using TRL, R&D^3, and an additional factor

called the technology need value (TNV). The TNV scale has five levels, with each corresponding to a weighting factor based on the importance of developing a given technology. The importance can be related to system-level improvement or the potential for the development efforts to support decisions. TRL, R&D^3, and TNV were fused into a summary of technology readiness and risk using a risk matrix. Here, Mankins [9] defined risk as the combination of the probability of technical failure and the consequences of failure. Probability of failure is related to R&D^3, and consequence is quantified as TNV multiplied by the difference between the current TRL and the TRL that must be achieved before the transition phase. Using the combination of quantified consequence and probability of failure, one can identify a location in the risk matrix to obtain a qualitative description of risk. This risk description can then be used to inform management decisions regarding future development activities. Moorhouse [7] established detailed TRL definitions to provide guidance on how to graduate technologies from one level to higher levels. In his definitions, he elaborated on the types of physical experimentation that should be conducted and the numerical models that should be constructed at each TRL. His definitions also provide qualitative descriptions of the risk and uncertainty associated with system-level integration of a technology as well as tasks for predicting technology-related costs at each level.

Largent [21] identified a need for a process focused on planning and managing technology development, and he hypothesized that it is possible to reduce uncertainty and programmatic risk in a technology development program by implementing such a process. He argued that the process should link performance, cost, and schedule uncertainty so that all three key dimensions can be accounted for when making decisions about development activities. He also claimed that technology-level development activities should focus on reducing system-level uncertainty to ensure that at the end of development, risk has been reduced to a point where the technology is

ready for system integration. Based on the identified need, Largent [21] formulated the Technology Development Planning and Management (TDPM) process. There are two foci of the TDPM process. The first focus is to provide a way to systematically identify technology performance uncertainties and to plan activities to reduce the uncertainties and maximize performance. The second focus is to provide a method for assessing risks in the initial development plan and to reassess the plan as development activities are completed. The TDPM process begins with defining the technology. The purpose of the first step is to gather existing information about the technology being developed through literature search, input from experts, etc. This information is divided into seven categories: defining the need for the technology, describing the physics that govern the operation of the technology, describing the way that the technology integrates with the system, identifying similar technologies, identifying previous development efforts, identifying analysis capabilities for modeling the technology and system, and identifying the current TRL for the technology. The second TDPM process step begins with identifying performance metrics at all levels of the system hierarchy, such as the technology level, the subsystem level, the system level, and the system of systems level. After identification of relevant metrics, the uncertainty associated with technology-level metrics is characterized using a probabilistic approach. The third step entails the propagation of technology-level uncertainties up to the system level and prioritization of uncertainties based on the contribution of each uncertainty source to higher level metrics. The fourth TDPM process step, mitigate technical uncertainty, is where activities such as numerical and physical experiments are planned to reduce uncertainty. Largent [21] claimed that development activities should be designed for the purpose of targeted uncertainty reduction, based on the results of step three, but there was no method identified for how this should be accomplished. However, he did provide some guidance as to what kind of development activities should occur at a given TRL. Steps five through seven

involve quantifying and assessing project risks to determine whether to proceed with the uncertainty mitigation plan or to modify the development activities so that the risks are acceptable. After devising a plan with acceptable risks, step eight is conducted. In the first phase of step eight, activities are completed up to a particular milestone or date. The second phase of step eight entails updating the uncertainties for activities that have been completed. In the last phase of step eight, a risk analysis is performed to determine whether another iteration is required that would begin at the planning stage (step four) or at the beginning of the TDPM process. Largent [21] explored the use of Bayesian inference for updating uncertainties, and he suggested further examination of updating techniques in future work. He also identified the use of joint probability distributions to capture correlation of performance metrics as possible future work.

Gatian [26] formulated a methodology to aid in risk-informed decision making during a technology development program. Her methodology addressed technology experimentation in addition to three other phases of technology development. Technology readiness assessment was combined with a probabilistic approach to quantify uncertainty to aid in the identification of experimental goals. The readiness assessment involves morphological analysis of the experiment characteristics and existing TRL definitions to identify the kind of experimentation that is appropriate for each TRL. The technologies that are part of the development program are prioritized according to the readiness risk and performance risk of each. Technologies that are preferred for experimentation are those with a combination of relatively high readiness risk and large contributions to the uncertainty surrounding the system-level metrics. Once one or more of the highest-priority technologies are selected for development activities, the next step in her methodology is to identify the objective of the experiments. The objective is selected based on which sources of uncertainty should be targeted and the current TRL of the technologies. As in Largent's [21] approach, the

details of the experiment design are left to technologists.

2.2 Overview of the Literature: Test and Experiment Selection for System Development

Wong [27] presented a methodology for providing decision makers with test resource allocation guidance by prioritizing the testing of subsystems within a large system. He argued that some subsystems should be tested more exhaustively than others and that the highest priority should be given to the subsystems that have the largest performance uncertainty and greatest effect on system performance. Two forms of the methodology were discussed. The first form, called the "extensive" form, has three phases. In the first phase, the sensitivities are modeled. The subsystems at all levels of the system hierarchy and their interactions are defined, and performance indices are established for each subsystem. A model of each performance index for every subsystem, as a function of lower level subsystem performance indices, must then be identified for all levels. In the second phase, probability density functions PDFs are used to characterize uncertainty in the performance indices. Wong [27] suggested eliciting the distributions from the developers of the system. The third phase entails determining test priorities. First, the total uncertainty in the system performance is quantified by propagating uncertainty from all of the lower level subsystems. Next, the uncertainty in system performance is computed as a PDF under the assumption that the ith subsystem operates at full capability with a probability of one. Then, the value of testing the ith subsystem is determined by calculating the difference between the expected performance of the system with the ith subsystem operating at full capacity and the expected performance of the system with total uncertainty. This is repeated for all subsystems. An alternative form of the methodology, called the "diminutive" form, offers multiple simplifications to reduce the cost and difficulty of using the "extensive" form. The focus of testing in his research is to reduce uncertainty surrounding the performance of a system that has already been developed.

Although this test resource allocation problem is different from the problem of interest in this dissertation, an important concept from his work is that the value of testing a subsystem depends on the uncertainty in the subsystem performance and the sensitivity of the total system performance to subsystem performance. His paper may be the first in the literature to mention this idea.

Thomke [28] studied how the economics of experimentation in product design have been affected by the use of multiple methods, or what he called "modes", for conducting experiments. Examples of different modes are physical experiments with prototypes, numerical experimentation, and thought experiments. He presented a generic, iterative experimentation cycle with four steps: designing the experiment, building the physical or virtual apparatus, running the experiment, and analyzing the results. At the end of each cycle, the analysis step indicates whether the quality of a design can be improved cost-effectively. If improvements can be made, then the design is modified based on what was learned from the experiment. However, at some point, the cost required to improve the design will outweigh the benefit of improvement. To help study the effectiveness of this experimentation cycle, Thomke [28] coined the term "experimentation efficiency", which is a ratio of the economic value of information gained during the experimental cycle and the cost of conducting the experiment. He proposed that as a design process progresses and experimental cycles are repeated using a given mode, the experimentation efficiency decreases because of diminishing returns from experimentation in that mode. And, he proposed that different modes can exhibit different rates of decay in experimentation efficiency. A notional illustration of these propositions that has been adapted for the technology development context is shown in Fig. 6. Note that this figure shows the overall value of information gained from the activities to the stakeholders. In Fig. 6, the point at which the curves cross was termed the "optimal switching point" by Thomke [28]. This is where one would ideally transition to the more efficient (higher value) mode

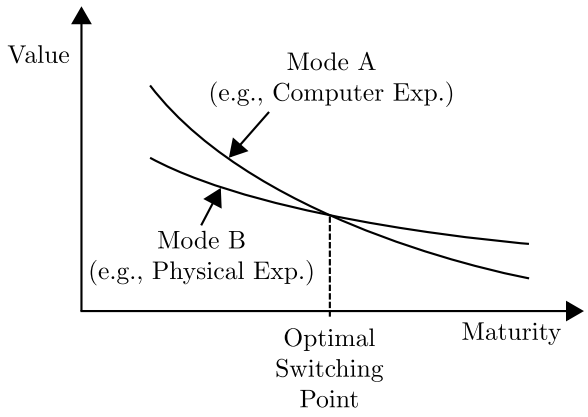


Figure 6: Notional relationships between the value of development activities and maturity for two different modes (types) of activities (adapted from Ref. [28]).

as an economic strategy. He proposed that finding this point can result in significant improvements in innovation cost and time. Thomke [28] conducted an empirical study that involved the collection of data on experimentation strategies from several hundred integrated circuit designers in the U.S., and the data strongly supported his propositions. His research findings are more relevant to the product development phase shown in Fig. 3 than they are to technology development, but the concepts are valuable for the design of technology development activities. In a technology development program, the modes of experimentation could be not only numerical and physical experimentation, but also the various levels of fidelity for each type. Also, one idea explored by him is that the optimal mode switching point can shift if the efficiency curve of a given experimentation mode changes. This idea can be useful in technology development as well. For example, the experimentation efficiency curve of numerical models can shift as data from physical experiments is used to validate the models, perhaps making the use of the numerical experimentation mode more appealing.

Thomke and Bell [29] developed mathematical models to study how to select the optimal frequency, timing, and fidelity of sequential testing activities for product design. In their work, test fidelity refers to the ability of a test to uncover problems

with a product design, and fidelity varies with the "completeness" of a test prototype and/or the realness of the testing environment. A full-fidelity test can detect all accrued problems at a given time in the development project, whereas a low-fidelity test can only detect a fraction of the problems. The mathematical models were designed to capture the behavior of total project cost due to testing and rework for solving problems identified from tests. The problems referred to by the authors are consequences of a lack of complete understanding of customer expectations, called customer uncertainty, or difficulty in predicting the feasibility of a design before testing, called technical uncertainty. These sources of uncertainty result in the accumulation of problems in a product design that are only revealed after physical or virtual testing. Three relationships between sequential low-fidelity tests were considered. One, called the fully overlapping case, is where the set of problems identified in an earlier test are also contained in the later test. Another, called the partially overlapping case, is for a scenario in which only a fraction of the problems identified in an earlier test could be "rediscovered" in a later test. The last, called the complementary case, is where none of the problems identified in the earlier test are detectable in the later test. The authors exercised their mathematical model to determine testing strategies that minimize total cost in different scenarios. Analysis of the model resulted in three key observations. The first is that the optimal testing strategy depends on the behavior of the redesign cost as a function of time, test cost as a function of fidelity, and the relationship between sequential tests. The second key observation was that the optimal number of tests is approximately the square root of the ratio of avoidable cost and the cost of a single test, where the avoidable cost is the difference between total cost with testing and rework only at the end of the development project and total cost with continuous testing and rework throughout the project. The third key observation is that the relationship between sequential tests affects the testing strategy. Thomke and Bell [29] found that few high-fidelity tests are optimal for sequential

tests that are refinements of one another (fully overlapping case), whereas a larger number of low-fidelity tests are ideal for complementary tests. They analyzed scenarios that are similar to a technology development program, where decision makers must select the fidelity of the experiments, when they occur, and how many to do. For this reason, the insights gained from their analysis are valuable for the research in this dissertation.

Loch et al. [30] studied the problem of how to determine the ideal amount of parallel and serial testing that is conducted during a product design process. In their work, the purpose of testing is to reduce uncertainty surrounding which design is the most preferred solution out of a set of multiple design alternatives. Loch et al. [30] based their work on the premise that parallel testing can require less time than serial testing but does not leverage the learning between tests that serial testing can, resulting in a larger number of required tests. They modeled this trade-off as a dynamic program in order to derive an optimal testing strategy that minimizes the total cost of testing and time. In their mathematical model, they accounted for test cost, lead time, prior knowledge of designers, and learning between tests. Acknowledging the fact that no one test can fully disclose whether a design alternative is the most preferred solution, Loch et al. [30] quantified the remaining uncertainty by employing a measure from information theory called entropy [31]. Using their model, they showed three important insights, two of which are relevant to this dissertation. First, they demonstrated that more expensive tests make sequential testing more economical, whereas slower tests make parallel testing more attractive. Second, they found that the selection of lower fidelity tests increases the appeal of sequential tests. Here, the term fidelity was defined as the ability of a test to identify a design alternative as the most preferable. The research by Loch et al. [30] is more applicable to evaluating alternative product designs than to technology development activity design, but the insights that they produced are pertinent. Also, their work is one of the earliest in the test selection

literature that uses the entropy measure to quantify uncertainty.

Urbina et al. [32] developed a methodology for allocating resources to increase confidence in performance predictions for high-consequence systems. Their definition of confidence was associated with the amount of uncertainty surrounding the probability of failure relative to an established threshold. In their work, resources could be related to physical experiments, model simulations, or model refinement. The type of problems that motivated their work involved systems with multiple components at various hierarchical levels, limited available data from expert input and physical experiments, and limited or no data at the system level. Urbina et al. [32] formulated the resource allocation problem as an optimization problem. They linked the objective function with measures of confidence in the system-level performance predictions. It is noteworthy that they specifically identified entropy as a potential metric to be used in the objective function as well. In their optimization problem, design variables could be related to increased budget for testing, refinement of computational models to reduce errors between model predictions and existing experimental data, or alternative models for representing the physics of the system. The only constraint included in the optimization problem was that total cost, as a function of the design variables, should be less than or equal to a given budget. They utilized a probabilistic model called a Bayesian network to quantify epistemic and aleatory uncertainties at all levels of the system hierarchy, including uncertainty in the performance metric that fed the objective function. A nice feature of Bayesian networks is that one can easily update probabilities in the network with new observations. However, the methodology proposed by Urbina et al. [32] hinges on how new observations are "virtually" generated when solving the optimization problem. For example, if the design variables are mapped to additional physical experiments, then their methodology requires simulation of the acquisition of data from the experiments. Specifying a PDF for a physical experiment that has not yet occurred could be problematic, particularly

in a technology development program.

Motivated by the difficulty of planning test programs months or years in advance of testing for a System of Systems, Hess and Valerdi [33] presented a framework for quickly and effectively planning and re-planning as information is obtained. Under the assumption that an organization is more likely to run out of time than money, they focused on the problem of scheduling. In order to select the best schedule with their framework, they needed to determine value for each test. Hess and Valerdi assumed that the purpose of testing is to determine whether a system under test has passed or failed each of its measures of performance, and they described the value of a test by its ability to reduce uncertainty surrounding this objective. The ideas presented by Hess and Valerdi are intriguing, but their paper left many elements of the framework for future work.

Sankararaman et al. [34], Sankararaman [35], and Sankararaman et al. [36] developed a methodology that extended the work of Urbina et al. [32] to multiple levels of models and tests. These authors also employed a Bayesian network for uncertainty quantification that linked computational models at all levels of the system hierarchy. An additional step was added to the methodology for selecting important types of tests to consider in the resource allocation problem. In this step, global sensitivity analysis was employed to identify important model parameters that significantly contribute to the system-level performance prediction uncertainty. The tests that could reduce uncertainty in these parameters were then used in an optimization problem for test resource allocation. Their optimization problem involved minimizing the expected variance in the system-level prediction by testing within a given total budget. This optimization problem answered the questions of which tests to do and how many repetitions of each test to do in order to capture inherent variability. Their approach also required the simulation of "virtual" test data to quantify the objective function. An important observation made by the authors after exercising their methodology

on example problems is that the uncertainty reduction achieved beyond a given investment may not be worth the additional resources. In other words, there could be diminishing return on investment.

Bjorkman [37] and Bjorkman et al. [38] formulated an approach for allocating test resources in a DoD test and evaluation program. In the first step of their methodology, test objectives are defined for each test required in a portfolio. The second step requires identification of multiple alternatives for each test. Step three entails determination of the portfolio cost constraint and estimation of technical uncertainty reduction for each test alternative. Based on a thorough review of the uncertainty measure literature, they decided to use entropy to quantify uncertainty reduction in this step. In step four, one alternative is selected for each test event that needs to occur. For a small portfolio, subject matter experts (SMEs) may be able to select the optimal combination of tests. But, for a large portfolio, they proposed resource allocation by solving an optimization problem. The last two steps of their methodology involve optional sensitivity analysis and further analysis. Using a case study, they showed that the methodology was easily applied and could generate optimized test portfolios with 5%–20% more total value than those selected by SMEs.

2.3 Analysis of the Literature

Most, if not all, of the research in the literature pertaining to guidance for creating technology development activities uses TRLs in some way. TRLs are an accepted way to capture technology maturity, and the definitions provide guidance for the fidelity, or realness of the experimental environment, of activities that must occur to graduate each level. Although TRL definitions provide some valuable guidance as to what kind of development activities should be conducted to mature a technology, they are still too vague to help decision makers and technologists design a set of development

activities in a defensible manner. Largent [21] and Gatian [26] formulated comprehensive technology infusion methodologies that include more specific guidelines for planning and managing technology development activities. A common theme in these processes is an emphasis on informing decisions about which development activities to pursue by prioritizing sources of technology-level performance uncertainty (i.e., integration impacts) and tracking the likelihood of meeting system-level performance goals. Prioritization of uncertainty sources is based on how sensitive the uncertainty surrounding system-level performance metrics is to technology-level uncertainties. Because building system prototypes integrated with immature technologies is not always viable, system-level performance effects and sensitivities are quantified with physics-based modeling and simulation (M&S) environments. Gatian [26] also identified sensitivities of the probability of meeting established system-level performance goals as being useful for prioritizing technologies for activities. This information is valuable as well because the purpose of integrating most advanced technologies is to close an established system-level performance gap.

The literature on test and experiment selection for system development provides important insights and approaches. A few of the papers present methodologies that come close to being directly applicable to selecting activities in a technology development program. The goal of testing in Refs. [32, 34, 35, 36, 37, 38] is to reduce uncertainty surrounding model parameters in order to reduce uncertainty in system-level performance predictions. Based on this objective, tests can be selected using optimization or experts to maximize the amount of uncertainty reduction obtained. An analogous problem is found in technology development. However, the types of systems tested in the aforementioned references are well defined and have been fielded or close to it, whereas in technology development program, the technology and the system that adopts the technology can start out with relatively low maturity. As TRL increases, the types of development activities may change in order to reduce

uncertainty, improve performance, and further mature the technology. For example, at a low TRL a sub-scale exploratory experiment may be the most appropriate choice for reducing performance uncertainty and increasing TRL, but at a high TRL a full-scale flight test with fewer available design degrees of freedom may be needed to reduce important uncertainties and increase TRL. The methodologies in Refs. [32, 34, 35, 36, 37, 38] do not capture these kinds of imperative considerations.

2.4 Current Practices for Designing Technology Development Activities

Although the planning and implementation details of technology development programs are not published due to proprietary concerns, there is evidence that current practices involve TRL-driven decision making. The GAO's knowledge-based acquisition practices require that technologies achieve TRL 7 before transition to the system development program [39]. With TRL maximization as an objective, a typical development activity design process involves establishing a goal for technology performance and designing a series of activities to demonstrate technology performance at multiple TRLs. The AFC-enhanced vertical tail technology discussed in Sec. 1.2 serves as a recent example. The goal of the development program implemented by NASA and Boeing "was to use AFC to achieve a substantial increase in the control authority of the vertical tail of a commercial transport airplane" [40]. With this goal in mind, the technologists designed three phases of activities to push the maturation of the technology from TRL 3 to TRL 6: sub-scale wind tunnel experimentation, full-scale wind tunnel experimentation, and flight experimentation.

The GAO's use of TRL in the definition of a knowledge point in its acquisition practices implies that TRL can be used as a measure of the degree of knowledge gained about a technology for supporting decisions. However, one of the criticisms of TRL scales is that they do not convey the uncertainty associated with graduating each level [41]. Thus, many alternative versions of development activities can

be designed to satisfy TRL definitions, but it is possible that the alternatives will be unequal in terms of uncertainty reduction and other characteristics, such as performance improvement. As an analogy, consider a student's development through schooling. The grades of school can be thought of as TRLs, and the use of TRL to represent the degree of knowledge gained is analogous to the use of grade level that a student passes as a measure of the student's knowledge. Not all school curricula are created equal, and despite the fact that a student can pass two different curricula for a specific grade, the student's knowledge and understanding of the subjects hinges on the quality of the learning process.

2.5 The Need for a Technology Development Activity Design Process

The primary problem with relying on TRL definitions for design decisions about future development activities for a technology is that TRL scales do not characterize the state of uncertainty surrounding the integration impacts of a technology. Ideally, epistemic uncertainty will be maximally reduced with each activity. The more epistemic uncertainty reduction achieved for a technology with integration benefits that clearly outweigh the costs, the more likely decision makers will be inclined to fund further development activities and to ultimately transition the technology to a system development program. With a focus on increasing TRL, the extent to which current practices for designing development activities target epistemic uncertainty is simply to meet the requirements of TRL definitions.

The quantitative techniques developed in the literature improve upon the current practices because they provide additional decision support information about what sources of uncertainty are most important and the impact of uncertainty on the knowledge of closing system-level performance gaps. However, this additional information can be difficult for decision makers to synthesize with their preferences to arrive at decisions, and much of the activity design decisions are delegated to technologists.

Hence, the methodologies from the literature fall short of providing a complete path to designing a portfolio of technology development activities.

The use of TRL as a representation of the degree of knowledge about a technology can result in misinformed decisions. Without a clear understanding of the state of knowledge about technology impacts, it is difficult for decision makers to determine what development activities to pursue to reduce uncertainty. A more serious consequence is that poor design of development activities may place the success of a promising technology in jeopardy. To increase the likelihood of development success, a novel approach is needed that incorporates methods from the literature for characterizing technology uncertainty and leverages this information to better inform the design of development activities. To fulfill this need, a framework is formulated in Chapter 3.

CHAPTER III

A NOVEL FRAMEWORK FOR DESIGNING TECHNOLOGY DEVELOPMENT ACTIVITIES

In technology development programs, decision makers must plan a series of activities that will contribute to achieving goals of uncertainty reduction, performance improvement, and maturation. There are three primary questions that must be answered to design technology development activities to achieve these goals:

1. *Which types of activities should be selected?*
2. *What is the best setup of the physical or computational environment for each activity?*
3. *How should each activity be executed to maximize the value of information that is generated?*

It is proposed that the design of technology development activities be divided into three phases that correspond with answering each question: (1) thought experimentation, (2) detailed definition of the activities, and (3) statistical design of experiments. Each of these three phases are discussed in the following sections. Then, the framework is applied to a case study to derive new insights about better ways an actual technology development program could have been conducted. Finally, opportunities to add rigor to the framework are discussed.

3.1 Phase 1: Thought Experimentation

It is ideal to be able to select the types of activities that will meet the goals of technology development while simultaneously minimizing the cost and time required

for the development program. Due to the significant resource expenditure that may be required for the candidate activities, this decision must be made before any of the candidate activities have been designed in detail and conducted. To make this decision systematically, defensibly, and rationally, a mental exercise called a *thought experiment* is necessary. In this section, thought experiments are characterized, and the concept is applied to the technology development context for selecting types of activities.

3.1.1 Literature Concerning the Characterization of Thought Experiments

Sorensen defined a thought experiment as an experiment that "purports to achieve its aim without the benefit of execution" [42]. Thought experiments have been proclaimed as important instruments for the discovery of multiple fundamental laws of physics, such as Archimedes' law of the lever and Einstein's theory of relativity. Some philosophers differentiate between thought experiments in general and scientific thought experiments, which is the type that is well known in physics. Gendler [43] characterized thought experiments as reasoning about a scenario, where the "mode of access to the scenario" is imagination in lieu of observation, with the purpose of testing a hypothesis or theory. She identified an additional feature for scientific thought experiments: the hypothesis or theory involves the physical world. Reasoning about a scenario entails mentally simulating alternative events or actions and their likely consequences. Shepard [44] identified cognitive characteristics of humans that are necessary for thought experiments to be effective with regard to producing new knowledge through mental simulation: (1) a motivation to understand our surroundings, (2) the ability to evaluate alternatives objectively, and (3) language for communicating and analyzing arguments. These observations from the literature can be synthesized to derive a process for a thought experiment. A motivation to gain new knowledge and understanding must first be established. Following the motivation, the

scenario or problem can then be defined. Next, the measures that characterize the consequences of interest need to be identified. Then, the alternative actions or events must be mentally generated. The next step is to evaluate each alternative in terms of the likely values of the measures. The final step is to draw conclusions from the mental observations.

3.1.2 An Example of a Thought Experiment

As an example of the implementation of the thought experiment steps, they are applied to a version of Galileo's law of falling bodies thought experiment presented by Shepard [44]. Prior to Galileo, the reigning claim attributed to Aristotle was that falling object drop toward the ground with velocities that are proportional to their weights. Galileo's motivation was to understand the motion of all objects in nature. The scenario defined in the modified version of his thought experiment was to simultaneously drop objects from the top of the leaning tower at Pisa and observe the motion of each one, while neglecting the effects of air resistance. The measure in this thought experiment that corresponds with the consequence of interest is the difference in the time at which each of the objects hits the ground. The alternative actions were defined to be either dropping three identical bricks at the same time or dropping one of the bricks and the other two bricks, glued to each other, at the same time. To evaluate these two alternatives, one must imaging dropping the bricks at the same time and observing whether there is a nonzero difference in the time of arrival at the ground between any of the objects. There is no reason why the three identical bricks should arrive at the ground at different times. When two of the bricks are glued to each other, they form a larger brick that is twice as heavy (neglecting the relatively insignificant weight of the glue). Would the larger brick fall faster than as the separate third brick? The answer is no, because the glue used to make the two bricks into a larger single brick would not cause the object to fall more quickly. Thus,

the correct conclusion was reached without performing the physical experiment, and Aristotle's claim was refuted.

3.1.3 Application of Thought Experimentation to Technology Development Activity Design

How does the thought experimentation procedure apply to technology development? The steps of a thought experiment can be used to arrive at a conclusion regarding which types of development activities should be pursued. The motivation for pursuing technology development activities in any case stems from the fact that many advanced technologies have the potential to benefit integrated systems, thereby improving profits for manufacturers and operators and ultimately the lives of the end users. Knowledge and understanding of the technology must be gained to build confidence in the benefits of integration to ensure success. Specifics of the problem formulation may vary from one development program to another, but some common elements can be identified. In any program, there will be one or multiple target systems for technology infusion. The purpose of infusing the technologies is to close performance or capability gaps that are present at the system level. By working toward the three primary goals of technology development, the confidence in the magnitude of the technology impacts should increase, and the technology should improve with maturation such that it meets the system-level requirements. Thus, the problem is to design the activities for simultaneous maximization of uncertainty reduction, performance improvement, and maturation, subject to resource limitations. The measures that characterize the consequences of interest naturally follow from the problem formulation. These measure serve as criteria for making a decision in the last step. Measures of potential performance improvement, uncertainty reduction, maturation, and costs or resources are needed. In the current practice, the primary measure of interest is TRL which is not sufficient because it can only satisfy the maturity measure; it does not explicitly characterize any of the other consequences. Next, the alternative actions need to be

generated. The alternatives may only be defined coarsely as a portfolio of activities, with many of the details undetermined. For example, one alternative may be a set of physical experiments at multiple scales for an individual technology. The alternative creation process requires the synthesis of multiple sources of information. The sensitivity analysis information from the techniques described in Chapter 2 should be used to identify the most important sources of uncertainty. There may be data from earlier development activities or even previous development programs that can be leveraged to inform the creation of alternatives. Because TRL is currently critical in development programs, the alternatives will likely be created to meet the constraints of TRL definitions so that the alternatives can be claimed as elevating the technology to a specific degree of maturation after the activities have been conducted. All of this information must be considered so that the alternatives are intelligently created to target the largest sources of uncertainty, leverage learning from previous activities, and achieve maturation goals. After the alternatives have been created, they all need to be evaluated by estimating the consequences associated with each. Since the activities have not been executed at this point, there will be uncertainty surrounding the measures of consequence. The final step of drawing conclusions is making a decision in this context, and the uncertainty surrounding the consequences of each alternative action increases the difficulty of the final step.

These steps form the first phase of the technology development activity design framework shown in Fig. 7. In the next phase, each of the selected activities must be defined in more detail. This process is described in the following section.

3.2 Phase 2: Detailed Definition of the Activities

In the second phase, many characteristics of each activity must be decided to determine the best setup of the physical or computational environment. To accomplish this, many questions need to be answered regarding the components shown in Fig. 8.

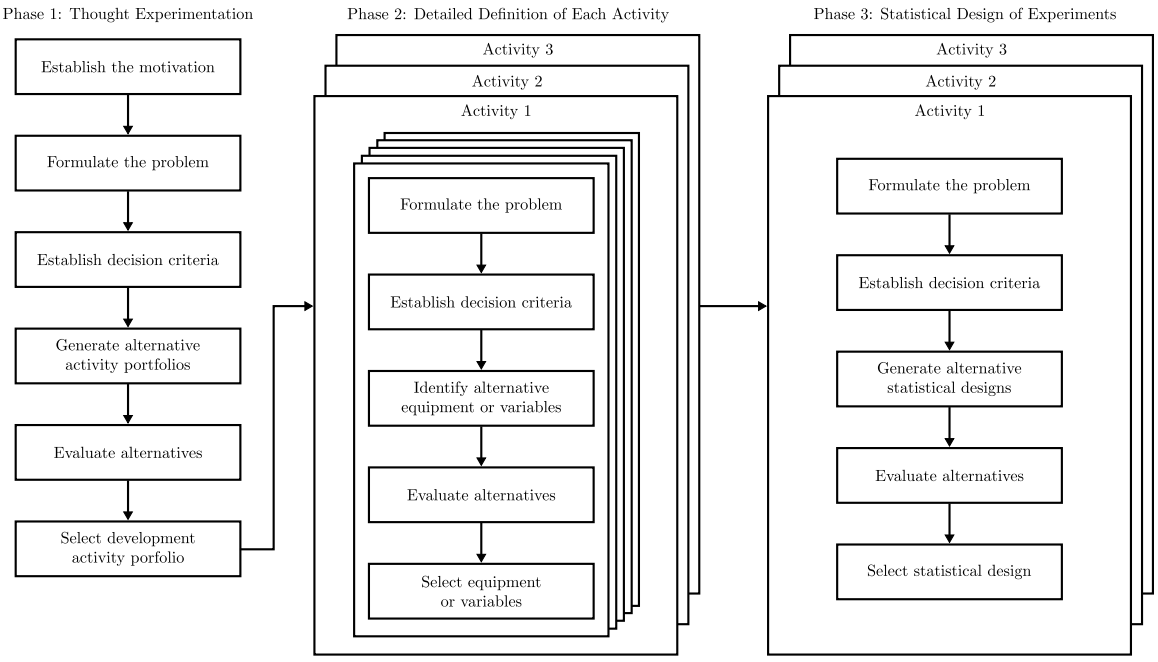


Figure 7: The proposed framework for designing technology development activities in three phases.

In general, any development activity consists of a physical or virtual system that produces outputs, given a set of inputs. This system is a combination of things such as a model of the technology, measurement devices, processes, people, hardware, and other resources that function altogether. The dependent variables are the outputs of the system that technologists wish to measure and observe. The inputs to this system are divided into three types of variables. The independent variables are under the control of technologists and can be set at target levels. Uncontrollable variables are either difficult or impossible to control for a given setup. The other inputs include any variables that must be decided to conduct the activity but are not of particular interest with regard to their effects on the dependent variables.

Development activities almost invariably concern learning the relationship between the independent variables and dependent variables for achieving objectives such as [45]:

- To determine which independent variables are most influential on the dependent

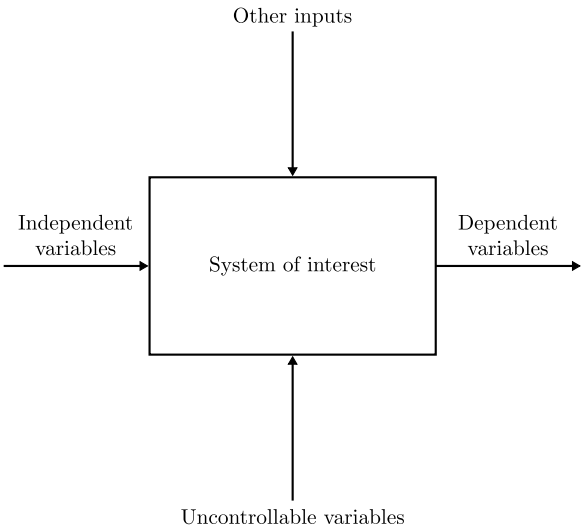


Figure 8: Conceptual model of a technology development activity.

variables

- To determine where to set the influential independent variables to achieve the best (minimum, maximum, or target) values of the dependent variables
- To determine where to set the influential independent variables to minimize the variability in the independent variables that is due to uncontrollable variables and other sources of uncertainty

With these objectives in mind, a series of decisions must be made to select the best components of the environment depicted in Fig. 8. Some of these decisions will involve simply selecting from existing devices, whereas other decisions will require a more creative approach. For example, consider choices that must be made for a physical wind tunnel experiment. Often times measurement devices such as pressure ports will be selected from off-the-shelf candidates, whereas the wind tunnel model will likely have to be uniquely designed and manufactured. Nevertheless, a rigorous approach to any kind of decision should follow a series of generic steps. For this purpose, the decision-making steps proposed by Mavris, Baker, and Schrage [46] are proposed for anchoring the decision process:

1. Establish the need.
2. Define the problem.
3. Establish value objectives.
4. Generate feasible alternatives.
5. Evaluate alternatives.
6. Make decision.

For the problem of interest here, the first step of establishing the need is not necessary because a motivation has already been identified in the first phase. In general, the definition of the problem in step two is to design a particular component of the activity to fulfill all requirements that have been defined. The value objectives in step three are criteria that will be used to evaluate the alternatives that are generated in step four for each component. Each alternative must then be mapped to these criteria, qualitatively or quantitatively, in step five so that the best one can be selected in step six. Each of these steps can be followed for every component of the activity, whether it be a concrete component such as a piece of hardware or a more abstract component such as the mathematical definition of a dependent variable. These decision processes can be followed for each activity, as depicted in the phase two portion of Fig. 7. Note that in the figure, the term "equipment" includes any component besides variables.

After the components of each activity have been determined, a plan of execution must be established. Many aspects of this plan are completely problem dependent, but a critical part of the plan that is common to all activities is the selection of settings for the independent variables at which the dependent variables will be measured. This is the focus of the third phase in the proposed framework.

3.3 Phase 3: Statistical Design of Experiments

Design of experiments (DoE) is a branch of applied statistics that aims to maximize the knowledge gained from experiments in an efficient way through strategic planning and execution. DoE began with the work of R. A. Fisher in the 1920s and 1930s for improving the way agricultural experimentation was conducted. Since then, DoE has become popular in many fields. Of particular relevance to the technology development context is the application of DoE to product and process (systems) development. Montgomery [45] identified three phases of systems design through experimentation: characterization, control, and optimization. Characterization is the process of learning the relationship between the inputs to the system of interest and the outputs, with a focus on identifying the inputs that drive the variability of the outputs. The control phase entails exploration of which variables affect the mean and/or variance of the outputs so that consistent performance of the system of interest can be achieved. The optimization phase is where the important input variables are manipulated to obtain the best compromise system performance. The word "compromise" is used here because there are typically multiple, conflicting outputs that characterize performance.

The phases of characterization, control, and optimization involve sequential experimentation to improve the state of knowledge about the system of interest. Each phase involves the selection of a type of experimental design, which consists of the independent variable settings at which the dependent variable measurements will be observed. For characterization, full factorial and fractional factorial designs are popular options. An example of a two-level fractional factorial design for three independent variables is shown in Fig. 9. Notice that the observations, marked by the large black circles, are only at four of the eight corners of the space, whereas the two-level full factorial design would include all eight corners. The fractional factorial design strategically places the observations to enable efficient estimation of main effects, where a

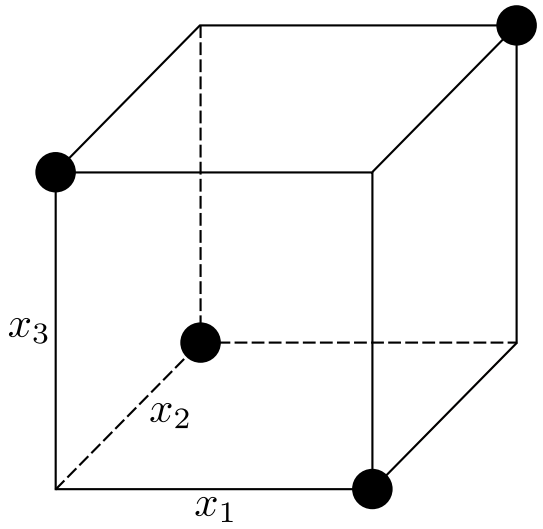


Figure 9: A notional two-level fractional factorial design for three independent variables.

main effect is the effect of an independent variable on the dependent variable averaged over the levels of other independent variables. Estimation of main effects provides knowledge of which independent variables are the most significant drivers of the dependent variables, and this information can be used to screen independent variables to reduce the dimensionality of the problem. A key feature of fractional factorial designs is their projection property. As an example, suppose that an analysis of main effects showed that the variable x_2 in Fig. 9 is not an important contributor to a dependent variable. If the x_2 dimension were to then be collapsed, then the resulting design in the two remaining variables would be a two-level full factorial in two dimensions. This can be mentally visualized by imagining the collapse of x_2 resulting in a design with large black circles at all four corners of the square.

If many independent variables are initially investigated, then the characterization phase may result in a reduced number of important independent variables. This reduced set can then be carried into the next two phases of control and optimization. The control and optimization phases are usually facilitated by representing the data

from an activity as a mathematical surface. Techniques for constructing this representation are often referred to as response surface methodology (RSM). In general, the idea of RSM is to fit a model to represent the dependent variables $\mathbf{y}$ as a linear combination of a function of the independent variables $f(\mathbf{x})$ and a measurement error term ϵ:

$$\mathbf{y} = f(\mathbf{x}) + \epsilon \tag{1}$$

A popular choice for the function of the independent variables is a first-order or second-order polynomial model that is fit to the data using least squares. Several experimental designs have been proposed for the polynomial models, such as the central composite design for second-order models. When the relationship between the dependent variables and independent variables is not well fit by a second-order model, transformations and/or higher-order terms are used in an attempt to improve the fit. When a polynomial model is not sufficient, other models are used, such as artificial neural networks and Gaussian processes. Special experimental designs, including the Latin hypercube, have been proposed for nonpolynomial regression models. Once a response surface model has been constructed with the available data, it can be used to query the value of the dependent variables at locations where observations are not available. This capability expedites the processes of exploring the relationship between the independent variables and dependent variables, searching for robust settings of the independent variables, and optimization.

For each technology development activity, all or a subset of the phases characterization, control, and optimization should be pursued to gain knowledge about the technology. Selecting appropriate experimental designs is a decision problem, and the generic decision-making steps used in phase two have also been applied for the last phase of activity design. The steps are shown at the far right in Fig. 7, and they must be implemented for each activity at least once. The theme of the problem formulation will virtually always be that an experimental design must be selected to maximize

the value of the activity while constrained by a resource budget. The decision criteria selected in the second step quantitatively define the value of a DoE. The criteria must be selected to correspond with the purpose of the activity. For example, a physical experiment may be conducted for the purpose of constructing a response surface with low prediction uncertainty. In this example, an appropriate criterion would be an estimate of prediction uncertainty obtained with an experimental design. Next, feasible DoEs need to be generated. Many considerations can limit the feasibility of the alternatives. For instance, a two-level fractional factorial is not a feasible design for a second-order polynomial response surface model because at least three levels are needed in each dimension for which second-order terms will be estimated. In the fourth step, each alternative is evaluated by quantifying the decision criteria. Finally, the preferred statistical design is selected in step five.

The proposed framework ends after a statistical design has been selected. Other problem-dependent decisions need to be made about the details of execution, but the critical characteristics have been determined by this point in the design of each activity.

3.4 Case Study: AFC-Enhanced Vertical Tail Technology Development

To provide an overview of how the proposed framework can be implemented, a case study based on the AFC-enhanced vertical tail technology introduced in Sec. 1.2 is presented here. The sequence of activities shown in Fig. 5 that were actually conducted were evaluated within the proposed framework and modifications are suggested based on the assessment.

3.4.1 Phase 1

It is apparent from Fig. 5 that the goal of developing the AFC technology was to reduce cruise drag of a future LTA vehicle by 1.5%, which would contribute to the

ERA fuel burn reduction goal. Thus, the motivation is directly linked with the motivation of ERA, which is to reduce the impact of aviation on the environment. The problem is to maximize uncertainty surrounding cruise drag, meet or exceed the 1.5% drag reduction goal, and mature the technology within resource constraints. The activities that were actually conducted indicate that a maturation goal of TRL 6 was used. Following this problem formulation, the decision criteria for thought experimentation would be cruise drag uncertainty reduction, cruise drag performance improvement, TRL increase, and cost. However, there are other important sources of uncertainty for this technology besides cruise drag reduction. A system integrator would be interested in other impacts of the AFC technology as well, such as the weight of the AFC architecture. The important impacts can be identified with a sensitivity analysis.

Boeing conducted a system integration study for a three-member, twin-aisle aircraft family where the family members with the shorter fuselages were assumed to be the only members with sweep jet AFC systems [47]. The value of infusing AFC technology was quantified as net present value (NPV) for the manufacturer, the operator, and total NPV. The impacts of the technology that were modeled included weight increase of the AFC architecture, weight reduction of a reduced size vertical tail, AFC system costs, vertical tail recurring cost reduction, nonrecurring costs, maintenance costs, drag reduction, and specific fuel consumption (SFC) increase. Uncertainty bounds were established for each of the impacts, and an NPV sensitivity analysis was used to rank the sources of uncertainty by their contribution to total NPV. It was clear from the results that the top four impacts, in order, were the drag impact, the vertical tail recurring costs, SFC increase, and the weight impact. Since NPV is a key figure of merit to the system integrator, a more appropriate problem formulation in the proposed framework includes objectives to maximize uncertainty reduction surrounding NPV, to meet or exceed one or more NPV goals (e.g., NPV $\geq$ 0), and

to mature the technology within resource constraints. Also, the decision criteria list would be expanded to include uncertainty reduction and performance improvement for each of the important impacts. As can be seen in Fig. 5, the Boeing system integration study was conducted after sub-scale and full-scale wind tunnel experiments had already been conducted. The study should have been conducted before either of these activities so that the results could be used to inform the selection of the activities.

The next step in the framework is to generate alternative activity portfolios. The ranking of uncertainty sources from Boeing's system integration study could have been used to identify activity portfolios that would target the most important impacts. The focus of the actual activities that were executed was to investigate vertical tail side force enhancement with sweeping jet actuators. Since the side force enhancement is directly linked with vertical tail area reduction, one could argue that uncertainty reduction of side force enhancement contributed to uncertainty reduction of the drag impact and weight reduction of the vertical tail. Additional activities should have been proposed for targeting the SFC, vertical tail recurring cost, and AFC architecture weight impacts. For example, computer-based studies could have been suggested to estimate these impacts accurately and precisely, and to investigate ways to improve performance for each of the impacts. The physical experiments that were conducted could have been modified to target AFC architecture weight uncertainty by weighing key components of the experimental equipment. Additionally, high-fidelity computer experiments could have been proposed to explore a larger technology design space than what was possible in the physical experiments. For instance, the impact of design variables such as actuator spacing, actuator geometry, and vertical tail geometry on side force enhancement could have been investigated computationally.

Once a set of alternative activity portfolios is generated, each alternative is evaluated by qualitatively or quantitatively assigning a measure of each decision criterion

to the portfolios. Alternatives that include computer-based studies in addition to the physical experiments that were actually conducted would likely have been seen by decision makers as adding value to the portfolio for a small cost penalty. Similarly, minor modifications to the physical experiments may have resulted in negligible consequences in terms of the resources required for the experiments. High-fidelity computer experiments may have resulted in more substantial costs to pursue, but the value of the knowledge gained from the activities would need to be weighed against the costs by decision makers. To make a decision in the final step of the framework, decision makers must balance these kinds of tradeoffs among the decision criteria to arrive at the most preferred alternative. The more alternatives and decision criteria that are involved, the more difficult this decision would be. Also, it is difficult to assign measures of uncertainty reduction and performance improvement to each alternative. Nevertheless, even qualitative considerations of the complete set of decision criteria is an improvement on the current practices that focus on TRL.

3.4.2 Phase 2

In the second phase of the framework, each of the selected activities must be designed in more detail. The wind tunnel experiments that were actually conducted were designed well because they included variations of multiple independent variables. A valuable but potentially expensive improvement for these activities would be to take measurements using at least one additional vertical tail geometry. The decision process in the framework could be applied to selecting a representative vertical tail configuration that is perhaps designed for use with the AFC system. However, the additional costs involved may have been prohibitive for physical experiment, but the investigation would likely be deemed appropriate for a computer experiment. For any computational activities that may have been selected in phase 1 using the framework, the phase 2 decision process would be followed to select the appropriate computer

M&S environments, independent variables, and dependent variables.

3.4.3 Phase 3

In phase three, experimental designs must be selected for each activity. A small subset of the data from the actual physical experiments that were conducted is published, so it is impossible to thoroughly evaluate the DoEs that were used. However, some suggestions can be made based on the DoE methodology. If at least one expensive, high-fidelity computer experiment had been selected, it could have been used for characterization with a fractional factorial design to identify the most important independent variables that affect the side force enhancement of a vertical tail. Then, the physical experiments could have been planned with statistical designs to facilitate response surface construction. The resulting response surfaces could be used for determining independent variable settings for the best performance of the AFC technology and for validating computer model predictions. The DoE approach encourages sequential experimentation, and these ideas could have been leveraged to better plan not just the physical experiments independently but rather to link the experimental designs to efficiently build knowledge as the complexity and scale of the experiments increased. Similarly, the DoE approach could have been applied to computer-based predictions for the SFC, vertical tail recurring cost, and AFC architecture weight impacts for the purposes of characterization, design space exploration, and robust design.

3.4.4 New Insights From the Framework

The overview of the implementation of the framework for the case study reveals important insights for how the AFC technology development activities should have been designed:

- The Boeing system integration study should have been conducted before the physical experiments so that the activities could have been designed to target

the most important sources of technology impact uncertainty.

- Additional computer-based development activities should have been proposed to improve the value of the portfolio by increasing the potential of gaining performance improvement and uncertainty reduction of the technology impacts.
- Additional vertical tail geometries should have been used in a physical or computer experiment to investigate the effect of the vertical tail design on side for enhancement, and the phase two decision process could have been applied for designing the vertical tail model used in the experiments.
- In phase three, the DoE methodology of sequential experimentation should have been leveraged to efficiently build knowledge of the technology impacts by planning the experimental designs of multiple activities simultaneously.

It is impossible to quantify the added value of these suggested changes to the AFC development program without the luxury of implementing them, but it is clear that these modifications would have resulted in additional uncertainty reduction and potentially more performance improvement while still graduating the technology to TRL 6.

3.5 Opportunities to Enhance the Proposed Framework

Although the proposed framework can be implemented as is by interpreting and applying each of the steps for a given technology development program, there are opportunities to enhance the framework by adding rigor to the decision making processes that comprise the framework. In phase one, decision makers must evaluate each of the activity portfolio alternatives by estimating the decision criteria measures and mentally balance tradeoffs to arrive at a decision. A quantitative decision aid would elucidate aspects of the decision process and provide a traceable tool for justifying

the selection. In Chapter 4, a decision analysis methodology is proposed for quantitatively evaluating technology development activity portfolios in phase one. Phase two is highly dependent on the type of activity being designed, but the selection process for equipment and variables could also benefit from quantitative evaluation of alternatives. In phase three, quantitative evaluation of experimental designs would be ideal. However, this would require estimation of measures such as performance improvement potential and uncertainty reduction potential. Also, the capability to quantify the uncertainty surrounding technology impacts would be useful for supporting decisions in phases one and three. Chapter 5 presents novel capabilities for characterizing the uncertainty surrounding technology impacts and estimating the uncertainty reduction associated with a statistical experimental design. A methodology for the special case of uncertainty characterization for a particular type of reliability development activity is presented in Chapter 6.

In the following three chapters, the novel capabilities that are presented are intended to provide quantitative components for the proposed framework to improve the decision making process. In Chapter 7, the contributions are summarized, limitations are enumerated, future research opportunities are discussed, and an overarching thesis statement is presented for the framework.

CHAPTER IV

MULTIATTRIBUTE UTILITY ANALYSIS FOR EVALUATING TECHNOLOGY DEVELOPMENT ACTIVITIES

In this chapter, the problem of how to inform decisions regarding the selection of technology development activity classes before details of the activities have been defined is confronted. Details of the problem are described in Sec. 4.1. Then, the state of the art is identified and the foundation for an improved approach is established in Sec. 4.2. Next, techniques from multiattribute utility analysis are incorporated and the proposed methodology is formulated in Sec. 4.3. The primary argument is as follows.

Argument 1: The proposed methodology improves upon the state of the art and is an appropriate way to evaluate technology development activity alternatives because

1. It aggregates decision makers' preferences, risk attitude, and system-level performance goals in the analysis
2. It quantitatively represents uncertainty surrounding the impacts of the alternatives
3. It enables the quantitative evaluation of alternatives under conditions of risk and uncertainty with a theoretically valid measure of value

An illustrative example is shown to support this claim in Sec. 4.4, and the chapter closes with a discussion and conclusions in Sec. 4.5.

4.1 Problem Definition

This section begins by introducing key assumptions that are made throughout this chapter. Then, features of the problem addressed in this chapter are discussed.

4.1.1 Overarching Assumptions

It is assumed that a technology has or a set of technologies have been chosen for and entered into a development program. If technology selection has been done properly, this assumption implies that a system has been identified for technology infusion, the important system-level metrics have been selected for defining the performance goals and describing the performance gap, and the uncertainty surrounding the integration impacts has been represented mathematically. With a mathematical model of the uncertainty surrounding technology impacts, an M&S environment, or surrogate models of the environment, can be used to propagate technology-level uncertainty up to system-level uncertainty. Given that such an M&S environment or surrogate models would have been built as part of the technology selection process, it is assumed that this environment is available to analysts. Lastly, it is acknowledged that there are proprietary best practices and methods for managing and planning technology development programs. An assumption is made that this type of process is being followed to manage programmatic risks. It is also assumed that any such process includes a taxonomy of development activities to select from. For an example of this type of process that is published in the open literature, the reader is referred to Ref. [21]. The methods presented in this chapter are not meant to compete with established procedures for managing the overall development program but rather to enhance them.

4.1.2 Technology Development Activity Portfolio Selection

An important overarching assumption mentioned in Sec. 4.1.1 is that a taxonomy of technology development activities exists in a given technology development program.

For generality, a specific taxonomy will not be used here. However, classes of activities that are likely to appear in any taxonomy will be mentioned throughout this chapter. The more extensive and detailed the taxonomy, the more alternatives that will be available to select from. Once a class of development activity is selected, there may be subclasses to choose from as well. Then, many characteristics of the activity need to be defined in later phases.

The component of the technology development activity design problem that is the focus of this chapter is the selection or prioritization of activity classes, not the detailed design of the activities. This corresponds with phase one in Fig. 7. *Why is this component the focus?* Although many characteristics of technology development activities must be nailed down to completely define them, the attributes of the activities that have important programmatic implications are largely determined when the class of activity is selected. Analogously, when an architect designs the floor plan of a building, he or she locks in a large percentage of construction costs before details such as flooring material have been decided. Based on the literature, it is clear that some of the most important attributes of technology development activities include uncertainty reduction, performance improvement, maturation, and required resources. To illustrate how the identification of activity class can bracket these attributes, consider the differences between a numerical design space exploration activity and a full-scale physical experiment. As notionally depicted in Fig. 10, the design space exploration activity would likely shed light on ways to improve the performance of the technology but not significantly reduce epistemic uncertainty, whereas the full-scale physical experiment may not result in changes to the technology for performance improvement but would primarily reduce epistemic uncertainty. Large differences are also present in maturation and the amount of resources required. It is possible that experts would decide that the physical experiment justifies a graduation of the technology to a higher TRL, whereas, depending on the TRL definitions used, the design space exploration

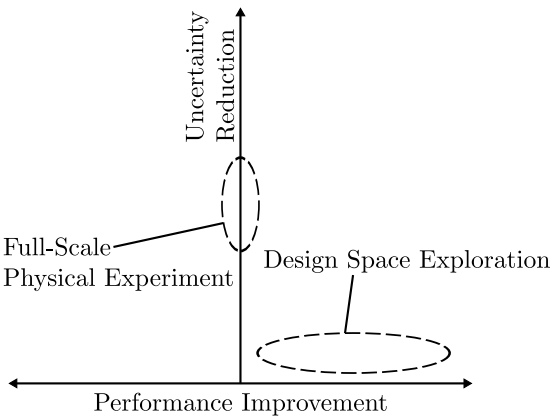


Figure 10: Notional depiction of where two types of technology development activities will lie in the attribute space of uncertainty reduction and performance improvement.

activity may not be considered as a contributor to TRL progression whatsoever. In terms of resources, costs associated with computational resources and man-hours required for design space exploration would almost certainly be dwarfed by the costs incurred to build and execute a full-scale physical experiment.

This notional comparison helps to highlight a key characteristic of the problem, which involves uncertainty surrounding the effects that a given technology development activity class will have on important attributes. Existence of this uncertainty is the reason Fig. 10 is drawn with ellipses instead of points for each activity. Design space exploration can lead to performance improvement, but before conducting this activity it is nearly impossible to know exactly how much performance improvement is attainable. Similarly, if the full-scale experiment is the first time the technology will be scaled up in an experiment, then technologists may discover that the technology performs slightly worse or better than when observed at a smaller scale. This additional epistemic uncertainty will be present in all of the activity attributes, including required resources. Thus, a solution to the activity selection problem must account for this uncertainty.

Based on the discussion up to this point, the problem addressed in this chapter can be summarized as follows. During the planning of a technology development

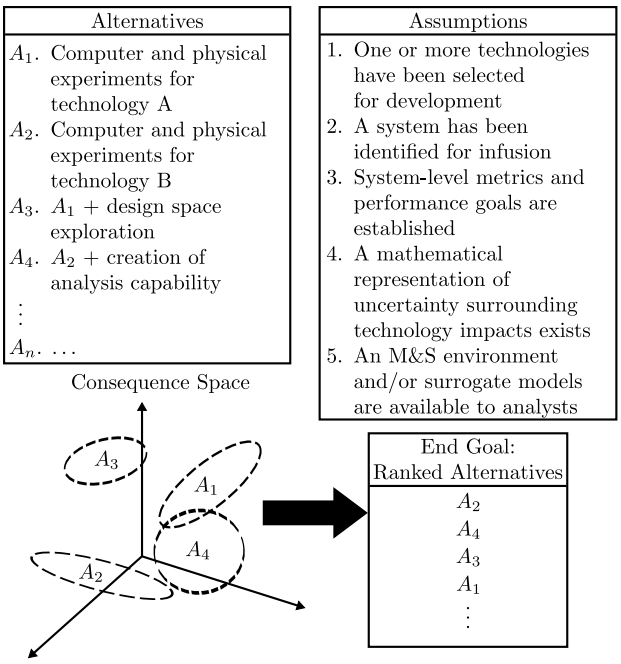


Figure 11: Components of the decision problem that is addressed in this chapter.

program, decision makers must choose a set of classes of technology development activities from an established taxonomy that will be associated with specific technologies. Each alternative will result in consequences for achieving the program goals of uncertainty reduction, performance improvement, and maturation, as well as consequences in terms of resource expenditure. However, decision makers cannot be certain of precisely what consequences will result from each alternative. As stated in Sec. 4.1.1, assumptions are made about the technology development program and what resources are available to the decision makers. The primary research question is:

Research Question 1.0: Given alternatives defined by combinations of technology development activity classes and technologies, what is an appropriate way for decision makers to evaluate the alternatives for downselection?

Figure 11 provides a graphical summary of the problem.

4.2 Establishing a Decision Framework

Although selecting technology development activity classes is difficult and prior researchers have pursued quantitative methods, the reader may still ask *Why should one expend the effort to use a quantitative decision aid?* With any decision, one can informally weigh tradeoffs in his or her mind, but it is believed that a more formal approach to prioritizing alternatives is indispensable for two reasons. First, as argued by Tversky and Kahneman [48] in their seminal paper, unaided humans use heuristic principles that reduce the complexity of difficult judgment tasks under uncertainty, which can lead to systematic errors in judgments. Similarly, in discussing the benefits of a quantitative approach to rank ordering design alternatives, Hazelrigg stated that "the comparison is generally too complex to make accurately and consistently without the use of a mathematical value model, particularly given the presence of uncertainty" [49]. Second, decisions regarding allocation of resources in a technology development program often must be justified to stakeholders, the public, and others, and quantitative analysis provides traceable, transparent decision support.

In this section, a foundation for the proposed methodology is formulated by synthesizing the current state-of-the-art approach to the activity downselection problem with additional elements from decision theory. First, the current state of the art is described and gaps are identified to motivate the need for a novel approach. Then, the decision-making process introduced as phase one in Sec. 3.1 is exploited to provide the foundation for the methodology. This process is built upon for the problem of interest by incorporating ideas from multiattribute utility theory (MAUT) and the current state of the art in Sec. 4.3.

4.2.1 The Current State of the Art

Although the methodologies presented in Refs. [21, 26] are each unique, pertinent steps for the activity selection problem have been extracted and combined. These

steps include:

1. Conduct readiness risk assessments for all technologies.
2. Conduct sensitivity analysis for the contribution of each technology to system-level metric probabilities of success.
3. Conduct sensitivity analysis for the contribution of each technology integration impact to system-level metric variability.
4. Select technologies for development activities.
5. Select technology integration impacts to target.
6. Select classes of development activities and proceed with detailed design of development activities.

The first step involves locating the technologies on a risk matrix with two axes: estimated number of years until highest TRL is achieved and current TRL. Technologies with the highest readiness risk are those that fall into a region of largest number of years until highest TRL is achieved and lowest current TRL, whereas those with the lowest readiness risk fall into the opposing corner of the matrix. The second step entails conducting a sensitivity analysis that quantifies the contribution of integrating each technology to variability of the probability of success (POS) for each system-level metric. POS is quantified by propagating uncertainty surrounding technology impacts to system-level metrics with an M&S environment, then calculating the probability of meeting or exceeding an established goal for each metric. A notional PDF representing uncertainty surrounding fuel burn reduction is shown in Fig. 12. The shaded region shows the area under the curve that meets or exceeds the established goal of 2%; this is POS. The POS is affected by which technologies are integrated with the aircraft, so sensitivities can be derived by calculating the POS with multiple combinations of the technologies in the development program. The third step

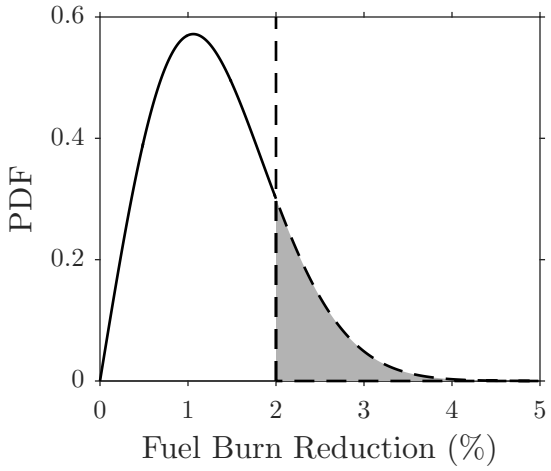


Figure 12: Notional fuel burn reduction PDF showing probability of successfully meeting to exceeding a goal of 2% as the gray area.

is a sensitivity analysis to quantify the contribution of each technology integration impact to the variability of the system-level metrics. This can be accomplished by using local or global sensitivity analysis methods. Local methods are typically based on partial derivatives of model outputs with respect to model inputs around nominal values, whereas global methods are based on statistical frameworks so that the entire range of the inputs is considered in the analysis. Global methods are preferred for the technology development context because they are capable of providing a decomposition of system-level metric variance for a given probabilistic model of technology integration impacts.

With sensitivity analysis results and readiness risk assessments available, technologies can be selected for development activities in step four. Technologies that are preferred are those with a combination of relatively high readiness risk and large contributions to POS and variability of the system-level metrics. Once technologies have been selected for development activities, integration impacts for each technology must be selected as targets for the activities in step five. Results from the sensitivity analysis in step three are used to inform this selection. The final step of the selection process is delegated to technologists, who must design the activities to target the

integration impacts within the constraints of TRL definitions.

The methods that comprise the state of the art are beneficial for informing a decision about which technologies and technology impacts to target with development activities, but there are some shortcomings. Decision makers must first consider tradeoffs between technologies based on a readiness risk description, the effect each technology has on improving or degrading the POS for all system-level metrics, and the effect each technology has on the uncertainty of all system-level metrics. Thus, the decision makers have to analyze a space with two dimensions for readiness risk, $\sum_{i=1}^{q} g_i$ dimensions for g_i POS sensitivities for each of the q system-level metrics, and q sensitivities for system-level metric variability. In total, that is a space of dimension $2+\sum_{i=1}^{q} g_i+q$. Even with only one system-level metric and one performance goal for that metric, the objective space would have four dimensions, and this neglects other metrics that may be important, such as the projected costs of development activities for each technology. Once technologies have been selected for development activities, the decision makers must then analyze a smaller q-dimensional space with sensitivities for system-level metric variability to decide which technology impacts to target with development activities. Although multidimensional criteria spaces can complicate the decision process, there are many techniques that have been proposed to handle these kinds of problems, such as multiobjective genetic algorithms. However, a more critical shortcoming of the state of the art is that it is not capable of quantitatively evaluating alternatives like the notional set shown in Fig. 11. This is because the criteria of readiness risk, POS sensitivity, and system-level metric uncertainty sensitivity are invariant under different development activity classes; these criteria only quantify the potential of *any* development action, targeting each technology and its impacts, to have value. Thus, any two development activity packages that target the same set of technology impacts would be considered to have equivalent value under the state-of-the-art approach, despite the fact that the two packages may result in very different

degrees of uncertainty reduction, performance improvement, etc.

4.2.2 A Decision-Making Process

Motivated by the need for a decision-based approach to selecting technology development activity classes, the novel methodology is anchored in the decision-making steps proposed in phase one of the novel framework for designing technology development activities, which are repeated here:

1. Establish the motivation.
2. Formulate the problem.
3. Establish decision criteria.
4. Generate alternative activity portfolios.
5. Evaluate alternatives.
6. Select development activity portfolio.

The important inputs to this process are described in Sec. 4.1.1, but there may be others depending on the technology. The first step of this process, establish the motivation, is assumed to be part of any technology management method and was discussed in the framework chapter. The second step is discussed in Sec. 4.1.2. The third step entails choosing measures that quantify value to the decision makers so that alternatives can be evaluated. This step is explored further in Sec. 4.3. The fourth step requires that decision makers select the alternatives that will be considered in the decision analysis. The approach described in Sec. 4.2.1 is viewed as a valid way to generate feasible alternatives. As part of this step, some technology development activity classes may be filtered out based on objectives at a particular point in a development program. For instance, if a decision maker wishes to only

consider development activity classes that correspond with TRLs 4–6, then the number of feasible alternatives could be greatly reduced. In this chapter, it is assumed that a set of feasible alternatives is available. Each alternative must be assigned a measure of value in step five so that the decision makers can quantitatively evaluate the alternatives to make an informed decision in the final step.

To better inform step six of the decision-making process, additional components are needed for steps three and five. Evaluation of alternatives in step five requires that a measure of value to the decision makers be established in step three. Additionally, quantitative evaluation of alternatives implies that a way to model the impacts of alternatives on the value measure is needed. These gaps in knowledge can be summarized with the following research questions:

Research Question 1.1: How should the value of technology development activities be quantified?

Research Question 1.2: What is an appropriate way to model the impact of technology development activities on a value measure before those activities have been performed?

4.2.3 Selecting an Enabler From the Literature

In answering RQs 1.1 and 1.2, some characteristics of a solution were identified in addition to those that are obvious from the problem definition. To answer RQ 1.1, concerning how value should be quantified, a function was needed such that it accurately represents the decision makers' preferences over the consequence space. For instance, a function could be defined to map uncertainty reduction to a measure of value. This function must not be restricted to a linear form, as decision makers' preferences may, for example, initially increase quickly with more uncertainty reduction and then level off. Regarding RQ 1.2, any solution must account for the uncertainty surrounding the impacts of the development activities. As discussed previously, this

implies that the consequences (attributes) cannot be known with certainty, and uncertain consequences propagate to an uncertain value measure.

The solution characteristics helped with selecting an enabler. Most decision making aids can be divided into multicriteria decision making (MCDM) methods and MAUT methods. MCDM methods are employed for alternative selection when the objective functions or attributes are deterministic and the decision maker's value function is implicit or not modeled at all, whereas MAUT is used when risk and uncertainty are critical to the evaluation of alternatives and when a value function is explicitly represented [50]. Because of these characteristics, MCDM techniques, such as multicriteria optimization, were ruled out. For the discrete alternative problem that is of interest here, analytic hierarchy process (AHP) [51]—which is sometimes classified as a MAUT approach—and the traditional MAUT method of Keeney and Raiffa [52] are two of the most mathematically-rigorous and commonly-used decision aids. Incorporating uncertainty and risk into the decision analysis is an integral part of MAUT, whereas the original formulation of AHP is deterministic. However, multiple ways of adjusting AHP to model uncertainty have been proposed. The reader is referred to Lafleur [53] for an example and a history of modifications to AHP to account for uncertainty. Loken et al. [54] applied MAUT and modified AHP methods to incorporate uncertainty in the process of local energy planning, and they concluded that MAUT is better at handling uncertainties than any of the modified AHP methods. Some researchers have published claims about the merits of both MAUT and AHP. With regard to decision-based engineering design, Thurston asserted that "multiattribute utility analysis is the tool best suited for making normative tradeoff decisions which exhibit one or both of the following features; nonlinearity of preference over an attribute range, and uncertainty which affects desirability (where that uncertainty can be modeled probabilistically)" [55]. Also within the engineering design context, Hazelrigg asserted that "Subject to the six axioms of [von Neumann-Morgenstern]

vN-M utility, not only does the expected utility theorem provide a valid utility measure (that is, a valid measure for rank ordering design alternatives), it is the only valid measure. All other measures are wrong (or equivalent)" [49]. On the opposing side, Gass [56] provided arguments to refute criticisms of AHP and claimed that it seems AHP has replaced MAUT and its variants in the realm of practical multicriteria decision-making problems. MAUT was selected as the enabler for the technology activity downselection problem primarily because MAUT provides a way to quantify value under conditions of risk and uncertainty, whereas consensus has not even been reached regarding how to augment AHP to handle uncertainties. For this reason, another originally deterministic decision making tool called the technique for order preference by similarity to ideal solution (TOPSIS) [57] was also judged to be an inferior option.

4.3 Evaluating Alternatives With Multiattribute Utility Analysis

Multiattribute utility analysis is a normative decision-making approach, meaning that its purpose is not to imitate an unaided human decision maker but rather to help recognize a decision that is better than what the unaided decision maker may have selected [55]. To overcome the problems of unaided human judgment, MAUT is based on axioms of rational behavior [58]. Instead of quoting these abstract axioms directly, an informal version from Keeney is presented here for ease of interpretation:

- (Generation of Alternatives). At least two alternatives can be specified.
- (Identification of Consequences). Possible consequences of each alternative can be identified.
- (Quantification of Judgment). The relative likelihoods (i.e., probabilities) of each possible consequence that could result from each alternative can be specified.

- (Quantification of Preference). The relative desirability (i.e., utility) for all the possible consequences of any alternative can be specified.

- (Comparison of Alternatives). If two alternatives would each result in the same two possible consequences, the alternative yielding the higher chance of the preferred consequence is preferred.

- (Transitivity of Preferences). If one alternative is preferred to a second alternative and if the second alternative is preferred to a third alternative, then the first alternative is preferred to the third alternative.

- (Substitution of Consequences). If an alternative is modified by replacing one of its consequences with a set of consequences and associated probabilities (i.e., a lottery) that is indifferent to the consequence being replaced, then the original and the modified alternatives should be indifferent. [59]

The key result obtained from these axioms, called the expected utility theorem, shows that the expected utility of an alternative is an indication of its desirability, and that expected utility is a valid measure of value for ranking alternatives under risk and uncertainty. The utility function is a scalar function that aggregates decision makers' preferences over all of the attributes as well as risk attitude.

To enable the decision-making process presented in Sec. 4.2.2 for the problem of interest here, techniques were incorporated from MAUT in five steps, as shown in Fig. 13. The first three steps correspond with step three of the decision-making process, and the last two steps correspond with step five. The notation and terminology used here is similar to that presented in one of the most popular references for multiattribute utility analysis, the text by Keeney and Raiffa [52].

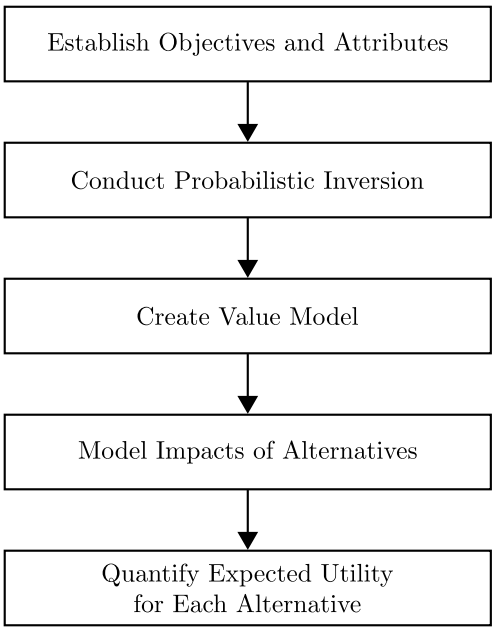


Figure 13: The proposed methodology for evaluating alternatives with multiattribute utility analysis.

4.3.1 Step One: Establish Objectives and Attributes

Establishing objectives typically involves an enumeration of the consequences of the alternatives and ultimately selection of low-level objectives from a hierarchy. Detailed guidelines for generating objectives are documented in the decision analysis literature (e.g., see Ref. [59] for an overview) and are not presented here. Established objectives must indicate direction of improvement, then attributes can be identified to measure degrees to which the objectives are attained. For instance, decision makers may wish to "maximize uncertainty reduction" as an objective. The attribute in this example is a measure of uncertainty reduction. Goals are different from objectives and attributes in that they are either achieved or not. For example, a goal would be to "reduce uncertainty by 10%".

Some researchers would contend that the primary objective of technology development is knowledge creation, with no specific objectives for improving the performance

of systems that are integrated with the technologies. However, many decision makers may view technology development similarly to how Bigwood views what he calls "new technology exploitation", which "lies in a gray area between [new product development] NPD and pure science, borrowing from the former the intent to produce something useful while exploiting the basic principles uncovered by the latter" [60]. Therefore, at a minimum, objectives concerning performance improvement, uncertainty reduction, and cost should be defined. In this formulation, attributes are suggested for uncertainty reduction, system-level performance, and cost.

4.3.1.1 Uncertainty Reduction

Two of the most commonly used measures of uncertainty for random variables are variance and entropy, defined for a random variable Y respectively as:

$$\mathrm{Var}(Y) = \mathrm{E}\left[(Y - \mathrm{E}Y)^2\right] \tag{2}$$

$$h(Y) = -\int_{\mathbb{R}} p(y)\, \log\, p(y)\, dy \tag{3}$$

where, $\mathrm{E}[\cdot]$ is the expectation operator, and $p(y)$ is the PDF of Y. Decision makers are likely to be concerned with the uncertainty surrounding a set of system-level metrics $(M_1, M_2, \ldots, M_q)$ rather than the lower-level technology impacts. The simplest approach to using variance as a measure of uncertainty in this case is to calculate the variance of all q marginal distributions using Eq. (2). A drawback of this approach is that q attributes would result, and decision makers would then need to create q utility functions. Equation (3) can be generalized to produce a single uncertainty measure for a random vector, but variance is likely to be a more intuitive measure for decision makers whom may not be familiar with information theory. In lieu of using q different uncertainty reduction attributes derived from the variance of the system-level metric marginal distributions, aggregate variance measures can be defined. In this chapter,

the following average variance reduction attribute is used:

$$\text{Average Variance Reduction } (\%) = 100\left(1 - \frac{1}{q}\sum_{i=1}^{q}\frac{\text{Var}(M_i)}{\text{Var}(M_{i_{\text{Present}}})}\right) \tag{4}$$

where, $M_{i_{\text{Present}}}$ represents the system-level metric random variable M_i defined under the present state of uncertainty.

4.3.1.2 System-Level Performance

Within the technology infusion literature, each system-level metric has an associated goal value used to calculate POS. Thus, an obvious choice for an objective is to maximize POS for all metrics. The corresponding attributes would simply be q POS probabilities. Depending on the nature of each metric, decision makers might wish to use more than one goal for some. As an example, consider aircraft fuel burn reduction, for which there is a goal of meeting or exceeding 2% with some probability. In addition to this goal, decision makers also want to see a very high probability of fuel burn reduction exceeding a lower bound, such as 0.5%. Again, as the number of attributes grows, the number of utility functions grows. In addition to lowering the difficulty of quantifying preferences, a minimal set of attributes diminishes the possibility of violating independence conditions of utility theory, which will be briefly described in the third step.

To provide decision makers with flexibility in establishing system-level performance objectives, an attribute was formulated that incorporates specific POS goals. It is assumed that decision makers wish to meet all POS goals simultaneously. The idea is to first determine a target joint probability distribution on technology impacts that can be propagated to the system-level metrics and will simultaneously meet all of the stated POS goals, then to calculate the probability that the joint distribution representing the state of uncertainty under the alternatives will meet or exceed the performance of the target distribution.

To produce a single attribute instead of one for each system-level metric, a composite function D called a desirability function is used [61]:

$$D = ({d_1}^{w_1} {d_2}^{w_2} \cdots {d_q}^{w_q})^{\frac{1}{\sum w_i}}, \qquad w_i > 0 \tag{5}$$

where, d_i is a desirability function corresponding with system-level metric M_i, and w_i is a weight that represents the relative importance of each system-level metric. Each desirability function is a transformation from a system-level metric to a desirability value between 0 and 1, with a value of 1 being the most desirable. The form of Eq. (5) follows a weighted geometric mean, and this form has the important property that $D = 0$ if any $d_i = 0$. Following the popular approach of Derringer and Suich [62], the transformation for a metric to be maximized is:

$$d_i = \begin{cases} 0 & M_i \leq M_{i*} \\ \left(\frac{M_i - M_{i*}}{M_i^* - M_{i*}}\right)^{r_i} & M_{i*} < M_i < M_i^* \\ 1 & M_i \geq M_i^* \end{cases} \tag{6}$$

where, M_{i*} is the minimum acceptable value of M_i, M_i^* is the value of M_i above which there is no additional value, and r_i is a parameter that controls the behavior of the desirability function in the interval (M_{i*}, M_i^*). Note that the case of minimization of M_i is equivalent to maximization of $-M_i$. There is another desirability function form for achieving a target value, but it will not be discussed here since system-level performance objectives are virtually always concerned with minimization or maximization. Figure 14 shows a plot of Eq. (6) for multiple values of r_i.

The interpretation of desirability functions is the same as utility functions, and some of their mathematical characteristics represent decision makers' risk attitude. One of the results of utility theory is that a strictly concave utility function represents a risk-averse attitude, a strictly convex utility function represents a risk-prone attitude, and a linear utility function represents a risk-neutral attitude. Therefore, the

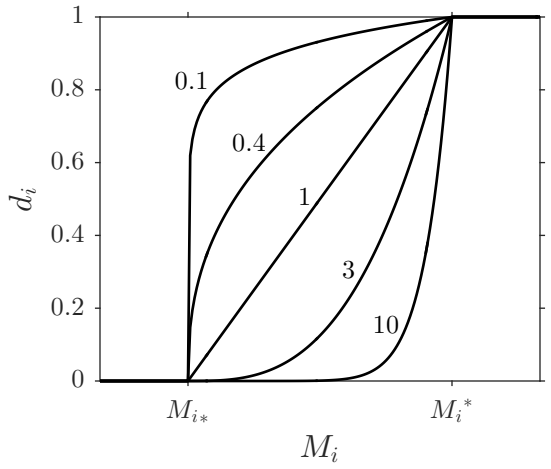


Figure 14: Plot of desirability functions for five values of r_i.

desirability function parameter r_i controls the implied risk attitude of the decision maker in the interval (M_{i*}, M_i^*).

Once the desirability functions d_i and weights w_i are specified, all POS goals must be propagated to the composite desirability D. Unfortunately, it is not possible to analytically propagate the POS goals to D. A solution to this problem is to find a joint distribution on the system-level metrics that meets the POS goals, then to propagate this distribution to composite desirability. For a finite number of POS goals, there is an infinite number of distributions that can meet the goals simultaneously. One way to constrain the search space is to require that this distribution be technically feasible within reasonable bounds of technology impact uncertainty. Here, a "feasible" distribution is one that is defined on the technology impact variables within a set domain such that it can be propagated, via M&S, to the system-level metrics and will simultaneously meet all of the POS goals. A technique for finding a feasible distribution is described in Sec. 4.3.2. Once this distribution is found, the system-level performance attribute can be defined as the probability of the desirability quantified for the state of uncertainty under the alternatives being greater than or equal to the desirability under the target distribution: $\mathrm{P}\left(D \geq D_{\mathrm{Target}}\right)$.

Other attributes can be derived that are also based on POS goals. For instance,

the probabilities of meeting or exceeding each individual system-level metric goal can be calculated and multiplied to produce a single attribute. Some alternative formulations may be simpler to implement, but ultimately the selection should be based on how easily the attribute can be interpreted by the decision makers. Note that the step described in Sec. 4.3.2 is not necessary if an alternative performance attribute is used.

4.3.1.3 Cost

There are many ways that the cost to conduct technology development activities can be mathematically represented. Any valid representation may be used in a utility analysis, so long as the resulting attribute is meaningful to decision makers. Throughout this chapter, the cost attribute used is percentage of the available budget.

4.3.2 Step Two: Conduct Probabilistic Inversion

The state of the art for finding a feasible target distribution is a set of algorithms for a generic problem called probabilistic inversion. Within the technology development context, the idea behind probabilistic inversion is as follows: given a random vector of system-level metrics $\mathbf{M} \in \mathbb{R}^q$ and an M&S environment $f : \mathbb{R}^l \to \mathbb{R}^q$, find a joint distribution on the technology impacts $\mathbf{k} \in \mathbb{R}^l$ such that $f(\mathbf{k}) \sim \mathbf{M}$, where $\sim$ indicates identical distributions. In practice the joint distribution on $\mathbf{M}$ is characterized with a set of quantiles for each of the marginal distributions on the q system-level metrics; these are the POS goals. Since it can be difficult or impossible to invert the model f, probabilistic inversion algorithms that do not require model inversion are preferred. Algorithms called Iterative Proportional Fitting (IPF) [63] and PARameter Fitting for Uncertain Models (PARFUM) [64] have been fused with an idea called sample reweighting to produce techniques for generic probabilistic inversion that do not require model inversion [65]. To illustrate the steps of probabilistic inversion, a notional AFC technology example is used.

4.3.2.1 Specify Uncertain Model Input Variables and Ranges

Uncertain model inputs that represent technology impacts have been referred to as "k-factors" by some authors (e.g., see Ref. [16]). As an example of how these impacts are implemented, consider the Breguet range equation rearranged to calculate aircraft fuel burn:

$$W_{\mathrm{F}} = k_{W_{\mathrm{E}}} W_{\mathrm{E}} \left[\exp\left(\frac{R k_{C_D} C_D c_t}{C_L V_\infty} \right) - 1 \right] \tag{7}$$

where, W_{E} is the empty weight of the aircraft, R is range, C_D is the aircraft drag coefficient, c_t is thrust-specific fuel consumption, C_L is the aircraft lift coefficient, and V_∞ is cruise velocity. Supposing that the AFC technology helps reduce wing drag, a variable is required for modeling this impact. Also, AFC technologies need a power supply architecture to supply flow or electricity to the actuators, and this additional equipment will add weight to the vehicle. Thus, the k-factors k_{C_D} and $k_{W_{\mathrm{E}}}$ were added to the equation to model AFC technology impacts of drag change and weight change, respectively. These are the uncertain model input variables for this example.

After specifying the input variables, the ranges of each must be defined. This is accomplished in the technology development context by considering physical constraints on the variables and determining reasonable domains of uncertainty. For the AFC technology example, suppose that the ranges have been defined as $[0.97, 1.0]$ for k_{C_D} and $[1.0, 1.03]$ for $k_{W_{\mathrm{E}}}$.

4.3.2.2 Specify Output Variables and Marginal Distribution Quantiles

In the technology development context, the output variables are the system-level metrics of interest $\mathbf{M}$. Quantiles are required for each model output variable, and these quantiles serve as constraints for the probabilistic inversion algorithms. As previously mentioned, for technology development the quantiles are defined by the system-level goals for POS. Specification of the quantiles entails enumeration of sets of pairs of values for each output variable. Each pair contains the quantile Q and a

probability π, where $Q(\pi) = \{m \,|\, \mathrm{P}(M \leq m) = \pi\}$.

For the AFC technology example, suppose that the output variables are fuel burn of three different aircraft, and the system-level metrics are percentage reduction of fuel burn for each. For simplicity, it was assumed that the k-factors are identical for all three aircraft. The constant variables in Eq. (7) are shown in Table 3 for each aircraft. The quantiles are identical for all three aircraft, and they include $\{0.01, 0.5\%\}$ and $\{0.5, 1.0\%\}$, where the first number in each pair is the probability π, and the second number is the quantile Q.

Table 3: Breguet range equation constants for the three aircraft in the notional AFC example

Variable	Aircraft 1	Aircraft 2	Aircraft 3
W_{E} (lb)	524,000	149,300	330,000
R (nmi)	5700	1600	3700
C_L/C_D	19	17	18
c_t (1/hr)	0.5	0.5	0.5
V_∞ (ft/s)	832	760	779

4.3.2.3 Generate Samples

Probabilistic inversion requires that samples be generated in the input variable domain, then propagated to a joint distribution on the output variables. To the best of the author's knowledge, there is not any guidance in the literature regarding how the samples should be generated. A common approach is to sample from independent uniform distributions on each input variable. With samples generated in the input variable domain, the propagation task is accomplished by uncertainty propagation to the system-level metrics. There are many uncertainty propagation techniques in the literature (e.g., see Ref. [66]), but when samples are inexpensive to generate with a rapidly executed M&S environment or surrogate models of it, Monte Carlo simulation provides accurate results with a relatively large sample size. Mathematically, the uncertainty propagation task entails computation of the joint cumulative distribution function (CDF) for the system-level metrics (the model outputs) given the

input variable joint distribution:

$$\begin{aligned} \mathrm{P}(\mathbf{M} \leq \mathbf{m}) &= \mathrm{P}(f(\mathbf{k}) \leq \mathbf{m}) \\ &= \mathrm{P}(f(\mathbf{k}) \in \mathcal{C}) \\ &= \mathrm{P}(\mathbf{k} \in f^{-1}(\mathcal{C})) \\ &= \int_{f^{-1}(\mathcal{C})} p(\mathbf{k})\, d\mathbf{k} \end{aligned} \tag{8}$$

where, $\mathcal{C} = \{\mathbf{x} \,|\, x_i \in (-\infty, m_i],\ i = 1, 2, \ldots, q\}$, $p(\mathbf{k})$ is the joint PDF on the k-factors, and $f^{-1}(\cdot)$ is the preimage of the M&S environment.

In the AFC example, the samples were generated from the following independent uniform distributions: $k_{C_D} \sim \text{Uniform}(0.97, 1.0)$ and $k_{W_\text{E}} \sim \text{Uniform}(1.0, 1.03)$. Then, Monte Carlo simulation was used to propagate the samples through Eq. (7) for all three aircraft.

4.3.2.4 Conduct Sample Re-Weighting

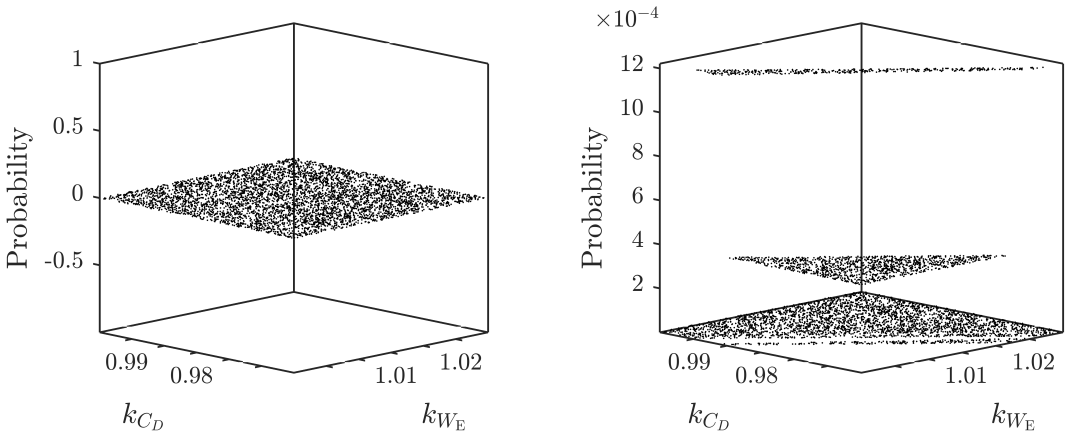
Once the samples are generated for the input variables and propagated to the output variables, they must be assigned initial probability weights. The approach used in the probabilistic inversion literature is to assign all samples weights according to a discrete uniform distribution, i.e., for N samples a weight of $1/N$ would be assigned to all samples. For technology development, another option is to use initial weights from the joint distribution on the k-factors that represents the present state of uncertainty. Then, PARFUM or IPF is used to re-weight the samples so that the quantile constraints are met. If the problem is feasible, then the quantile constraints will be met within a specified tolerance. If the problem is infeasible, then PARFUM will provide a solution such that the quantile constraints are met as closely as possible. To measure "closeness", Kullback-Leibler (KL) distance between two probability mass functions $R(y)$ and $K(y)$ is used:

$$D_{\text{KL}}(R||K) = \sum_{y \in \mathcal{Y}} R(y) \log \frac{R(y)}{K(y)} \tag{9}$$

where, $\mathcal{Y}$ is the support of the random variable Y. The value of $D_{\mathrm{KL}}(R||K)$ is always nonnegative and is zero if and only if $R = K$ [67]. When the logarithm to base 2 is used in this equation the units are bits, whereas the units are nats when the natural logarithm is used. Eq. (9) is used to measure the "distance" between the desired quantiles and those obtained from the sample. A feasible probabilistic inversion problem will result in KL distance at or close to zero. In the case of an infeasible problem, PARFUM will minimize the KL distance, whereas the convergence behavior of IPF is not as predictable. IPF has been show to converge more quickly than PARFUM, but the speed of both is not a practical concern with modern computers.

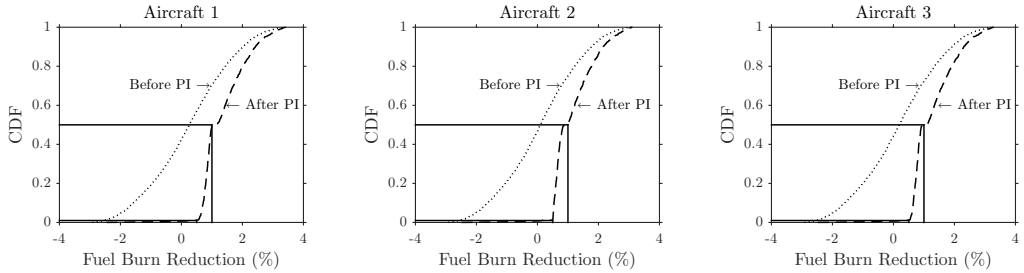
For the application of interest here, it is important that the probabilistic inversion problem be a feasible problem so that the quantile constraints are met. If it is not feasible, then the bounds on the k-factors and/or the desired quantiles for the system-level metrics may need to be adjusted. An exploratory analysis of the system-level metric samples can be used to quickly determine whether the desired quantiles are feasible given the k-factor bounds. Also, increasing the sample size can help in some cases. Once a feasible problem is established, it is possible that IPF and PARFUM will produce different solutions. If this is the case, then the solution that is closer to that which characterizes the present state of uncertainty may be preferred by an analyst. Csiszar [68] showed that IPF is capable of converging to the distribution that has minimum KL distance relative to the initial distribution out of the set of distributions that meet the quantile constraints. PARFUM has not been shown to share this property with IPF. For a thorough illustration and comparison of the two algorithms, the reader is referred to Ref. [69].

The k-factor samples for the AFC example are shown in Fig. 15a. Since $N =$ 5,000 was used, each sample had an initial weight of 1/5,000. The resulting discrete distribution after applying the IPF algorithm for sample re-weighting is shown in Fig. 15b. This figure shows how the weights were changed after running IPF until



(a) Initial sample probability weights (b) Probability weights after IPF

Figure 15: k-factor samples before and after probabilistic inversion for the notional AFC example.



(a) (b) (c)

Figure 16: CDFs for the notional AFC example before and after probabilistic inversion.

the KL distance between the desired quantiles and the solution was 1E-8. When propagated to the fuel burn metrics, this proabilistic inversion solution meets the quantile constraints. Evidence of this is shown for all three aircraft in Fig. 16. The CDFs before probabilistic inversion do not meet the quantile constraints, which are drawn as dotted lines, whereas the solution after probabilistic inversion aligns with the constraints.

4.3.3 Step Three: Create Value Model

This step comprises the creation of a model of the decision makers' value that can be used to evaluate the alternatives. To accomplish this, Keeney [59] proposed a

five-step process. Each of these steps is briefly described here.

4.3.3.1 Introduce Nomenclature and Concepts

To conduct a decision analysis properly, decision makers must be educated to understand terms and concepts that are necessary for communicating their preferences. Of particular importance is that the decision makers have a thorough understanding of the attributes, the corresponding objectives, and the approach to modeling impacts of the alternatives on the attributes. The analyst must also ensure that the decision makers know there are no right or wrong preferences and that the model of their preferences can be modified at any time during the analysis.

4.3.3.2 Determine the Form of the Multiattribute Utility Function

The analyst has to determine the form of the multiattribute utility function by determining which of three value independence conditions hold: preferential independence, utility independence, and additive independence. Preferential independence means that the decision makers' rank ordering of preferences for any single attribute is independent of the fixed values of all other attributes. Preferential independence is implied by utility independence, so preferential independence need not be tested if utility independence is satisfied. Utility independence between two attributes X_1 and X_2 means that the degree of risk aversion encoded in the utility function of X_1 for a fixed setting of X_2 does not depend on the value of that fixed setting. The test for utility independence is shown in Fig. 17. In this figure the subscripts A, B, and C indicate different levels of the attributes. Note that utility independence and preferential independence conditions lack a reflexive property, meaning, for example, that X_1 being utility independent of X_2 does not imply that X_2 is utility independent of X_1. All attributes in a decision analysis are mutually utility independent if all subsets of $\{X_1, X_2, \ldots, X_l\}$ are utility independent of the complement of each [52]. If attributes $X_1, X_2, \ldots, X_l$ are mutually independent, then the appropriate form of

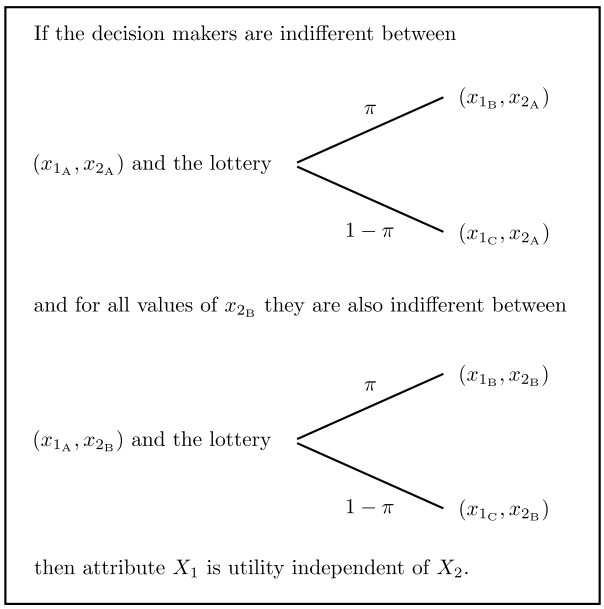


Figure 17: A test to determine if X_1 is utility independent of X_2 (adapted from Ref. [55]).

the multiattribute utility function is the multiplicative form, defined as:

$$U(\mathbf{x}) = \frac{1}{K}\left[\left[\prod_{i=1}^{l}(K\lambda_i U_i(x_i)+1)\right]-1\right] \tag{10}$$

where, $U(\mathbf{x})$ is the overall utility, scaled from 0 to 1, for the attribute vector $\mathbf{x} = (x_1, x_2, \ldots, x_l)$; x_i is the level of attribute X_i; the $U_i(x_i)$ are single-attribute utility functions, also scaled from 0 to 1; the λ_i are single-attribute scaling constants; and K is a normalizing constant that ensures the range of $U(\mathbf{x})$ is 0–1. By enforcing that $U(\mathbf{x}^*) = 1$ and all $U_i(x_i^*) = 1$ when the attributes are at the best levels $\mathbf{x}^*$, Eq. (10) reduces to an equation that can be used to solve for K:

$$K+1 = \prod_{i=1}^{l}(K\lambda_i+1) \tag{11}$$

A simpler form of the multiattribute utility function is used when the additive independence condition in Fig. 18 is also satisfied. Again, in this figure the subscripts A and B indicate different levels of the attributes. The additive form is

$$U(\mathbf{x}) = \sum_{i=1}^{l}\lambda_i U_i(x_i) \tag{12}$$

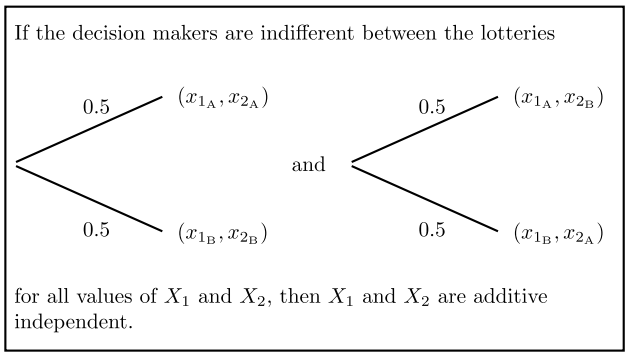


Figure 18: A test for additive independence between X_1 and X_2 (adapted from Ref. [55]).

Note that the single-attribute scaling constants are denoted by λ_i instead of k_i, which is the notation commonly found in the literature. This is done here to avoid confusion with k-factors.

One of the reasons that as few attributes as possible should be used is to minimize the number of independence conditions that must be checked. Practitioners of utility analysis have documented claims that in most practical problems attributes will fail the test for additive independence. This is likely to be the case for technology development activity downselection as well. It is more difficult to make general statements about mutual utility independence for all attributes. For example, some decision makers' preference for cost risk aversion may be dependent on the level of uncertainty reduction achieved, and the test in Fig. 17 for these two attributes may be negative. If this is the case, it may still be possible to construct a multiattribute utility function, but the form of the function will be more complex and require elicitation of many more preferences over the consequence space. Keeney and Raiffa [52] proposed multiple options for a case of utility dependence. One of the simplest methods is to aggregate the attributes into one. This could be done with the attributes of cost, uncertainty reduction, and system-level performance by, for example, multiplying or summing normalized versions of the three. Then, there would not be any

independence conditions to check and only one utility function to assess, but the attribute might be more difficult for decision makers to understand and make preference statements for.

After selecting a multiattribute utility function, one might be tempted to interpret the scaling constants λ_i as indicators of relative attribute importance. Keeney and Raiffa [52] stressed that the scaling constants cannot be interpreted this way. For example, if $\lambda_{\text{Uncertainty}} = 0.75$ and $\lambda_{\text{Cost}} = 0.25$, it cannot be concluded that uncertainty reduction is three times more important than cost. This is because changing the ranges of one attribute could result in scaling constants that would lead to a completely different interpretation of importance. What *can* be said about the scaling constants is that they indicate which attributes the decision makers would prefer to see improvements in. For instance, if the decision makers would prefer to see the level of uncertainty reduction shift from the lower bound to the upper bound than cost shift from the lower bound to the upper bound, then $\lambda_{\text{Uncertainty}} > \lambda_{\text{Cost}}$.

4.3.3.3 Elicit Single-Attribute Utility Functions

Procedures for eliciting decision makers' preferences over single attributes and multiple attributes are abundant in the literature. A summary of the general process is discussed here.

Before beginning the assessment, the analyst must specify bounds on the attributes. The bounds can be global, best and worst expected, or acceptable. As an example, global bounds of 0 and 1 are appropriate for the proposed system-level attribute $\text{P}(D \geq D_{\text{Target}})$ because it is a probability. For cost as a percentage of budget, the analyst may decide to limit the range to 60%, for example, if none of the alternatives are expected to reach that level of cost. Once the attribute bounds have been established, the next task is to assess the decision makers' risk attitude. To be clear, risk aversion in MAUT means that decision makers always prefer a consequence

$\frac{x_A+x_B}{2}$ with probability 1 to a lottery yielding x_A with probability 0.5 and x_B with probability 0.5, where A and B indicate low and high levels of the attribute X. This means that the decision makers would rather accept the average of the two attribute values than participate in a 50-50 gamble that could result in either the better or worse consequence. Risk prone decision makers would prefer the opposite.

The standard approach to eliciting risk attitudes is to use a series of questions regarding lotteries. Decision makers can exhibit different degrees of risk aversion/proneness depending on what region of the attribute range the lottery questions pertain to. In his seminal paper, Pratt [70] argued that decision makers' risk attitude over the attributes restricts the functional form of the single-attribute utility functions. To measure the local risk aversion of a risk averse utility function, he proposed the function

$$\gamma(x) = -\frac{U_i''(x)}{U_i'(x)} \tag{13}$$

Lottery questions form the basis for building utility functions, but analytical functions are typically selected and then fit to particular points over the attribute range. Although many functional forms for utility functions can be considered, an exponential form is often used because it models constant risk aversion/proneness over the attribute range. An example is the form

$$U(x) = a + be^{cx} \tag{14}$$

where, c captures the degree of risk aversion/proneness, and a and b are constants that ensure $U_i(x)$ is normalized between 0 and 1. Note that for this exponential form, $\gamma(x) = -c$. If $c > 0$ and utility increases with the attribute, then $\gamma(x)$ is negative for all x indicating a convex utility function with constant risk proneness. If $c < 0$ and utility increases with increasing attribute level, then $\gamma(x)$ is positive for all x indicating a concave utility function with constant risk aversion. For a utility function that decreases with increasing attribute level, the negative is dropped in

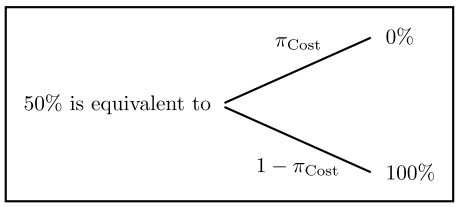


Figure 19: An example of a lottery question for building the single-attribute utility function for cost.

Eq. (13), and the inequalities for c are reversed.

After selecting the form of the utility function, the constants must be determined. For instance, the constants in Eq. (14) would be found by eliciting three points on the utility function to provide three independent equations, then simultaneously solving the three equations. Suppose that this process is followed for the attribute cost, as a percentage of total budget, and bounds have been set at 0% and 100%. The utility function would be anchored at the lower and upper bounds as $U_{\text{Cost}}(0\%) = 1$ and $U_{\text{Cost}}(100\%) = 0$ to provide two equations. The third equation is then found by first asking the decision makers a lottery question about what their indifference probability π_{Cost} is, which is illustrated in Fig. 19. The indifference probability is that at which the decision makers are indifferent between a cost of 50% of the budget with probability 1 and a lottery in which there is a π_{Cost} probability of the cost being 0% and a $1 - \pi_{\text{Cost}}$ probability of the cost being 100%. Then, because the decision makers are indifferent between the two options, the expected utilities are set equal: $U_{\text{Cost}}(50\%) = \pi_{\text{Cost}} U_{\text{Cost}}(0\%) + (1 - \pi_{\text{Cost}})\, U_{\text{Cost}}(100\%)$.

4.3.3.4 Determine Scaling Constants

To solve for the scaling constants λ_i in Eq. (10) or (12), a system of l independent equations is needed. Certainty scaling, probabilistic scaling, or a combination of the two can be used to generate the set of equations. Certainty scaling entails identification of two levels of all attributes that are considered indifferent and equating the utilities at those levels. For two attributes, this would mean that levels A and B need to be found

such that $U(x_{1_A}, x_{2_A}) = U(x_{1_B}, x_{2_B})$. Probabilistic scaling involves a lottery question with all attributes to find an indifference probability. In the case of two attributes, an indifference probability π_1 must be found so that (x_{1_B}, x_{2_A}) with certainty is indifferent to a lottery with π_1 probability of (x_{1_B}, x_{2_B}) and $1-\pi_1$ probability of (x_{1_A}, x_{2_A}). Then, the expected utilities are equated: $U(x_{1_B}, x_{2_A}) = \pi_1 U(x_{1_B}, x_{2_B}) + (1 - \pi_1)\, U(x_{1_A}, x_{2_A})$. If Eq. (10) is used for the multiattribute utility function, then the normalizing constant K is solved for using Eq. (11) after the scaling constants have been determined.

4.3.3.5 Check for Consistency

After the multiattribute utility function has been constructed, the final step is to test for consistency of the utility function. Tests for consistency are capable of revealing whether the utility function properly represents decision makers' preferences. This is an important step because the efficacy of evaluating alternatives with MAUT hinges on the accuracy of the utility function. Keeney and Raiffa [52] suggested three consistency checks. One check entails pairwise comparisons of points in the consequence space to make sure that the preferred points have higher utility than the less preferred points. Another check involves lottery questions to determine whether the decision makers are risk averse/prone along multiple vectors in the consequence space. The third consistency check is to use lottery questions to ensure the appropriate sign of the scaling constants.

If any consistency checks are failed, the decision makers should be made aware of this and at least part of the elicitation procedure repeated. If the decision makers are unsure about some of their preference statements, sensitivity analysis can be used to study the effects of their uncertainty on the valuation of alternatives.

At this point in the formulation, it is appropriate to mention how the preference elicitation process can be implemented when multiple decision makers are involved. One relatively complex approach is to aggregate the utilities of multiple decision

makers with a group utility function that requires additional value assessments to quantify the relative importance of each decision maker. The reader is referred to Refs. [71, 72, 52] for details of the elicitation process for a group utility function. A simpler approach is to elicit utility functions from the group of decision makers, treating the group as an individual decision maker. The potential issue with this technique is that consensus may not be reached with the group. Finally, the utility analysis can be conducted for each decision maker individually. In the ideal scenario, the top alternatives will be common to all decision makers, or at least some dominated alternatives can be removed from the set. With this approach, sources of conflict between the assessments of each decision maker can be identified to stimulate discussion and adjust the assessments, ultimately converging on a single ranking of alternatives.

4.3.4 Step Four: Model Impacts of Alternatives

Up to this point in the formulation, RQ 1.1 has been addressed. Steps four and five address RQ 1.2. As previously mentioned, one of the overarching assumptions is that a probabilistic model of technology impacts, also referred to as k-factors, exists. To model the effects of technology development activities on the probabilistic model, mathematical operations that map these effects to changes in the characteristics of the joint distribution are needed. Following the rationale for the objectives and attributes defined in Sec. 4.3.1, the primary effects that are of interest at the technology level are uncertainty reduction and performance change, as these will directly affect the attributes for system-level uncertainty reduction and system-level performance. Additionally, the cost for each activity needs to be estimated.

If a parametric distribution is used to characterize the technology-level uncertainty, then it is possible to model the effects of technology development activities by modifying the parameters of the distribution. As an example, suppose that the k-factor uncertainty is represented by a multivariate normal distribution: $\mathbf{k} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

where, $\boldsymbol{\mu}$ is the vector of mean values and $\boldsymbol{\Sigma}$ is the covariance matrix. To implement a change in the mean of any technology impact k_i, one could add a constant δ_i to the corresponding element in the mean vector μ_i. For a change in variance, a new variance value Σ_{ii} would be substituted in the covariance matrix. Modeling these impacts is simple for the multivariate normal distribution, but the majority of parametric distributions do have parameters that are as easily interpreted. Besides, it is possible that a *non*parametric distribution characterizes uncertainty. This would be the case if the k-factor distribution is quantified by Monte Carlo simulation. Then, there would not be any parameters to vary the characteristics of the distribution with but rather one would have to resort to operations on the sample.

To build a more widely applicable methodology it was decided that an approach for modeling the effects of technology development activities on the k-factors was needed that can accommodate both parametric and nonparametric distributions. From probability theory, it is know that adding a constant δ to a random variable k results in a translation of the mean by that constant amount: $\mathrm{E}[k+\delta] = \mathrm{E}[k] + \delta$. Based on this result, it was decided to model performance change as a translation of the distribution on the k-factors. Variance change is more complicated. If a random variable k is multiplied by a constant α, the variance is changed to $\mathrm{Var}(\alpha k) = \alpha^2 \mathrm{Var}(k)$, but the mean is affected as well: $\mathrm{E}[\alpha k] = \alpha \mathrm{E}[k]$. Ideally, the translation of mean and change in variance can be implemented independently, so a method called the mean-preserving transformation [73] was borrowed from the operations research literature for modeling variance change. To simultaneously model performance translation and variance change for a joint k-factor distribution, the following equation has been derived:

$$\mathbf{k}_{\text{Transformed}} = \underbrace{\boldsymbol{\alpha} \circ \mathbf{k} + (\mathbf{1} - \boldsymbol{\alpha}) \circ \mathrm{E}[\mathbf{k}]}_{\text{mean-preserving transformation}} + \underbrace{\boldsymbol{\delta}}_{\text{mean translation}} \tag{15}$$

where, $\mathbf{k}_{\text{Transformed}}$ is the transformed distribution on the k-factors, $\boldsymbol{\alpha}$ is a vector of variance-scaling parameters, $\boldsymbol{\delta}$ is a vector of mean-translating parameters, and the

symbol $\circ$ denotes the Hadamard product (element-wise multiplication). The effect of $\boldsymbol{\alpha}$ is found by deriving the variance of the marginal distributions of $\mathbf{k}_{\text{Transformed}}$: $\text{Var}\left(k_{i_{\text{Transformed}}}\right) = \alpha_i^2 \text{Var}(k_i)$. Thus, if $\alpha_i < 1$, the variance of k_i is reduced, and the variance is increased for $\alpha_i > 1$. The means, however, are translated by $\boldsymbol{\delta}$ and not affected by $\boldsymbol{\alpha}$: $\text{E}\left[k_{i_{\text{Transformed}}}\right] = \text{E}[k_i] + \delta_i$.

Note that the dependence characteristics of the joint distribution will not necessarily be maintained after scaling variance. As an example of this, consider a joint distribution on two k-factors with the variance of one of the variables scaled. It can be shown that the covariance of the two k-factors after transforming one of them is: $\text{cov}\left(k_{1_{\text{Transformed}}}, k_2\right) = \alpha_1 \text{cov}\left(k_1, k_2\right)$. Hence, if variance reduction is implemented by setting α_1 to a value less than one, then the covariance will be reduced as well. This effect is one of the drawbacks of the proposed approach, but it is a necessity if each component of $\mathbf{k}$ is to be transformed independently.

4.3.4.1 Probability Distribution Elicitation Methods From the Literature

Determining the mapping between technology development activities and $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and cost requires input from technologists who are familiar with the technology and the alternatives that are being considered. Since the purpose of this methodology is to support decisions before the activities have been designed in detail, it is unlikely that any of the mean-translation parameters, variance-scaling parameters, or costs can be specified with certainty. To represent the epistemic uncertainty surrounding these variables, probability distributions should be elicited from technologists. When the variables are treated independently, elicitation would typically entail the technologist first identifying multiple probabilities over intervals or multiple quantiles of their subjective distributions. Then, either parametric or nonparametric distributions would be fit to the summaries provided by the technologist. Finally, consistency checks are used to determine how well the fitted distributions agree with the technologist's

opinions. The elicitation process is more complicated when there are dependencies between the activity impacts. For instance, performance translations could be positively correlated with cost. A vast literature exists regarding the elicitation of probabilities. The reader is referred to the paper by Garthwaite et al. [74] for a review of the state of the art. When probabilities must be elicited from multiple technologists, the approach taken depends on whether the technologists interact or not during elicitation. If they do not interact, then separate elicitation sessions occur for each technologist, and the results are aggregated using weighted combinations of each technologist's probabilities. If the technologists interact, then the typical elicitation approach is to facilitate a discussion between the technologists in an attempt to reach a consensus view. The reader is referred to the seminal paper by Genest and Zidek [75] for a review and critique of techniques for combining probability distributions.

It should be noted that scaling the variance of k-factors to model epistemic uncertainty reduction is not entirely appropriate if the joint distribution is composed of aleatory and epistemic uncertainties. If it is possible to decompose the distribution into epistemic and aleatory sources, then one way around this is to scale only the epistemic component.

4.3.5 Step Five: Quantify Expected Utility for Each Alternative

With a multiattribute utility function constructed and the effects of all alternatives mathematically represented as cost and changes to the technology impact distributions, distributions on overall utility must be characterized for each alternative. This is an uncertainty propagation problem with multiple layers, as shown notionally with two k-factors and two system-level metrics in Fig. 20. First, distributions on $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ are sampled to generate multiple distributions on $\mathbf{k}$. Each sample from the $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ distributions corresponds with a joint k-factor distribution that has translated means and scaled marginal variances. Next, each of the distributions on $\mathbf{k}$ is

propagated through the M&S environment to produce a series of joint distributions on the system-level metrics **M**. Then, each joint distribution on **M** is evaluated in terms of performance and variance reduction to ultimately generate distributions on $\mathrm{P}\left(D \geq D_{\text{Target}}\right)$ and average variance reduction. The distributions on all three attributes are then propagated to distributions on each of the single-attribute utility functions. Finally, the uncertainty surrounding the single-attribute utility functions is propagated to the multiattribute utility, and the expected utility is computed:

$$\mathrm{E}[U] = \int_0^1 u\, p(u)\, du \tag{16}$$

The value of Eq. (16) can be used to rank the alternatives. At this point, sensitivity analyses should be carried out to determine the effect of preferences elicited from the decision makers and distribution assumptions on the expected utility of the alternatives. The sensitivity analyses may reveal that the ranking of the top alternatives is sensitive to certain parameters that the decision makers and/or technologists are unsure of. This kind of result can help identify the parameters that are most important to establish conclusively.

4.4 Illustrative Example: Technology Development Activity Evaluation

An example problem was built to illustrate the merits of the proposed methodology. Three modern technologies have been selected for the example. The problem entails the evaluation of four technology development activity alternatives, each of which targets uncertainty reduction and performance improvement for one of the three technologies. For simplicity, it is assumed that there is a single decision maker. In this section, the problem setup is described, the methodology implementation details are explained, and the results are presented and discussed.

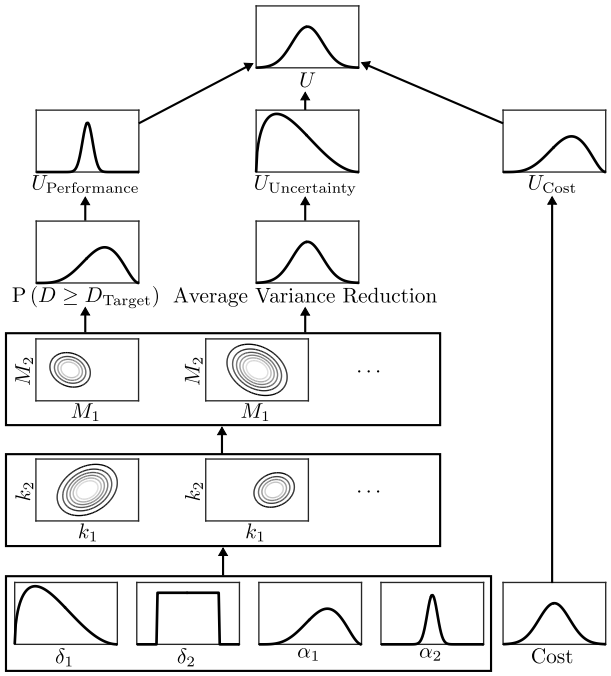


Figure 20: Propagation of uncertainty to multiattribute utility.

4.4.1 Problem Setup

Setting up the problem involved identifying the elements that are assumed to exist, which are described in Sec. 4.1.1, and generating alternatives.

4.4.1.1 Integrated System and Advanced Technologies

The integrated system that was identified for technology infusion is an LTA commercial aircraft, similar in baseline technology and performance to the Boeing 777. For the example, the goal of technology infusion for this aircraft is to simultaneously reduce block fuel burn and sideline noise while restricting the increase of takeoff field length (TOFL) with one engine operative. These goals were characterized with sets of values and associated minimum probabilities of success. The baseline values for each system-level metric, desired lower bound of reduction, and target reduction, are listed in Table 4. The probabilities of meeting or exceeding the goals are in parentheses next to the values.

Table 4: LTA aircraft system-level metric baseline values and goals for the example

Metric	Baseline Value	Lower Bound	Target
Sideline Noise	97.6 dB	0.1% (0.99)	0.7% (0.6)
Block Fuel	229,567 lb	1.0% (0.99)	5.0% (0.6)
TOFL	8979 ft	-2.5% (0.99)	N/A

The technologies that have been selected for development and are slated for infusion with the LTA vehicle include: (1) an AFC technology for enhancing the side force generated by the vertical tail, (2) an adaptive compliant trailing edge (ACTE) technology for wing gust-load alleviation, and (3) a fan vertical acoustic splitter to suppress the aft fan discharge noise during takeoff. The aim of the AFC technology is to control flow separation over the vertical tail to increase side force during critical one-engine-operative low-speed conditions. By increasing the side force of a vertical tail, the AFC technology enables the design of smaller vertical tails, resulting in drag and weight reduction for the vehicle. However, the addition of an AFC power supply architecture on board an aircraft will result in additional weight, and the subsystem supplying pressurized flow to the fluidic actuators, such as an auxiliary power unit (APU), would burn additional fuel. The idea behind the ACTE technology is to actively reduce wing bending moments during gusts in flight, thereby enabling wing design with a lighter structure. The fan vertical acoustic splitter technology is simply a splitter plate mounted in the engine bypass flow aft of the fan.

4.4.1.2 M&S Environment

In order to map the impacts of the three technologies to the three system-level metrics, a credible M&S environment was required. The Environmental Design Space (EDS) was selected for this purpose. EDS was developed for the U.S. Federal Aviation Administration (FAA) Office of Environment and Energy to enable thorough assessment of the environmental effects of aviation [76]. EDS is physics-based, integrated, and multidisciplinary. It consists of core modules originally developed by

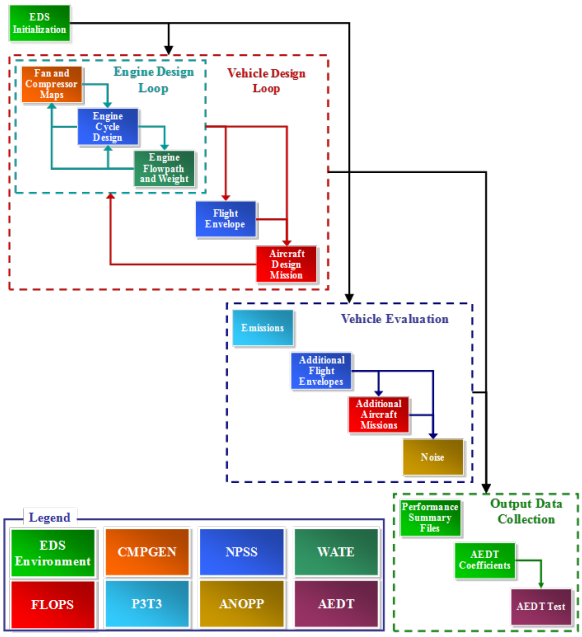


Figure 21: Diagram of the M&S environment used in the example.

NASA, and the modules are coupled with design rules and user-defined propulsion and airframe design parameters to generate and analyze aircraft designs with engine-to-airframe matching. A flow chart of EDS execution for a single aircraft is shown in Fig. 21. Propulsion system design modules include CMPGEN for compressor map generation, Numerical Propulsion System Simulation (NPSS) for thermodynamic cycle analysis, and Weight Analysis of Turbine Engines (WATE++) for engine flow path analysis and weight estimation. Vehicle sizing and synthesis is accomplished with the FLight OPtimization System (FLOPS) code, and vehicle noise is predicted with Aircraft Noise Prediction Program (ANOPP). EDS has been vetted through its use in multiple programs of record.

An existing EDS baseline LTA vehicle model was used. To model the three technologies, EDS k-factors were identified to represent the impacts. After the k-factors were identified, Weibull probability distributions were constructed to model the baseline uncertainty surrounding the technology impacts. The distributions are notional,

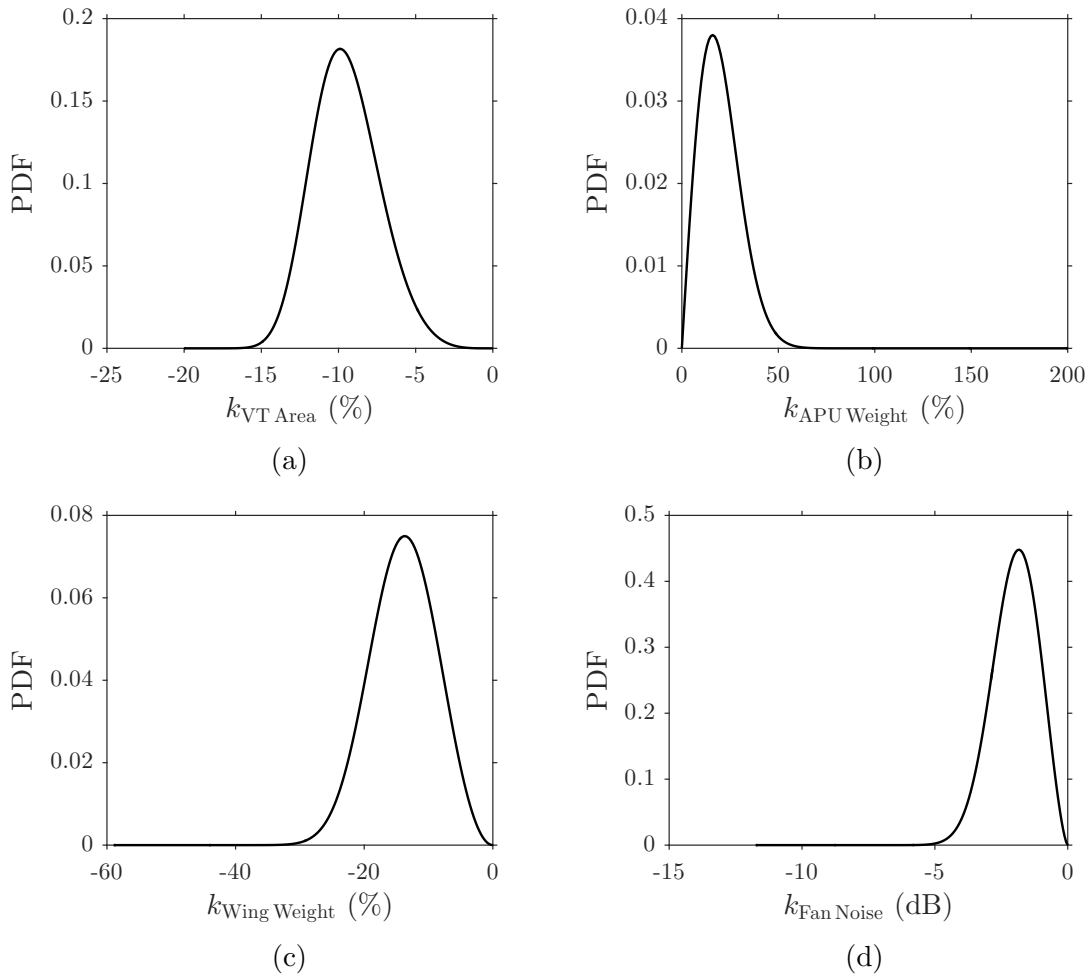


Figure 22: Baseline PDFs for the k-factors used in the example.

but their selection was partly informed by documented performance of the real technologies. For the purposes of illustration, each k-factor distribution is independent of all others, and they are plotted in Fig. 22. For the AFC technology, $k_{\text{VT Area}}$ represents vertical tail area reduction due to design with enhanced side force, and $k_{\text{APU Weight}}$ represents increased vehicle weight due to the installation of an AFC power supply architecture that delivers flow from an APU bleed point to the actuators. The ACTE technology wing weight reduction impact was represented by $k_{\text{Wing Weight}}$. Noise reduction at takeoff for the fan vertical acoustic splitter was modeled with $k_{\text{Fan Noise}}$. Note that there are performance enhancements and penalties associated with the use of all three technologies, but only the salient impacts were used in this example.

The EDS environment runs relatively quickly on a desktop computer, but it was decided to expedite the uncertainty propagation process by using surrogate models to approximate EDS predictions. Surrogate models were available for the system-level responses of interest here. These surrogate models were created by generating a 15,000-case design of experiments sample, filtering out failed cases, and fitting artificial neural network regression models. Although four EDS inputs were used as k-factors for the example, the surrogate models were built with over 200 input variables. An assessment of the predictive accuracy of the surrogate models is presented in Appendix A.

4.4.1.3 Alternatives

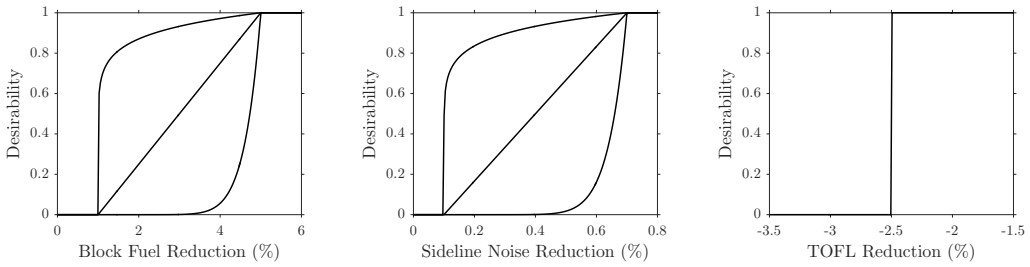
Four alternatives were generated for evaluation, and they are enumerated in Table 5. The first activity involves a computer experiment with a high-order, physics-based model to characterize the relationship between control variables for the fan vertical acoustic splitter and engine noise. The second activity entails a full-scale wind tunnel experiment to characterize the relationship between AFC system control variables and side force enhancement for a vertical tail model. The third alternative is a computer model-based design study to estimate wing weight reduction due to the use of ACTE on a clean-sheet wing design. The last alternative is a full-scale flight test to measure the effectiveness of ACTE for gust-load alleviation. Note that it is unlikely that these alternatives would be compared in a real technology development setting. They have been selected solely for the purpose of demonstrating the mechanics of the proposed methodology with a diverse set of alternatives. This is discussed in more detail in Sec. 4.5.

4.4.2 Implementation of the Proposed Methodology: Expected Utility

The steps described in Sec. 4.3 were followed for this example, and the details of each step are presented here.

Table 5: Technology development activity alternatives for the example problem

Alternative	Description
A_1	Computer experiment to investigate the effects of control variables on noise reduction for the fan vertical acoustic splitter technology
A_2	Full-scale wind tunnel experiment to investigate the effects of control variables on AFC effectiveness for the AFC-enhanced vertical tail technology
A_3	Design study to predict wing weight reduction for a clean-sheet wing design with the ACTE technology
A_4	Full-scale flight test to measure gust-load alleviation effectiveness of the ACTE technology



(a) (b) (c)

Figure 23: Desirability functions for the system-level metrics used in the example problem.

4.4.2.1 Step One: Establish Objectives and Attributes

The objectives used in the example included maximization of uncertainty reduction, maximization of system-level performance, and minimization of cost to execute the activities. The corresponding attributes formulated in Sec. 4.3.1 were used. For the performance attribute, desirability functions were created that correspond with the goals listed in Table 4, and these functions are shown in Fig. 23. The block fuel reduction and sideline noise reduction desirability functions are shown for exponent parameter r values of 0.1 (concave), 1 (linear), and 10 (convex). The TOFL desirability is a step function because there is only a lower bound for this metric.

4.4.2.2 Step Two: Conduct Probabilistic Inversion

The steps for probabilistic inversion described in Sec. 4.3.2 were applied to the example problem and are discussed here.

The uncertain input variables for the EDS M&S environment were specified in Sec. 4.4.1.2. The ranges on each of the k-factors were selected to encompass the probability distributions shown in Fig. 22, and these ranges are listed in Table 6.

Table 6: k-factor ranges for probabilistic inversion

Technology Impact	Range for Probabilistic Inversion
$k_{\text{VT Area}}$	$[-20, 0]$ (%)
$k_{\text{APU Weight}}$	$[0, 75]$ (%)
$k_{\text{Wing Weight}}$	$[-40, 0]$ (%)
$k_{\text{Fan Noise}}$	$[-6, 0]$ dB

The output variables and quantiles for the marginal distributions are enumerated in Table 4. Note that the quantile probabilities in the table are defined such that $\pi = 1 - \mathrm{P}(M_i \leq m_i)$ for $i = 1, 2, 3$.

The k-factor space was sampled with 500,000 draws from independent uniform distributions with the bounds listed in Table 6. Then, the joint k-factor distribution was propagated through the EDS surrogate models to produce samples for the three system-level metrics.

As an experiment, probabilistic inversion was conducted using two methods for generating the initial weights. One method is the business-as-usual approach of assigning weights associated with a discrete uniform distribution, and the other method is to assign weights from the baseline distribution on $\mathbf{k}$. The iterative PARFUM and IPF algorithms were used for sample re-weighting with both weighting methods to investigate which algorithm would provide a solution that is closer to the initial distribution. The probabilistic inversion problem was found to be feasible, and both algorithms were converged to a KL distance of 1E-8. For the experiment, 100 replications of 500,000 draws from the k-factor space were used to capture variability due to pseudo-random sampling. A measure often referred to as J distance is the sum of two KL distances, as they are defined in Eq. (9): $J(R||K) = D_{\text{KL}}(R||K) + D_{\text{KL}}(K||R)$.

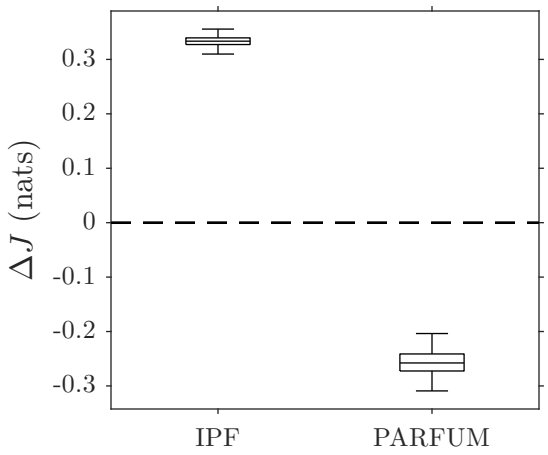
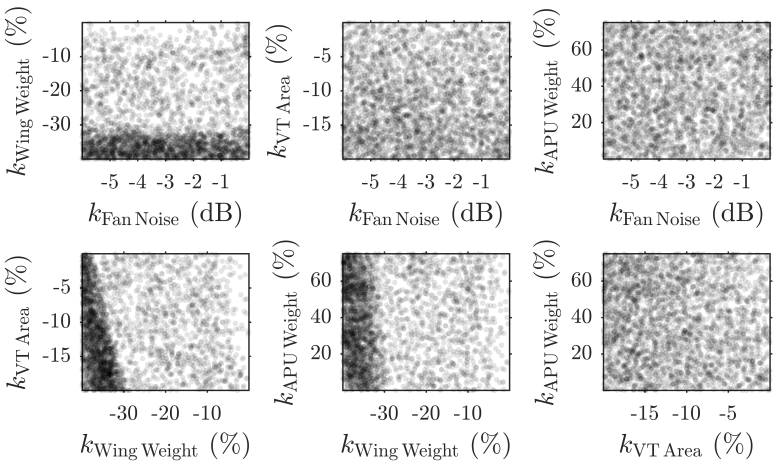


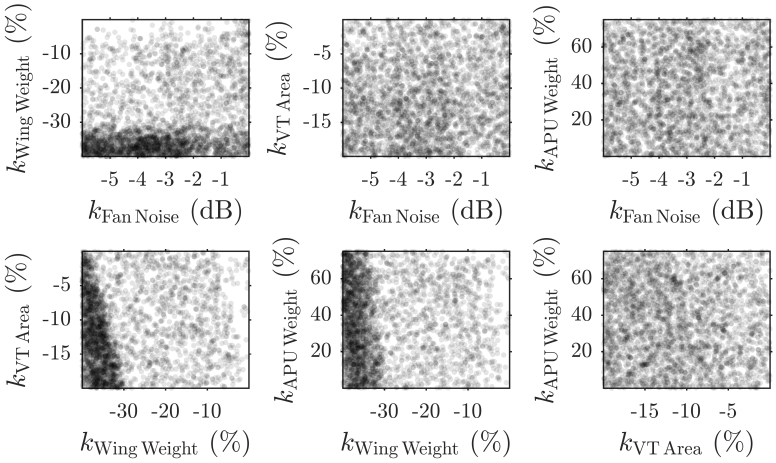
Figure 24: Effectiveness of IPF and PARFUM at producing solutions similar to the baseline k-factor distribution.

J distance was used to estimate the distance between: (1) the solutions that corresponded with the uniform initial sample weights and the k-factor distribution, and (2) the distance between the solutions that corresponded with the k-factor initial sample weights and the k-factor distribution. Then, ΔJ was calculated by subtracting the second distance from the first to quantify whether using initial weights from the k-factor distribution would result in a solution that is closer to the k-factor distribution. It was hypothesized that IPF would produce relatively closer solutions than PARFUM because of the theoretical result from Csiszar [68] that was discussed in Sec. 4.3.2.4. The results of this experiment support this hypothesis and are summarized in Fig. 24. IPF consistently resulted in $\Delta J > 0$, indicating that IPF produced solutions that were relatively closer to the baseline k-factor distribution. $\Delta J < 0$ for all of the PARFUM solutions, indicating that PARFUM consistently produced solutions that were relatively farther from the baseline k-factor distribution.

The IPF solution with initial weights specified according to the baseline k-factor Weibull distributions was carried through the rest of the example. For comparison, solutions from both uniformly distributed initial weights and k-factor distribution initial weights were computed. Samples of size 3,000 were drawn from the solutions



(a) Uniform initial weights



(b) Weibull initial weights

Figure 25: Scatterplots showing two solutions from probabilistic inversion.

and are shown in Fig. 25. The plots illustrate that higher probability density was placed at large wing weight reductions to meet the quantile constraints, indicating the significant sensitivity of the system-level metrics to the wing weight k-factor. Minor differences in density of the samples can be found when visually comparing the two solutions.

4.4.2.3 Step Three: Create Value Model

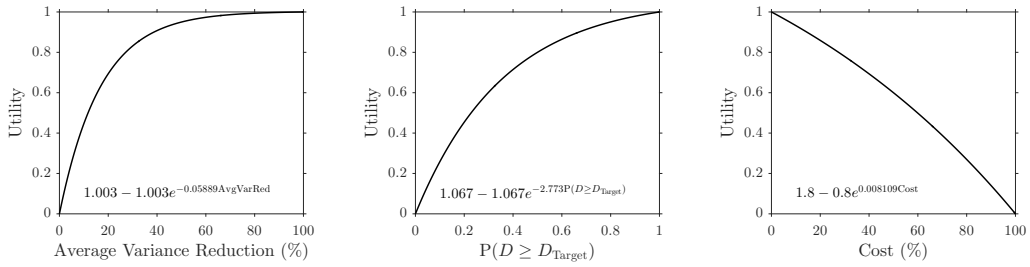
The value model was created by the author, who acted as decision maker and analyst simultaneously. Thus, the steps for introducing nomenclature and concepts and checking consistency were not necessary. However, sensitivity analyses were completed and are presented in Sec. 4.4.4.

Mutual utility independence was assumed for all three attributes, so the multiplicative utility function defined in Eq. (10) was used. The single-attribute utility functions were defined with global bounds. The author decided to approach the problem with a constantly risk-averse attitude over all of the attributes, and the exponential utility function form in Eq. (14) was selected to reflect this attitude. The constants in Eq. (14) were determined by anchoring the lower and upper bounds of each utility function with a value of 0 or 1, then using lottery questions to assess single-attribute indifference probabilities. The results of the lottery questions are shown in Table 7. Note that the indifference probabilities correspond with the probability of the best value for each attribute in the lotteries. With three points on each utility curve, the MATLAB fsolve function was used to solve the system of three equations and three unknowns, and the resulting single-attribute utility functions are shown in Fig. 26.

Table 7: Results of lottery questions to determine single-attribute indifference probabilities

Lottery Value	Cost	Average Uncertainty Reduction	$\mathrm{P}\left(D \geq D_{\mathrm{Target}}\right)$
Certainty	50%	50%	0.5
Best	0%	100%	1
Worst	100%	0%	0
Indifference Probability	0.6	0.95	0.8

With the single-attribute utility functions determined, the next task was to specify the scaling constants in Eq. (10). This was accomplished by implementing probabilistic scaling. The results of the lottery questions are shown in Table 8. Having used probabilistic scaling, the λ_i values are interpreted as indifference probabilities for the



(a) (b) (c)

Figure 26: Single-attribute utility functions used in the example problem.

certainty attribute vector versus a lottery in which the best attribute vector has a probability of λ_i and the worst attribute vector has a probability of $1 - \lambda_i$. The order of attributes in the vectors follows the left-to-right order of the attributes at the top of the table. Finally, the normalizing constant K was calculated using Eq. (11) to be -0.9699.

Table 8: Results of lottery questions to determine multiattribute scaling constants

Lottery Value	Cost	Average Uncertainty Reduction	$\text{P}(D \geq D_{\text{Target}})$
Certainty	(0%,0%,0)	(100%,100%,0)	(100%,0%,1)
Best	(0%,100%,1)	(0%,100%,1)	(0%,100%,1)
Worst	(100%,0%,0)	(100%,0%,0)	(100%,0%,0)
λ_i	0.8	0.7	0.6

4.4.2.4 Step Four: Model Impacts of Alternatives

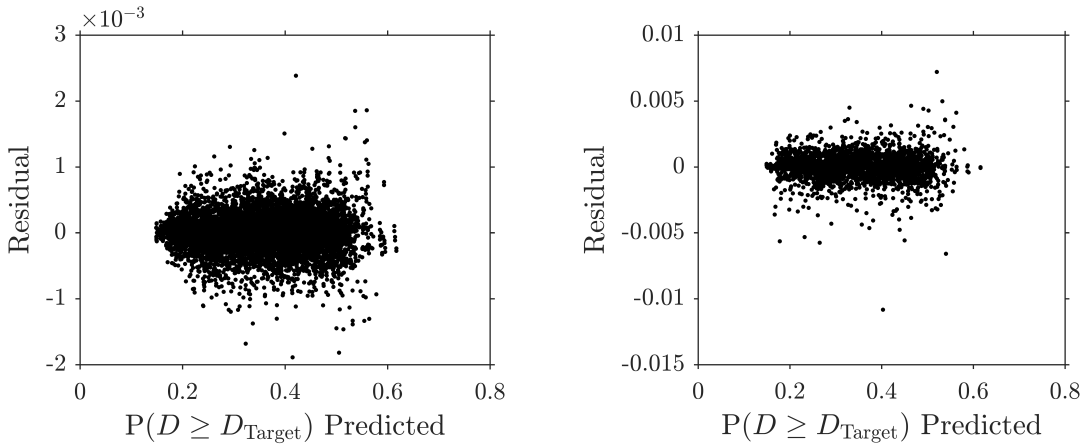
Following the approach proposed in Sec. 4.3.4, variance-scaling distributions, mean-translation distributions, and cost distributions were established for each alternative. Uniform distributions were used to model uncertainty surrounding all of the activity impacts. The lower and upper bounds of the uniform distributions are listed in Table 9. Blank entries in the table correspond with no change in the variable.

Table 9: Uniform distribution bounds for **k** mean translation, **k** variance scaling, and cost of each alternative

Variable	A_1	A_2	A_3	A_4
$\delta_{\text{Fan Noise}}$ (dB)	(-2,0)	-	-	-
$\delta_{\text{Wing Weight}}$ (%)	-	-	(-5,0)	(-10,0)
$\delta_{\text{VT Area}}$ (%)	-	(-3,0)	-	-
$\delta_{\text{APU Weight}}$ (%)	-	-	-	-
$\alpha_{\text{Fan Noise}}$	(0.9,0.95)	-	-	-
$\alpha_{\text{Wing Weight}}$	-	-	(0.95,1)	(0.8,0.9)
$\alpha_{\text{VT Area}}$	-	(0.7,0.85)	-	-
$\alpha_{\text{APU Weight}}$	-	-	-	-
Cost (%)	(5,10)	(20,30)	(3,5)	(50,60)

4.4.2.5 Step Five: Quantify Expected Utility for Each Alternative

Before propagating uncertainty to multiattribute utility, it was decided to expedite the computations by building surrogate models for $\mathrm{P}\left(D \geq D_{\text{Target}}\right)$ and uncertainty reduction. A design of experiments was conducted with $\mathrm{P}\left(D \geq D_{\text{Target}}\right)$ as the response and $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and desirability parameters r_i as the factors. A two-level full-factorial with 1,024 points was generated in addition to 8,976 Latin hypercube points, for a total sample of 10,000. The Latin hypercube design was optimized with a maximin criterion for 5,000 iterations using the MATLAB lhsdesign function. For each sample, 100,000 draws from the probabilistic inversion solution and the k-factor distribution were used to compute $\mathrm{P}\left(D \geq D_{\text{Target}}\right)$. A MATLAB Gaussian process regression model was fit to 7,500 randomly-selected points from the results, and the remaining 2,500 points were used for validation of the regression model. The training and validation residuals are shown in Fig. 27. A similar procedure was used to build a surrogate model for the sum of nondimensional variances for all of the system-level metrics as a function of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$. A total of 8,000 samples were generated. The Gaussian process model was trained using 6,000 randomly selected points from the results, and the remaining 2,000 points were used for validation. The training and validation residuals are shown in Fig. 28. Due to the lack of any obvious patterns in any of the residual plots and low errors for the predictions of the validation data, the surrogate models were deemed



(a) Training data

(b) Validation data

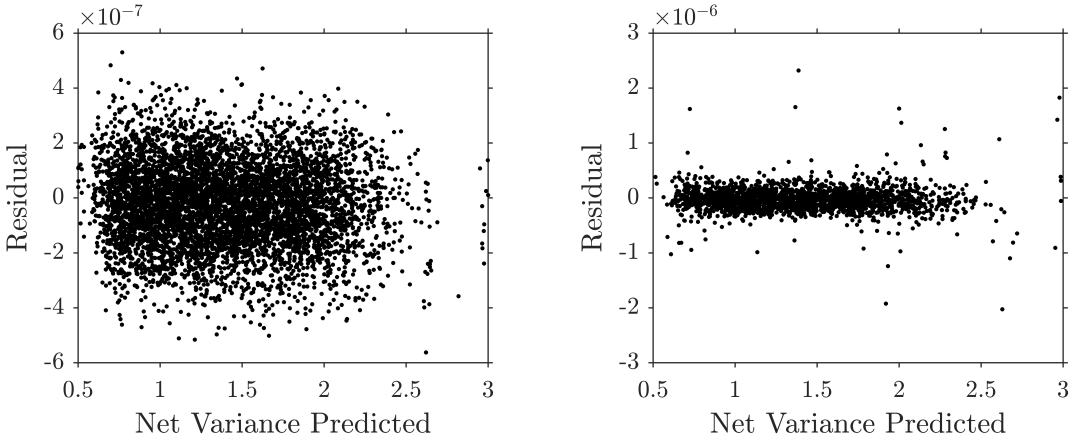
Figure 27: Residuals for the $\mathrm{P}\left(D \geq D_{\text{Target}}\right)$ Gaussian process regression model.

appropriate for use.

To propagate uncertainty from the impacts of the alternatives to multiattribute utility, 10,000 samples were drawn from the $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and cost uniform distributions, and utility was computed for each sample, similar to the flow of data shown in Fig. 20. Expected utility of each alternative was computed using the sample mean from 10,000 samples, and the results are shown in Fig. 29. For a detailed illustration of the propagation process for A_1, see Appendix B. It may be surprising to the reader that the top two activities are both computational. In Sec. 4.4.4, results from sensitivity analyses are presented to probe for further understanding of the relationship between the parameters of the utility model and the rankings.

4.4.3 The Current State of the Art

Two sensitivity analyses were conducted to represent the state of the art. The first sensitivity analysis quantified the effect each technology had on the POS goals listed in Table 4. The sensitivities were quantified by calculating the change in each POS between the LTA vehicle operating with all three technologies and operating with each technology removed one-at-a-time. The other sensitivity analysis quantified the percentage contribution of uncertainty surrounding performance of each technology



(a) Training data (b) Validation data

Figure 28: Residuals for the nondimensional net system-level metric variance Gaussian process regression model.

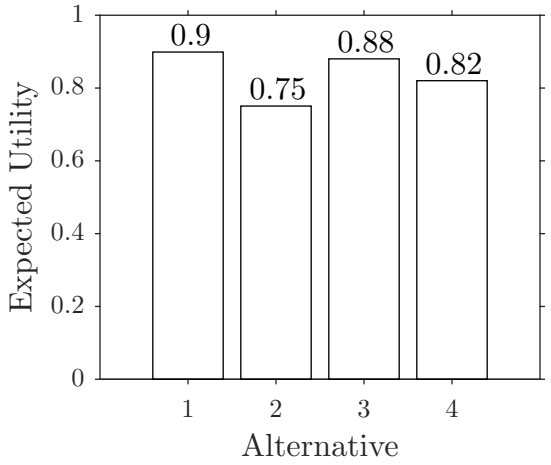


Figure 29: Expected utilities of the four alternatives in the example.

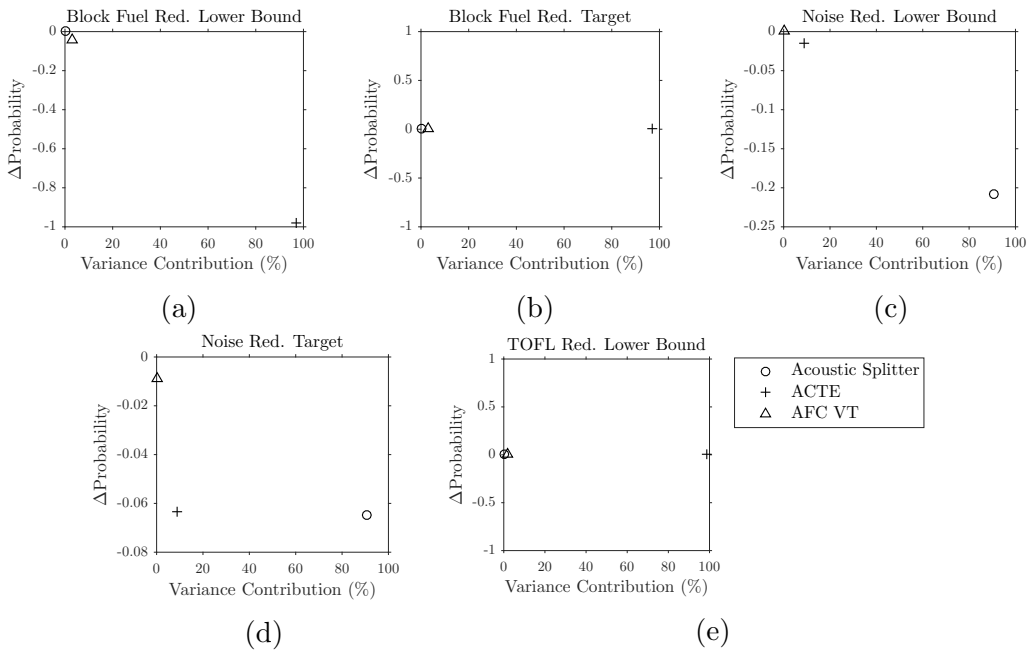


Figure 30: Sensitivity analysis using the state-of-the-art approach.

to the variance of the system-level metrics. These contributions are called first-order sensitivity indices and were calculated with 10,000,000 samples using the variance decomposition global sensitivity analysis technique described in Ref. [77]. The results are shown in Fig. 30. Positive changes in probability indicate that removing a particular technology had a performance benefit, whereas a negative probability change means that removing the technology degraded performance.

This sensitivity study provides information that can support a selection of the alternatives. It is assumed that the readiness risk is the same for all three technologies, so that dimension does not need to be considered for this example. Due to the fact that the ACTE technology is the largest contributor to variance in two out of three system-level metrics and positively affects three out of five POS measures, a decision maker would likely prioritize any activity that targets uncertainty reduction and/or performance improvement of that technology. However, the sensitivity results are indifferent between A_3 and A_4. Next in priority would come the fan acoustic splitter technology since it drives sideline noise POS changes and variability. The AFC

technology is clearly not contributing to performance improvement or uncertainty reduction as much as the other two technologies. These observations may lead one to believe that alternatives that target ACTE should be preferred. The expected utility assessment resulted in alternative 1, which targets a noise technology, having the highest quantified value. This result may be surprising to a decision maker, even if he or she has an approximate estimate of cost.

Prioritization of the alternatives using the results in Fig. 30 may not be overwhelming for many decision makers with only three technologies and four alternatives to consider, but the difficulty would increase greatly if technologies, system-level metrics, POS goals, and/or alternatives were added to the problem. Also, an additional dimension of cost would need to be considered. An important advantage of the utility-based approach is that it quantitatively incorporates the decision makers' preferences and risk attitudes over the consequence space to aid in the decision process, rather than decision makers qualitatively synthesizing this information in their minds.

4.4.4 Implementation of the Proposed Methodology: Sensitivity Analysis

To demonstrate sensitivity analysis with the utility-based approach, three scenarios were devised. In the first scenario, the decision maker was unsure of the indifference probability $\pi_{\text{Performance}}$ elicited for the system-level performance utility function. The second scenario was similar to the first with the only difference being that the cost indifference probability π_{Cost} was of interest. In the third scenario, the decision maker was unsure of the multiattribute utility function scaling constant $\lambda_{\text{Uncertainty}}$ for the uncertainty reduction attribute. In addition to the analysis of these scenarios, visualization of the attribute and utility space are presented as additional tools for conducting sensitivity analyses.

4.4.4.1 Scenario 1: Uncertain Degree of Performance Risk Aversion

For the first scenario, the notional decision maker was confident of being risk averse with regard to system-level performance improvement, but the specific degree of risk aversion was not established conclusively. Thus, it was decided to do an experiment to investigate the importance of the precise specification of the indifference probability. The indifference probability $\pi_{\text{Performance}}$ was varied between values of 0.6 and 0.95, and expected utility was calculated for 20 levels of $\pi_{\text{Performance}}$. The single-attribute utility functions corresponding with these bounds are shown in Fig. 31a. The expected utilities of all four attributes over the range of values for $\pi_{\text{Performance}}$ are shown in Fig. 31b. The expected utility magnitudes were affected, but the ranking of the alternatives did not change. This result means that for this scenario, the decision maker's specification of the indifference probability was not critical.

One may study the $\pi_{\text{Performance}}$ sensitivity results and question why the expected utilities for all alternatives increased with increasing degree of risk aversion. The concept of risk premium helps to explain the causality. For an increasing utility function, risk premium is defined as the expected value of a lottery minus the certainty equivalent of that lottery. For example, the lottery used to elicit the baseline indifference probability for the performance attribute had a consequence of 1 with probability $\pi_{\text{Performance}}$ and a consequence of 0 with probability $1 - \pi_{\text{Performance}}$, and the certainty equivalent was 0.5. The baseline indifference probability was set at 0.8, so the expected consequence of the lottery was $1 \cdot 0.8 + 0 \cdot 0.2 = 0.8$. The risk premium for this example is $0.8 - 0.5 = 0.3$. As shown in Fig. 31c, the risk premium increases as the indifference probability increases. According to Keeney and Raiffa, the risk premium can be interpreted as "the amount of the attribute that the decision maker is willing to 'give up' from the average (i.e., the amount less than the expected consequence) to avoid the risks associated with the particular lottery" [52]. As $\pi_{\text{Performance}}$ increases, the utility of lower performance attribute values increases, as the decision maker is

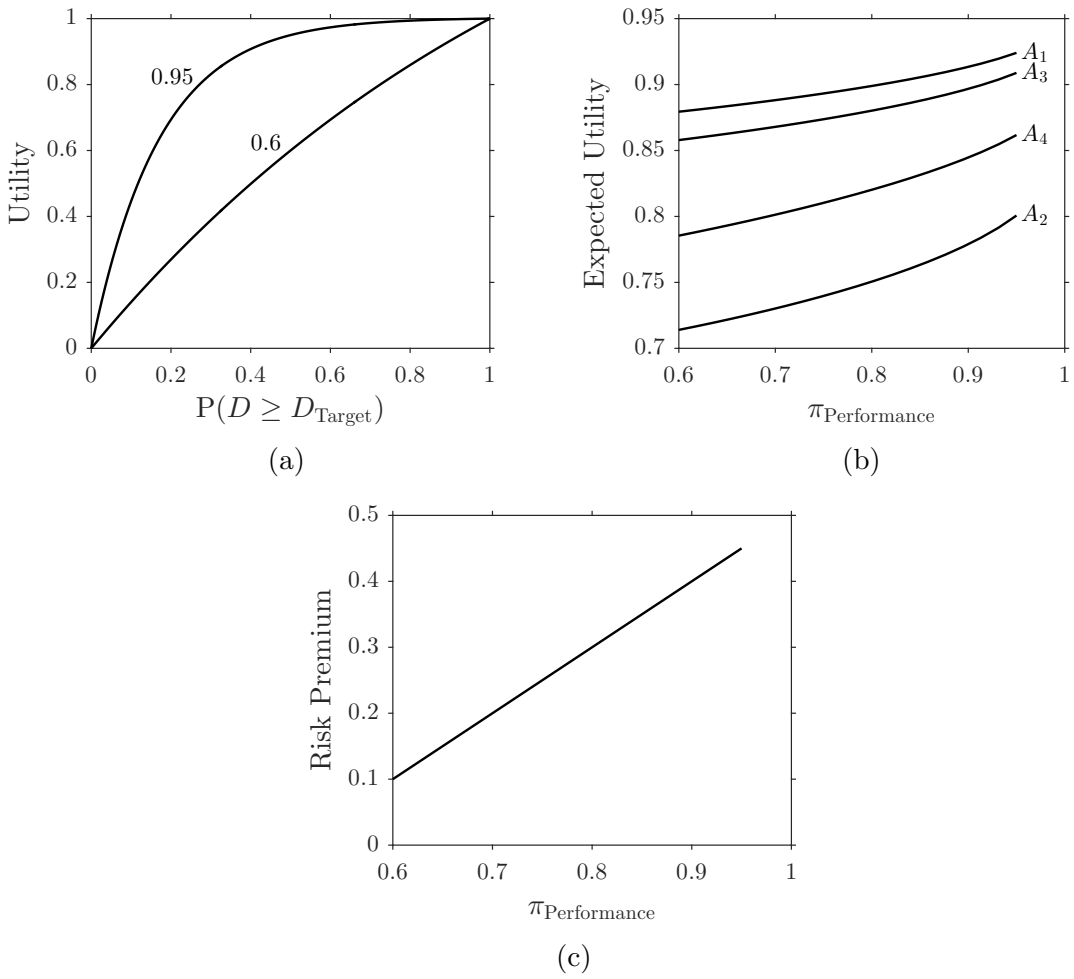


Figure 31: Scenario 1 utility function range (a), expected utilities (b), and risk premium (c).

more willing to budge on performance to avoid the risk of a lottery. This is why the expected utility values of all alternatives increases with $\pi_{\mathrm{Performance}}$.

4.4.4.2 Scenario 2: Uncertain Cost Risk Attitude

For the second scenario, the notional decision maker was unsure about risk attitude with regard to cost. An experiment was conducted to investigate the importance of the indifference probability for cost. The indifference probability π_{Cost} was varied between values of 0.1 and 0.9, and expected utility was calculated for 20 levels of π_{Cost}. The single-attribute utility functions corresponding with these bounds are shown in

Fig. 32a. The expected utilities of all four attributes over the range of values for π_{Cost} are shown in Fig. 32b. For this scenario, the ranking of the alternatives was affected. A_1 and A_3 switched rankings for highly risk prone cost utility functions, and A_4 surpassed the expected utility of A_3 and approached A_1 for highly risk averse cost utility functions.

The causality behind the trends of the rankings can be explained by considering the shape of the cost utility function and the risk premium. The estimated range of cost for A_1 is higher than A_3, so as π_{Cost} decreased, the cost utility of the more expensive A_1 decreased. This result may not agree with the reader's intuition, as one's concept of risk might lead one to conclude that highly risk-prone cost preference would lead to the more expensive alternative having higher expected utility. The risk premium behavior in Fig. 32c is contradictory, as the decision maker was willing to "give up" negative cost to avoid the lottery. In other words, the risk-prone decision maker prefers the risk of the lottery to the expected consequence of the lottery. The behavior of the expected utility curves for A_2 and A_4 can be explained with similar logic. As π_{Cost} increased, risk premium increased and the utility curve changed shape such that larger cost percentages had relatively higher utility. One might intuitively expect that a highly risk-averse cost preference would lead to higher value of the least expensive alternatives. On the contrary, as utility over the cost range of 20% to 60% increased rapidly with π_{Cost}, the expected utility of A_2 and A_4 rose quickly as well. For A_4, this change in expected utility was so aggressive that it surpassed the much less expensive A_3. If the decision maker had a risk-prone attitude, then a determination of which side of $\pi_{\text{Cost}} = 0.18$ is preferred would have needed to be made, as this was the point at which A_1 and A_3 had the same expected utility. If the decision maker had a risk-averse attitude, then a determination of which side of $\pi_{\text{Cost}} = 0.83$ is preferred would have needed to be made, as this was the point at which A_3 and A_4 had the same expected utility. A similar determination would have

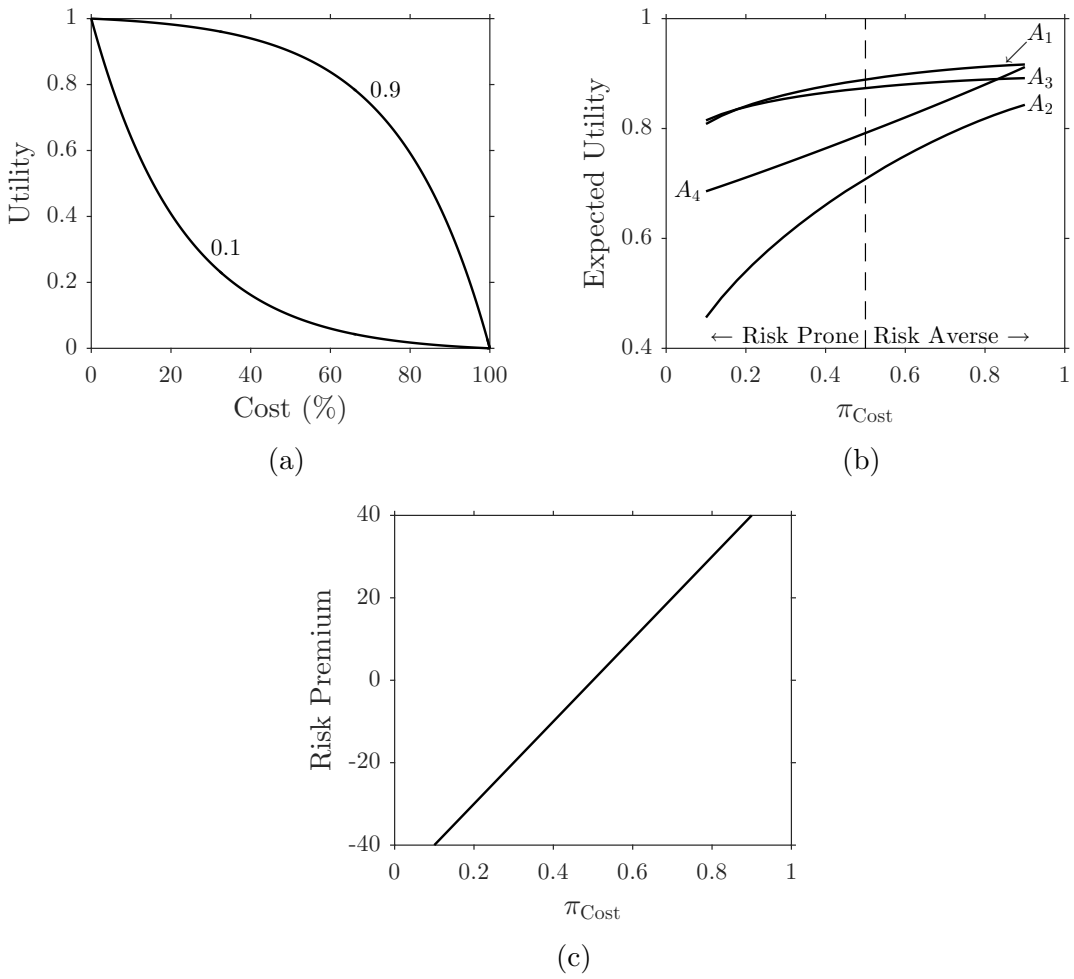


Figure 32: Scenario 2 utility function range (a), expected utilities (b), and risk premium (c).

been made if indifference probabilities greater that 0.9 were considered as well.

4.4.4.3 Scenario 3: Uncertain Strength of Preference for Uncertainty Reduction

For the final scenario, the notional decision maker was unsure about the strength of preference for increasing the attribute for uncertainty reduction. An experiment was conducted to investigate the importance of the mutliattribute utility scaling constant for average variance reduction. The scaling constant $\lambda_{\text{Uncertainty}}$ was varied between values of 0.1 and 1, and expected utility was calculated for 20 levels of $\lambda_{\text{Uncertainty}}$. The expected utilities of all four attributes over the range of values for $\lambda_{\text{Uncertainty}}$ are shown

in Fig. 33. The ranking of the alternatives was also affected in this scenario. In a similar behavior as seen for cost, A_1 and A_3 switched rankings for high values, and A_4 surpassed the expected utility of A_3 and approached A_1 for low values. Additionally, A_2 and A_4 switched rankings.

The causality behind these trends was different than for the other two scenarios. $\lambda_{\text{Uncertainty}}$ can be interpreted as a probability, as it was elicited in this example through probabilistic scaling. The higher the probability value, the more risk averse the decision maker was with regard to selecting between the certainty equivalent and the lottery shown in Table 8. A higher value of $\lambda_{\text{Uncertainty}}$ resulted in a larger contribution of uncertainty reduction utility to the multiattribute utility. This effect combined with the fact that A_4 reduced uncertainty in wing weight, which is a large contributor to overall system-level uncertainty, is why the expected utility increased rapidly. For $\lambda_{\text{Uncertainty}} \leq 0.24$, A_1 had lower expected utility that A_3 because the uncertainty reduction attribute contributed less to overall utility. A_1 was sensitive to this effect, whereas A_3 was relatively more robust. The implications for the decision maker were similar to what was found in scenario 2; a determination would need to be made as to which $\lambda_{\text{Uncertainty}}$ interval between crossover points in expected utility was preferred, not what the specific value was.

4.4.4.4 Visualization of Utility and Attributes

Another kind of sensitivity analysis that can be performed with the proposed methodology is to explore the effects of each development activity impact on utility. One approach to accomplish this is by visualizing the multiattribute utility function, as shown in Fig. 34. This visualization technique plots slices of utility as a function of the variables that were used to model the impacts of the technology development activity alternatives. Each plot shows how utility varies as a function of the independent variable, with all other variables fixed. When this type of plot is used on a

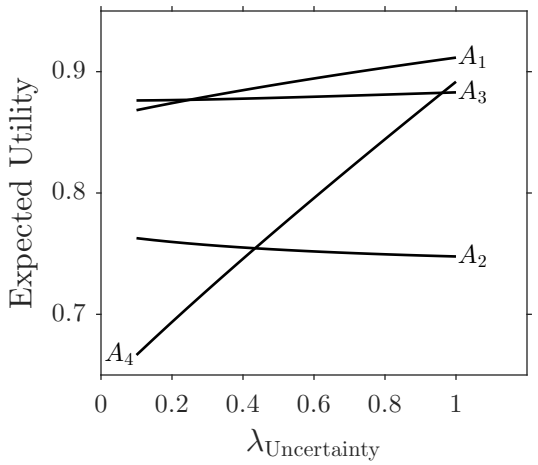


Figure 33: Scenario 3 expected utilities.

computer, the slider bars enable dynamic exploration of the utility function. When a slider bar is changed for one variable, the plots for all other variables are updated. The plots indicate that utility is sensitive to translations and variance scaling of the wing weight and fan noise k-factors, whereas the impacts for APU weight and vertical tail area are small contributors to changes in utility. It is also evident that utility is largely affected by cost.

Decision makers may also wish to see a similar presentation of the attributes. Similar plots are shown for system-level performance and average variance reduction in Figs. 35 and 36, respectively. The performance plots include the log base 10 of the desirability function exponents for block fuel reduction and sideline noise reduction on the far right. Two interesting observations from the performance plots are: (1) there are diminishing returns for noise impact translation, whereas $\delta_{\text{Wing Weight}}$ improves performance over its entire range, and (2) performance is relatively insensitive to variance scaling of any impacts. Note that these observations are only valid at the region of the space shown in the figure. Fascinating observations can also be made from the uncertainty reduction plot. For instance, it may be surprising to some that translations of the fan noise and wing weight variables contribute to average variance reduction.

A potential application of the multidimensional sensitivity plots is the generation of alternatives. For example, decision makers and analysts can perform quick analyses with the utility function plots to roughly determine what kind of activities might be appropriate to maximize utility.

4.5 Discussion and Conclusions

This chapter explored the problem of how to inform decisions regarding the selection of technology development activity classes before details of the activities have been defined. Through an analysis of the literature, the current state of the art was identified, and it was argued that there was still a need to close research gaps in order to enable a formal, systematic approach to supporting activity selection decisions. A decision process provided the foundation for a novel methodology, and techniques from MAUT were incorporated to address the research gaps.

The proposed methodology for prioritizing development activity alternatives was applied to a notional problem in the illustrative example. The key steps of the methodology were illustrated, and the results were compared with sensitivity analyses from the state-of-the-art approach. The notional decision maker's preferences, risk attitudes, and system-level performance goals were synthesized to produce a valid measure of value for the alternatives. Uncertainty surrounding the impacts of the technology development activity alternatives was explicitly modeled with probabilities. The proposed approach was shown to be capable of quantitatively evaluating the set of alternatives using expected utility, rather than only providing measures of potential for each technology. Sensitivity analyses were conducted to demonstrate the flexibility of the novel methodology for studying the impacts of the decision maker's preferences and risk attitude on the expected utilities.

In the example problem, the alternatives were all defined as an individual activity class coupled with a technology. The problem was set up this way to facilitate

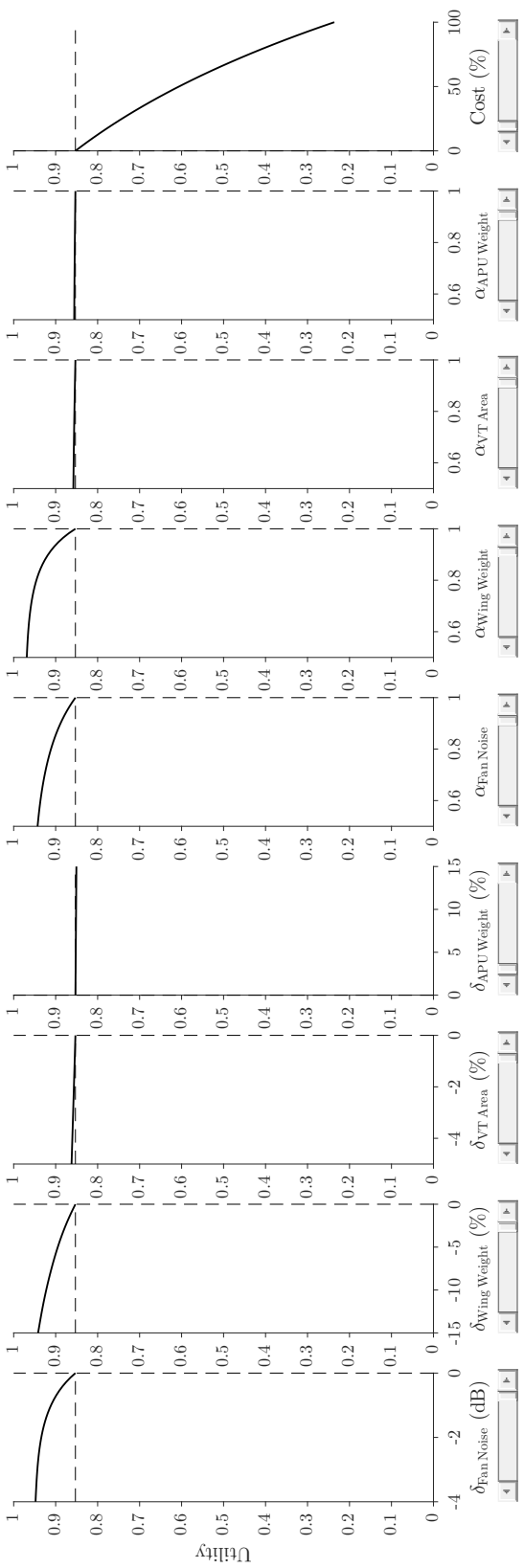


Figure 34: Multiattribute utility function slices.

Figure 35: System-level performance attribute slices.

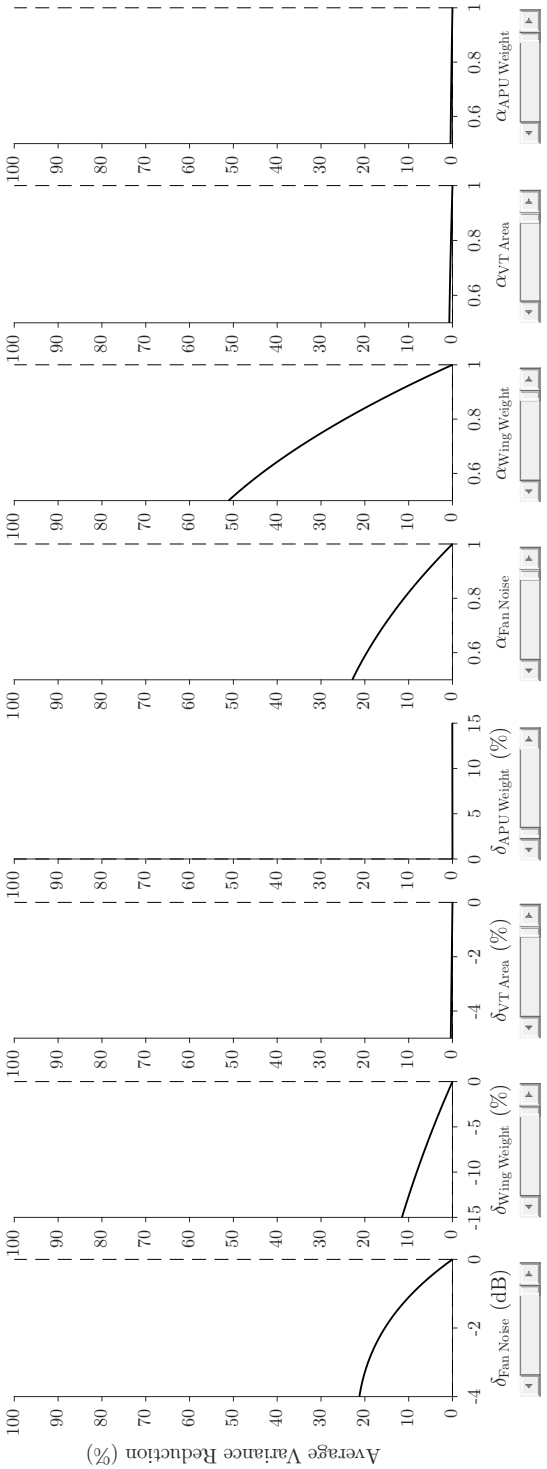


Figure 36: Uncertainty reduction attribute slices.

demonstration and understanding of the approach. In a more realistic problem, some of the alternatives would most likely not be compared as they are defined in the example problem. For example, A_3 and A_4 may both be necessary to conduct at some point during technology development, but the example expected utility results might lead one to conclude that A_3 "beats" A_4. This is not the intended use of the capability in a practical setting. A more appropriate use of the methodology in practice would be to compare alternatives comprised of sets of activities. For instance, decision makers may wish to compare multiple portfolios of development activities that are aimed at maturing a single technology from TRL 4 to TRL 6. For this type of assessment, sets of activities can be modeled by eliciting distributions on $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and cost for each activity and adding the distributions together to obtain the net impacts of the portfolios.

Compared with the state of the art method for development activity selection, the proposed methodology requires significant modeling effort on the part of the analyst and the decision makers. Besides the reasons to use a quantitative decision aid that were presented in the beginning of Sec. 4.2, another reason the effort is worthwhile is that a formal decision analysis is intended to provide insight and stimulate deeper thinking to help decision makers make better decisions. The additional modeling effort also results in more degrees of freedom for mathematically representing the alternatives. Similar to the adage "with great power comes great responsibility", with the flexibility provided by the novel methodology comes the responsibility to take great care when modeling the decision problem. Although a benefit of evaluating alternatives with expected utility is the aggregation of information that is pertinent to the problem, this characteristic can also make it difficult to identify errors or poor assumptions in the analysis.

The author also acknowledges that there may be many other considerations in the decision problem. For example, political and social considerations may significantly

affect a decision maker's preferences. Decision makers must assess the implications of a formal decision analysis with other considerations that are not included in the analysis. This observation highlights the fact that the results of a decision analysis should not be interpreted as a dogma but rather as an additional input to the decision problem.

CHAPTER V

UNCERTAINTY QUANTIFICATION WITH MULTITASK GAUSSIAN PROCESSES FOR TECHNOLOGY DEVELOPMENT EXPERIMENTS

In this chapter, two problems are addressed: (1) how to quantify technology integration impact uncertainty in light of data from multiple, heterogeneous experiments and (2) how to quantitatively estimate the uncertainty reduction that a planned experiment will achieve. These problems are characterized in Sec. 5.1. Then, a novel methodology for solving these problems is formulated in Sec. 5.2. The primary arguments are as follows.

Argument 2: The proposed methodology provides an appropriate way to quantify the uncertainty surrounding technology integration impacts in light of data from multiple, heterogeneous technology development experiments because

1. It is anchored in proven machine learning methods for making predictions under uncertainty
2. It provides a flexible, quantitative approach to model the epistemic uncertainty associated with extrapolating technology impacts to the future

Argument 3: The proposed methodology provides an appropriate way to quantitatively estimate uncertainty reduction for a planned experiment because

1. It implements a rigorous information theoretic framework that is the state of the art in experiment design
2. It aggregates prediction uncertainty from a probabilistic regression model and the additional layer of epistemic uncertainty associated with technology maturity in the estimation process

Due to a gap in knowledge regarding identified enablers, called multitask Gaussian process models, an experiment was conducted, and this experiment is described in Sec. 5.3. Next, an illustrative example is presented in Sec. 5.4 to demonstrate the key contributions of the methodology. Finally, the chapter closes with a summary in Sec. 5.5.

5.1 Problem Definition

In this section, the problems addressed in this chapter are presented. First, the need for methods to model technology impact uncertainty and the reduction of uncertainty with technology development activities are discussed. Then, characteristics of the problems are described, and two research questions are presented to concisely define

them.

5.1.1 Quantifying Technology Impact Uncertainty

The technology development management processes from the literature that were discussed in Chapter 2 assumed that the combined epistemic and aleatory uncertainty surrounding technology impacts can be mathematically represented using probability theory. These uncertainty models are important for tracking technical progress, as shown for a real development program in Fig. 5; quantifying system-level uncertainty; and informing decisions regarding the design of future activities. After a set of activities has been conducted, the uncertainty models must be updated to reflect the changes in epistemic uncertainty that are associated with the acquisition of new knowledge. As previously discussed, uncertainty reduction is one of the attributes that constitutes the overall value of technology development activities. If decision makers were capable of estimating the epistemic uncertainty reduction due to a planned activity, then this information could be used to guide the design of the activity.

Each class of technology development activity effects a change in particular sources of uncertainty through different mechanisms. A diverse set of examples based on the taxonomy of development activities in Ref. [21] helps to illustrate this idea. One class of activity is called a feasibility study, and its purpose is to demonstrate whether the technology functions as intended or at the minimum level of performance that has been established. In other words, the result of a feasibility study is binary: the technology works for its intended purpose or it does not. The creation of an analysis capability is another type of technology development activity. The analysis capability is often realized in the form of a physics-based M&S environment that is used to predict technology performance. If the M&S environment is shown to be more accurate than what already exists, then this activity reduces model form uncertainty, which

is uncertainty due to assumptions and the selection of the mathematical model [12]. The practical consequence is technology performance prediction with more credibility. Design studies are another type of computational activity. Their purpose is to numerically investigate the design space of the technology, with the goal of finding regions of the space with the best performance. Uncertainty is impacted by the reduction of ranges on the design variables after decisions are made regarding the settings of these variables. Physical experiments are the final example. The traditional purpose of experiments is to improve the understanding of physical phenomena. This usually entails measuring dependent variables (responses) at various settings of independent variables (factors) with the goal of characterizing the relationship between them. The uncertainty surrounding technology performance in the proximity of the measurements is directly reduced as a result.

Different approaches are required for quantifying the change of epistemic uncertainty due to each type of technology development activity. The focus of this chapter was limited to activities involving physical and computer experimentation. Experimental activities were selected as the focus due to their importance in technology development. These are the types of activities that are an integral part of all TRL definitions; technologies cannot be considered as maturing without them.

5.1.2 Characteristics of the Problem

There are multiple aspects of the problem that make uncertainty quantification with technology experimental data a difficult task. One issue is accounting for the maturity dimension. Some experiments are more realistic and credible than others. For example, the uncertainty associated with extrapolating sub-scale wind tunnel data to the future performance of an aircraft is likely going to be larger than extrapolation uncertainty from data collected during a full-scale flight experiment. Another aspect that makes the uncertainty quantification task difficult is that data may be sparse

and only partially relevant. For an instance of this, consider the vertical tail AFC data shown in Fig. 37. It was not possible to take measurements in flight for the AFC technology in the shaded regions, and because of this, flight test data was not available in the "critical β range" which was of primary interest. The solid black lines in the figure show the result of flight simulation predictions after correcting for the flight test setup and the measurements that were available. A related challenge is how to quantitatively represent learning from previous experiments. For the AFC technology example, knowledge about AFC effectiveness was gained through sub-scale and full-scale wind tunnel experimentation before the flight experiment. Although data from multiple experiments can be similar, there are usually differences between the experimental setups such that the results are not identical. For example, dynamic similarity may not be satisfied across multiple fluids experiments due to different geometries, constraints of the facilities, etc. In the vocabulary of statistics, the data from multiple heterogeneous experiments may be *nonexchangeable*. These observations led to the following research question:

Research Question 2.0: What is an appropriate way to quantify the uncertainty surrounding technology integration impacts in light of data from multiple, heterogeneous technology development experiments?

Once an uncertainty model has been constructed, it would be valuable to have the capability to then estimate how much uncertainty reduction a given experimental plan will achieve. This capability could be used to quantify one of the attributes needed to evaluate alternatives. The corresponding research question investigated in this chapter is as follows:

Research Question 3.0: What is an appropriate way to quantitatively estimate expected uncertainty reduction for a planned technology experiment?

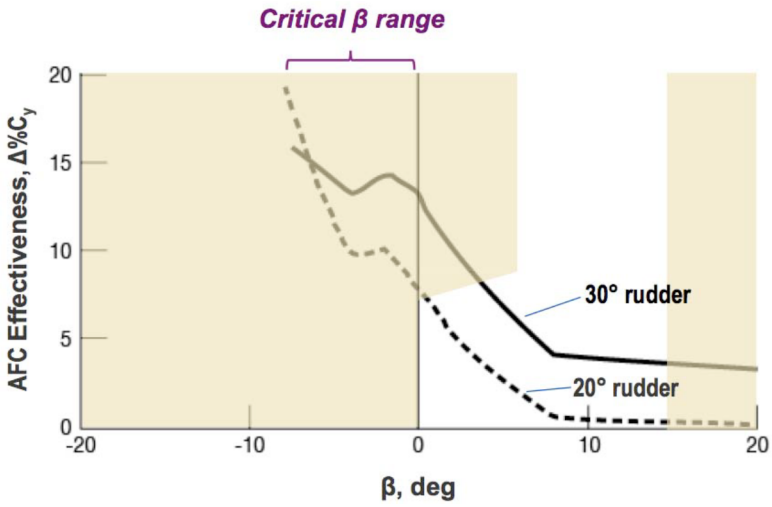


Figure 37: Vertical tail AFC effectiveness predictions for a range of sideslip angles (from Ref. [40]). The shaded area indicates regions where no flight test data was available, including the "critical *beta* range", which was of primary interest.

5.1.3 Literature Review

The most relevant existing work that addresses RQ 2.0 was produced by Largent [21]. He assumed that probability distributions on technology impacts could be built from a combination of sources, and he proposed the use of Bayesian inference to update the technology impact uncertainty once development activity data are collected. However, he acknowledged that sequential updating using data from multiple, heterogeneous development activites could result in misleading inferences. There are also large bodies of literature in statistics and machine learning that are relevant to learning from multiple, heterogeneous sources of information. A statistical data fusion approach used in many disciplines is called meta-analysis, which has been defined by Christine Anderson-Cook as "information synthesis using multiple data sources to answer a global question(s) by leveraging knowledge and statistical power through understanding data connections" [78]. As an example, a typical application of meta-analysis is to combine data from multiple, similar medical studies that aim to test the same hypothesis. In the machine learning literature, the broad research area of transfer learning involves transferring knowledge from source tasks to target tasks in

order to improve the learning process for the target tasks [79]. Transfer learning is motivated by the fact that humans can apply knowledge from one domain to another to find better solutions to new problems and do it more efficiently. For instance, learning to play a guitar may help one with learning to play a banjo.

There are also large bodies of literature that pertain to answering RQ 3.0. In the test resource allocation literature summarized in Chapter 2, the authors of Refs. [32, 34, 35, 36] used simulated test data combined with Bayesian inference to estimate uncertainty reduction. In the statistics literature, Lindley [80] proposed the idea of designing experiments to maximize expected information gain, which is an information theoretic representation of uncertainty reduction. Several other authors have used a similar framework to evaluate experiments based on the information contained in them (e.g., see Refs. [81, 82, 83]).

5.2 Methodology Formulation

In general, the measured dependent variables of an experiment depend on the settings of independent variables that are under the control of the experimenter and environmental effects that are not. In addition to characterizing the uncertainty surrounding experimental data, it is ideal to have the capability to characterize the relationships between dependent variables and independent variables. With models of these relationships, predictions can be made to estimate technology performance at locations in the independent variable space where data do not exist; this is called generalization. There are many benefits of having this ability, including facilitating understanding of physical phenomena and informing decisions regarding the design of future experiments.

The disciplines of machine learning and statistics provide the tools to characterize uncertainty surrounding both the data obtained from experiments and predictions for future data. The types of tools that are applicable to data from most technology

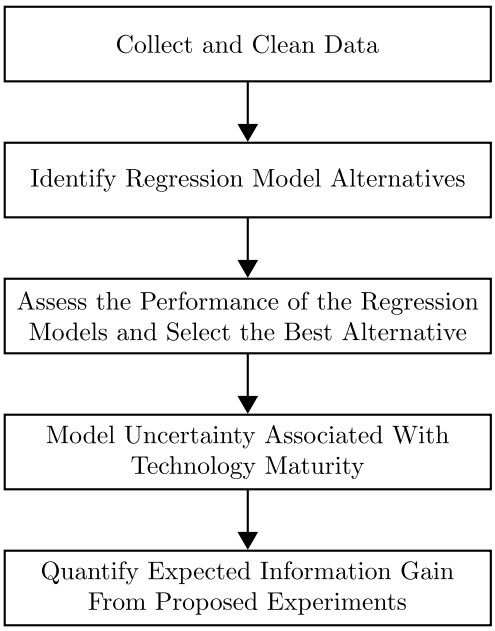


Figure 38: The proposed methodology for quantifying technology impact uncertainty and estimating uncertainty reduction for future experiments.

development experiments are categorized as supervised learning methods. The goal of these methods is to learn a mapping from inputs (independent variables) to outputs (measured dependent variables) from data. To address RQ 2.0, the proposed novel methodology shown in Fig. 38 begins with three steps to construct a supervised learning model. These steps were derived by synthesizing best practices from the literature. It is assumed that the outputs are real-valued, and this assumption restricts the focus to a learning problem called regression.

Once a regression model has been built by training it with experimental data for a technology, uncertainty surrounding its predictions can be quantified. However, there is an additional layer of epistemic uncertainty that these models do not capture: uncertainty associated with the maturity of the technology. A method for accounting for this uncertainty in forecasts produced by regression models is proposed in step four of the methodology. The final product is a predictive model that can be used to forecast technology impacts with quantified aleatory and epistemic uncertainty.

The purpose of the last step of the methodology is to use the predictive model to

estimate uncertainty reduction for proposed experiments that have not yet been performed. To implement this, ideas from the test resource allocation literature and the information theoretic framework are borrowed. Uncertainty reduction is quantified by estimating the expected information gain from each of the proposed experiments.

To help the reader locate this proposed methodology in the overall solution to the motivating question from the introduction chapter, it is useful to consider where the methodology fits in Fig. 7. This methodology provides a predictive model for forecasting technology performance at a point in the future when the technology has been fully matured. This capability can be used to establish k-factor distributions to enable the evaluation of development activity alternatives in the decision process from phase one, which was discussed in Chapter 4. Having the capability to estimate uncertainty reduction from proposed experimental designs is crucial for enabling the evaluation of alternatives in the decision process from phase three.

The steps of the methodology are described in the following sections. Although the methodology can be adapted—with some effort—to virtually any type of regression model, the descriptions of steps two, four, and five are specific to Gaussian process (GP) models. GPs were selected for three reasons: (1) they are nonparametric models that are flexible enough to fit highly nonlinear data, (2) they naturally provide a probabilistic representation of aleatory and epistemic uncertainty, and (3) there is substantial prior research in modifying GPs for transfer learning. According to Wolpert's "no free lunch" theorem [84], there cannot be a universally best learning model. Hence, by this theorem GPs cannot be the best type of regression model for all problems. However, the intent of selecting GPs is to provide a methodology that is flexible enough to be directly applied to experiments for a wide variety of technologies.

5.2.1 Step One: Collect and Clean Data

Once raw data are available from one or more experiments, the raw data must be collected and processed before they can be used for training regression models. Processing the raw data may involve the use of data reduction equations and other transformations to arrive at desired measures. The details of data processing are dependent on the technology and the best practices of the disciplines involved. The end goal of this step is a set of tables conforming to what Wickham defined as a tidy data set, where

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table [85].

Note that the term "observation" is used in this chapter, but observations are usually referred to as "examples" in the machine learning literature and as "data points" in statistics.

5.2.2 Step Two: Identify Regression Model Alternatives

Although the scope has been limited to GP models, there are still many models that can be selected as feasible alternatives. A brief overview of GP regression with a single training data set (single-task setting) is presented first. Then, GPs that are designed for learning with multiple data sources simultaneously (multitask setting) are discussed. The notation and terminology loosely follows Ref. [86].

5.2.2.1 Single-Task Gaussian Process Regression

The tidy data from step one serves as the training data for the GP models. This training set with n observations is denoted by $\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|\, i = 1, 2, \ldots, n\}$, where $\mathbf{x}$ is a D-dimensional input vector (independent variables) and y is a scalar output

(dependent variable). Note that the output is often referred to as the "target", and this terminology is used here. The data for the n observations of the column vector inputs are collected in the design matrix $\mathbf{X}$ of dimension $D \times n$, and the n observations of the target are assembled in the vector $\mathbf{y}$. With this notation, the training data can be written $\mathcal{D} = (\mathbf{X}, \mathbf{y})$. The goal is to make inferences about the underlying target function f that maps the inputs to the targets.

According to Rasmussen and Williams, a GP is defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" [86]. For GP regression, the idea is to specify a GP prior distribution over the target function $f(\mathbf{x})$ and infer a posterior distribution $p(f|\mathcal{D})$ after observing the training data. This posterior distribution over functions can then be used to make predictions with uncertainty. The prior distribution over the target function is

$$f(\mathbf{x}) \sim \mathrm{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')) \tag{17}$$

where, $\mathbf{x}$ and $\mathbf{x}'$ are any two locations in the input space, and the mean function $m(\mathbf{x})$ and covariance function (or kernel) $\kappa(\mathbf{x}, \mathbf{x}')$ are defined respectively as

$$m(\mathbf{x}) = \mathrm{E}[f(\mathbf{x})] \tag{18}$$

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathrm{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \tag{19}$$

Note that the covariance function must be positive definite. Given any finite set of n input points, the GP specifies a joint Gaussian distribution:

$$\mathbf{f}|\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K}) \tag{20}$$

where, $\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)))^\top$, $\boldsymbol{\mu} = (m(\mathbf{x}_1), m(\mathbf{x}_2), \ldots, m(\mathbf{x}_n)))^\top$, and $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$.

A common practice is to use a mean function of $m(\mathbf{x}) = 0$ in Eq. (17) because the GP regression model is flexible enough to model the mean quite well. Parametric models for the mean function can also be used. When a set of fixed basis functions is

used, the regression model behaves as a global linear model with the residuals being modeled by a GP. Murphy referred to this approach as semi-parametric modeling because it "combines the interpretability of parametric models with the accuracy of non-parametric models" [87]. Note that throughout the remainder of the formulation a mean function of zero is used.

It is assumed that the observations of the target are noisy realizations of the underlying function: $y = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \sigma_y^2)$. With the addition of the noise term, the covariance of the targets is

$$\mathrm{cov}(y_p, y_q) = \kappa(\mathbf{x}_p, \mathbf{x}_q) + \sigma_y^2 \delta_{pq} \tag{21}$$

where, δ_{pq} is the Kronecker delta that is equal to one if $p = q$ and zero otherwise. Written in matrix form:

$$\mathrm{cov}(\mathbf{y}|\mathbf{X}) = \mathbf{K} + \sigma_y^2 \mathbf{I} \tag{22}$$

Note that the second matrix in Eq. (22) is diagonal because of an assumption of independent noise terms.

Given a set of prediction locations $\mathbf{X}_*$ of size $D \times n_*$, generating predictions from the GP regression model begins with constructing the joint distribution of the training targets and the latent target function $\mathbf{f}_*$ at the prediction locations:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_y^2 \mathbf{I} & \mathbf{K}_* \\ \mathbf{K}_*^\top & \mathbf{K}_{**} \end{bmatrix} \right) \tag{23}$$

where, $\mathbf{K} + \sigma_y^2 \mathbf{I}$ is an $n \times n$ matrix, $\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*)$ is an $n \times n_*$ matrix, and $\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*)$ is an $n_* \times n_*$ matrix. By conditioning using standard rules for multivariate Gaussian distributions, the posterior distribution is obtained:

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathrm{N}(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)), \quad \text{where} \tag{24}$$

$$\bar{\mathbf{f}}_* = \mathbf{K}_*^\top \left[\mathbf{K} + \sigma_y^2 \mathbf{I} \right]^{-1} \mathbf{y}, \tag{25}$$

$$\mathrm{cov}(\mathbf{f}_*) = \mathbf{K}_{**} - \mathbf{K}_*^\top \left[\mathbf{K} + \sigma_y^2 \mathbf{I} \right]^{-1} \mathbf{K}_* \tag{26}$$

For noiseless observations, the noise variance term $\sigma_y^2\mathbf{I}$ is simply removed from these equations. This the approach that is often used when regressing data from computer experiments [88, 89]. However, Gramacy and Lee [90] argued that the noise variance term should be included for better statistical properties of the regression model.

Before a GP regression model can be used as a prediction tool, decisions must be made about the form of the covariance function and its hyperparameters; this process is referred to as "training" a GP. Many covariance functions have been proposed, such as the squared exponential, polynomial, and Matérn forms. The squared exponential covariance function is a common choice and can be parameterized as follows:

$$\kappa(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^\top \mathbf{M}(\mathbf{x}_p - \mathbf{x}_q)\right) + \sigma_y^2 \delta_{pq} \tag{27}$$

where, $\boldsymbol{\theta} = (\{M\}, \sigma_f^2, \sigma_y^2)^\top$ is a vector of hyperparameters, and $\{M\}$ is the set of hyperparameters contained in the symmetric matrix $\mathbf{M}$ (not to be confused with the symbol for the vector of system-level metrics in Chapter 4). When $\mathbf{M} = \text{diag}(\boldsymbol{\ell})^{-2}$, the hyperparameters $\ell_1, \ell_2, \ldots, \ell_D$ are characteristic length-scales. These parameters govern how far apart two points in input space must be for the regression function values at those points to become uncorrelated. This type of covariance function structure implements automatic relevance determination (ARD) because the inverse of the length-scale indicates how relevant each input is; if the length-scale is relatively large for an input, then the covariance is virtually independent of that dimension. The hyperparameters $\boldsymbol{\ell}$ control the smoothness of the regression function, whereas σ_f^2 controls the magnitude of the regression function.

One the covariance function form has been selected, the values of all hyperparameters must be determined. This is typically done by selecting $\boldsymbol{\theta}$ to maximize the natural logarithm of the marginal likelihood: $\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$. For a description of this technique and others, the reader is referred to Ref. [86].

5.2.2.2 *Multitask Gaussian Process Regression*

The machine learning concept of transfer learning is intriguing for the problem of interest, where there are multiple data sources to learn from. If regression models that implement transfer learning can be used to improve predictive accuracy and quantitatively represent the effects of knowledge transfer on uncertainty, then they are preferred. According to the transfer learning taxonomy of Pan and Yang [79], the regression problem of interest here is classified as a type of inductive transfer learning called multitask learning. Here, the word "task" refers to two components: (1) the underlying target function and (2) the label space, which is the set of possible values for the targets. The idea behind multitask learning is to learn the target functions of multiple related data sets simultaneously, while sharing information across the different tasks, in order to improve the generalization performance of the models [91].

Multitask variants of multiple classes of regression and classification models have been proposed, including GP models. One of the earliest methods proposed was to learn a set of common hyperparameters using data from all tasks [92]. This method included an informative vector machine (IVM) [93] to select the most informative training observations from all tasks to lower computational expense. Borrowing the concept of hierarchical Bayesian modeling commonly used for meta-analysis, hierarchical GP models have also been proposed for multitask learning (e.g., see Refs. [94, 95]). Motivated by the need to borrow strength from multiple computer codes, Kennedy and O'Hagan [96] proposed an autoregressive multitask model. Two other examples are learning a covariance matrix to model inter-task dependencies [97] and constructing covariance functions with convolution processes [98, 99].

An open issue in transfer learning is knowing when transfer should occur. In some situations, transfer can degrade performance, which is referred to as negative transfer. Caruana [91] showed empirical evidence of negative transfer with multitask artificial neural networks, and he concluded that the benefits of a multitask

approach are problem-dependent. He also identified characterization of what related tasks are as an open problem. Rosenstein et al. [100] showed that the performance of a hierarchical naive Bayes classifier was hindered when the tasks were dissimilar. More recently, Toal [101] empirically demonstrated negative transfer with an autoregressive GP model using analytical functions that represented computationally expensive and cheap deterministic computer codes. Based on the results of his experiments, he derived a set of guidelines for when a multitask approach should be used. There are other examples in the literature where transfer has been shown to improve and degrade learning performance through experimentation with data from practical problems. For the GP regression class of models, there is still a lack of knowledge and understanding regarding when a multitask approach will outperform a single-task GP and which multitask techniques are robust to dissimilarity between tasks. Toal's observations partially fill this gap, but he restricted his investigation to one type of multitask GP and noiseless observations. The existing research gap is summarized with the following research question:

Research Question 2.1: Under what conditions will a multitask GP regression model provide better generalization performance than a single-task GP regression model?

An experiment has been conducted to investigate RQ 2.1. The setup of the experiment and the results are presented in Sec. 5.3.

5.2.3 Step Three: Assess the Performance of the Regression Models and Select the Best Alternative

This step involves selecting the best regression model from the set of alternatives. The objective of regression analysis is to accurately infer the underlying target function, not to fit the noise, which has no predictive value. Thus, the model that exhibits the highest generalization performance (predictive accuracy) should be selected. The

accuracy of the fit to the noisy training data, called postdictive accuracy, is not as important as predictive accuracy because it is possible for a model to fit the training data perfectly but have poor predictive accuracy for new data.

5.2.3.1 Model Assessment Methods From the Literature

In an ideal data-rich scenario, the best method for evaluating the alternatives is to divide the available data into three separate parts: (1) a training set, (2) a validation set, and (3) a test set [102]. The training set is then used to train all of the models. Once they are trained, predictive accuracy is estimated using the validation set to inform model selection. Finally, the test set is used to estimate the generalization performance of the chosen model.

In most practical situations, the data are too scarce to set aside validation and test data sets. Analytical criteria have been proposed to approximate the validation step, such as the Akaike information criterion and Bayesian information criterion. Re-sampling methods called cross-validation (CV) and bootstrapping are also used to estimate predictive accuracy. These re-sampling methods are both capable of estimating the average generalization error when models are used to predict independent test observations. CV and bootstrap require more computation than analytical criteria, but re-sampling methods are universally applicable to any learning method and have been shown to provide better estimates of generalization error (e.g., see Ref. [102]).

5.2.4 Step Four: Model Uncertainty Associated With Technology Maturity

The nature of the uncertainty surrounding future technology impacts was described in Sec. 1.1.1. There are two key characteristics of this forecasting uncertainty that change as the technology matures: (1) the uncertainty reduces, which is usually represented as reductions in variability of the PDFs that characterize the uncertainty,

and (2) technologies tend to change in performance, typically demonstrating improvements over time. As shown notionally in Fig. 2, one approach to model technology forecasting uncertainty is to use a probability distribution with parameters that are a function of TRL. A similar idea is used in this step.

The prediction uncertainty from the GP posterior predictive distribution in Eq. (24) is Gaussian. Hence, the uncertainty surrounding the GP predictions is modeled with a symmetric distribution. This is appropriate when there is no justification for skewing the predictive distribution. There may be compelling arguments for why the predictive distributions on technology impacts should be skewed, even when experimental data is available that suggests otherwise. Nevertheless, the approach taken here is to model the additional layer of epistemic uncertainty by augmenting the variance of the symmetric predictive distribution.

The magnitude of epistemic uncertainty inflation due to technology maturity is inherently subjective. What is needed is the mathematical machinery to map the degree of technology maturity to uncertainty inflation. The measure of technology maturity that was selected for this purpose is TRL because it is already in widespread use. Since most TRL scales are ordinal, cardinal versions must be defined for use in mathematical operations. Conrow [103] formulated an approach to establish cardinal TRL scales by using expert opinion combined with AHP. As an example, he estimated the cardinal TRL coefficients, adjusted to a maximum TRL of 9, shown in Table 10. He also provided a cubic regression equation that can be used to map the ordinal TRLs to the cardinal coefficients:

$$\text{Cardinal TRL Coefficient} = 0.346 + 0.012(\text{TRL})^3 \tag{28}$$

This relationship between the ordinal and cardinal values is shown in Fig. 39. Note that Eq. (28) can be used to calculate cardinal TRL coefficients for noninteger TRLs. With a set of cardinal TRL coefficients, ratios and differences of the coefficients can be used to quantify the differences in maturity between ordinal TRLs. For example,

Table 10: Ordinal TRLs and the corresponding cardinal TRL coefficients, adjusted to 9.0 (data from Ref. [103])

Ordinal TRL Value	Cardinal TRL Coefficient
1	0.26
2	0.53
3	0.71
4	1.14
5	1.97
6	2.74
7	4.26
8	6.81
9	9.00

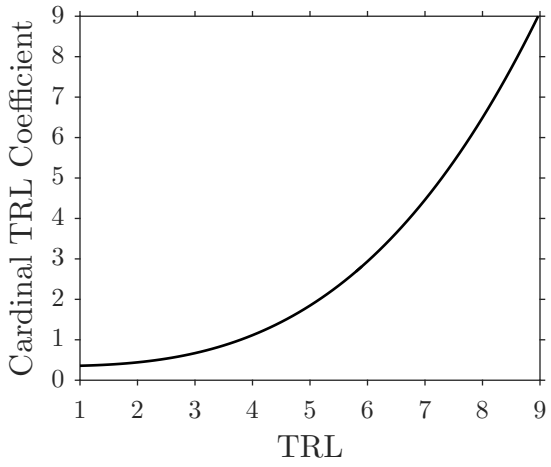


Figure 39: The relationship between ordinal TRLs and the cardinal TRL coefficients dictated by Eq. (28).

a technology at TRL 8 is likely to be more than twice as mature as a technology at TRL 4. The corresponding ratio of cardinal TRL coefficients from Table 10 is $6.81/1.14 = 5.97$, indicating a much larger gap in maturity than the ratio of the ordinal TRLs.

The mean and 95% prediction intervals are shown in Fig. 41 for a single-task GP fit to noisy data generated from the equation $y = x^2$. This example illustrates how GP prediction uncertainty changes throughout the independent variable space. When $x \in [0, 10]$, the prediction uncertainty is tight around the data, but as x moves away from the data regime, the prediction uncertainty grows rapidly. When modeling additional epistemic uncertainty due to maturity level, it is desirable to maintain this

behavior. However, the prediction uncertainty does not account for the additional uncertainty associated with forecasting technology performance.

One option for modeling the additional uncertainty is to scale the prediction uncertainty as a function of TRL. This could be implemented by multiplying the variance of the predictions, at each input location, by a function of a maturity measure. Three problems arise if scaling is used. Scaling can lead to unnecessarily large prediction uncertainty when extrapolating. For example, adding additional uncertainty to predictions when x approaches -10 in Fig. 41 may not be of any practical use because the prediction uncertainty of the GP is already so large. Another problem is that scaling operates on a distribution with combined epistemic and aleatory uncertainty. A more appropriate approach would not scale the aleatory contribution because the maturity component of uncertainty is epistemic in nature, but scaling is not capable of this since the two components of uncertainty cannot be separated in the predictive distribution. The third problem is that the scaled uncertainty will be zero if the prediction uncertainty is zero. This scenario can occur if a GP is used to fit data from a computer experiment that are modeled as noiseless observations. The prediction uncertainty collapses to zero at the location of the observations.

Another option for modeling the additional technology maturity uncertainty is to add variance to the predictive distributions. This approach does not have the same problems as the use of scaling does. To add variance, it was decided to take advantage of the fact that the sum of independent normal distributions is also normally distributed, both in the univariate and multivariate cases. The additional epistemic uncertainty is modeled by adding a multivariate normal random vector $\boldsymbol{\tau}$ to Eq. (24).

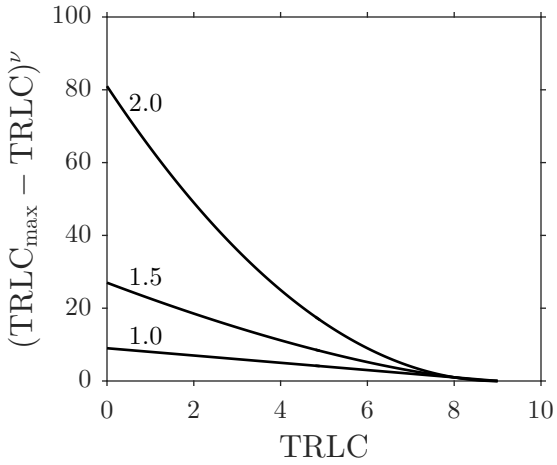


Figure 40: Behavior of the variance term in Eq. (30) for three settings of ν and $\sigma_\tau^2 = 1$.

The random vector has perfectly correlated elements to ensure that the correct behavior of the GP model is maintained. $\boldsymbol{\tau}$ is defined as

$$\boldsymbol{\tau} \sim \text{N}(\mathbf{0}, \text{cov}(\boldsymbol{\tau})), \quad \text{where} \tag{29}$$

$$\text{cov}(\boldsymbol{\tau})_{ij} = (\text{TRLC}_{\text{max}} - \text{TRLC})^{\nu}\sigma_\tau^2, \qquad \forall i, j \tag{30}$$

In Eq. (30), TRLC denotes the cardinal TRL coefficient corresponding with the experimental data, TRLC_{max} is the highest achievable cardinal TRL coefficient (9.0 in the Table 10 example), ν governs the rate at which the maturity uncertainty grows, and σ_τ^2 is the characteristic variance. The value of σ_τ^2 must be specified to properly model the scale of the additional epistemic uncertainty. This variable has the units of the target variable squared. Specification of the exponent ν must also be selected according to how one wishes to model the rate of uncertainty change as TRLC is varied. The behavior of Eq. (30) for three different settings of ν and $\sigma_\tau^2 = 1$ is shown in Fig. 40. As can be seen in the figure, the multiplier on the characteristic variance grows quickly with increasing values of ν, particularly for small values of TRLC.

The additive uncertainty model has some desirable characteristics. One is that it provides a lower bound for the forecasting uncertainty when $\text{TRLC} < \text{TRLC}_{\text{max}}$. If the diagonal terms of Eq. (26) approach zero in regions of the input space that

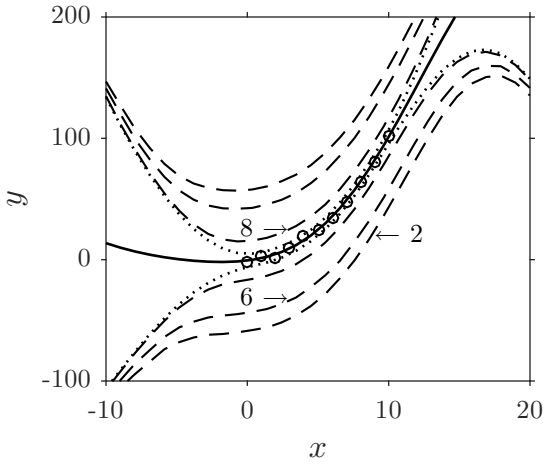


Figure 41: GP predictions for the mean (solid line) and 95% prediction intervals for TRLs of 2, 6, and 8 (all dashed lines) and TRL 9 (dotted line) for notional data (◯ symbols).

are densely populated with data, the additional maturity uncertainty modeled with Eq. (29) ensures that the prediction uncertainty cannot shrink below the specified bound. As TRLC $\rightarrow$ TRLC$_{\text{max}}$, the maturity uncertainty approaches zero and the GP prediction uncertainty approaches that governed by Eq. (26). This behavior is illustrated in Fig. 41. The dashed lines show prediction bounds for multiple TRLs. As TRL increases, the prediction intervals shrink toward the TRL 9 interval shown as the dotted lines. Another desirable characteristic is the behavior of prediction uncertainty in sparsely-populated regions of the input space and when extrapolating. As shown in Fig. 41, the prediction intervals for all TRLs converge with the dotted lines as x moves away from the data. Thus, the uncertainty for scenarios where TRLC $<$ TRLC$_{\text{max}}$ does not grow unnecessarily large in sparse regions of the input space.

In the case that training data from computer experiments are used, one may wish to include model form uncertainty in addition to maturity uncertainty. This can be accomplished by specifying another multivariate normal random variable, similar to Eq. (29), that has a covariance matrix with variance terms that are estimated based

on validation of the computer model. Note that the model form uncertainty will not be constant throughout the input space, as it will be relatively small in regions where validation data exist and grow larger with distance from the validation data.

As indicated in Fig. 38, after this step is complete, a predictive GP model has been built. This model can then be used for additional modeling tasks, such as propagating uncertainty to system-level metrics or building k-factor distributions. If another round of experimentation is being planned, then the next step of the methodology should be used to evaluate the uncertainty reduction of the proposed experiments.

5.2.5 Step Five: Quantify Expected Information Gain From Proposed Experiments

To estimate the uncertainty reduction that can be expected from a proposed experiment, an information theoretic framework has been implemented. In this framework, uncertainty of GP model predictions $\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j$ with PDF $p(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j)$ is measured using the differential entropy, which is defined as follows:

$$h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j) = \mathrm{E}[-\log\, p(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j)] \tag{31}$$

where, $\mathcal{T}_j = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_j\}$ is the set of all training data available from previous experiments 1 through j, and $\mathbf{X}_*$ denotes prediction locations of interest. After a proposed experiment $j+1$ has been conducted, the information gained by collecting the data is $h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j) - h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1})$. The posterior entropy after experiment $j+1$ cannot be quantified until the experiment has been conducted and a GP regression model trained on the data. Before the proposed experiment, the targets at $\mathbf{X}_*$ are uncertain and can be treated as random. As an estimate, the average posterior uncertainty $\mathrm{E}_{y_{j+1}|\mathbf{x}}[h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1})]$, where $\mathrm{E}_{y_{j+1}|\mathbf{x}}[\cdot]$ is the expectation with respect to the distribution of the target variable conditioned on the inputs, can be computed before the proposed experiment has been conducted. With this approach, the expected

information gain from experiment $j+1$ is

$$I(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j\,;\,\mathcal{D}_{j+1}) = h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j) - \mathrm{E}_{y_{j+1}|\mathbf{x}}[h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1})] \tag{32}$$

This quantity is called mutual information, and it can be interpreted as the reduction in the uncertainty of $\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j$ due to the knowledge of $\mathcal{D}_{j+1}$ [67]. Ideally, a planned experiment will maximize mutual information.

An assumption that is made in this methodology is that the prior entropy $h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_j)$ is not a function of the design of the planned experiments. With this assumption, maximization of mutual information is equivalent to minimizing the expected posterior uncertainty $\mathrm{E}_{y_{j+1}|\mathbf{x}}[h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1})]$. Thus, in practice it is not necessary to quantify the prior entropy; the expected posterior entropy can be used in place of mutual information.

Steps for estimating the expected posterior entropy for a given planned experiment are presented here. It is assumed that there are locations in the input space that are of interest. A simulation-based approach is used to estimate the joint posterior entropy at these points, for each realization of experimental observations. Then, the results are averaged.

5.2.5.1 Establish Points of Interest

For any technology experiment, there will be points in the independent variable space where technologists want to measure performance. These points may be contained in the convex hull defined by the training data or outside of it. The former is referred to as interpolation and the latter as extrapolation. As a notional example, consider the data shown for two inputs in Fig. 42. The circle symbols represent locations for measurements in an experiment, and the solid line is the corresponding convex hull. The + symbols represent points of interest. Prediction for the four points of interest inside of the convex hull require interpolation using a regression model, whereas the other 12 points of interest require extrapolation with the regression model. The

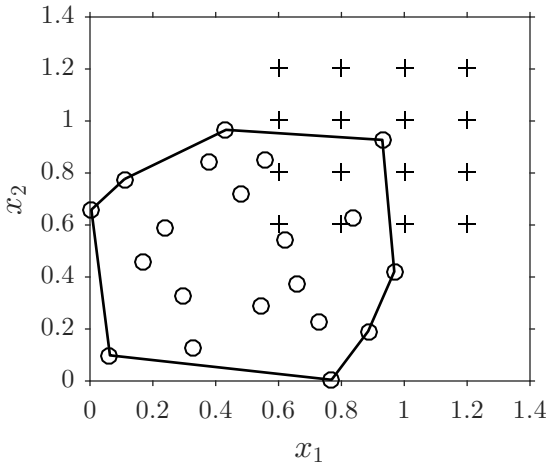


Figure 42: Notional training data (◯ symbols), the corresponding convex hull (solid line), and notional points of interest (+ symbols).

farther the points of interest are outside of the convex hull, the larger the uncertainty in predictions due to extrapolation.

For a real example where points of interest lie outside of the domain of the training data, consider Fig. 37. The points of interest for this example lie in the range $\beta \in [-7.5°, 0°]$. If only data from a flight experiment were to be used to train a regression model for $\beta \in [0°, 15°]$, then predictions for the points of interest would require extrapolation.

5.2.5.2 Simulate Observations From the Proposed Experiment

Before simulating data from a proposed experiment, the design matrix $\mathbf{X}_{j+1}$ must be specified. It is assumed that this information is given. To completely define a simulated experiment, the targets are needed as well. To simulate the targets at locations in the design matrix, a target function is drawn from Eq. (24), then targets are sampled with or without noise. If noisy data are simulated, then the noise variance $\sigma^2_{y_{j+1}}$ must be selected. The most accurate way to do this is to estimate the precision that can be achieved in the proposed experiment. Nevertheless, it is assumed that the noise variance can be specified. The process of simulating data is repeated N_{sim}

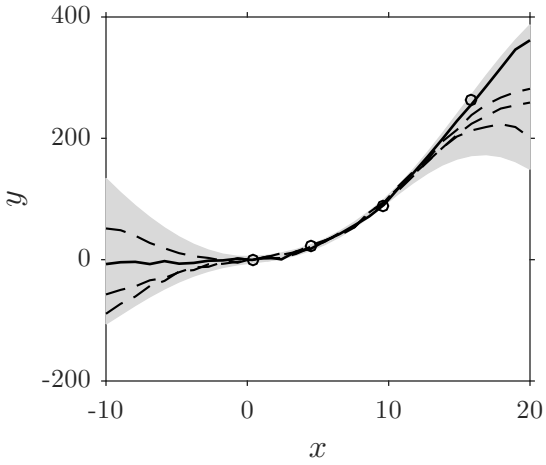


Figure 43: Target function realizations (dashed lines and solid line), simulated data (◯ symbols), and the GP 95% prediction intervals (gray area).

times.

To illustrate the sampling method, multiple random target function draws are plotted in Fig. 43. These functions were sampled from a single-task GP trained with the data shown in Fig. 41. The four observations, designated with circle symbols, were then sampled for one of the random target function realizations (solid line) from a normal distribution with mean defined as in Eq. (25) and variance $\sigma^2_{y_{j+1}} = 10$. The 95% prediction intervals are shown as the gray area for reference.

5.2.5.3 Train Regression Models for Each Simulated Experiment

For each of the N_{sim} sets of simulated training data, a regression model must be trained. A single-task or multitask GP model can be used. The value of N_{sim} should be selected based on computational budget, as GP models can be expensive to train and use for predictions when the number of observations is large. For a fair comparison between proposed experiments, the same GP model architecture should be used for all of the experiments. The selected architecture may not be ideal for all data sets. However, for this step the primary objective is not predictive accuracy but rather to measure uncertainty in the predictions at the points of interest.

5.2.5.4 *Estimate Posterior Entropy*

In order to estimate $\mathrm{E}_{y_{j+1}|\mathbf{x}}[h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1})]$, predictions at the points of interest must be made with all N_{sim} GP models. Of particular interest is the covariance matrix for the joint distribution of the predictions. The covariance matrix of the predictions is the sum of Eqs. (26) and (30). The TRLC value corresponding with the proposed experiment should be used in Eq. (30). With the covariance matrix computed, the posterior joint entropy for a single realization of training data is that of a multivariate normal distribution [67]:

$$h_q(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1}) = \frac{1}{2}\log\left((2\pi e)^{n_{j+1}}|\mathrm{cov}(\mathbf{f}_*) + \mathrm{cov}(\boldsymbol{\tau})|\right) \tag{33}$$

where, n_{j+1} is the number of observations for the proposed experiment, $|\cdot|$ denotes the determinant, and the entropy is in units of nats. Note that differential entropy for continuous variables, unlike entropy for discrete variables, can be negative. To include model form uncertainty, an additional covariance term can be used in Eq. (33). Finally, the posterior entropy mean is estimated as follows:

$$\mathrm{E}_{y_{j+1}|\mathbf{x}}[h(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1})] \approx \frac{1}{N_{\text{sim}}}\sum_{q=1}^{N_{\text{sim}}} h_q(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1}) \tag{34}$$

where, the $\approx$ symbol indicates that the quantity is an estimator for the true mean.

On a practical note, many GP regression model software packages do not explicitly produce the full posterior predictive covariance matrix but rather the diagonal of this matrix. This is because the diagonal of the covariance matrix is all that is needed to estimate prediction intervals. If one prefers not to perform the matrix algebra to obtain the full covariance matrix, there is an alternative uncertainty measure that can be used in lieu of Eq. (33). The alternative measure is the upper bound for joint differential entropy, which is the sum of the entropies of each random variable in the joint distribution [67]. For GPs, the distribution at each point of interest is Gaussian, so the upper bound of the posterior joint entropy for a single realization of training

data is

$$h_q(\mathbf{f}_*|\mathbf{X}_*, \mathcal{T}_{j+1}) = \sum_{d=1}^{N_{\text{POI}}} \frac{1}{2} \log(2\pi e(\text{cov}(\mathbf{f}_*)_{dd} + \text{cov}(\boldsymbol{\tau})_{dd})) \tag{35}$$

where, N_{POI} is the number of points of interest.

5.3 Gaussian Process Comparison Experiment

In Sec. 5.2.2.2 a research gap was identified. For convenience, the research question is repeated here:

Research Question 2.1: Under what conditions will a multitask GP regression model provide better generalization performance than a single-task GP regression model?

An experiment was conducted to investigate this research question. Four different multitask GP models and one single-task GP were compared based on generalization performance for analytical functions under different scenarios. The setup and results are presented here.

5.3.1 Setup of the Experiment

The baseline single-task GP regression model selected for the experiment is the MATLAB built-in implementation [104]. Special options used for training the single-task GPs include a squared exponential ARD covariance function, a lower bound on σ_y^2 of 1E-15, and no basis functions. All other options were left at the default settings.

The training data for all cases were standardized before regression. For the inputs, the simulated observations were standardized by subtracting the mean of the observations from the data, then dividing by the standard deviation of the observations. The same approach was used to standardize the targets.

The four multitask GP models were selected from the literature to provide results for different approaches to inductive transfer learning. Each of the models and their implementation settings are presented here. Then, the analytical functions used

in the experiment are presented. Finally, the training data generation process and performance measures are described.

5.3.1.1 MT-IVM

The multitask IVM (MT-IVM) GP model [92] was selected to represent a model structure where all tasks communicate only by sharing the same set of hyperparameters. To illustrate the model architecture, a graphical model for MT-IVM is shown in Fig. 44b for three tasks. For comparison, a single-task GP architecture is shown in Fig. 44a. No information is shared between tasks for the single-task GP model; GP models for all three tasks are trained independently. Information sharing across tasks for MT-IVM is achieved through the common set of hyperparameters $\boldsymbol{\theta}$. These hyperparameters are determined through maximization of the marginal likelihood. Note that independent noise variance parameters σ_y^2 are used for each task.

The IVM component of the method enables the use of only the most informative observations during training. The size of this "active point" set is an input to the IVM algorithm, and the selection process occurs both within and across tasks. For fair comparison, all of the observations were used in this experiment.

To implement MT-IVM, a software package written by one of the authors of Ref. [92] was used in MATLAB (see Ref. [105]). A squared exponential ARD covariance function and a Gaussian noise model were used for all experiments. Hyperparameter optimization was limited to a maximum iteration number of 30,000. All other options were left at the default settings.

5.3.1.2 MTGP

Bonilla et al. [97] formulated a multitask GP with the goal of obtaining regression models that have improved performance when tasks are related and do not suffer performance degradation when tasks are unrelated. They proposed the use of a "free-form" task-similarity matrix to model inter-task dependencies. To help minimize

the possibility of over-fitting in situations where observations are sparse, a common covariance function is used for all tasks. This model was selected for the experiment because it uses shared hyperparameters for the tasks and also explicitly models the relatedness of the tasks. Assuming zero-mean GP priors on the target functions for all tasks, the key feature of the model is the form of the GP covariance:

$$\text{cov}(f_l(\mathbf{x}), f_m(\mathbf{x}')) = K^f_{lm}\kappa(\mathbf{x}, \mathbf{x}') \tag{36}$$

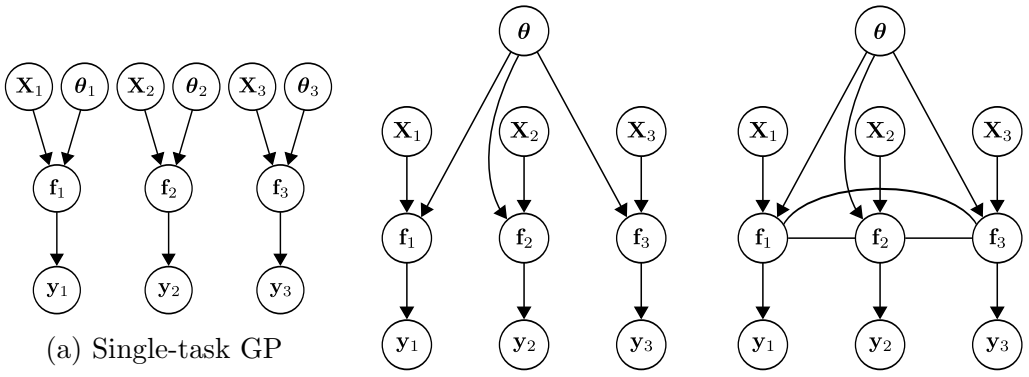
where, f_l and f_m are target functions from two different tasks, and $\mathbf{K}^f$ is a positive semi-definite matrix that defines inter-task similarities. The graphical model for this architecture is shown in Fig. 44c. The undirected edges connect all of the latent target functions because the inter-task relationships are explicitly modeled. Note that independent noise variance parameters σ_y^2 are used for each task in this model. An interesting property of this model is that when noiseless observations are all at the same input locations for all tasks, there is no inter-task transfer.

This multitask model was implemented using a software package called "Multi-task Gaussian process" (MTGP) [106], which was written by one of the authors of Ref. [97]. A squared exponential ARD covariance function was used, and the option to use full rank $\mathbf{K}^f$ matrix was selected. Hyperparameter optimization was limited to a maximum iteration number of 30,000. All other options were left at the default settings.

5.3.1.3 Co-Kriging

An autoregressive multitask GP model referred to as co-Kriging (CK) was also used in the experiment. It was selected due to its success in multifidelity optimization. The seminal work for this application domain is that of Forrester et al. [107]. The CK model was built on the following assumption regarding two levels of computer codes $Z_t(\cdot)$ and $Z_{t-1}(\cdot)$, where $Z_t(\cdot)$ is the higher level (fidelity/order/expense) code:

$$\text{cov}(Z_t(\mathbf{x}), Z_{t-1}(\mathbf{x}')|Z_{t-1}(\mathbf{x})) = 0 \qquad \forall \mathbf{x}' \neq \mathbf{x} \tag{37}$$



(a) Single-task GP (b) MT-IVM (c) MTGP

Figure 44: Graphical models representing three different ways to learn three tasks.

The interpretation of Eq. (37) is that given the nearest observation of $Z_{t-1}(\mathbf{x})$, no more can be learned about $Z_t(\mathbf{x})$ from any other observation $Z_{t-1}(\mathbf{x}')$ for $\mathbf{x}' \neq \mathbf{x}$. Note that the notation found in Ref. [96] is used here, where the task index t is equivalent to $j+1$ used in Sec. 5.2.5. The assumption in Eq. (37) led to the autoregressive form

$$Z_t(\mathbf{x}) = \rho_{t-1} Z_{t-1}(\mathbf{x}) + \delta_t(\mathbf{x}), \quad \text{where } t = 2, 3, \ldots, s \tag{38}$$

where, $Z_{t-1}(\mathbf{x})$ is a GP modeling the lower level code, $\delta_t(\mathbf{x})$ is a GP that represents location adjustment that is independent of $Z_{t-1}(\cdot), \ldots, Z_1(\cdot)$, and ρ_{t-1} is a scale adjustment parameter. The CK model has the most explicit knowledge transfer mechanism of the multitask models selected for this experiment; the regression model of one task is built as a corrected model of a GP for a lower level task. One of the disadvantages of CK is that it requires nested training observations $\mathcal{D}_t \subseteq \mathcal{D}_{t-1}$. This facilitates estimation of the adjustment terms in Eq. (38). For data sets that are not nested, a GP model can be used to estimate the $Z_{t-1}(\mathbf{x})$ observations at the locations in $\mathcal{D}_t$. This is the approach used in the experiment.

A CK model was implemented in the R programming language [108] with the MuFiCoKriging package [109]. A "Gauss" (squared exponential ARD) covariance function was used for the experiment. Constant basis functions were used for both tasks. The nugget (noise variance) estimation option was turned on for all cases. All

other options were left at the default settings. The DiceKriging package [110] was used to predict observations for task j at locations of the data for task $j+1$ with a single-task GP. The settings used in DiceKriging were the same as those used for MuFiCoKriging.

5.3.1.4 MULTIGP

Álvarez and Lawrence [98, 99] proposed the representation of each task as the convolution of a smoothing kernel and a latent function. With certain restrictions, their model reduces to the MTGP architecture. The convolution process approach was selected for the experiment because it provides a sophisticated class of covariance structures. As with MTGP, the key ingredient is the GP covariance form:

$$\mathrm{cov}(f_l(\mathbf{x}), f_m(\mathbf{x}')) = \sum_{s=1}^{S} \int_{-\infty}^{\infty} \kappa_{ls}(\mathbf{x}-\mathbf{z}) \int_{-\infty}^{\infty} \kappa_{ms}(\mathbf{x}'-\mathbf{z}') \kappa_{u_s u_s}(\mathbf{z},\mathbf{z}')\, d\mathbf{z}' d\mathbf{z} \qquad (39)$$

where, $\kappa_{ls}(\cdot)$ and $\kappa_{ms}(\cdot)$ are smoothing kernels for tasks l and s, respectively, and $\kappa_{u_s u_s}(\mathbf{z},\mathbf{z}')$ is the covariance function for the latent function $u_s(\mathbf{z})$. Note that independent noise variance parameters σ_y^2 are used for each task in this model.

A convolution process model was implemented using the Multi-output Gaussian Processes (MULTIGP) software package [111]. The full GP model was used, rather than one of the approximations offered in the software. One latent function was used for the experiment. A squared exponential ARD form was used for the latent function covariance function and the smoothing kernel. Hyperparameter optimization was limited to a maximum iteration number of 30,000. All other options were left at the default settings.

5.3.1.5 Analytical Functions

Since the behavior of technology performance measured in experiments can vary from one technology to another due to differing governing physics and setups of the experiments, it was decided to use analytical functions in lieu of measurements from specific

technology experiments. Another reason analytical functions were used is that they provide a truth function for comparison with predictions from the regression models. Three analytical functions were used to simulate a scenario in which a data set is available from an "expensive" higher-TRL experiment and another data set is available from a "cheap" lower-TRL experiment. Each set of functions provides a range of task similarities.

The first analytical function is the Branin function [112], which has been used by many researchers to test the performance of optimization algorithms and regression models. The Branin function served as the expensive target function, and the equation is

$$f_e(\mathbf{x}) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10 \tag{40}$$

For the cheap experiment, a parametric function formulated by Toal [101] was used:

$$f_c(\mathbf{x}) = f_e(\mathbf{x}) - (A_1 + 0.5)\left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 \tag{41}$$

where, $A_1 \in [0, 1]$ is a parameter that governs the degree of similarity in behavior between f_e and f_c. The domain used for Eqs. (40) and (41) is $x_1 \in [-5, 10]$, $x_2 \in [0, 15]$. The similarity between the two functions was quantified with the squared sample correlation r^2 and the root mean square error (RMSE), which are defined respectively as

$$r^2 = \left(\frac{\sum_{i=1}^{n}(y_{e_i} - \overline{y}_e)(y_{c_i} - \overline{y}_c)}{\sqrt{\sum_{i=1}^{n}(y_{e_i} - \overline{y}_e)^2 \sum_{i=1}^{n}(y_{c_i} - \overline{y}_c)^2}}\right)^2 \tag{42}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{e_i} - y_{c_i})^2} \tag{43}$$

where, y_e and y_c are targets for the expensive and cheap functions, and $\overline{y}_e$ and $\overline{y}_c$ are the expected values of the n observations from each. The correlation and RMSE between the expensive Branin function and the cheap function for a range of A_1 values are shown in Fig. 45. As can be seen, varying A_1 simulates multiple types

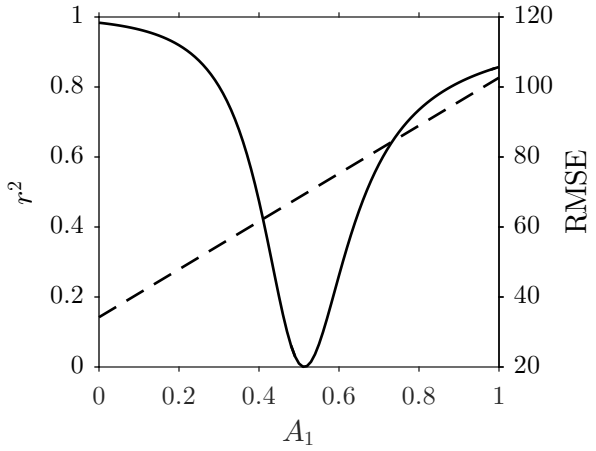
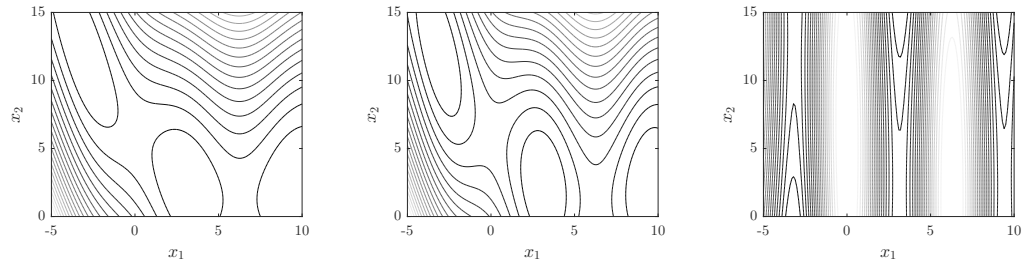


Figure 45: Behavior of correlation (solid line) and RMSE (dashed line) between the expensive and cheap Branin functions as A_1 varies.



(a) $f_e(\mathbf{x})$ (b) $f_c(\mathbf{x})$, $A_1 = 0$ (c) $f_c(\mathbf{x})$, $A_1 = 0.514$

Figure 46: Contour plots of the expensive Branin function and the cheap function with two different settings of A_1. Lighter gray indicates high magnitude of f.

of scenarios. Low values of A_1 correspond with high correlation between the two functions and relatively low error. As A_1 approaches values in the neighborhood of 0.5, the correlation approaches zero. Higher values of A_1 dictate high correlation with relatively large error.

To further facilitate understanding of how A_1 affects the similarity between the expensive and cheap functions, contour plots are shown in Fig. 46. Comparing Figs. 46a and 46b, one will see that the contours are similar but not identical. When these figures are compared with Fig. 46c, the effect of A_1 is apparent. At a value of 0.514, where the correlation reaches a minimum, A_1 almost nullifies the entire squared term in Eq. (41), and the cosine term dominates.

The second analytical function used was also of input dimension two, and it is referred to as the Paciorek function [101]. The expensive and cheap versions are defined respectively as

$$f_e(\mathbf{x}) = \sin\left(\frac{1}{x_1 x_2}\right) \tag{44}$$

$$f_c(\mathbf{x}) = f_e(\mathbf{x}) - 9A_2^2 \cos\left(\frac{1}{x_1 x_2}\right) \tag{45}$$

where, $A_2 \in [0, 1]$ is a parameter that governs the degree of similarity between f_e and f_c. The domain used for Eqs. (44) and (45) is $x_1 \in [0.3, 1]$, $x_2 \in [0.3, 1]$. The correlation and RMSE between the two functions is plotted in Fig. 47. The figure illustrates that the Paciorek function with $A_2 = 1$ provides a scenario where the correlation and error are both at their respective maxima. Also, the Paciorek function with $A_2 = 0$ provides a situation in which the expensive and cheap functions are identical. Contours of the cheap Paciorek function for two different settings of A_2 are shown in Figs. 48b and 48c. Comparing these with the expensive function contours in Fig. 48a, it can be seen that the correlation reduction is due the cosine term shifting the contours and introducing a local minimum in the top right corner of the cheap function plots.

The third analytical function used in the experiment was the Trid function with 10 input dimensions [113]. This function is popular for testing unconstrained optimization algorithms. A cheap version of the Trid function was also used from Ref. [101]. The equations for the expensive and cheap functions, respectively, are

$$f_e(\mathbf{x}) = \sum_{i=1}^{10}(x_i - 1)^2 - \sum_{i=2}^{10} x_i x_{i-1} \tag{46}$$

$$f_c(\mathbf{x}) = \sum_{i=1}^{10}(x_i - A_3)^2 - (0.65 - A_3)\sum_{i=2}^{10} i x_i x_{i-1} \tag{47}$$

where, $A_3 \in [0, 1]$ is a parameter that governs the degree of similarity between f_e and f_c. The domain used for Eqs. (46) and (47) is $x_i \in [-100, 100]$ for all i. The correlation and RMSE between the two Trid functions are plotted in Fig. 49. Note

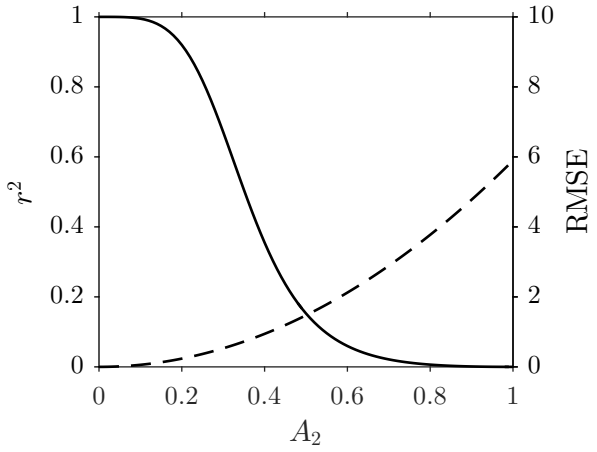
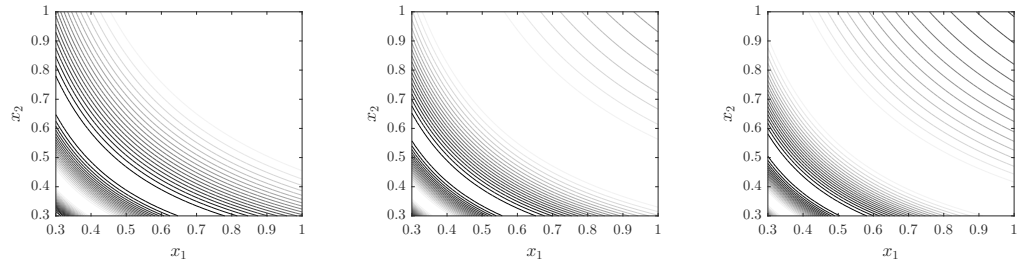


Figure 47: Behavior of correlation (solid line) and RMSE (dashed line) between the expensive and cheap Paciorek functions as A_2 varies.



(a) $f_e(\mathbf{x})$ (b) $f_c(\mathbf{x})$, $A_2 = 0.35$ (c) $f_c(\mathbf{x})$, $A_2 = 1$

Figure 48: Contour plots of the expensive Paciorek function and the cheap function with two different settings of A_2. Lighter gray indicates high magnitude of f.

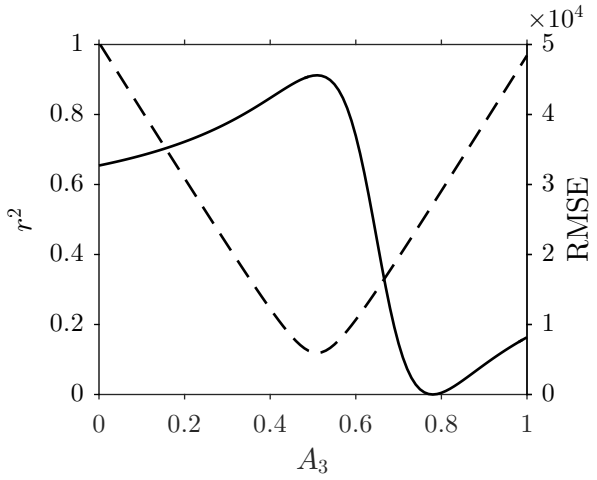


Figure 49: Behavior of correlation (solid line) and RMSE (dashed line) between the expensive and cheap Trid functions as A_3 varies.

that the term $(0.65 - A_3)$ in Eq. (47) was published as $(A_3 - 0.65)$ in Ref. [101]. The author believes that this was a mistake, as the correlation and RMSE behavior did not follow what was published by Toal. The Trid function was included in the experiment to investigate whether the results would differ significantly from the two-dimensional functions. Since $D = 10$ for the Trid function, it is not easily visualized.

5.3.1.6 Training Data Generation

To simulate a variety of scenarios, data was generated from the analytical functions a number of ways. The training data size for the expensive functions was fixed, but it was varied for the cheap functions. All input points were determined by the MATLAB Latin hypercube generator lhsdesign with 100 iterations to improve the designs according to the maximin criterion. The Latin hypercube design strategy was selected due to its demonstrated superiority over other techniques, within the context of GP regression, in the literature. To simulate both physical and computer experiments, training data were generated using Gaussian noise with variance corresponding to two signal-to-noise ratios (SN) or using deterministic observations. Data were generated with 11 settings of the A parameters for all functions. For each combination of sample sizes, SN values, and A parameter settings, $N_{\text{rep}} = 50$ replicates were generated to capture variability in the results due to the random nature of Latin hypercube sampling and noisy observations.

In all cases, five points per input dimension, $\text{ND}_e = 5$, were sampled from the expensive functions. In other words, 10 data points were used for the Branin and Paciorek functions, and 50 points were sampled from the Trid function. The specification of $\text{ND}_e = 5$ was partly motivated by the arguments presented by Loeppky et al. [114] that 10 observations per input dimension is a sufficient rule for the initial sampling of computer experiments. They demonstrated that this sampling rule typically provides sufficient accuracy in GP predictions. For this experiment, it was

desirable for the single-task GP predictive accuracy to be relatively low so that the effect of knowledge transfer would be evident. This situation is aligned with a technology development setting where there may be sparse data from expensive experiments. Thus, it was decided to halve the suggested sample size. If a much denser sampling of the input space had been used, there may not have been clear differences between the regression models. Another reason five points per dimension was used is that Toal [101] observed significant differences between a single-task model trained with this sampling rule and a CK model trained with a variety of different sampling rules.

Samples for the cheap functions were generated using 5, 10, and 15 observations per input dimension, ND_c. For the Branin and Paciorek functions, this resulted in 10, 20, and 30 observations, respectively. For the Trid function, the sampling rules resulted in 50, 100, and 150 observations, respectively. For each sample size rule, the data were generated from the target cheap functions using A values ranging from 0 to 1 by 0.1 increments.

The cheap and expensive functions were sampled using three SN values: 100, 400, and ∞. The setting $\mathrm{SN} = \infty$ is a deterministic sampling, where the observations were drawn directly from the analytical function without added noise. For the other two SN settings, noise was generated with a normal distribution having zero mean. The variance of the normal distribution was determined by dividing the squared difference between the maximum and minimum truth function values by SN: $\frac{(f_{\max}-f_{\min})^2}{SN}$. Hence, the noise variance for the expensive functions was fixed for a given SN, whereas the noise variance also depended on A for the cheap functions. All four combinations of noise on/off for the cheap and expensive functions were run. In all cases where both cheap and expensive functions were sampled with noise, the same SN was used for the cheap and expensive functions to reduce the number of cases. This was done to simulate scenarios where both are physical experiments, both are computer experiments, or one of the experiments is a physical experiment and the other is a

Table 11: Summary of the data generation scenarios investigated in the GP comparison experiment

Independent Variable	Settings
ND_e	5
ND_c	5, 10, 15
SN_e	100, 400, ∞
SN_c	100, 400, ∞
A	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1

computer experiment.

The data generation scenarios investigated in the experiment are summarized in Table 11. In total, 11,550 scenarios were simulated for each of the three analytical functions.

5.3.1.7 Performance Measures

A multitask GP was trained for each of the data sets that were generated. For comparison, a single-task GP was trained only with the data from the expensive function. The generalization performance of each regression model was estimated using Eqs. (42) and (43). In these equations, the mean of the GP predictions at validation points were used in lieu of y_c. The validation points were selected using a 32-level full factorial design for the Branin and Paciorek functions, resulting in a total of 1,024 validation cases. The Trid function validation points were selected using a combination of a 2-level full factorial design and 18,976 Latin hypercube samples, resulting in a total of 20,000 validation cases. A Latin hypercube design was used for the Trid function to provide space-filling coverage. The 2-level full factorial provided samples of the performance measures at the corners of the input space. A large full-factorial design was not possible due to the combinatorial explosion in 10 dimensions. All validation cases were generated in MATLAB using the fullfact function for the full factorial designs and the lhsdesign function for the Latin hypercube designs.

5.3.2 Hypotheses

A set of hypotheses was established before the experiment to guide the analysis of the results. The rationale for the hypotheses is developed first, then the list of hypotheses is presented.

Caruana [91] identified the lack of a particular way to characterize task relatedness as an important open problem in inductive transfer learning. Chai suggested that "two tasks are related to each other when they benefit mutually under meta-learning" [115]. This is the definition that is used here. There is an assumption based on this definition that is implicit in all of the following hypotheses: higher r^2 correlation between the cheap and expensive functions implies more relatedness between the functions, and more relatedness between tasks improves the performance of multitask learning. Toal [101] showed empirical evidence that supports this supposition. In the experiment, r^2 between tasks was not an independent variable, but the A parameters were used to control it through functional dependence.

In the context of technology development, the goal of a multitask modeling approach is to improve the predictive capability for the expensive or higher-TRL experiment. With this in mind, the hypotheses that were formulated pertain to the generalization capability for the expensive function only.

With the work of Toal [101] as a precedent, an interaction effect between the number of cheap function observations and the correlation between the expensive and cheap functions was anticipated. It was expected that the multitask model predictive performance would increase as the r^2 correlation between the functions increased, for any fixed setting of the other independent variables in this experiment. Similarly, for fixed correlation between the target functions beyond a critical value, it was also anticipated that the multitask models would perform better as more cheap training data became available. However, for fixed correlation between the target functions below a critical value, the multitask models were expected to perform worse as more

cheap training data became available. This is because the more cheap data that are available, the more heavily the cheap function influences the training process. These predictions were expected to be exhibited regardless of the SN for the cheap and expensive functions. The corresponding hypothesis is listed first in the enumeration below.

The effect of SN_c was also expected to exhibit an interaction effect with the correlation between the cheap and expensive functions. For correlation between functions beyond a critical value, higher values of SN_c were anticipated to improve generalization performance. The rationale is as follows. The less noisy the cheap function observations are, the less masked the cheap target function is. If the cheap function were to be highly correlated with the expensive function, then data sampled from the cheap function with higher SN_c should help more with learning the highly correlated expensive function. The opposite effect was expected when the correlation between functions was below a critical value. In this case, the less masked the cheap target function, the more likely inductive transfer would degrade performance with increasing SN_c. These predictions were expected to be exhibited regardless of SN_e and ND_c. Hypothesis 2 follows from this rationale.

With regard to comparison of the multitask GP models described in Sec. 5.3.1, hypotheses were formulated by considering the inductive transfer approach of each model. Predictions of which model would perform best when the correlation between functions was high were not justifiable. However, expectations of the generalization performance of each model in difficult scenarios was possible. Scenarios that were considered to be particularly difficult were any where the correlation between tasks was close to zero. The effects of other independent variables were expected to amplify this effect. Due to the explicit relationship between tasks that the CK model implements, it was predicted that the generalization performance of this model would

decrease below that of the single-task model as the correlations between tasks approached zero. It was reasoned that this would be the case because the CK model for the expensive function is a correction of the cheap model regression. Also, the results from Toal [101] support this prediction. A similar prediction was established for MT-IVM because of the possibility that the hyperparameters would be biased by the cheap data in such a way that predictions for the expensive function would be poor. MTGP and MULTIGP were designed with sophisticated covariance functions to avoid negative transfer. Thus, it was anticipated that the generalization performance of these two models would be less sensitive to the correlation between tasks. The term used here for this insensitivity, regardless of the predictive performance, is "robust". Hypothesis 3 summarizes this prediction.

Hypotheses:

1. If r^2 of the target functions is above a critical value and ND_c is increased with SN_e and SN_c held fixed at any settings, then the generalization performance of the multitask GP regression models will increase.

2. If r^2 of the target functions is above a critical value and SN_c is increased with SN_e and ND_c held fixed at any settings, then the generalization performance of the multitask GP regression models will increase.

3. If r^2 of the target functions is decreased with SN_c, SN_e, and ND_c held fixed at any settings, then the generalization performance of MTGP and MULTIGP will decrease at a slower rate than the generalization performance of CK and MT-IVM.

Evidence that either supports or refutes each of the hypotheses is presented in the following sections for all three analytical functions.

5.3.3 Branin Function Results

To investigate the validity of the hypotheses, the results were summarized by plotting the RMSE and r^2 between the regression predictions and the true expensive Branin function at the validation points. These measures were computed for the multitask and single-task models using the same training data sampled from the expensive function. Note that there were some cases where the training algorithms failed and predictions could not be made.

Evidence supporting hypothesis 1 is shown in Fig. 50 for the lowest SN for both expensive and cheap functions. In these plots, the points represent the median of all 50 replicates (minus any failures) at each value of A_1. The horizontal dashed line in each plot indicates the median for the single-task model. The median was used as a summary instead of mean because of certain cases in which some of the training algorithms had difficulty converging. For a small number of these cases, the RMSE was much larger than the bulk of the other replicates, and these outliers would have heavily biased a mean estimate. Instead of discarding these cases, the median was used, which is more robust to outliers. As observed by Toal [101], the behavior of r^2 and (-)RMSE mimic r^2 in Fig. 45. For certain values of A_1, all models showed improvement in r^2 as ND_c increased, with CK having the largest change. The sensitivity of CK performance to the sample size of the cheap data is not a surprising result because of the dependence structure the model uses. The critical values of A_1 appeared to be near 0.4 and 0.6, which correspond with r^2 of the target functions of approximately 0.45 and 0.3, respectively. Similar behavior was observed for RMSE.

Comparable results are shown in Fig. 51 for the case where SN of both functions was set to 400. However, the large critical value of A_1 appeared to extend beyond 0.6 for MTGP and CK. The plots in Fig. 52 exhibit similar trends for the case when both functions were sampled deterministically. These observations suggest that the performance of MTGP and CK were either unaffected or degraded by an increasing

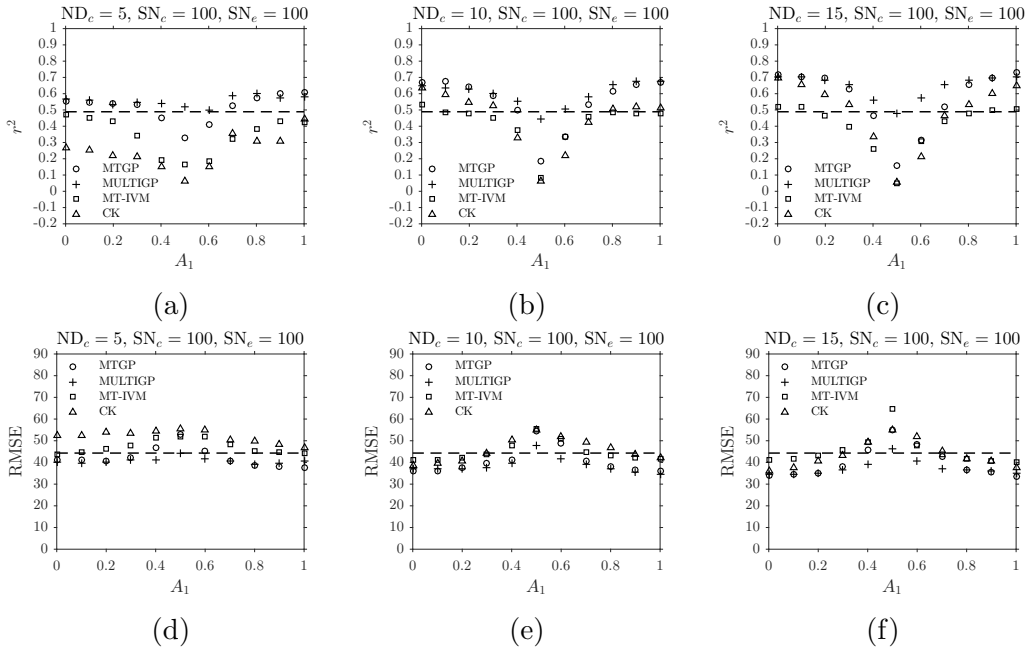


Figure 50: Branin function prediction r^2 and RMSE for multiple values of ND_c, with $\mathrm{SN}_c = 100$ and $\mathrm{SN}_e = 100$.

amount of cheap data even when correlation between target functions was near 0.8. Nonetheless, the plots shown here support hypothesis 1. Also, none of the cases that are not plotted here refuted hypothesis 1.

At the highest correlation between target functions, where $A_1 = 0$, the case where $\mathrm{SN}_c = 400$ and $\mathrm{SN}_e = 400$ showed small improvements in r^2 and RMSE for all of the models, relative to the other noise settings. To more precisely probe the results regarding the validity of hypothesis 1, box plots are shown in Fig. 53. The lower and upper edges of all boxes are the 25th and 75th percentiles of the replicates, respectively. The solid horizontal line within each box was placed at the median of the replicates. As before, the horizontal dashed line indicates the median of the measures for the single-task model. The whiskers were drawn to a maximum length of 1.5 times the vertical length of the boxes, and any data that fell outside of the whiskers is shown with a circle marker. Note that one CK case at $\mathrm{ND}_c = 5$ produced an RMSE value on the order of 10^{16}, but this point is not shown. The contraction

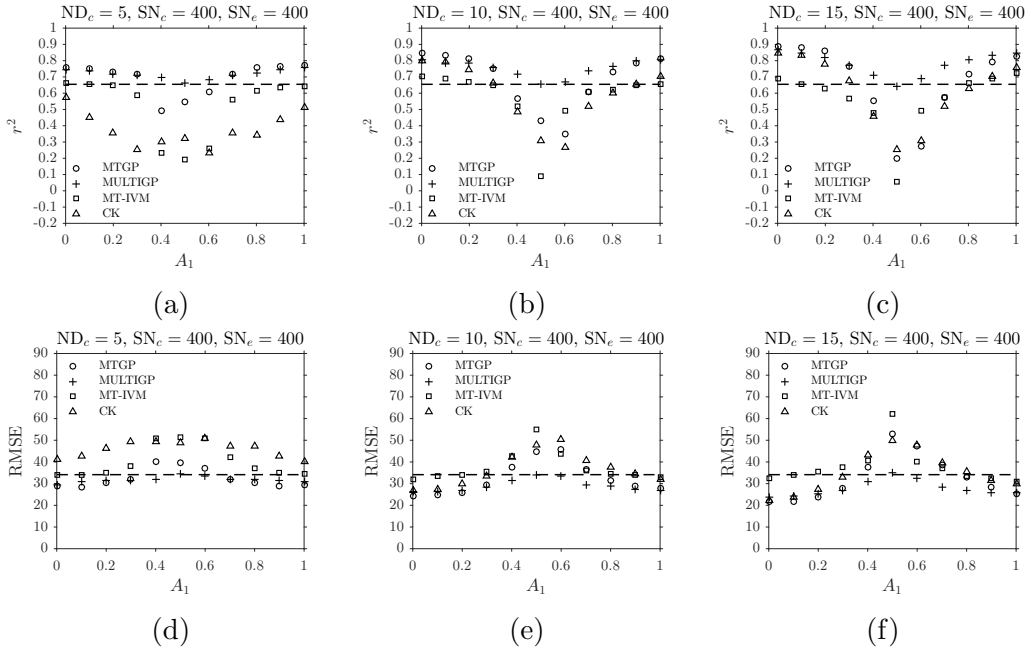


Figure 51: Branin function prediction r^2 and RMSE for multiple values of ND_c, with $\mathrm{SN}_c = 400$ and $\mathrm{SN}_e = 400$.

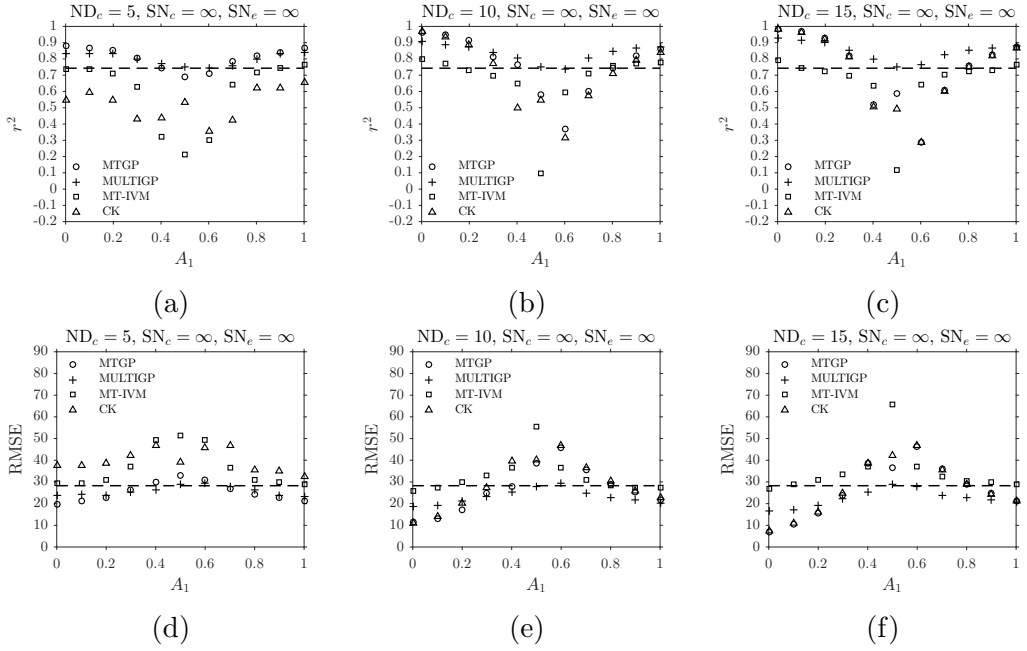


Figure 52: Branin function prediction r^2 and RMSE for multiple values of ND_c, with $\mathrm{SN}_c = \infty$ and $\mathrm{SN}_e = \infty$.

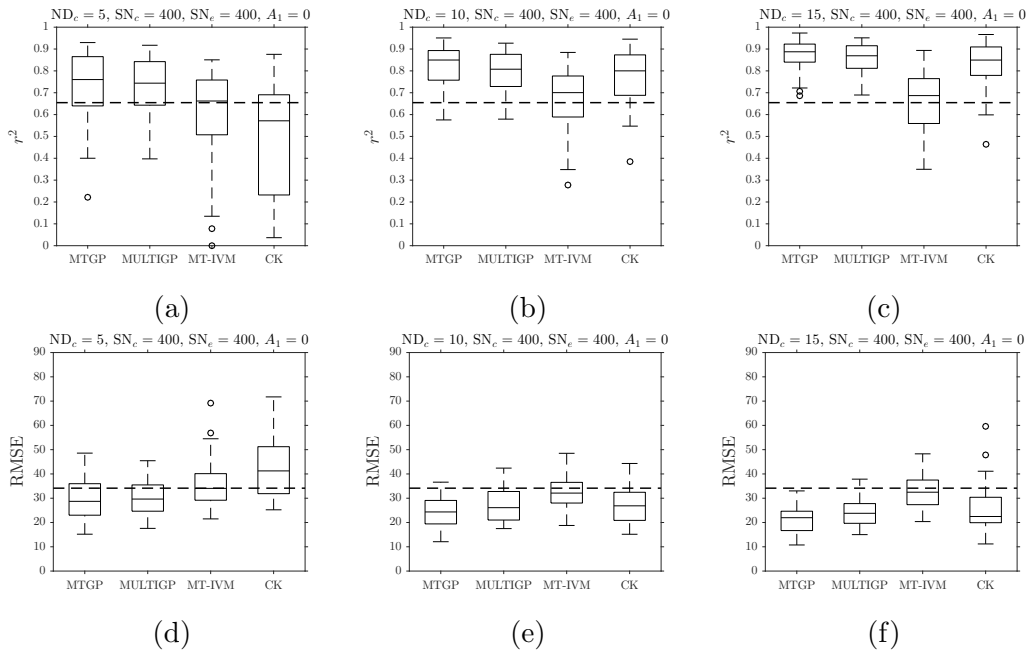


Figure 53: Box plots of Branin function prediction r^2 and RMSE for multiple values of ND_c, with $\mathrm{SN}_c = 400$, $\mathrm{SN}_e = 400$, and $A_1 = 0$.

of the box plots with increasing ND_c indicates that the performance of all multitask models besides MT-IVM did improve with more cheap data, despite the small shift in the medians of certain models. Interestingly, MT-IVM performance appeared to improve as ND_c changed from 5 to 10, but moving to $\mathrm{ND}_c = 15$ did not make a large difference. With the exception of MT-IVM performance, these results support hypothesis 1.

Data for assessing the validity of hypothesis 2 is shown in Fig. 54. As SN_c increased, all of the models showed small gains in performance for certain settings of A_1 that corresponded with high correlation between the target functions. CK was the only model that appeared to have been virtually unaffected by increasing SN_c at $A_1 = 0$.

The effect of increasing SN_c with $\mathrm{ND}_c = 15$ is illustrated in Fig. 55. An interaction effect between ND_c and SN_c is not clearly observed by comparing Figs. 50, 51, and 52. When Fig. 54 is compared with Fig. 55, the interaction is clearly seen. The gains in

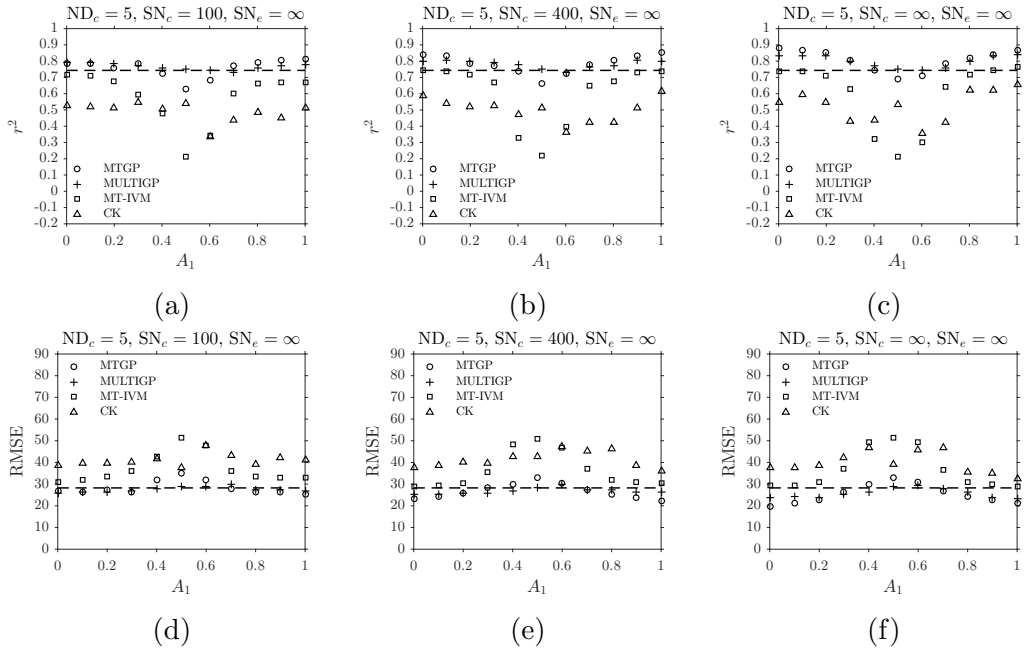


Figure 54: Branin function prediction r^2 and RMSE for multiple values of SN_c, with $\mathrm{ND}_c = 5$ and $\mathrm{SN}_e = \infty$.

performance for more cheap data were larger than for less cheap data. This observation is confirmed by the box plots in Fig. 56. Interestingly, the performance measures of MT-IVM and MULTIGP were not as significantly affected as the other two models. This was particularly true for MT-IVM, as changes in performance for the model were possibly due to variability in the experiment rather than an underlying effect due to increasing SN_c. Although the performance of MULTIGP was not as sensitive to SN_c as MTGP and CK, the majority of replicates outperformed the median performance of the single-task model. Overall, the results suggest that hypothesis 2 is valid.

The validity of hypothesis 3 was not apparent from information found in the figures presented thus far. MULTIGP appeared to be the least sensitive to changes in the correlation between the target functions, but all three of the other models showed varying degrees of sensitivity in different scenarios. As additional evidence, RMSE and r^2 values are plotted in Fig. 57 for the median of all Branin function results for different A_1 settings. The relative insensitivity of MULTIGP is clear in both plots,

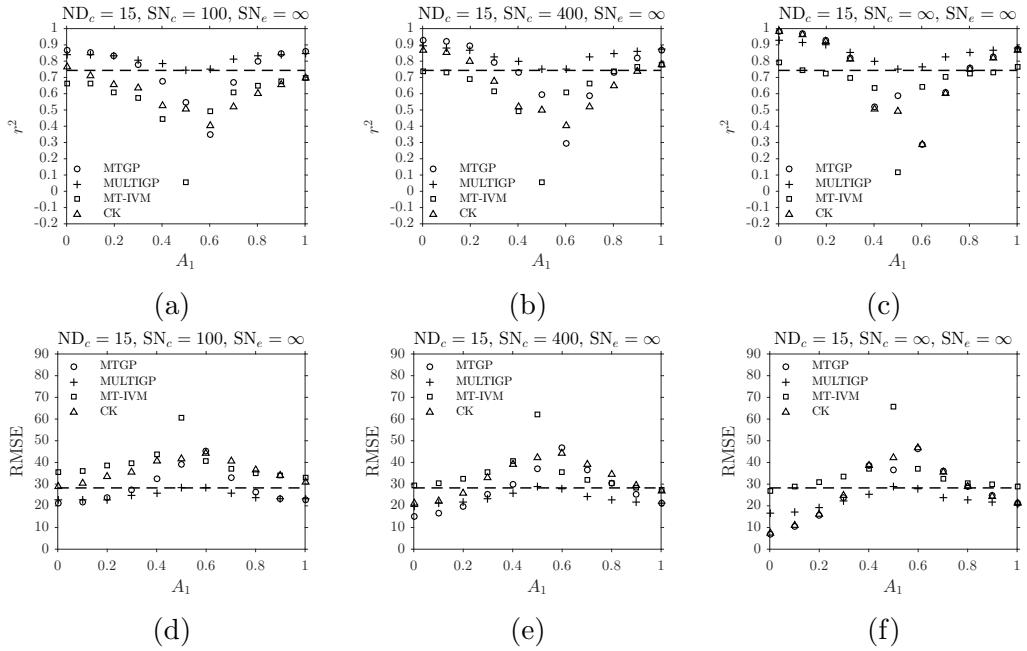


Figure 55: Branin function prediction r^2 and RMSE for multiple values of SN_c, with $\mathrm{ND}_c = 15$ and $\mathrm{SN}_e = \infty$.

whereas MTGP exhibited relatively large changes in RMSE and r^2 as the correlation between target functions decreased. To more clearly visualize the robustness of each model, second-order finite differences were computed for the performance measures and are shown in Fig. 58. In these figures MTGP and CK behaved similarly in terms of local partial derivatives with respect to A_1. For the majority of A_1 values, MT-IVM had smaller local derivatives than CK and MTGP. The derivatives plots serve to confirm that MULTIGP was the most robust overall. Thus, the evidence supports hypothesis 3 with regard to MULTIGP, but the evidence does not support hypothesis 3 with regard to MTGP.

5.3.4 Paciorek Function Results

Figure 59 shows evidence that hypothesis 1 is valid for the Paciorek function as well. The performance of all of the models besides MT-IVM noticeably improved with more cheap data at low values of A_2. As seen with the Branin function, the behavior of r^2 and RMSE in the prediction results followed a similar pattern to r^2 of the target

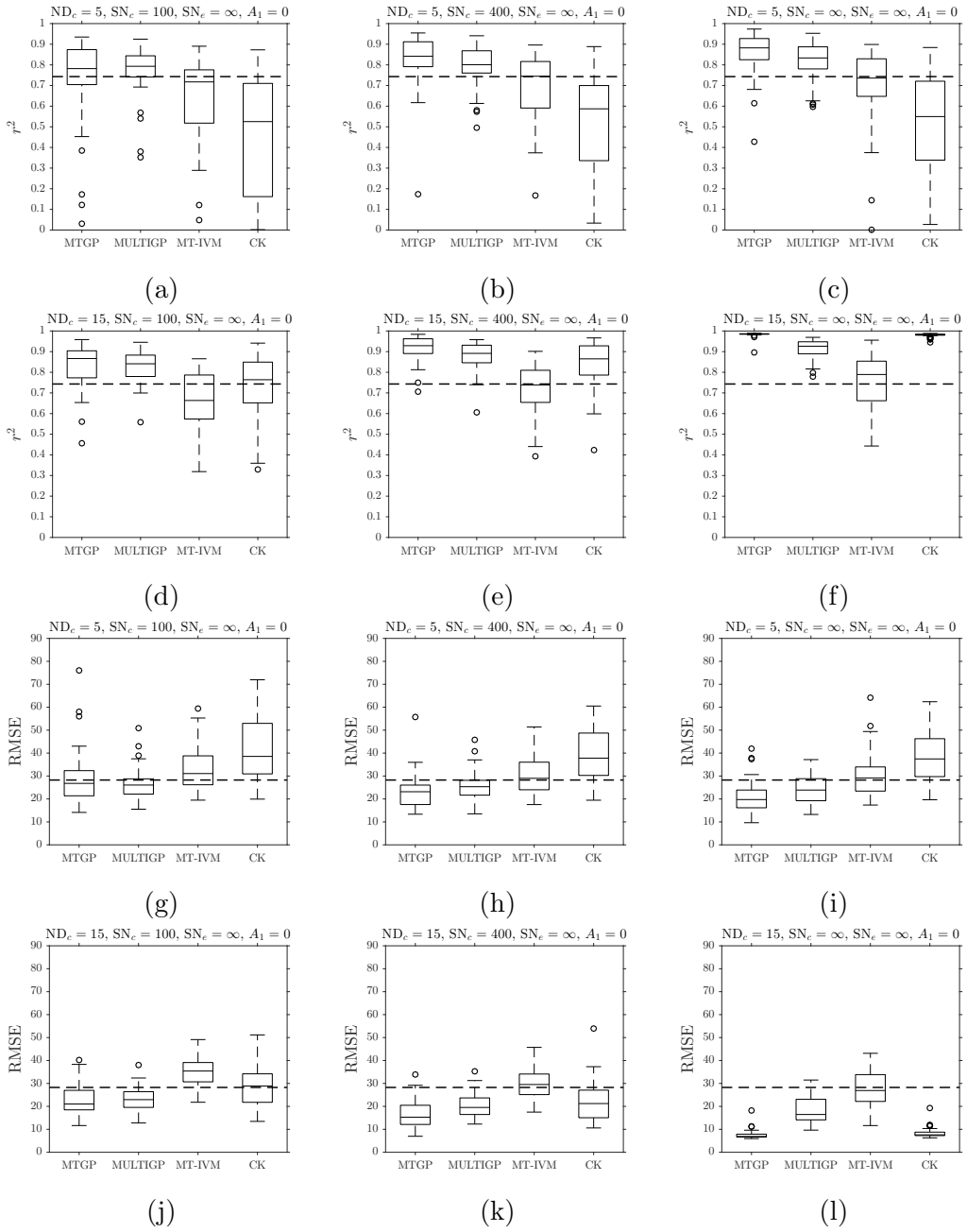
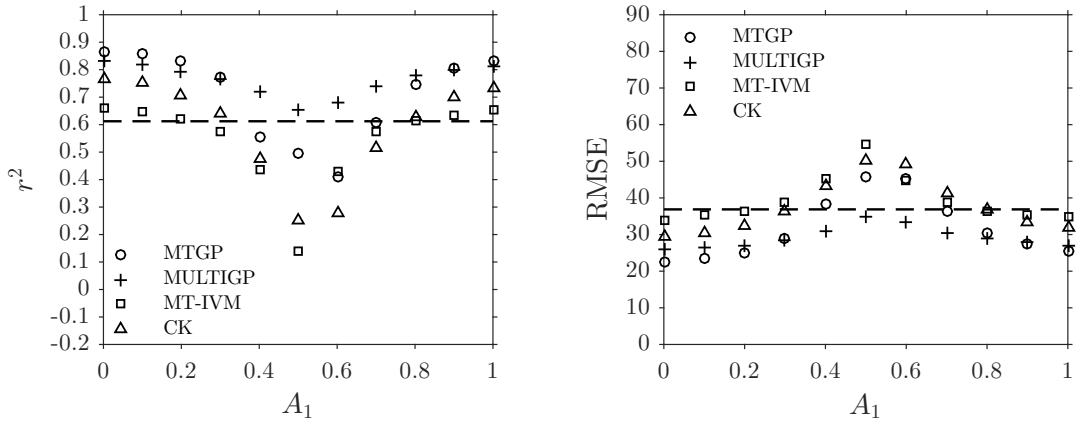
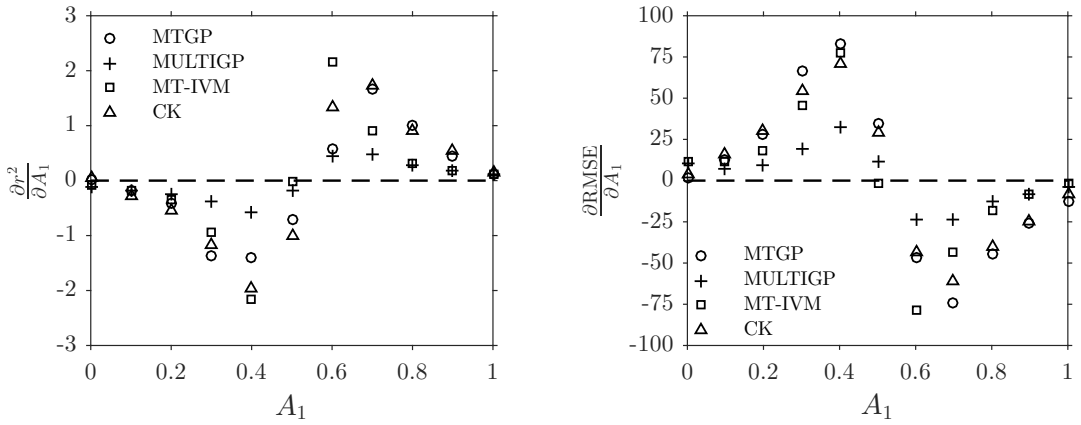


Figure 56: Box plots of Branin function prediction r^2 and RMSE for multiple values of SN_c and two levels of ND_c, with $SN_e = \infty$, and $A_1 = 0$.



(a) (b)

Figure 57: Branin function prediction r^2 and RMSE for all results.



(a) (b)

Figure 58: First partial derivatives of Branin function prediction r^2 and RMSE for all results.

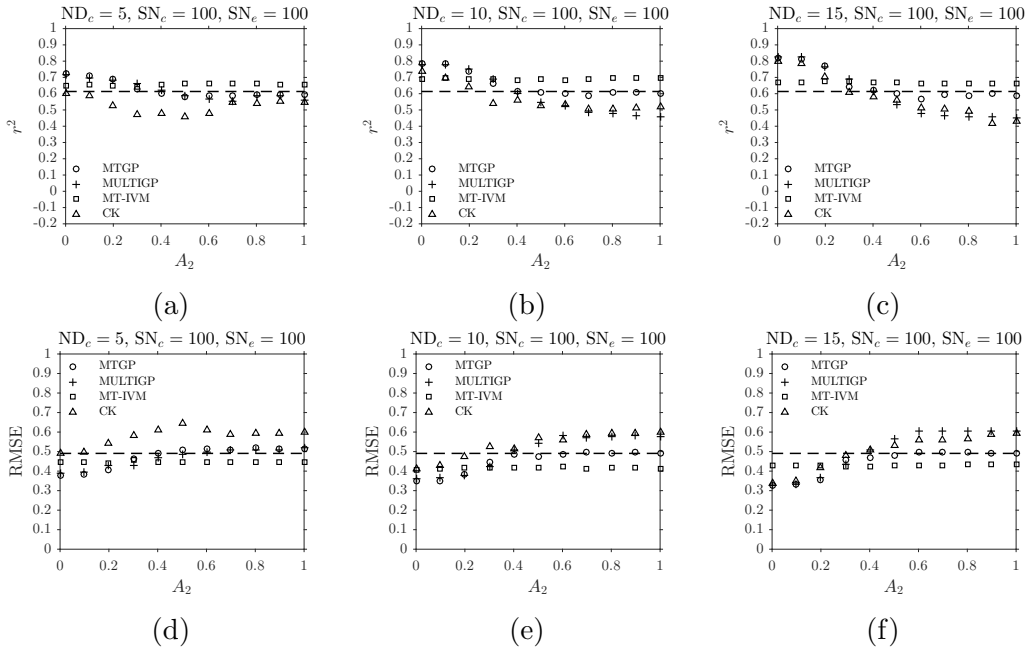


Figure 59: Paciorek function prediction r^2 and RMSE for multiple values of ND$_c$, with SN$_c$ = 100 and SN$_e$ = 100.

functions in Fig. 47. An unexpected result is that the performance of MT-IVM was virtually unaffected by the correlation between target functions or the addition of cheap data. Another unanticipated result is that MULTIGP was more sensitive to reduced correlation between the target functions for the Paciorek function than for the Branin function. The critical value of A_2 is near 0.6, which corresponds with r^2 between the target functions of approximately 0.06.

The behavior of generalization performance for high SN values is shown in Figs. 60 and 61. As with the Branin function, the critical value of A_2 made a slight shift with higher SN, in this case toward 0.5. MT-IVM remained nearly insensitive to SN and A_2, whereas the performance of MULTIGP degraded rapidly with increasing values of A_2 for all of the SN settings.

The performance increase for all of the models at $A_2 = 0$ was comparable for all SN values. The case where the cheap and expensive were both sampled deterministically had some of the smallest performance changes, most noticeably for MTGP

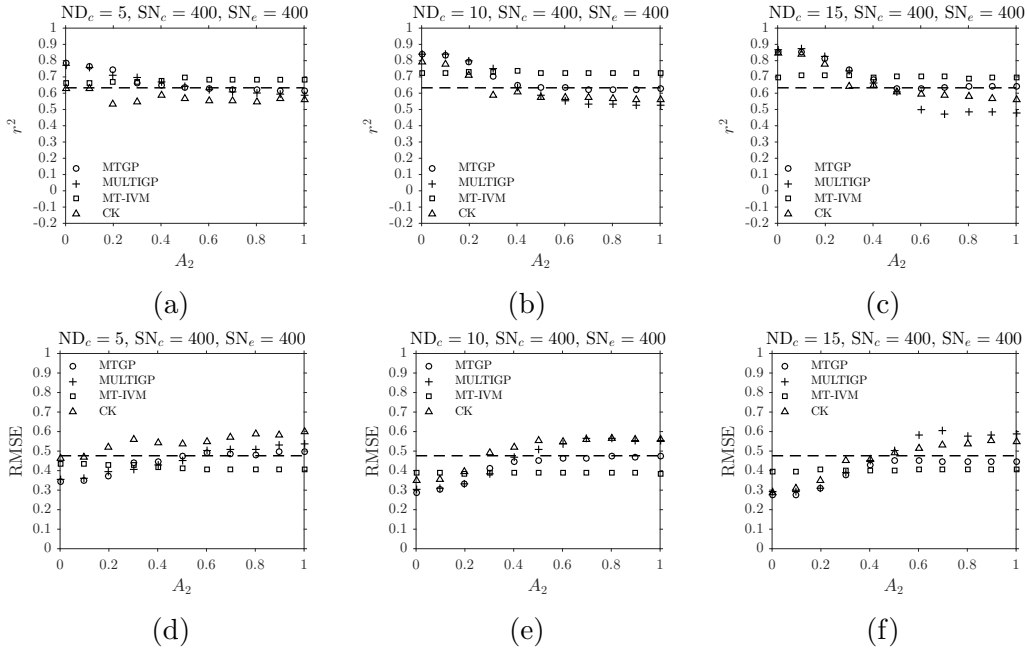


Figure 60: Paciorek function prediction r^2 and RMSE for multiple values of ND$_c$, with SN$_c$ = 400 and SN$_e$ = 400.

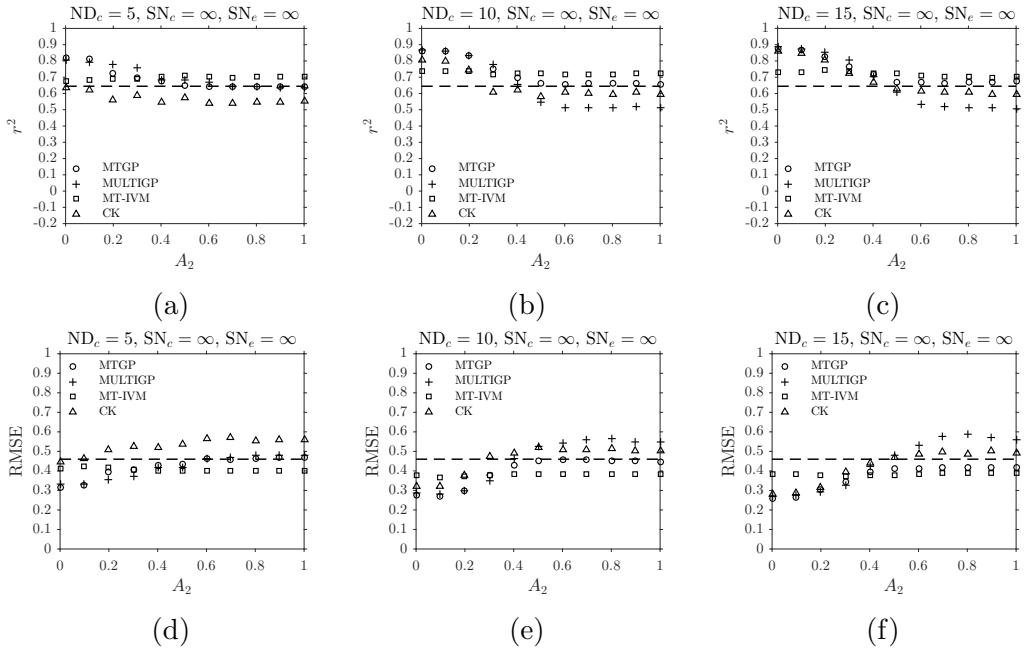


Figure 61: Paciorek function prediction r^2 and RMSE for multiple values of ND$_c$, with SN$_c$ = ∞ and SN$_e$ = ∞.

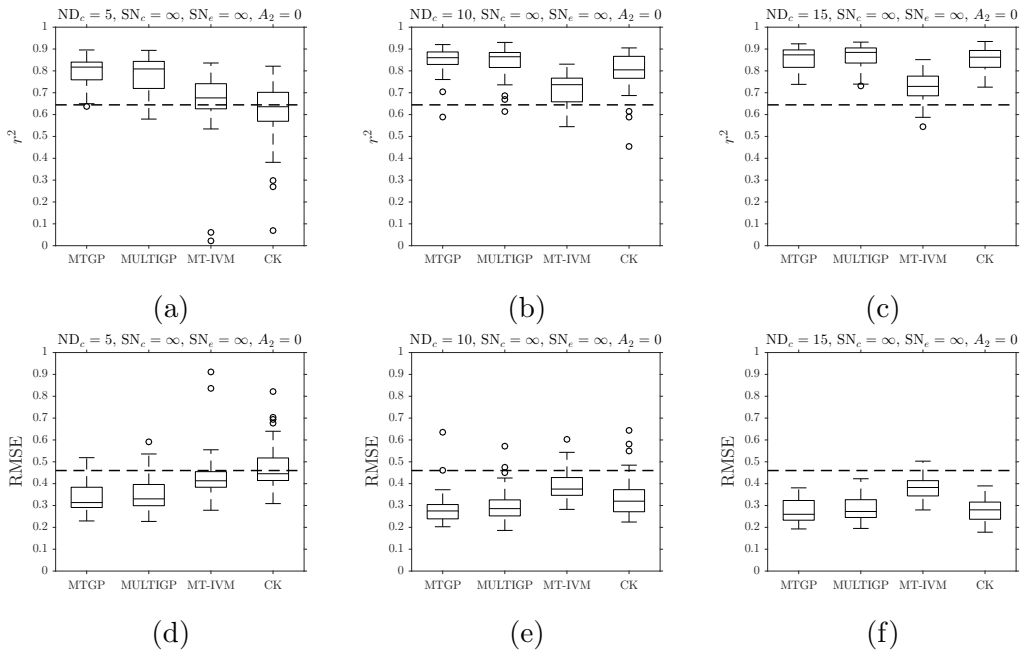


Figure 62: Box plots of Paciorek function prediction r^2 and RMSE for multiple values of ND$_c$, with SN$_c$ = 400, SN$_e$ = 400, and $A_2 = 0$.

and MULTIGP. To more thoroughly investigate the validity of hypothesis 1 for the Paciorek function, box plots are shown in Fig. 62. The r^2 boxes for MTGP did not consistently contract with increasing ND$_c$ as the boxes for the other models did, but the the median and the box shifted upward, and the whisker length also shortened. The same trend was exhibited for RMSE. All of the evidence suggests that hypothesis 1 is appropriate for the Paciorek function.

The evidence shown in Fig. 63 does not strongly support hypothesis 2. The shift in performance for all of the models was small, and the strong interaction effect between SN$_c$ and ND$_c$ observed for the Branin function was not evident for the Paciorek function. The box plots in Fig. 64 also show that the evidence does not strongly support or refute the hypothesis. The changes in performance for all of the models were small enough that they may have been due to variability in the experiment rather than an underlying effect.

The evidence presented thus far for the Paciorek function refutes hypothesis 3

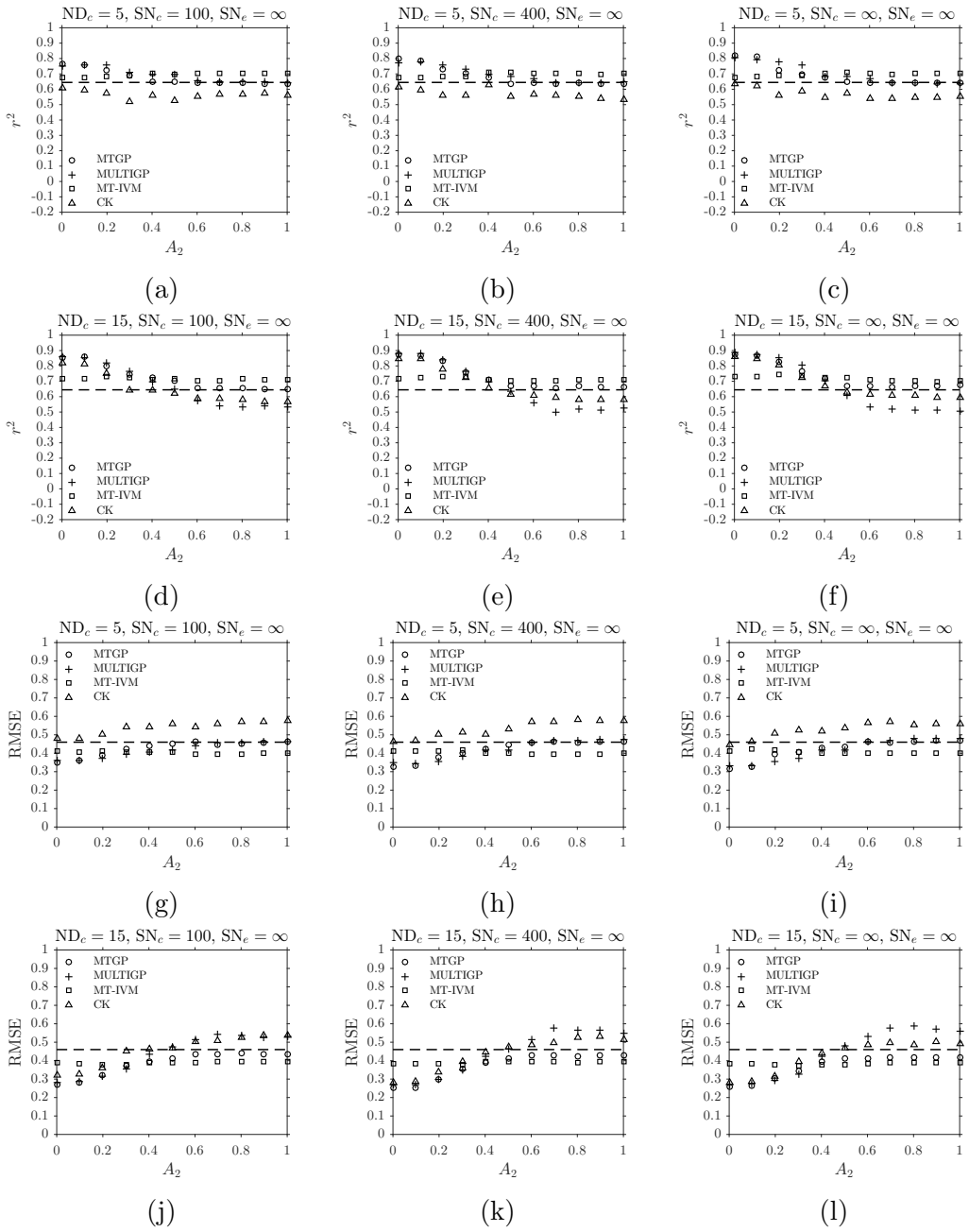


Figure 63: Paciorek function prediction r^2 and RMSE for multiple values of SN$_c$ and two levels of ND$_c$, with SN$_e$ = ∞.

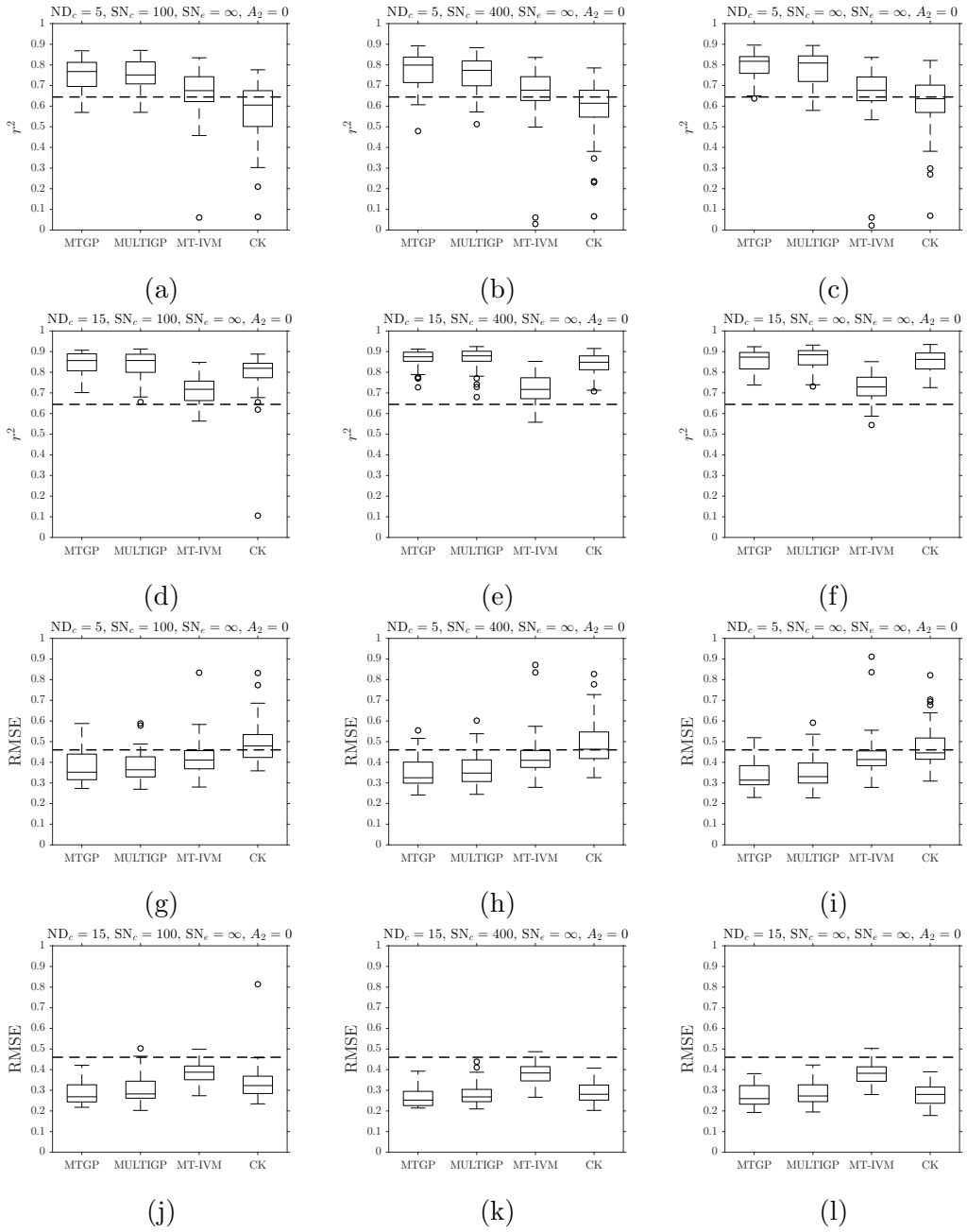
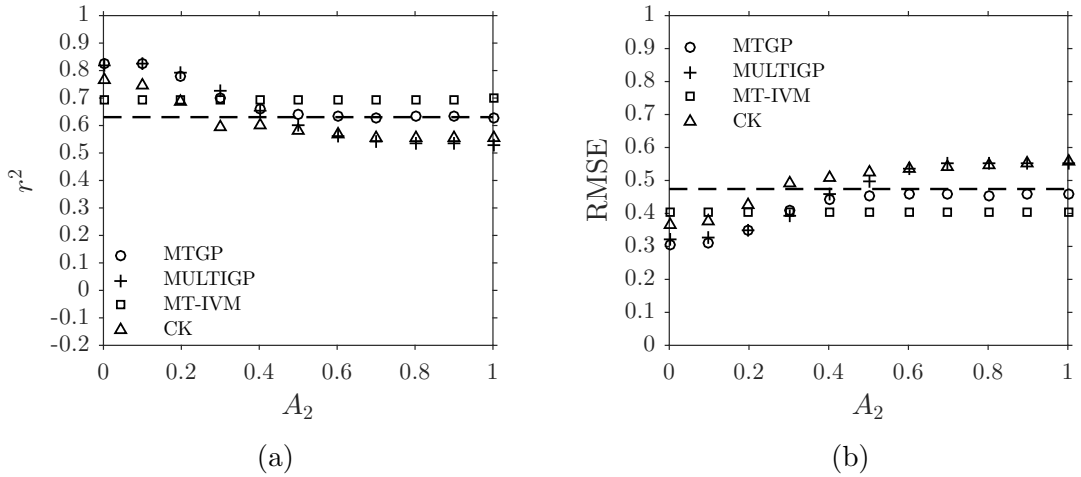


Figure 64: Box plots of Paciorek function prediction r^2 and RMSE for multiple values of SN$_c$ and two levels of ND$_c$, with SN$_e = \infty$, and $A_2 = 0$.



(a) (b)

Figure 65: Paciorek function prediction r^2 and RMSE for all results.

because MT-IVM clearly was the most robust to changing correlation between the target functions. The MTGP and CK models were the next best in terms of robustness, but MULTIGP was the worst. Additional evidence supporting these claims is shown in Figs. 65 and 66. As found in the other plots, the median performance for MT-IVM taken over all cases was not sensitive to A_2, and MULTIGP performance degraded rapidly as A_2 increased from 0 to 1. The derivative plots show this behavior more clearly. The derivatives for MT-IVM remained close to 0 for all values of A_2. MULTIGP r^2 derivatives were consistently larger than all other models in the A_2 range of [0.3,0.9], and the RMSE derivatives were consistently larger than all other models in the A_2 range of [0.3,0.7]. The derivatives plots confirmed that MTGP and CK were similar in terms of robustness, with MTGP performing slightly better at the majority of A_2 values.

5.3.5 Trid Function Results

The results shown in Fig. 67 are not conclusive regarding the validity of hypothesis 1 for the Trid function. The performance of all GP models appears to improve with increasing ND_c at the highest correlation point $A_3 = 0.5$, but this may have been due to experimental variability rather than an underlying effect. Similarities between the

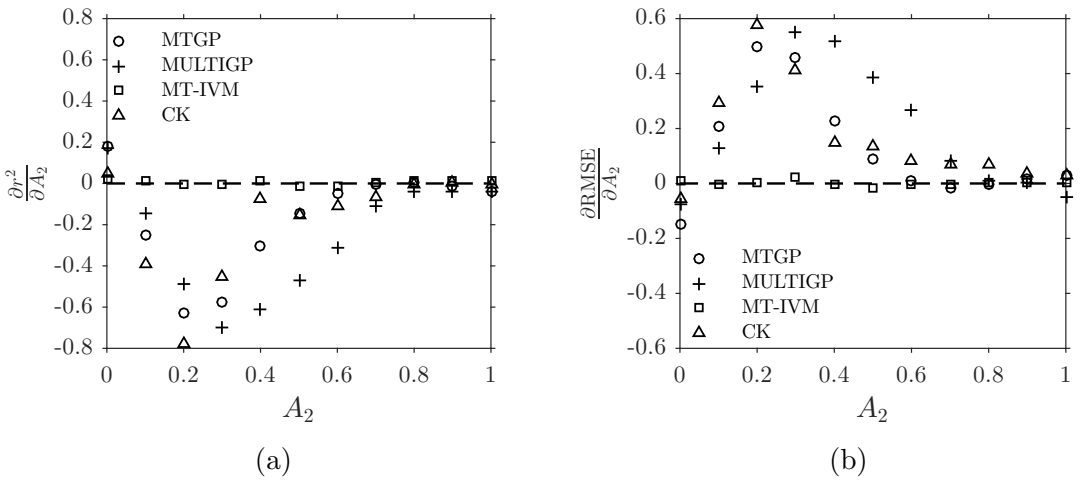


Figure 66: First partial derivatives of Paciorek function prediction r^2 and RMSE for all results.

behavior of the r^2 and RMSE performance and the correlation between the target functions in Fig. 49 were not as apparent as they were for the other two analytical functions. The critical value of A_3 was in the vicinity of 0.8, which corresponded with a correlation between the target functions near 0. Note that all RMSE results presented for the Trid function were been divided by 1E4.

The performance results in Fig. 68 support hypothesis 1. At A_3 values near 0.5, the performance of all models improved, which aligns with the hypothesis. But, there was not a single critical value of A_3 that corresponded with a correlation between target functions below which the performance of all models remained the same or decreased. The performance at $A_3 = 0.8$, which corresponded with the lowest correlation between target functions in the range of discrete A_3 values, remained similar for all models besides MT-IVM. Also, the performance of MTGP degraded with increasing cheap data at A_3 values of 0.9 and 1. Thus, the critical value of correlation differed for each model. Another interesting observation is that CK performance followed the correlation between the target functions as A_3 varied more closely than any of the other GP models. This was most likely because of the explicit relationship between the two data sources that is built into the autoregressive structure. Comparing the

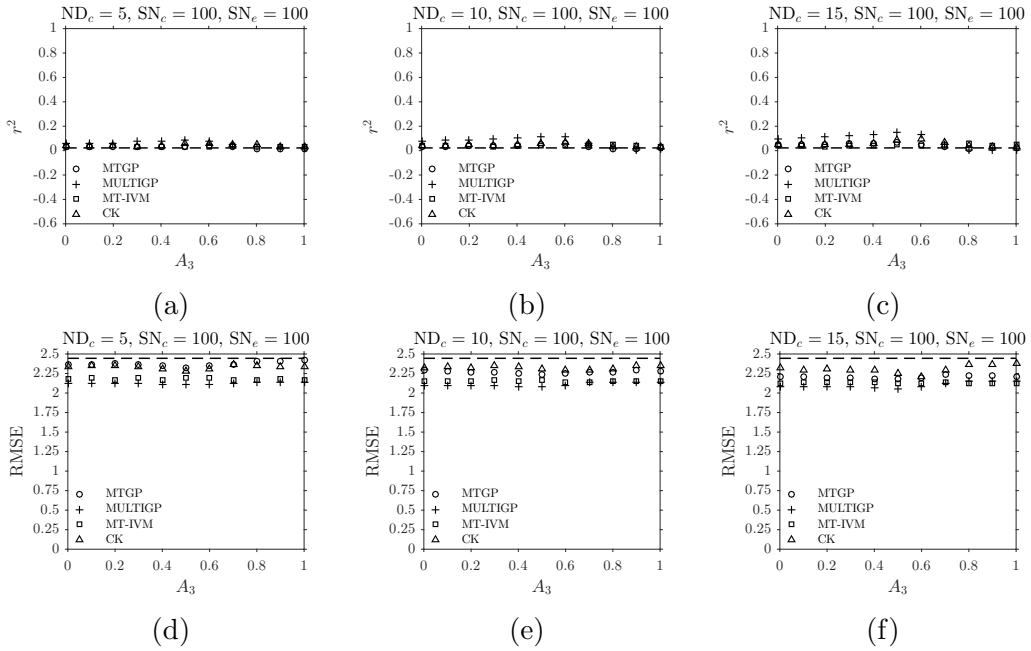


Figure 67: Trid function prediction r^2 and RMSE for multiple values of ND$_c$, with SN$_c$ = 100 and SN$_e$ = 100.

figures for SN values of 100 and ∞, one will also notice that the interaction effect between SN and ND$_c$ is more apparent for the Trid function than the other two analytical functions. This interaction effect resulted in larger changes in performance as ND$_c$ increased for the higher SN value.

At the highest correlation between target functions, where $A_3 = 0.5$, the case where SN$_c$ = 100 and SN$_e$ = 100 showed small improvements in r^2 and RMSE for all of the models, relative to the other noise settings. To more precisely probe the results regarding the validity of hypothesis 1, box plots are shown in Fig. 69. Although the improvements were small, the boxes and medians did shift toward higher r^2 and RMSE performance as ND$_c$ increased. This observations implies that hypothesis 1 is valid.

The evidence shown in Fig. 70 strongly supports hypothesis 2 for all models beside MT-IVM. This is because the performance of MT-IVM did not appear to be affected by increasing SN$_c$ for ND$_c$ = 5. To visualize the performance more clearly at the

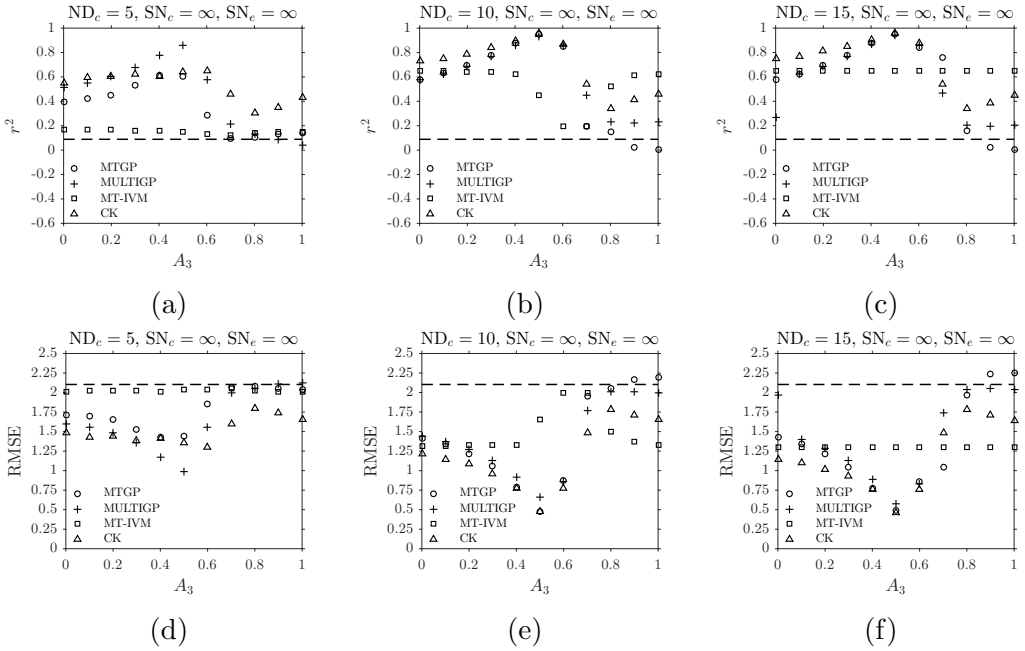


Figure 68: Trid function prediction r^2 and RMSE for multiple values of ND_c, with $\mathrm{SN}_c = \infty$ and $\mathrm{SN}_e = \infty$.

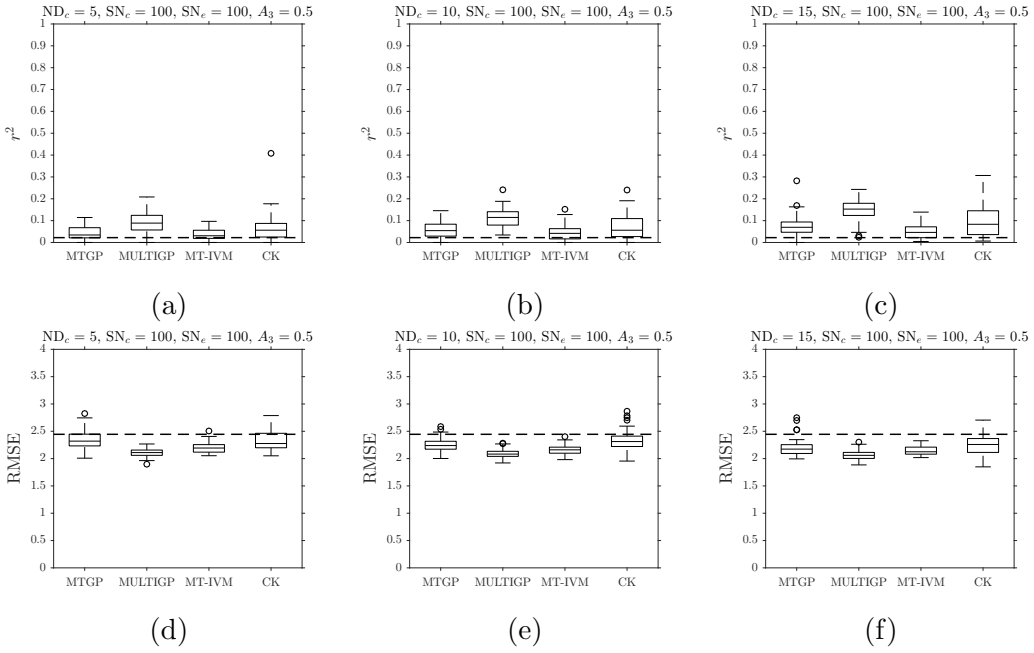


Figure 69: Box plots of Trid function prediction r^2 and RMSE for multiple values of ND_c, with $\mathrm{SN}_c = 100$, $\mathrm{SN}_e = 100$, and $A_3 = 0.5$.

highest correlation between the target functions, box plots are in Fig. 71. The change in MT-IVM r^2 performance was small, but the bottom whisker shifted upward as SN_c increased. However, MT-IVM RMSE performance was virtually the same across all levels of SN_c. Performance improvements for all three of the other models was obvious. For this reason, the validity of hypothesis 2 is inconclusive.

The evidence presented thus far for the Trid function refutes hypothesis 3 because MT-IVM is the most robust to changing correlation between the target functions. Additional evidence supporting this claim is shown in Figs. 72 and 73. The median performance of MT-IVM taken over all cases was nearly unchanged, and the derivatives plots confirmed this. Although CK has the best performance overall, it was the worst in terms of robustness to the correlation between the target functions. MTGP was the second in robustness next to MT-IVM but third in overall predictive performance.

5.3.6 Discussion and Conclusions

With the exception of some of the MT-IVM performance results, the evidence supports hypothesis 1. When the correlation between the target functions was above a critical value, the generalization performance of the multitask models CK, MULTIGP, MTGP, and (sometimes) MT-IVM improved with increasing ND_c. Below the critical value of correlation, there were instances where the generalization performance either did not change significantly or degraded as ND_c increased. The critical value is not necessarily identical for each multitask model, as evident in the Trid function results. The results for all of the analytical functions suggest that diminishing returns are reached once ND_c increases beyond a certain level. This was observed for the cases in which a noticeable improvement in predictive performance was achieved when ND_c was changed from 5 to 10 but minimal improvement was found for ND_c increasing from 10 to 15.

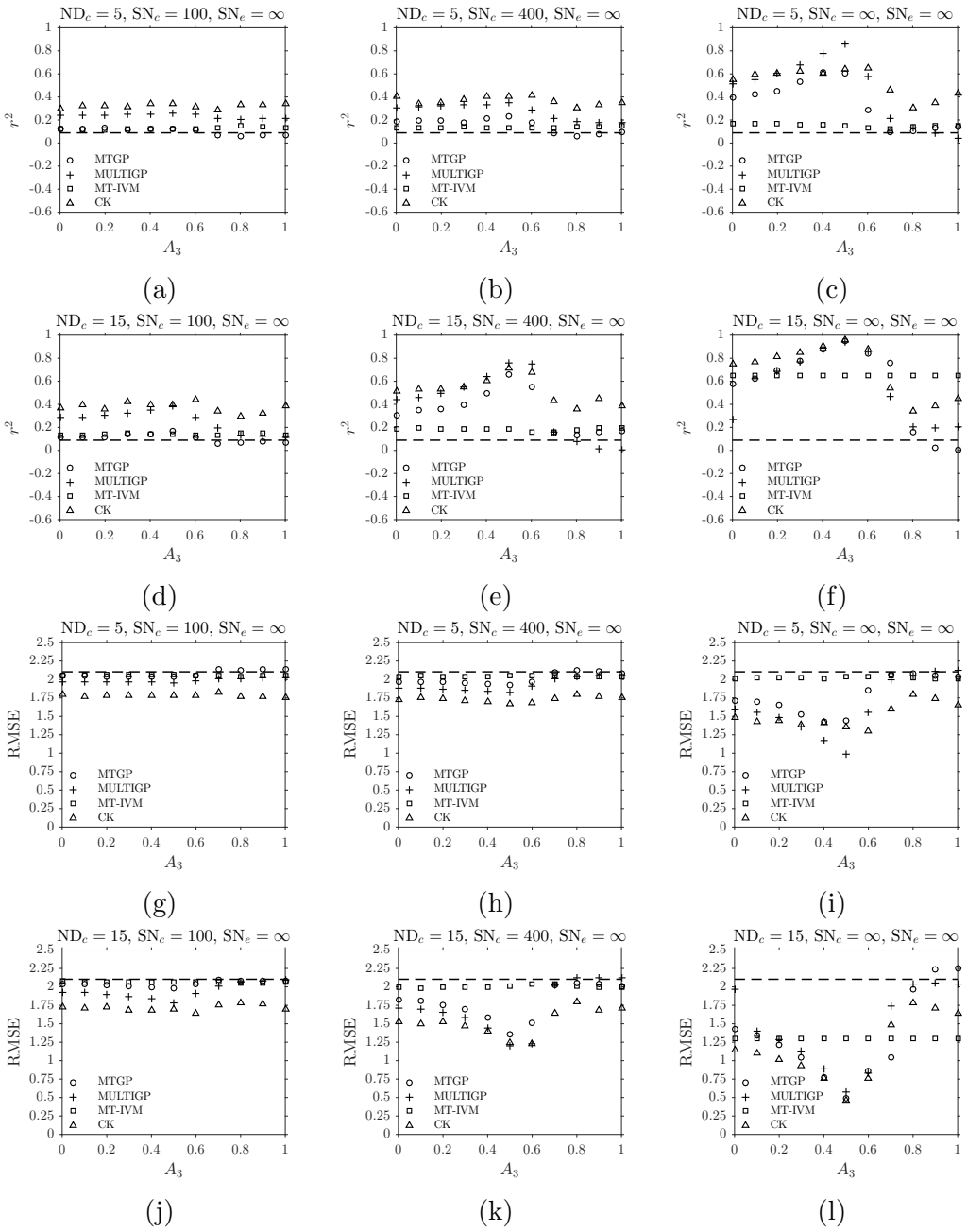


Figure 70: Trid function prediction r^2 and RMSE for multiple values of SN$_c$ and two levels of ND$_c$, with SN$_e$ = ∞.

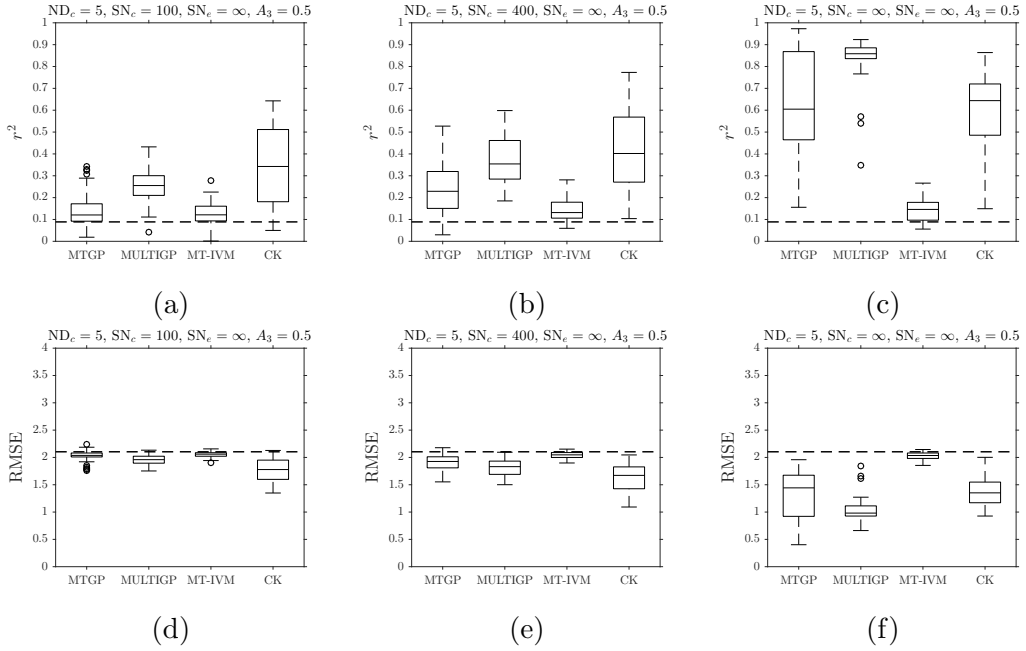


Figure 71: Box plots of Trid function prediction r^2 and RMSE for multiple values of SN_c and $ND_c = 5$, with $SN_e = \infty$, and $A_3 = 0.5$.

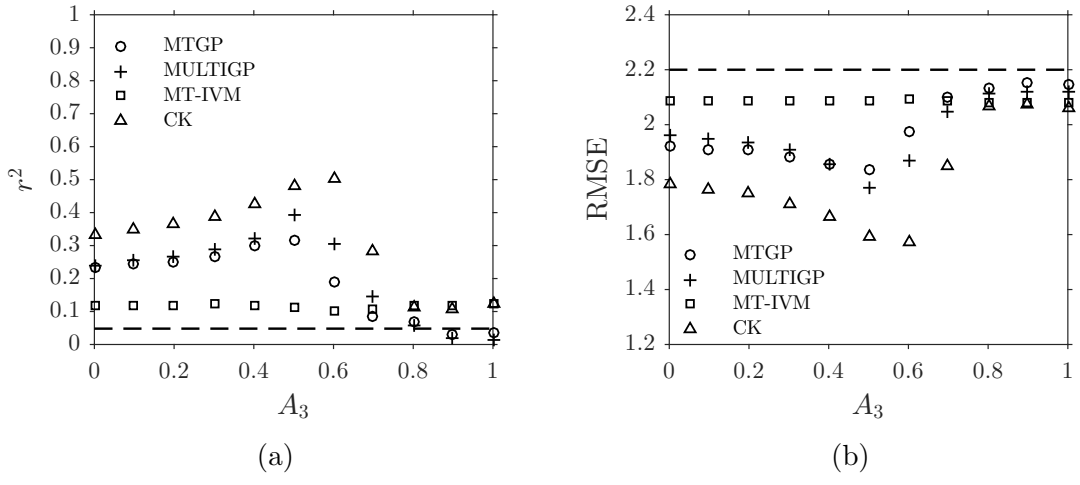
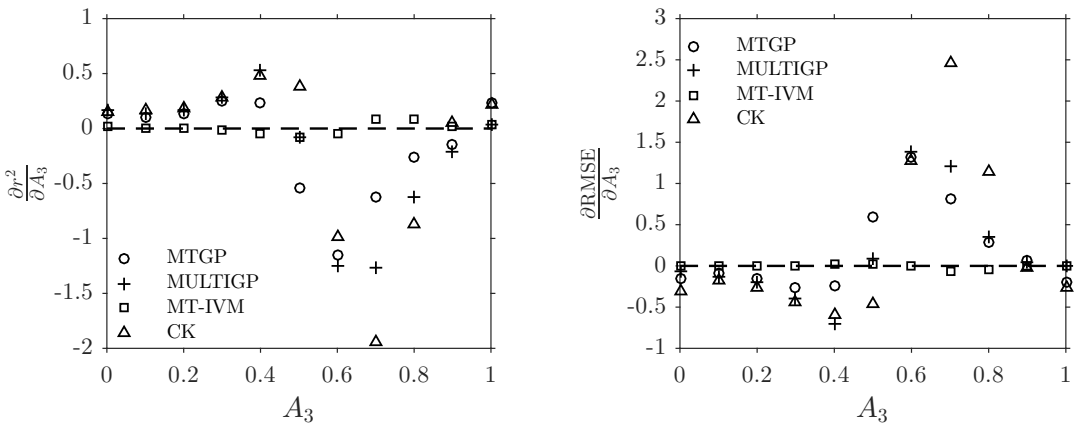


Figure 72: Trid function prediction r^2 and RMSE for all results.



(a) (b)

Figure 73: First partial derivatives of Trid function prediction r^2 and RMSE for all results.

The validity of hypothesis 2 is not clearly defined by the results presented here. Although increasing SN_c resulted in generalization performance improvements in many cases, there were instances where performance was not noticeably affected at the highest settings of correlation between the target functions.

Hypothesis 3 is refuted for all three of the analytical functions. Although MT-IVM was outperformed in many cases by at least one other multitask model, it was more robust to changes in the correlation between the target functions than MTGP was for all three analytical functions, and MT-IVM was more robust than MULTIGP for the Paciorek and Trid functions.

Due to the limited number of GP model architectures and data scenarios that were used in this experiment, an answer to RQ 2.1 for all regression problems is not possible. However, conclusions can be derived with a degree of uncertainty. Also, the conclusions for the three hypotheses imply how conditions can be changed to better ensure when a multitask GP regression model will outperform a single-task GP regression model.

When the correlation between the underlying cheap and expensive target functions is above the critical value, this experiment indicates that a multitask GP can

outperform a single-task GP when the number of cheap observations is equivalent to the number of expensive observations, despite the level of noise in the data. The validity of hypothesis 1 suggests that increasing the number of cheap data will increase the likelihood that a multitask GP will outperform a single-task GP. However, there are diminishing returns. The opposite effect is true as well; if the correlation between the target functions is below the critical value, then increasing the number of cheap observations can degrade generalization performance. Although hypothesis 2 cannot be definitively verified with the results from this experiment, it is possible that increasing the signal-to-noise ratio of the cheap data will also increase the likelihood that a multitask GP will outperform a single-task GP, if the correlation between the target functions is above the critical value. The opposite effect was observed as well; if the correlation between the target functions is below the critical value, then increasing the signal-to-noise ratio of the cheap observations can degrade generalization performance. Hence, it is best to use cheap data from a low-noise experiment, such as a computer experiment, as long as the correlation between the target functions is above the critical value. The interaction effect between the cheap data signal-to-noise ratio and sample size that was observed for the Branin function and Trid function suggests that the improvements in generalization performance gained by increasing these parameters can be much greater if both are increased simultaneously.

The sample size of the expensive data was not varied in this experiment. For the five points-per-dimension used in this experiment, there were cases in which the r^2 and RMSE performance measures were significantly improved. As the number of observations increases, there will be a point at which a multitask model will result in diminishing returns in terms of generalization performance. When the generalization performance of a multitask GP is close to but not worse than a single-task GP, it should only be used if it results in justifiably less prediction uncertainty.

In the case that the correlation between the target functions is low, the complexity of the functions may have an effect on the success of a multitask GP. This was suggested by Toal [101], and the results presented here further support this observation. The Branin and Paciorek functions are both multimodal with multiple local minima, whereas the Trid function is convex. This may be part of the reason that the multitask GPs outperformed the single-task GPs for the Trid function in many cases when the correlation between the target functions was close to zero. Also, the robustness of a given multitask GP for a given problem determines when it should be used under the condition of low correlation between the target functions. These observations imply that the critical value of correlation between the target functions can vary depending on the problem complexity and the multitask architecture used. Toal [101] suggested a critical value of $r^2 = 0.9$, but this may be conservative for relatively simple functions such as the Trid function.

The observations that the behavior of the generalization performance measures as the A parameters were varied is similar to the correlation between the target functions suggests that correlation is a valid measure of task relatedness, or degree of homogeneity, for regression. However, in practice the true correlation between the target functions will generally not be known. The correlation can be estimated for two functions, but the final decision regarding whether to use a multitask GP or not should be informed by an appropriate model selection process.

5.4 Illustrative Example: AFC Technology Experiments

An analysis of notional AFC technology experiments was conducted to demonstrate the proposed methodology. The setup of the example problem is described first. Then, the implementation of the proposed methodology is presented. Implementation of the first three steps of the methodology are not described in detail here because the focus of this example is on the primary contributions in steps four and five.

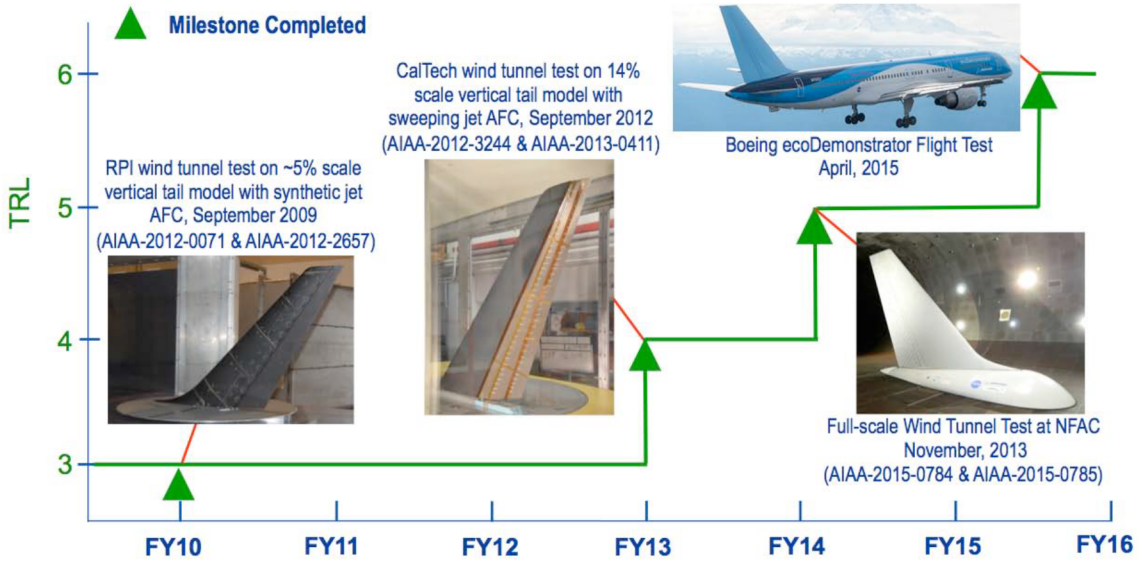


Figure 74: Summary of AFC-enhanced vertical tail technology development activities (from Ref. [40]).

5.4.1 Problem Setup

The technology that motivated this example is the AFC-enhanced vertical tail technology that was described in Sec. 1.2. This technology was part of a development program operated by NASA and Boeing, and the experiments that were conducted are shown with corresponding TRLs in Fig. 74. For the illustrative problem, it was assumed that only data existed from the sub-scale wind tunnel experiment that is associated with TRL 4. The objectives were (1) to characterize the uncertainty surrounding this experiment, (2) to estimate the uncertainty reduction from a proposed full-scale wind tunnel experiment corresponding with TRL 5 and a proposed full-scale flight experiment corresponding with TRL 6, and (3) to compare the estimated uncertainty reduction values with "truth" values.

Because of the proprietary nature of the data for the real experiments, few data have been published in the open literature. Thus, synthetic data were created for all three of the notional experiments. To ground the example in the real physics of the technology, "truth" functions were extracted from the published results shown

in Fig. 37. The sub-scale wind tunnel truth function is the "30° rudder" line in the figure, and the truth function for the two proposed experiments is the "20° rudder" line in the figure. These choices of truth functions were motivated by the fact that the real technology showed reduced AFC effectiveness as the scale and fidelity of the experiments increased over time. The actual behavior of AFC effectiveness in the full-scale wind tunnel and flight experiments were not identical, but they were assumed to have the same truth function in this example. Notional sub-scale wind tunnel observations were generated by adding Gaussian noise to the truth function at 11 evenly-spaced points in the β interval $[-7.5°, 7.5°]$. The Gaussian noise distribution had an SN value of 5,000. The synthetic data are shown as circle markers in Fig. 75. All notional full-scale wind tunnel data were generated at 7 evenly-spaced points in the β interval $[-7.5°, 7.5°]$. To simulate the real constraints of flying the technology on an aircraft, the notional full-scale flight experiment data were generated at 7 evenly-spaced points in the β interval $[0°, 15°]$.

5.4.2 Implementation of the Proposed Methodology

The first step of the methodology regarding cleaning of the data was not necessary since all data were synthetic. The Gaussian Processes for Machine Learning (GPML) toolbox v3.6 [116] was used to build a regression model for the sub-scale wind tunnel data. The squared exponential covariance function was used, and a linear basis function for the mean was selected because of the global linear trend of the observations. Predictions from the GP model are shown in Fig. 75. The prediction mean approximately followed the trend of the true target function over the β interval $[-8°, 8°]$, but it diverged from the truth function beyond 8°. The additional epistemic uncertainty due to immaturity of the AFC technology was modeled using $\sigma_\tau^2 = 0.1$, $\nu = 1$ and $\nu = 2$, and TRLC = 1.14 from Table 10. The 95% prediction intervals for the sub-scale wind tunnel GP regression model with the additional maturity uncertainty are

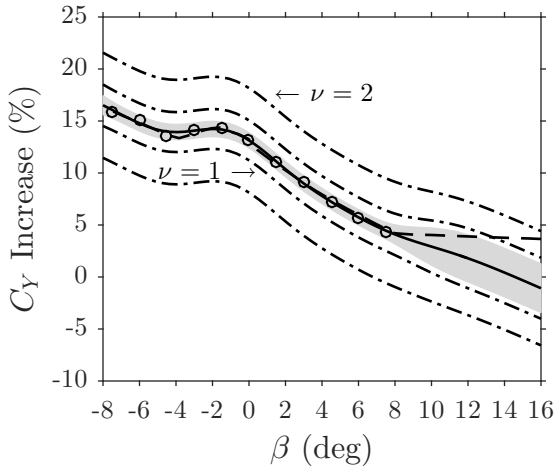


Figure 75: Notional sub-scale wind tunnel experiment data (◯ symbols), underlying true target function (dashed line), mean GP prediction (solid line), GP 95% prediction intervals (gray area), and GP 95% prediction intervals inflated with technology maturity uncertainty (dash-dotted lines).

plotted as dash-dotted lines in Fig. 75.

The next implementation step was to estimate the uncertainty reduction from the proposed experiments. The points of interest were selected to coincide with the critical β range in Fig. 37: 20 evenly-spaced points in the β interval $[-7.5°, 0°]$. Then, observations from the proposed experiments were simulated by drawing $N_{\text{sim}} = 1{,}000$ random functions from the sub-scale wind tunnel GP model. Two of the random function realizations are plotted as dotted lines in Fig. 76 for the $\nu = 2$ scenario. These random functions were drawn from the GP model that included the covariance matrix that models the epistemic uncertainty due to immaturity of the technology. If the standard GP model had been used, the random function realizations would have been distributed more closely to the gray area shown in Fig. 75. For each of the 1,000 random functions, observations were simulated at the design points for both of the proposed experiments. Observations were generated by sampling from normal distributions with means at the true target function location and variances equal to the noise variance estimated by the single-task sub-scale wind tunnel GP model. The noise variance was estimated by the GP training process to be 0.09. Examples of the

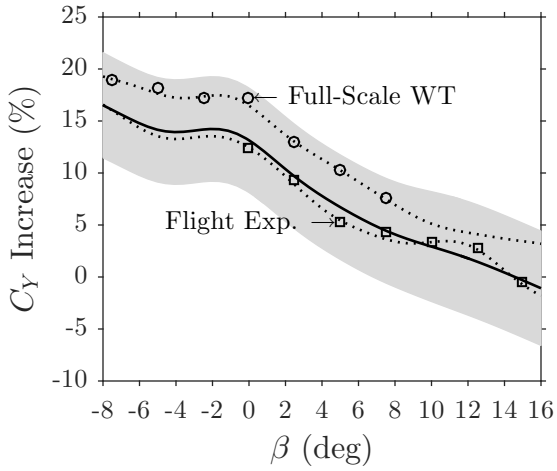


Figure 76: Simulating observations for the proposed experiments with the sub-scale wind tunnel GP model for $\nu = 2$. The solid line and gray area show the GP mean predictions and 95% prediction intervals, respectively. Two random function draws from the GP are shown (dotted lines), and simulated observations from the flight experiment (◯ symbols) and full-scale wind tunnel experiment (□ symbols) are shown.

generated observations for the proposed experiments are shown as square and circle markers in Fig. 76. It is clearly seen in the figure that, with the exception of $\beta = 0°$, the flight experiment observations were outside of the region of interest, whereas the full-scale wind tunnel experiment included four observations in the region of interest.

The next steps for estimating uncertainty reduction from the proposed experiments were to train regression models with the simulated observations and to estimate posterior entropy at the points of interest. For comparison, MTGP and the MATLAB built-in GP capability were both used to perform regression of the simulated data. The options used for both models were the same as in the GP comparison experiment, except a constant basis function was implemented for the single-task GP. After each regression model was trained, the sum of entropies of the marginal distributions was estimated at the points of interest using Eq. (35). Entropy estimation was conducted both with the addition of the $\text{cov}(\boldsymbol{\tau})_{dd}$ term and without it. The TRLC values in Table 10 corresponding with TRL 5 and TRL 6 were employed for modeling maturity uncertainty surrounding the full-scale wind tunnel and flight experiments,

respectively.

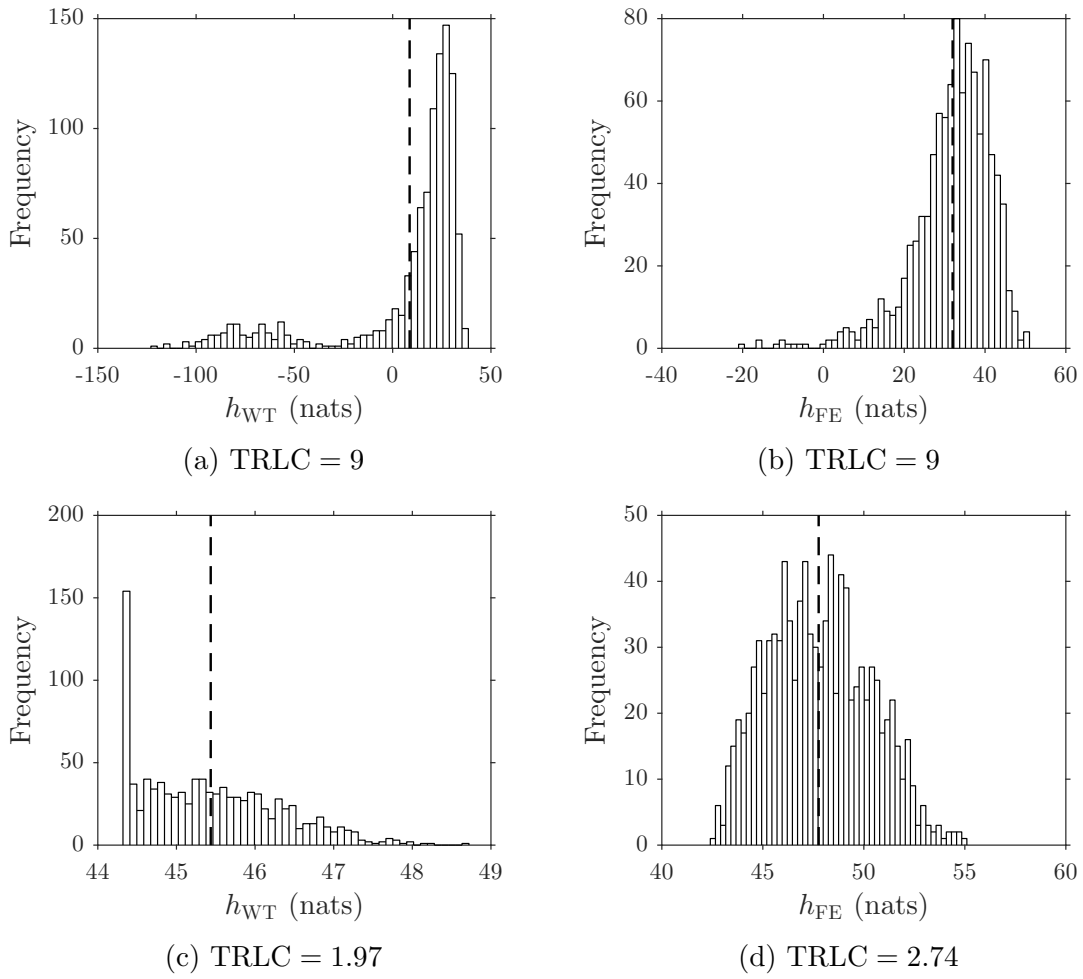
Since the notional true target functions were known in this example, the "true" posterior uncertainties were also computed to compare with the estimates. This was accomplished by simulating observations from the notional experiments after execution, training GP models with the data, and calculating posterior entropy at the points of interest. The observations from the wind tunnel and flight experiments were generated by sampling from normal distributions centered at the truth function with SN values of 2,500 and 1,000, respectively. The SN was decreased as the TRL of the experiments progressed to model more noise in the data due to a decreasing degree of control in the experiments. Also, in the real flight experiment, it was not possible to measure the vertical tail side force directly; side force was calculated using the flight experiment data as an input to proprietary models. Single-task and multitask GP models were trained with 1,000 realizations of observations from the two experiments. Posterior entropy was calculated at the points of interest, and predictive performance of the models was quantified at the points of interest with RMSE and r^2 between the mean predictions and the underlying truth function.

5.4.3 Results

First, results and observations are presented for the $\nu = 2$ scenario. Then, results are presented for the $\nu = 1$ scenario to demonstrate the effect of the maturity uncertainty growth rate parameter on posterior entropy. The truth entropies are presented, and the predictive performance of the single-task GP model is compared with MTGP. Finally, the evolution of uncertainty with maturation is demonstrated for this example problem.

5.4.3.1 Scenario 1: $\nu = 2$

The posterior entropy results from the single-task GP predictions for the two proposed experiments are plotted as histograms in Fig. 77. For brevity, h_{WT} and h_{FE} denote



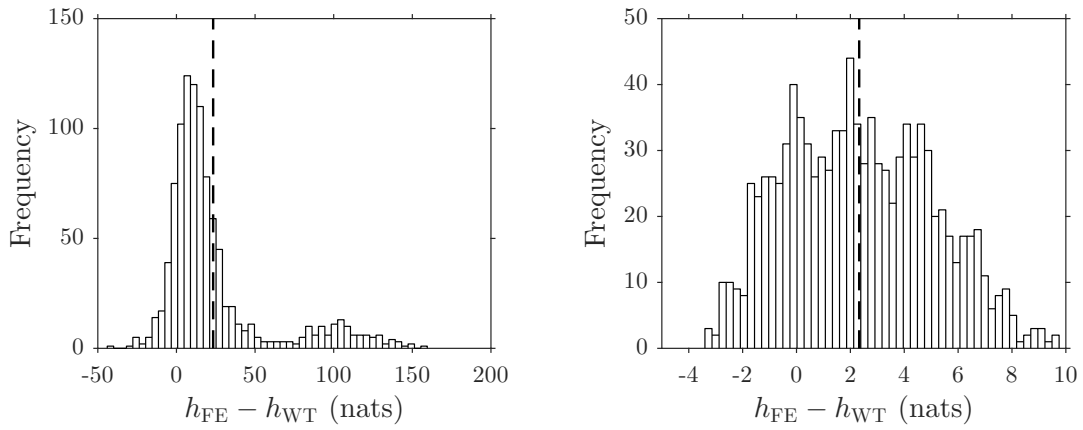
(a) TRLC = 9 (b) TRLC = 9

(c) TRLC = 1.97 (d) TRLC = 2.74

Figure 77: Posterior entropies from the single-task GP predictions with and without maturity uncertainty and $\nu = 2$. The sample mean is plotted as a vertical dashed line.

posterior entropy for the full-scale wind tunnel experiment and the flight experiment, respectively. The vertical dashed lines are plotted at the location of the mean entropy. Entropy without maturity uncertainty is shown in the top two plots (TRLC = 9), whereas entropy with maturity uncertainty included is shown in the bottom two plots. Comparing the two scenarios, one will immediately notice that the inclusion of maturity uncertainty resulted in larger entropies.

Histograms of the differences of the posterior entropies for the two proposed experiments from the single-task GP predictions are plotted in Fig. 78. Any positive

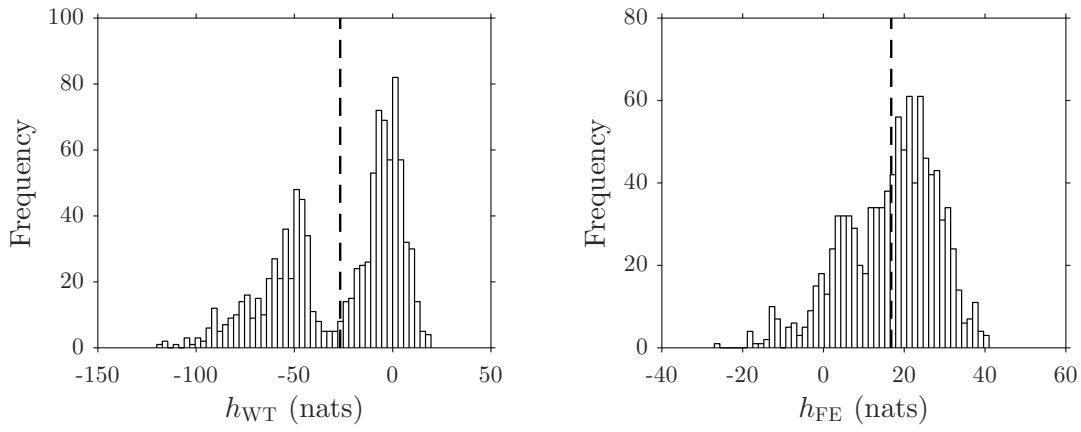


(a) TRLC = 9 (b) TRLC ≠ 9

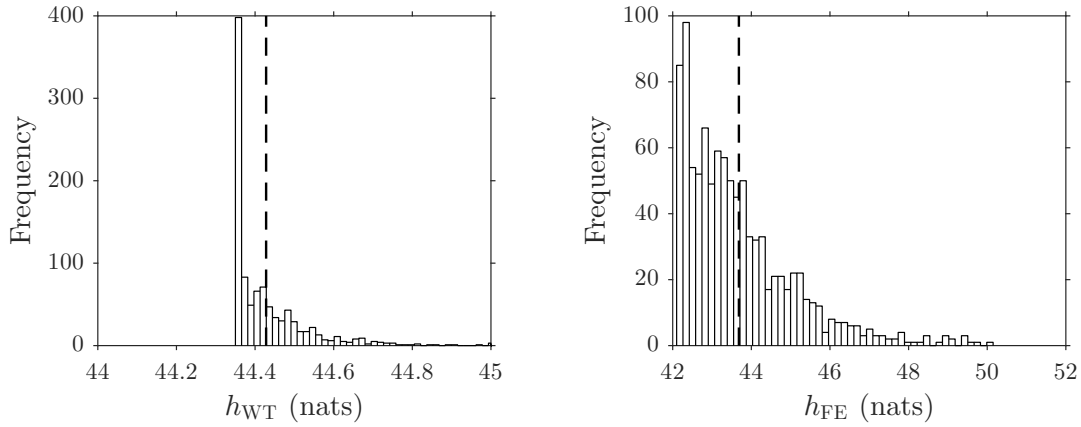
Figure 78: Differences of the posterior entropies for the two experiments from the single-task GP predictions with and without maturity uncertainty and $\nu = 2$. The sample mean is plotted as a vertical dashed line.

values in these plots indicate larger posterior uncertainty from the flight experiment and vice versa. The expected values in both scenarios were positive, so the conclusion for this example was that the full-scale wind tunnel experiment would result in more uncertainty reduction than the flight experiment. This is not surprising considering that the GP model trained with the simulated flight experiment data had to extrapolate to make predictions in the region of interest. Comparing the cases with and without maturity uncertainty, the effect of including maturity uncertainty is apparent; the difference in expected entropy between the two experiments was smaller when maturity uncertainty was included.

The conclusion that the wind tunnel experiment would result in more uncertainty reduction than the flight experiment was derived from an entropy estimation process in which no knowledge transfer had been captured in the regression predictions. The posterior entropies from the multitask GP predictions for the two proposed experiments are plotted as histograms in Fig. 79. Comparing the expected values in Figs. 77 and 79, one will notice that the multitask cases were lower. This is evidence that the multitask GP predicted lower uncertainty than the single-task GP.

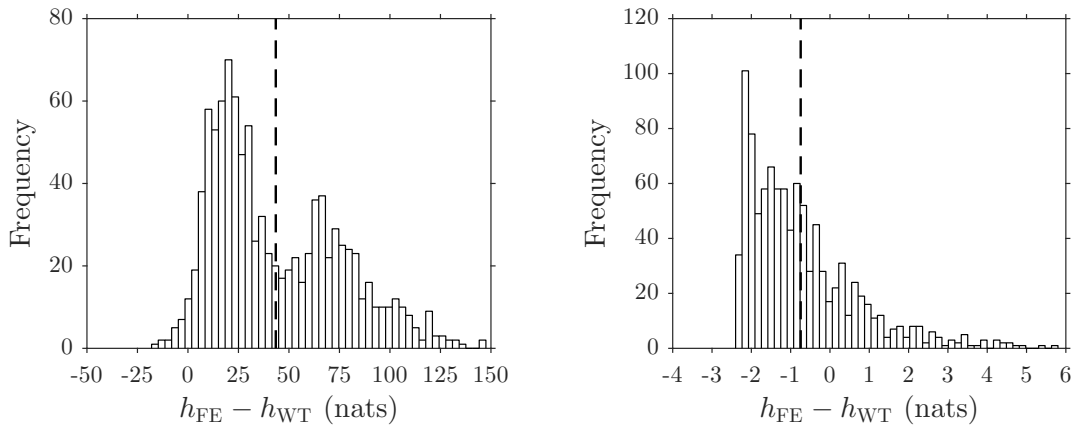


(a) TRLC = 9 (b) TRLC = 9



(c) TRLC = 1.97 (d) TRLC = 2.74

Figure 79: Posterior entropies from the multitask GP predictions with and without maturity uncertainty and $\nu = 2$. The sample mean is plotted as a vertical dashed line.

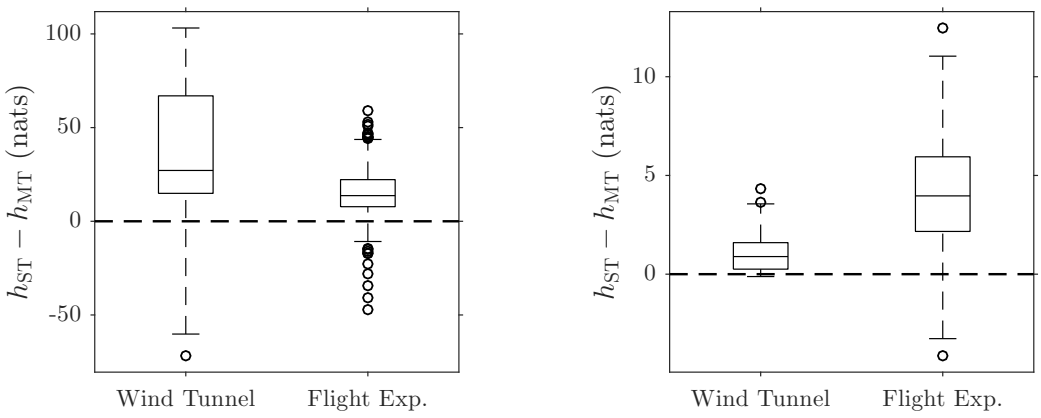


(a) TRLC = 9 (b) TRLC ≠ 9

Figure 80: Differences of the posterior entropies for the two experiments from the multitask GP predictions with and without maturity uncertainty and $\nu = 2$. The sample mean is plotted as a vertical dashed line.

The differences of the posterior entropies for the multitask case are shown in Fig. 80. Although the individual entropies were lower for the multitask model, the expected value of the differences for the case without maturity uncertainty was higher than it was in the single-task results. However, when maturity uncertainty was accounted for, the expected value was less than zero, which indicated that the flight experiment would reduce uncertainty more than the wind tunnel experiment. This is evidence that transfer learning combined with modeling maturity uncertainty can result in nontrivial predictions. Despite the extrapolation uncertainty surrounding predictions at the points of interest, the lower maturity uncertainty of the flight experiment combined with knowledge transfer from the sub-scale wind tunnel data resulted in higher predicted uncertainty reduction.

To present clearer evidence of the differences between the single-task GP model and MTGP results, box plots are shown in Fig. 81. In the figure, the subscripts ST and MT are abbreviations for "single-task" and "multitask", respectively. Any values above zero corresponded with the multitask model predictions having less uncertainty than the single-task predictions. The evidence suggests that in the majority of the



(a) TRLC = 9 (b) TRLC ≠ 9

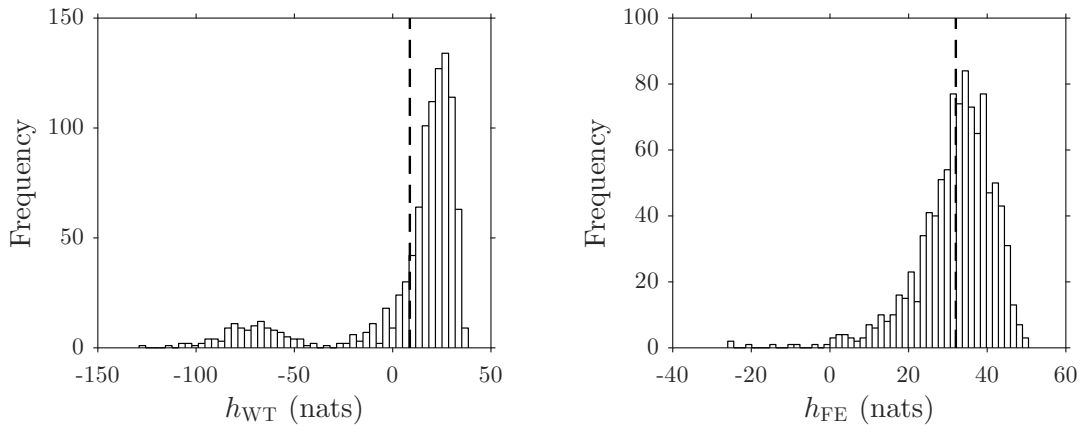
Figure 81: Differences of the posterior entropies for the two experiments from the multitask GP and single-task GP predictions with and without maturity uncertainty and $\nu = 2$.

cases, the MTGP prediction had less uncertainty at the points of interest than the single-task GP.

5.4.3.2 Scenario 2: $\nu = 1$

As expected, the entropies for $\nu = 1$ plotted in Fig. 82 for the cases with TRLC $\neq 9$ were lower than the $\nu = 2$ case in Fig. 77. The histograms for the cases without maturity uncertainty are similar. This demonstrates the effect of lowering the maturity uncertainty growth rate with a fixed characteristic variance σ_τ^2. The differences of the posterior entropies are plotted in Fig. 83. A larger margin between the entropies is shown for the $\nu = 1$ scenario, which is not surprising because of the smaller shrinkage in maturity uncertainty from TRL 5 to TRL 6 relative to the $\nu = 2$ scenario. Once again, the results for the case without maturity uncertainty are similar for both scenarios.

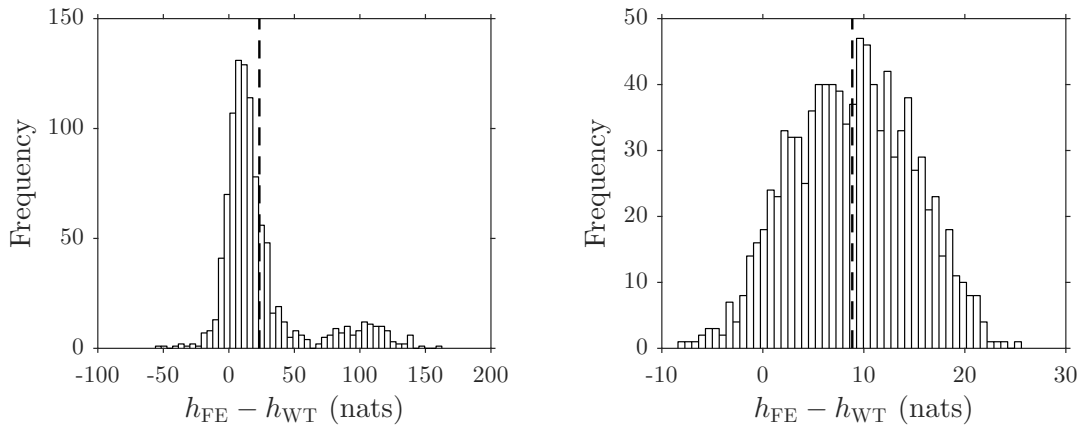
The individual posterior entropy histograms for the multitask model are in Fig. 84. As with the single-task GP case, the multitask entropies were lower for the $\nu = 1$ scenario with maturity uncertainty accounted for. The only noticeable difference for the case without maturity uncertainty was the shift of the sample mean to a



(a) TRLC = 9 (b) TRLC = 9

(c) TRLC = 1.97 (d) TRLC = 2.74

Figure 82: Posterior entropies from the single-task GP predictions with and without maturity uncertainty and $\nu = 1$. The sample mean is plotted as a vertical dashed line.



(a) TRLC = 9 (b) TRLC ≠ 9

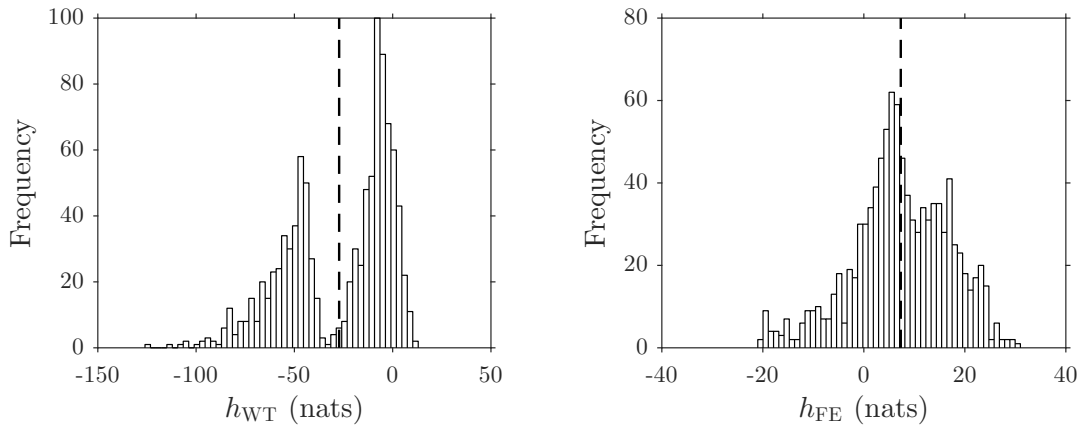
Figure 83: Differences of the posterior entropies for the two experiments from the single-task GP predictions with and without maturity uncertainty and $\nu = 1$. The sample mean is plotted as a vertical dashed line.

lower value for the flight experiment. However, the differences histogram in Fig. 85 indicates that the posterior entropy of the flight experiment was higher than the full-scale wind tunnel experiment. Hence, by changing the maturity uncertainty growth rate parameter from two to one, the conclusion was the opposite for the multitask case.

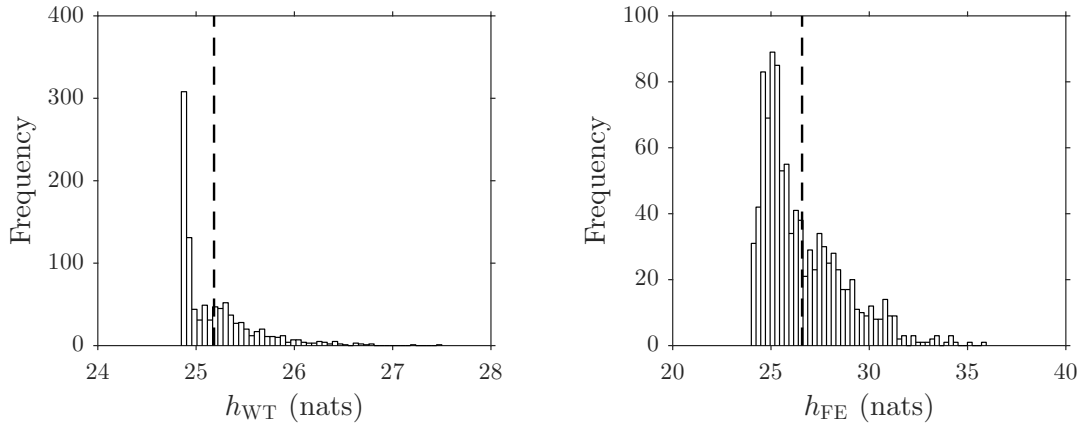
As in scenario 1, the box plots shown in Fig. 86 indicate that the majority of multitask GP predictions had less uncertainty than the single-task GP.

5.4.3.3 Truth Results

The true differences in posterior entropy between the two proposed experiments for the single-task GP model are shown in Fig. 87. These results indicate that the simulated flight experiment reduced uncertainty more than the wind tunnel experiment in all three cases. None of the predictions made using the single-task GP model predicted that this would be the result. The causality behind the higher entropy in predictions for the wind tunnel experiment was that the single-task GP model attributed the variability in the observations to noise rather than signal and thus the uncertainty surrounding predictions for the points of interest was relatively large.

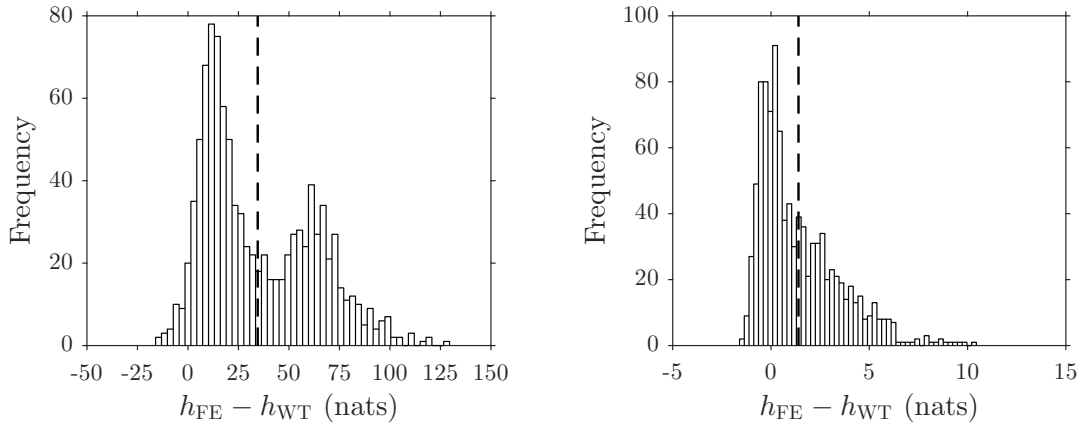


(a) TRLC = 9 (b) TRLC = 9



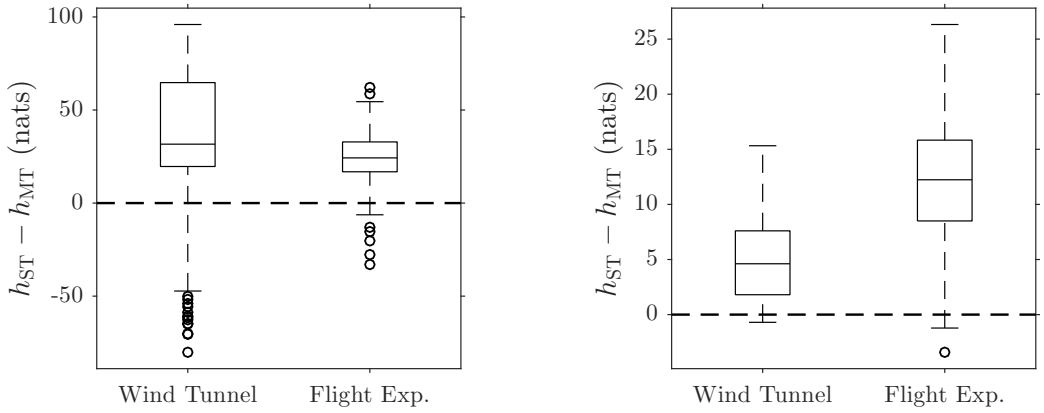
(c) TRLC = 1.97 (d) TRLC = 2.74

Figure 84: Posterior entropies from the multitask GP predictions with and without maturity uncertainty and $\nu = 1$. The sample mean is plotted as a vertical dashed line.



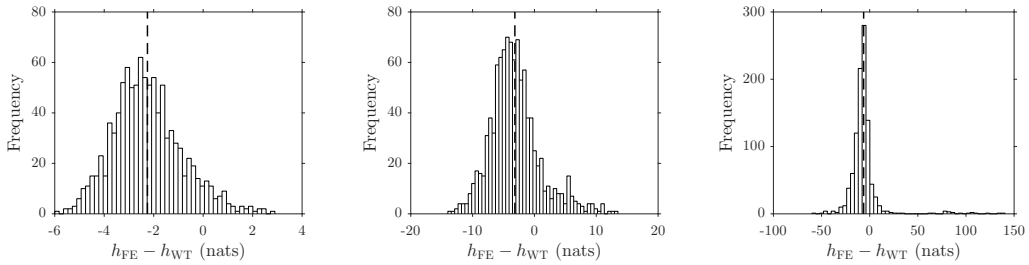
(a) TRLC = 9 (b) TRLC ≠ 9

Figure 85: Differences of the posterior entropies for the two experiments from the multitask GP predictions with and without maturity uncertainty and $\nu = 1$. The sample mean is plotted as a vertical dashed line.



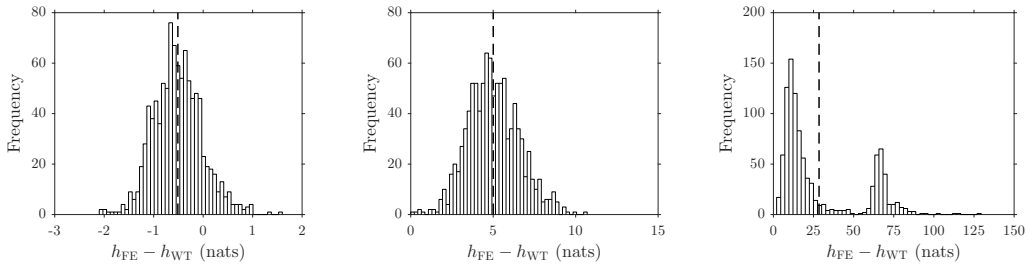
(a) TRLC = 9 (b) TRLC ≠ 9

Figure 86: Differences of the posterior entropies for the two experiments from the multitask GP and single-task GP predictions with and without maturity uncertainty and $\nu = 1$.



(a) $\nu = 2$ (b) $\nu = 1$ (c) TRLC = 9

Figure 87: True differences of the posterior entropies for the two experiments from the single-task GP predictions with and without maturity uncertainty. The sample mean is plotted as a vertical dashed line.



(a) $\nu = 2$ (b) $\nu = 1$ (c) TRLC = 9

Figure 88: True differences of the posterior entropies for the two experiments from the multitask GP predictions with and without maturity uncertainty. The sample mean is plotted as a vertical dashed line.

The true differences in posterior entropy between the proposed experiment for the multitask GP model are plotted in Fig. 88. These results were consistent with the predictions made in both scenarios; the scenario where $\nu = 2$ was the only one in which the flight experiment had less uncertainty than the wind tunnel experiment. Hence, the transfer of knowledge with the multitask GP models resulted in predictions that led to the correct conclusions about the relative uncertainty reduction between the two experiments.

The differences in accuracy of predictions at the points of interest of the two GP models for the simulated truth data are shown in Fig. 89. In both box plots, any results that are above the zero line correspond with a case where the multitask model

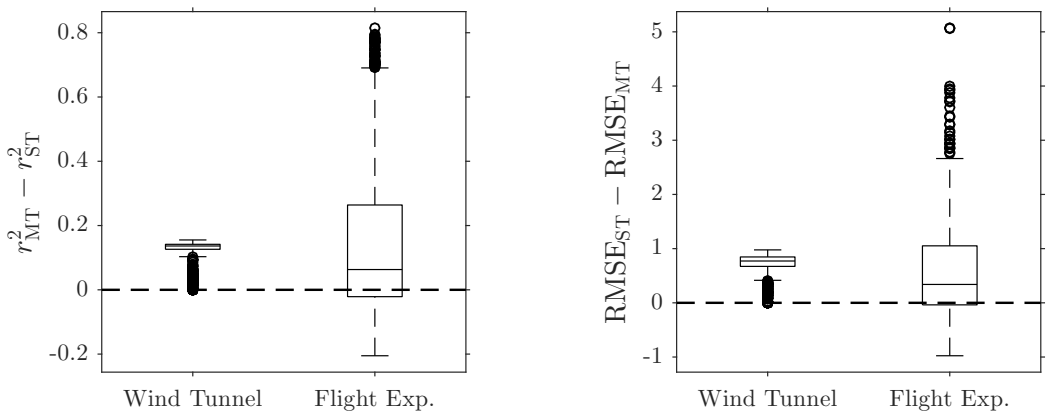
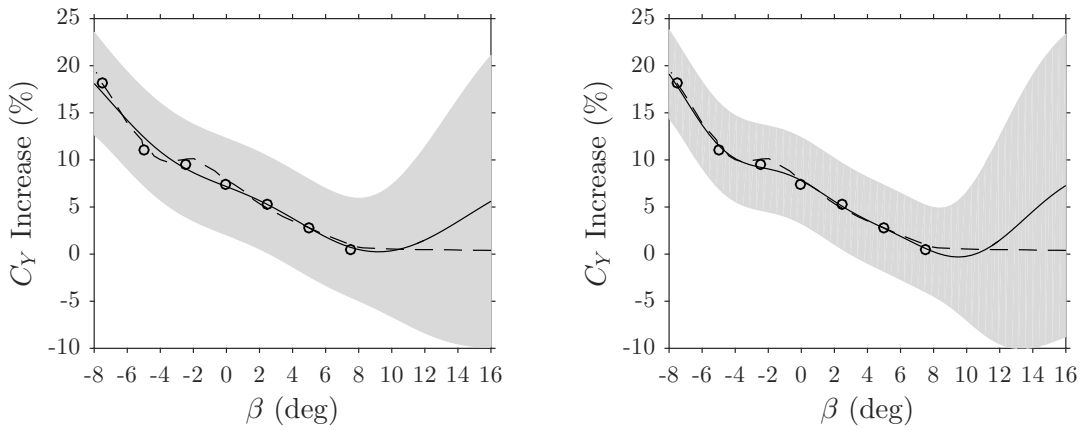


Figure 89: Differences in RMSE and r^2 between the single-task GP and MTGP predictions at the points of interest for the simulated truth data.

outperformed the single-task model. Clearly the multitask model had better predictive performance for the wind tunnel data. A larger variability in the performance results was exhibited for the flight experiment data because of the fact that extrapolations were made to all but one point of interest. Nevertheless, the evidence suggests that the multitask model performed better than the single-task model for the flight experiment as well.

To visualize the differences between MTGP and the single-task model, predictions for a single realization of the full-scale wind tunnel experiment and the flight experiment are shown in Figs. 90 and 91, respectively. For both experiments, the MTGP prediction intervals were observed to be shorter at certain points in the region of interest, particularly near $\beta = -7.5°$. For the wind tunnel experiment, the single-task GP attributed variability in the training data to aleatory noise, and the predictions were smoother than for MTGP. This difference was likely due to the fact that the MTGP model leveraged the trend of the sub-scale wind tunnel data during training. A similar effect was observed for the flight experiment predictions, where the mean predictions for MTGP were closer to the true target function than the single-task GP mean predictions. However, as shown in Fig. 89, there were some realizations of the



(a) Single-task GP

(b) MTGP

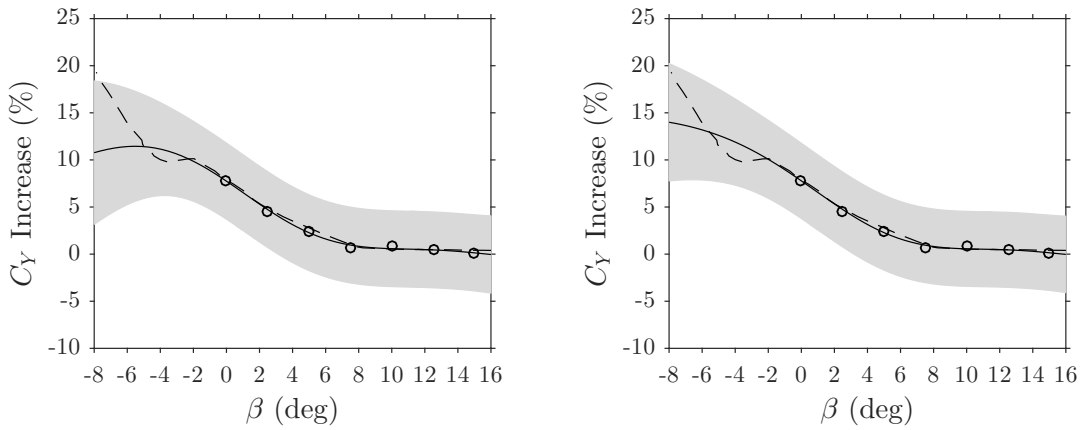
Figure 90: Predictions from MTGP and the single-task GP for a single realization of wind tunnel experiment truth data.

flight experiment data for which the predictive accuracy of the multitask regression model did not perform as well as the single-task model.

5.4.3.4 Evolution of Uncertainty With Maturation

In addition to estimating uncertainty reduction for the proposed experiments, entropy at the points of interest can be used to track the evolution of uncertainty as experiments are conducted and the technology matures. This is possible with single-task or multitask predictive modeling, and both techniques were used for this illustrative example to compare the two. Also, the predictive models were built with and without explicitly characterizing the additional layer of epistemic maturity uncertainty. The exponent ν was set to 2 for this demonstration.

Single-task and multitask GPs were constructed for the full-scale wind tunnel and flight experiments using the same settings described in Sec. 5.4.2 and a single realization of observations for the experiments. For the full-scale wind tunnel predictive model, the multitask GP was trained with the sub-scale wind tunnel data as well, for a total of two data sources. The multitask model for the full-scale flight experiment was trained with the sub-scale wind tunnel data and the full-scale wind tunnel data,



(a) Single-task GP

(b) MTGP

Figure 91: Predictions from MTGP and the single-task GP for a single realization of flight experiment truth data.

for a total of three data sources. The sum of entropies at the points of interest were computed with the predictive models built for all three experiments, and the results are plotted in Fig. 92. The $\times$ symbols in the plot at TRL 4 are the entropies at the points of interest for the sub-scale wind tunnel GP predictions. The higher-entropy point is the case where maturity uncertainty was accounted for, and the lower point is the case where it was not accounted for. The circle symbols are the entropies from the single-task GPs without maturity uncertainty for the full-scale wind tunnel experiment (TRL 5) and the flight experiment (TRL 6). In this case, the uncertainty increased from TRL 4 to TRL 5, then decreased slightly at TRL 6. The triangle symbols mark the entropies of the multitask GP predictions without maturity uncertainty for the two experiments. Notice that the entropies at TRLs 5 and 6 were much lower than the single-task case because of transfer learning with information from the two previous data sets. Once again the uncertainty increased from TRL 4 to TRL 5, but it decreased more noticeably at TRL 6. The square and + symbols represent the same single-task and multitask predictions, respectively, with maturity uncertainty accounted for. The multitask GP with maturity uncertainty was the only model that exhibited uncertainty reduction consistently as TRL increased.

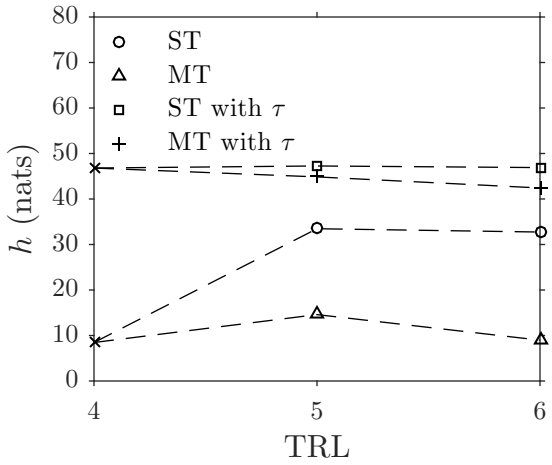


Figure 92: Evolution of entropy with maturation for various GP predictive models in the illustrative example.

At this point one may ask, *Is the lower uncertainty exhibited by the multitask GP appropriate in this example problem?* In this example, it is appropriate because of the improved generalization performance. For the full-scale wind tunnel predictions at TRL 5, the RMSE of the single-task predictions at the points of interest was 0.87, and the RMSE of the multitask predictions was 0.49. A difference was also found for r^2 correlation between the predictions and the target function at the points of interest, with a value of 0.93 for the single-task model and 0.99 for the multitask model. For the flight experiment predictions, the single-task model predictions had an RMSE of 2.56, whereas the multitask model predictions had an RMSE of 0.58. The r^2 correlations for the single-task and multitask predictions were 0.41 and 0.98, respectively.

5.4.4 Discussion and Conclusions

For this example problem, the results indicate that entropy estimation with the multi-task GP model was more accurate than with the single-task model, but this conclusion was based on the established underlying target functions. The r^2 correlation between the target functions shown in Fig. 37 is 0.91. For a problem in which the correlation

is much lower, it is possible that the multitask GP model would not provide any entropy-estimation benefits over the single-task model.

The GP comparison experiment presented in Sec. 5.3 focused on the predictive accuracy of the GP models and did not measure differences in predictive uncertainty. The AFC technology illustrative example discussed in this section demonstrated that a multitask approach is capable of modeling decreased prediction uncertainty, relative to a single-task approach, due to knowledge gained from an auxiliary data source. This result follows the intuitive behavior of one's epistemic uncertainty as data is generated from multiple, heterogeneous experiments. However, the reduced uncertainty is only appropriate when a multitask model has better generalization performance. A larger uncertainty band that contains the truth is preferred to a smaller uncertainty band that does not. This is why step three of the methodology is important.

5.5 Summary

This chapter explored the problems of how to quantify technology integration impact uncertainty in light of data from multiple, heterogeneous experiments and how to quantitatively estimate the uncertainty reduction that a planned experiment will achieve. These problems have not been satisfactorily addressed in the technology development literature, but key elements of a solution were identified in the statistics and machine learning literature. These elements were synthesized and adapted for the technology development context to formulate a methodology that addresses the research gaps.

The use of a maturity measure in the proposed methodology can be viewed as a necessary disadvantage. However, there is no way around somehow incorporating subjective judgments in the epistemic uncertainty characterization process. The modeling approach proposed in this chapter includes a more traceable and defensible way to incorporate subjective judgments than experts simply applying an inflation

based on their opinion. Despite the disadvantage of using a maturity measure, the proposed methodology provides an appropriate way to quantify the uncertainty surrounding technology integration impacts in light of data from multiple, heterogeneous technology development experiments because: (1) it is anchored in proven machine learning methods for making predictions under uncertainty and (2) it provides a flexible, quantitative approach to model the epistemic uncertainty associated with extrapolating technology impacts to the future. The methodology also provides an appropriate way to quantitatively estimate uncertainty reduction for a planned experiment because: (1) it implements a rigorous information theoretic framework that is the state of the art in experiment design and (2) it aggregates prediction uncertainty from a predictive model and the additional layer of epistemic uncertainty associated with technology maturity in the estimation process.

Although the proposed methodology is generally applicable with any kind of regression model, the scope was limited to GP regression models. Multitask GPs were identified as enabling techniques that are capable of borrowing strength from multiple, potentially heterogenous data sources for improving generalization performance and reducing epistemic prediction uncertainty. The primary contribution of the methodology is an approach for incorporating epistemic technology maturity uncertainty in GP predictions and estimates of uncertainty reduction for proposed experiments.

Due to limited empirical evidence in the literature, a gap in knowledge and understanding was identified regarding when a multitask GP will have better generalization performance than a single-task GP. An experiment was conducted, and it was determined that the conditions under which a multitask GP is the best option vary with the way transfer learning is accomplished and the complexity of the regression problem. Also, some guidelines regarding the characteristics of the regression problem were established for how to increase the likelihood that transfer learning will be beneficial.

The proposed methodology was implemented for a notional AFC technology experimentation example. It was shown that the multitask GP provided benefits in terms of generalization performance and reduced prediction uncertainty due to transfer learning with an auxiliary data source. The methodology was shown to provide nontrivial conclusions for which of the two notional proposed experiments would provide more uncertainty reduction. This is because the proposed approach aggregates prediction uncertainty from the GP model and the additional epistemic uncertainty associated with the anticipated maturation level of the proposed experiments.

A possible application of the methodology is to use the uncertainty reduction estimates as a component of an objective function to optimize the placement of observations for a planned experiment. This was not pursued in this work, but there is a large literature in adaptive sampling that can be leveraged toward this end.

CHAPTER VI

MATURITY-WEIGHTED BAYESIAN INFERENCE FOR RELIABILITY ANALYSIS OF SUCCESS/FAILURE DATA

The problem of how to quantify probability of failure for success/failure reliability data when data are potentially heterogeneous. First, the characteristics of the problem are described and the state of the art is identified in Sec. 6.1. Then, a maturity-weighted Bayesian inference methodology is formulated in Sec. 6.3. The primary argument is as follows.

Argument 4: The proposed methodology improves upon the state of the art and is an appropriate way to modify the Bayesian inference process because it provides analysts with the flexibility to incorporate epistemic uncertainty associated with technology or design maturity in the Bayesian reliability analysis process.

An illustrative example problem involving a rocket engine reliability analysis is presented in Sec. 6.4 to support this claim. Finally, the chapter closes with a summary in Sec. 6.5.

6.1 Problem Definition

Reliability, which can be defined as the ability of an item to perform a required function under given conditions for a stated period of time [117], is an important evaluation criterion in the engineering design decision-making process for complex systems. For high-consequence systems, such as space launch vehicles, reliability can be as crucial as performance and other considerations for evaluating design alternatives. As depicted in Fig. 93, decisions are made during the early design phases that lock in the life-cycle cost committed for the system, and it is desirable to intelligently

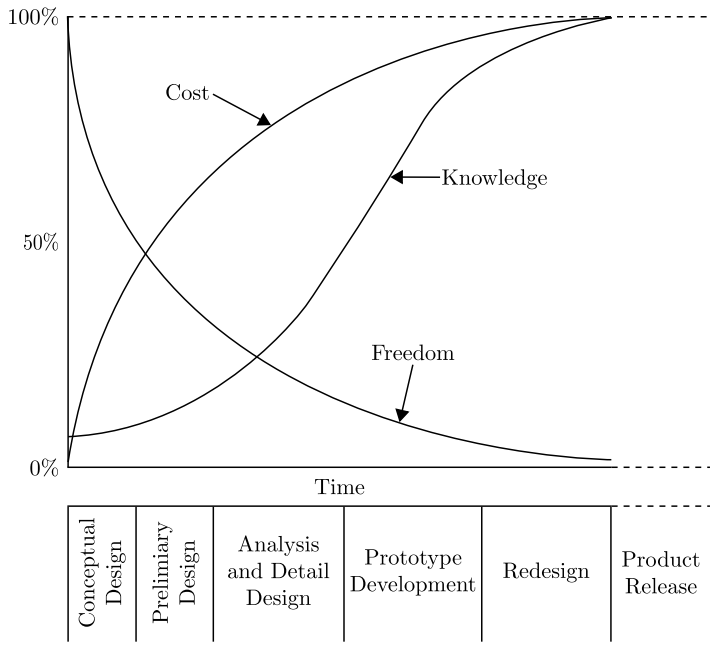


Figure 93: Notional changes in knowledge about the system design, cost committed, and design freedom over time (adapted from Ref. [118]).

select design options to avoid costly redesign in later phases. Thus, it is especially critical to credibly predict system reliability, among other system characteristics, in the early design phases. When advanced technologies are integrated with design options, the uncertainty surrounding reliability will grow. The focus of this chapter is how to properly quantify the uncertainty surrounding subsystem and component reliability predictions during technology development and the early design phases and how to fuse these predictions with data from multiple stages of maturity.

An integrated system is a collection of subsystems, each of which is composed of lower-level subsystems and ultimately components at the lowest level in the system hierarchy. In order to assess the reliability of an integrated system, the reliability of individual components and subsystems must be estimated. Reliability analysis frequently depends on multiple data sources such as physics-based modeling, expert elicitation, historical data from similar subsystems or components, and physical testing. Data are generated from these types of sources at multiple phases of the design or technology development process as uncertainty surrounding the system reliability

is reduced. The reducible component of this uncertainty is epistemic in nature. For example, the probability of failure of a given component or subsystem is a source of epistemic uncertainty. One can reduce this uncertainty by observing the outcomes of reliability tests. The outcome of a reliability test (e.g., the number of failures in a series of Bernoulli experiments), for a fixed failure probability, is aleatory in nature. This type of uncertainty can only be reduced by modifying the system that generates the test observations.

As a design or technology development program progresses, components and subsystems evolve as knowledge is accumulated and decisions are made. Bayesian and frequentist inference techniques can be used to quantify the uncertainty surrounding reliability of components throughout the design process. However, unlike frequentist inference, the Bayesian approach naturally enables one to explicitly represent both kinds of uncertainty with probability; the prior distribution represents one's epistemic uncertainty for the present state of knowledge about a parameter before acquisition of data, the likelihood characterizes aleatory uncertainty associated with observations, and the posterior distribution reflects the updated epistemic uncertainty after observing the data. A Bayesian approach is the focus in this chapter because of two important advantages of Bayesian inference in the context of reliability estimation as technologies and designs mature. One advantage is that when reliability data are scarce, the Bayesian approach can produce more realistic reliability estimates with uncertainty. As an example, if no failures occur during a reliability test, the binomial failure probability quantified using the frequentist maximum likelihood estimate (MLE) will be zero, which is not realistic. For a nonpathological prior distribution, Bayesian inference will produce a nonzero posterior point estimate of failure probability. The other advantage is that because posterior distributions from Bayesian inference are "true probability statements", they can be directly propagated through system reliability models, such as fault trees [119].

During the progression of a design process or technology development program, test articles, test conditions, and physics-based models will almost certainly change. Although it is desirable to employ Bayes' theorem to infer reliability based on all available data collected throughout the design process, it is likely that this approach would violate a necessary assumption called exchangeability. This exchangeability assumption is only appropriate when all characteristics of the experiments or processes that generate the data have been judged to be similar. The potential consequence of violating this assumption is poor inference of reliability, which could negatively affect the decision-making process.

Success/failure reliability data are the focus here, and there are three options for proceeding with a Bayesian reliability analysis given possibly nonexchangeable data: (1) continue with inference under the assumption of exchangeability, (2) perform inference without data pooling, or (3) modify the prior and/or likelihood to account for the data heterogeneity. The first option could lead to the consequences described in the previous paragraph. For the second option, inference could be conducted after each test, without the use of previous data. But, this approach would require the elicitation of a new prior for each test and would not leverage the information gained during earlier design stages. The third option, while lacking negative attributes of the first two options, begs the following question.

Research Question 4.0: What is an appropriate way to modify the Bayesian inference process to enable the proper representation of epistemic uncertainty when the success/failure reliability data are potentially nonexchangeable?

Answering this question is difficult because of the subjective nature of epistemic uncertainty; an approach is needed that enables analysts to quantitatively represent their uncertainty according to their lack of knowledge.

6.2 Literature Review

Multiple researchers have proposed solutions to similar reliability problems. Whitmore et al. [120] and Young [121] formulated three approaches to accumulate life test data for multiple versions of a given device in order to reduce the burden on manufacturers to demonstrate reliability after each design modification. This was achieved by either modifying the prior distribution on failure rate for a given design, which was the posterior distribution for the previous design, based on engineering judgment or modeling the relationship between the failure rates of different designs. Modification of the prior distribution was implemented by multiplying the distribution parameters with constants that represent similarity of the failure rates between designs. Huang and Jin [122] proposed a "consistency" measure, based on a χ^2 statistic, that quantifies the level of consistency among success/failure data sets from various sources. It should be noted that this consistency measure quantifies differences in the data sets, not differences in the data sources. Their approach requires the selection of a mapping from the consistency measure to a "data adjustment score" that multiplies the number of successes and failures in a selected data set. An algorithm is used to determine which data sets contribute to inconsistency and the data adjustment score discounts these data during the Bayesian inference process in order to achieve statistical consistency. Peng et al. [123] developed an approach to assess reliability throughout the life cycle of new products that is based on a Bayesian updating method for combining reliability data at all life cycle stages. A key feature of their Bayesian updating method is that they use a "reliability improvement factor" that quantifies experts' judgments about the difference in reliability between the new product and similar, existing products. A distribution on the reliability improvement factor is elicited from experts and is included as subjective information in the uncertainty updating process.

The prior and data validation and adjustment scheme (PDVAS) proposed by Huang and Jin [122] has been identified as the current state of the art for answering

RQ 4.0. However, the potential weakness of their approach is that data from a test of a more mature component or subsystem could be inconsistent with earlier test data and would then be erroneously discounted. A novel approach is proposed here that incorporates concepts that have been established by previous authors. In particular, the concept of a maturity weight with uncertainty is used that is akin to the method introduced by Peng et al. [123], and this maturity weight is used to modify the prior distribution at each development phase in a manner that is similar to Whitmore et al. [120] and Young [121].

6.3 A Maturity-Weighted Bayesian Inference Approach

According to Gelman et al., any Bayesian analysis follows three generic steps:

1. Setting up a *full probability model*—a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

2. Conditioning on observed data: calculating and interpreting the appropriate *posterior distribution*—the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

3. Evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? In response, one can alter or expand the model and repeat the three steps [124].

These steps serve as a foundation for the Bayesian inference approach that is proposed here and is shown in Fig. 94. Evaluation of the model fit is not explicitly included in this methodology because it is not anticipated that one would modify the probability

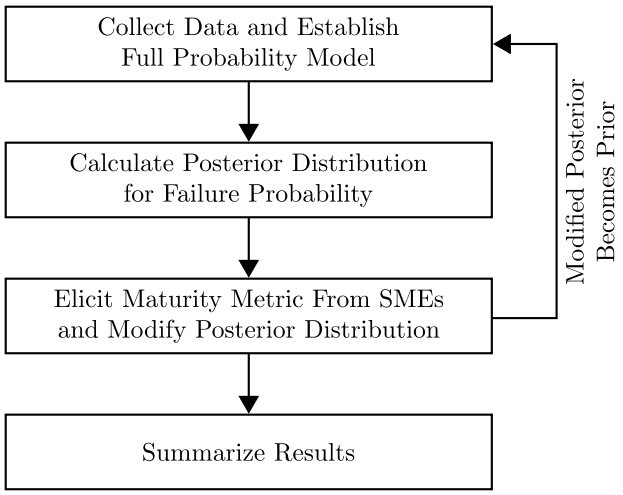


Figure 94: Steps of the proposed Bayesian inference methodology for success/failure reliability data.

model based on how well the model fits the data at a particular point in the design process. The approach has been designed such that the model fit may be poor for certain data sets. The maturity weight that controls this effect is the salient feature of the proposed methodology. Steps 1, 2, and 4 are all typical of a Bayesian reliability analysis of success/failure data. The novel components are described in the following subsections.

This proposed methodology fits in phase one of the overall solution shown in Fig. 7. This methodology provides a predictive model for forecasting technology reliability at a point in the future when the technology has been fully matured. This capability can be used to establish k-factor distributions to enable the evaluation of development activity alternatives. The methodology is also applicable to any system design process.

6.3.1 The Traditional Beta-Binomial Model

To construct a full probability model for inference of success or failure probability, three elements are needed: a sampling model, a prior distribution for success or failure probability, and the number of failures and successes from each data source. Under

the conditions of a fixed number of test articles or trials n and tests that are assumed to be conditionally independent given success or failure probability θ, the binomial distribution is an appropriate sampling model for the success/failure data [119]. The binomial probability mass function (PMF) can be written as

$$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x \in \mathbb{Z}_{\geq 0} \tag{48}$$

where, x can represent the number of successes or failures; in this chapter, x denotes the number of failures and θ denotes the failure probability. A commonly selected prior distribution for the unknown failure probability is the beta distribution, with probability density function (PDF)

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \alpha, \beta \in \mathbb{R}_+ \tag{49}$$

where, α and β are hyperparameters that can be interpreted as the prior number of failures and successes, respectively, and $\Gamma(\cdot)$ is the gamma function [119]. The beta distribution is a convenient choice of failure probability prior because it provides a conjugate structure with the binomial sampling model and the distribution support, the interval [0,1], is appropriate for a probability measure. By constructing a likelihood function with the sampling model and the available data and applying Bayes' theorem

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{50}$$

it can be shown that the posterior failure probability distribution, given data x, is also beta:

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \end{aligned} \tag{51}$$

For brevity, the posterior distribution of θ will be written in the form $\theta|x \sim \text{Beta}(\alpha + x, n - x + \beta)$.

In step 1 of Fig. 94, the prior distribution will either be an initial prior, with the form shown in Eq. (49), or a posterior distribution from a previous iteration through the first three methodology steps. For example, suppose that data (x_1, n_1) from tests conducted during a given development phase and an initial prior distribution on θ_0 have been pushed through Bayes' theorem to infer failure probability θ_1. Also, additional test data (x_2, n_2) are generated after some modifications of the component or subsystem. If the posterior failure probability θ_1 from the first inference is employed as the prior for the second inference, then the posterior distributions for θ_0, θ_1, and θ_2 would be defined as follows:

$$\theta_0 \sim \text{Beta}(\alpha, \beta) \tag{52a}$$

$$\theta_1|x_1 \sim \text{Beta}(\alpha + x_1, n_1 - x_1 + \beta) \tag{52b}$$

$$\theta_2|x_1, x_2 \sim \text{Beta}(\alpha + x_1 + x_2, n_2 + n_1 - x_2 - x_1 + \beta) \tag{52c}$$

An implicit assumption that is made in performing such an inference is that the data sets are exchangeable. An exchangeability assumption means that one can "express uncertainty as a joint probability density $p(y_1, \ldots, y_n)$ that is invariant to permutations of the indexes" for uncertain quantities y_i [124]. Thus, the indexes on the uncertain quantities in the joint distribution do not convey any information about the outcomes of the quantities y_i. In practice, this means that the characteristics of all tests that produced the observations have been judged to be similar. A desirable consequence of exchangeability is that it justifies the treatment of the reliability data as conditionally independent given the failure probability. However, when reliability data are generated by different types of sources and for multiple evolutionary versions of a particular component or subsystem, the assumption of exchangeability is questionable at best.

6.3.2 Adaptation of the Traditional Beta-Binomial Model to Account for Maturity

In order to account for potentially nonexchangeable success/failure data and to properly capture the epistemic uncertainty surrounding failure probability as a design or technology matures, the traditional beta-binomial model has been adapted with a maturity weight γ. This weight modifies the parameters of all posterior distributions that are computed throughout the development process. To illustrate the implementation for two data sets obtained sequentially, the maturity weights have been placed at the appropriate locations in Eqs. (52b) and (52c) as follows:

$$\theta_1|x_1,\gamma_1 \sim \text{Beta}(\gamma_1(\alpha+x_1),\gamma_1(n_1-x_1+\beta)) \tag{53a}$$

$$\theta_2|x_1,x_2,\gamma_1,\gamma_2 \sim \text{Beta}(\gamma_2(\gamma_1(\alpha+x_1)+x_2),\gamma_2(n_2-x_2+\gamma_1(n_1-x_1+\beta))) \tag{53b}$$

Because the parameters of the beta distribution must be greater than zero, the maturity weight must also be greater than zero. Mathematically, there is no need to establish an upper bound on the maturity weight. However, moments of the modified posteriors in Eqs. (53a) and (53b) have been examined to determine a practical upper bound. The variances for these distributions are

$$\text{Var}(\theta_1|x_1,\gamma_1) = \frac{(\alpha+x_1)(n_1-x_1+\beta)}{(\alpha+n_1+\beta)^2(\gamma_1(\alpha+n_1+\beta)+1)} \tag{54a}$$

$$\text{Var}(\theta_2|x_1,x_2,\gamma_1,\gamma_2) = \frac{(\gamma_1(\alpha+x_1)+x_2)(n_2-x_2+\gamma_1(n_1-x_1+\beta))}{(\gamma_1(\alpha+n_1+\beta)+n_2)^2(\gamma_2(\gamma_1(\alpha+n_1+\beta)+n_2)+1)} \tag{54b}$$

By inspecting Eq. (54a) one will see that for $\gamma_1 \in (0,1)$ the variance of the posterior distribution is increased relative to the traditional case where $\gamma_1 = 1$. This is a desired effect when using data that is produced at an early phase of the development process; when the system or technology is immature, one's epistemic uncertainty is higher than it is during later design phases. Mathematically, larger variance in the posterior distribution on failure probability represents this higher uncertainty that is due to lack of knowledge. γ_2 has the same effect on the variance in Eq. (54b), and

the impact of γ_1, which is not as apparent, also increases variance in this equation. The expected values of the failure probabilities in Eqs. (53a) and (53b) are

$$\mathrm{E}[\theta_1|x_1,\gamma_1] = \frac{\alpha + x_1}{\alpha + n_1 + \beta} \tag{55a}$$

$$\mathrm{E}[\theta_2|x_1,x_2,\gamma_1,\gamma_2] = \frac{\gamma_1(\alpha + x_1) + x_2}{\gamma_1(\alpha + n_1 + \beta) + n_2} \tag{55b}$$

One will immediately notice that γ_1 is not present in Eq. (55a) and γ_2 is not present in Eq. (55b). Thus, the expectation after the first inference is no different than the traditional result. However, because γ_1 is present in Eq. (55b), it affects the expectation after the second inference iteration. As $\gamma_1 \to 0$, the hyperparameters α and β and the first dataset (x_1, n_1) are discounted in Eq. (55b), and $\mathrm{E}[\theta_2|x_1,x_2,\gamma_1,\gamma_2]$ approaches the MLE result for the binomial distribution: $\hat{\theta}_2 = x_2/n_2$. This is also a desired effect, as the expected value should be affected less by data from previous design phases and more by data from recent design phases.

Examination of variance and expected values of the posterior distributions in Eqs. (53a) and (53b) has revealed the behavior of these important moments, given a range of values for the maturity weights. It was deduced that γ must be greater than zero; that $\gamma \in (0, 1)$ results in a discounting of the data, relative to the traditional approach; and that $\gamma = 1$ produces the traditional Bayesian inference results. As the development progresses, the author believes that γ should approach a value of 1 and only a value of greater than 1 in certain scenarios. For example, if critical hazards have been mitigated after a given test, the analyst may strongly believe that a maturity weight greater than 1 is appropriate to shrink the uncertainty surrounding failure probability. Using values greater than 1 could result in misleading inferences that are too optimistic with regard to the uncertainty surrounding failure probability.

6.3.3 Specification of Maturity Weight Values

The purpose of the maturity weight is to provide a way to quantitatively represent the differences between the characteristics of the experimental apparatus that generates

success/failure data at a given point in the development process and the characteristics of the fully matured system or technology. This "experimental apparatus" can be anything that is used during the design process to predict reliability. For example, reliability data from an existing system may be used as an estimate during conceptual design, an M&S environment can be used to predict reliability during preliminary design, or a sub-scale physical laboratory test could be conducted during any stage of the development process. The "differences" between characteristics of the experimental apparatus and the fully matured system or technology can be any discrepancies that could result in dissimilar failure probabilities. Given that all differences are impossible to enumerate, due in part to the fact that the final product is unknown until maturation, the maturity weight must be estimated. As indicated in step 3 of the methodology in Fig. 94, the maturity weight should be elicited from SMEs.

A practicing engineer that would use the proposed updating approach is obviously free to use any process to establish the maturity weight γ. However, some suggestions for doing so are provided. In general, there are two primary sources that result in differences between a given experimental apparatus and a mature system or technology. The first source of differences is due to dissimilarities between the conditions and physical characteristics of the experimental apparatus and the mature operational device, which can be due to limited resources and lack of knowledge. The second source of differences pertains to the fidelity of M&S environments that are used to generate reliability data.

It is suggested that the degree to which the first source of differences affects the Bayesian inferences be quantified with a metric that captures lack of component maturity during the development process. An example of this type of metric that is widely used in the systems engineering community for technology development is the TRL scale. A limiting attribute of many maturity scales, such as TRL scales, is that they are defined on ordinal scales. For example, a technology that is rated with a

TRL of 3 may be much less than half as mature as a technology at TRL 6. Thus, ordinal maturity scale values must be converted to a cardinal scale and mapped to the interval (0,1] before they are used in Bayesian calculations. An example of one approach to converting TRLs to cardinal coefficients process is presented in the work of Conrow [103].

The effects of the second source of differences can be quantified through model verification and validation processes. Verification quantifies the uncertainties due to numerical approximations in the M&S environment, whereas validation quantifies model accuracy by comparing simulation results with credible experimental data [15]. Based on verification and validation results, a determination should be made about how much the reliability data generated by a modeling and simulation environment should be discounted. As with the maturity metric, this model fidelity metric should be mapped to a value in the interval (0,1].

The components of the maturity weight that are attributed to design maturity and fidelity of M&S environments are denoted here as w_M and w_F, respectively. Once these values are established, they can be combined to form the maturity weight by simply multiplying the two: $\gamma = w_M w_F$. If the reliability data are generated from a physical experimental apparatus, then w_F should be set to a value of 1. The author acknowledges that specifying point values for w_M and w_F may be difficult due to epistemic uncertainty inherent in an SME's belief about these quantities. This uncertainty can be accommodated by placing probability distributions on these quantities. Uncertainty surrounding γ can then be integrated out of the posterior distributions that are computed using the procedure presented in Section 6.3.2. For instance, suppose that uncertainty surrounding γ_1 in Eq. (53a) is represented with distribution $p(\gamma_1)$. Then the posterior distribution on failure probability θ_1, with γ_1 marginalized out, is

$$p(\theta_1|x_1) = \int_{\mathbb{R}} p(\theta_1|x_1, \gamma_1) p(\gamma_1) d\gamma_1 \tag{56}$$

Table 12: Rocket engine reliability data for the example problem (data from Ref. [122])

Design stage	Data category	Number of failures	Number of successes
Concept exploration	Demonstrated reliability from heritage engine A	0	69
	Demonstrated reliability from heritage engine B	0	13
Conceptual design	Combination of SCA and PBMS	1	999
Embodiment design	Laboratory test result	1	4
Development	Subscale development test results	2	18
	Full scale development test results	3	147
Certification	Certification test results	0	120

6.4 Illustrative Example: Rocket Engine Reliability

In this section, the proposed Bayesian reliability analysis methodology is compared with the traditional approach and PDVAS [122]. First, the problem setup is described. Then, the application of the proposed methodology is presented. Finally, results of the three methods are presented and compared.

6.4.1 Problem Setup

The example problem entails reliability analysis for a rocket engine. Notional data have been obtained chronologically from multiple points in the design process. These data were extracted from the paper by Huang and Jin [122] and are shown in Table 12. The first two data sets are from two heritage engines that are fully mature. The third data set is from a combination of a similarity and comparative assessment (SCA) and physics-based modeling and simulation (PBMS). The remaining data set categories are self-explanatory. The problem is that estimates of failure probability are needed to support design decisions.

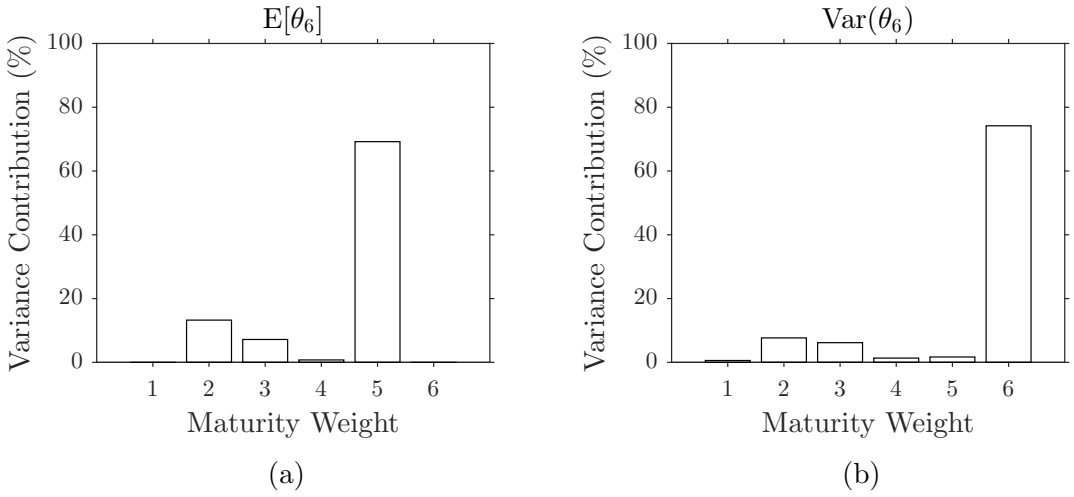


Figure 95: First-order sensitivity indices for failure probability mean and variance.

6.4.2 Global Sensitivity Analysis

Before selecting maturity weights to use in the proposed methodology, suppose that the analyst wishes to understand the impact of the weights on the inference with all of the data. A global sensitivity analysis can provide this information, and it was implemented by specifying distributions on each of the maturity weights and calculating first-order sensitivity indices using the method described in Ref. [77]. The first two data sets were lumped into one with a corresponding maturity weight. Thus, there were six data sets and maturity weights used. Uniform distributions were used for all of the maturity weights, with a lower bound of 0.1 and an upper bound of 1. An initial noninformative beta prior distribution with parameters $\alpha = 0.5$ and $\beta = 0.5$ was used. First-order sensitivity indices were calculated using 100,000 samples for the mean and variance of the final beta distribution on θ_6. The sensitivity indices for the mean and variance of failure probability are shown in Figs. 95a and 95b, respectively.

Comparing the mean sensitivities with Eq. (55b), it is not surprising that the maturity weight used to perform inference with the fifth data set was the largest contributor to variability. The mean sensitivities for data sets two, three, and four were also relatively significant. The weight for the second data set was a large driver

because the test involved a large number of trials, and any data set with significant evidence will have a large impact on the posterior distribution. The impact tapered off with the third and fourth data sets, as these had smaller trial sizes. However, these data sets pulled the posterior distributions toward failure probabilities on the order of 10^{-2}, away from the inference for the second data set of the order 10^{-4}. The behavior of the variance sensitivities was similar, with the most noticeable difference being that the weight corresponding with the sixth data set was the largest driver. This difference was not unexpected and follows from the effect observed in Eq. (54b).

6.4.3 Comparison of the Inference Methods

The traditional Bayesian inference technique is straightforward to implement by following the analytical updating scheme shown in Eqs. (52b) and (52c). PDVAS is more complicated to implement because it requires the selection of multiple parameters. To ensure that the intended implementation of PDVAS was reflected in the comparison, the results published in Ref. [122] for the example problem were used. The proposed methodology was applied using the maturity weights shown in Fig. 96a. These weights represent the judgment of the author and are notional. For all of the methods, an initial noninformative beta prior distribution with parameters $\alpha = 0.5$ and $\beta = 0.5$ was used. For each of the methods, the mean and 95% credible sets were calculated from the posterior distributions at all chronological stages, and the results are shown in Fig. 96b.

The impact of the low maturity weight on the first inference step was apparent; the means of all three methods were identical, but the maturity weight clearly inflated the uncertainty relative to the traditional method and PDVAS. The large trial size of the second data set pulled the posteriors of all methods toward a lower mean and reduced variance. The third data set indicated that failure probability may be much higher than predicted by the M&S results, but PDVAS and the traditional method

maintained low uncertainty and means relative to the weighted method. The fourth data set also indicated a higher failure probability than the first two data sets, yet the traditional approach and PDVAS had low means and uncertainty, whereas the proposed approach exhibited a higher mean and larger uncertainty. The fifth and sixth tests indicated lower failure probability, and the proposed approach estimated lower means and uncertainty accordingly. The weighted method and traditional method were more sensitive to the last two data sets than PDVAS. An important observation from these results is that even with the third, fourth, and fifth tests having MLEs of 0.2, 0.1, and 0.02, respectively, the final PDVAS posterior had a mean of 0.0018 and a standard deviation of 0.0012. It is possible that the true engine failure probability decreased over the final phases of design, but conservative reliability engineers might not have been convinced. If mitigation actions had been taken, then it may have been appropriate use higher maturity weights. The creators of PDVAS explained that the low final posterior values were due to the discounting of the "inconsistent" data sets that had relatively high failure probabilities. This is an optimistic approach that can be misleading. The weighted method does the opposite; the tests with higher failure probability had higher weights than earlier tests, and the weighted method estimated more conservative failure probabilities as a result.

Another option that the analyst has is to specify distributions for the maturity weights to reflect the analyst's epistemic uncertainty surrounding the appropriate weighting. As a comparison, the proposed approach was implemented with uniform distributions on the maturity weights for all data sets. The upper and lower bounds of the uniform distributions were 1 and 0.1, respectively. The means and 95% credible sets of the proposed inference approach are plotted along side the traditional Bayesian approach and PDVAS in Fig. 96c. Comparing the cases with deterministic maturity weights and uniformly-distributed maturity weights, some interesting observations can be made. One of the most apparent differences is that the length of the credible

sets for data sources 1–4 were shorter when uncertainty surrounding the maturity weights was modeled, whereas the credible sets increased in length for data sources 5 and 6. This was due to the significant jump in the deterministic maturity weights between data sources 4 and 5, as can be seen in Fig. 96a. Another key observation is that modeling the uncertainty in how much each data source should have been discounted resulted in lower means, particularly for data sources 4, 5, and 6. This was due to the propagation of high maturity weight samples for the more optimistic test results, such as data source 2, to the later stages. A takeaway from this example is that modeling uncertainty in the maturity weights will not necessarily lead to more uncertainty surrounding the failure probabilities at all stages of development, compared to a deterministic setting.

6.5 Summary

This chapter investigated the problem of how to incorporate multiple data sources in the Bayesian reliability analysis of success/failure data during the design or technology development process. Through an analysis of the literature, the current state of the art was identified, and it was argued that a novel approach was needed to properly represent epistemic uncertainty when the success/failure reliability data are potentially nonexchangeable. An adaptation of the traditional beta-binomial probability model was formulated to address the research question.

The proposed methodology was applied to a rocket engine reliability example problem and compared to traditional Bayesian inference and the state-of-the-art method PDVAS. A global sensitivity analysis was conducted with the weighted method to demonstrate the effects of the maturity weights on the final posterior distribution mean and variance. Notional maturity weights were established, and results of the three methods were presented. PDVAS and the traditional methods consistently estimated lower variance than the weighted method, indicating that the weighted

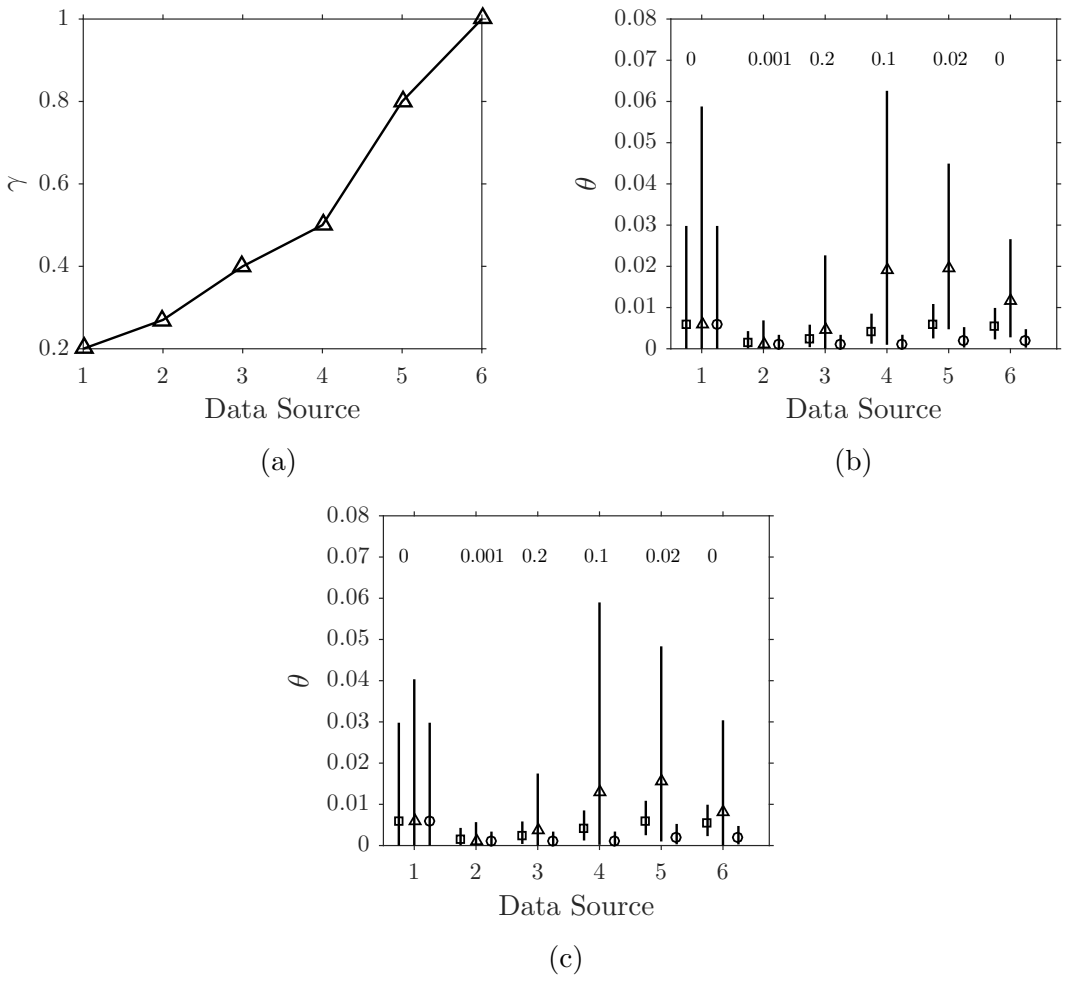


Figure 96: Plots of (a) the maturity weights used in the comparison with deterministic weights; (b) the means and 95% credible sets of the traditional Bayesian approach (□ symbols), the proposed methodology with deterministic maturity weights (△ symbols), and PDVAS (◯ symbols); and (c) the means and 95% credible sets of the traditional Bayesian approach (□ symbols), the proposed methodology with uniformly-distributed weights (△ symbols), and PDVAS (◯ symbols). The maximum likelihood estimates for each data set are shown above the credible sets.

method more conservatively represented the epistemic uncertainty surrounding failure probability. The final failure probability posteriors of the traditional method and PDVAS were highly influenced by tests from early design phases, where low failure probabilities were predicted. Thus, the final mean and variance predictions were potentially too optimistic. Due to relatively large maturity weights for the later stages of design, the weighted method demonstrated higher failure probability means. A scenario in which large uncertainty surrounded the maturity weights was also investigated, and the results showed more optimistic failure probabilities in the late stages of design. Also, it was shown that incorporating additional uncertainty through the maturity weights will not always result in increased uncertainty surrounding failure probabilities because of the influence of the weights on variance of failure probability. Although the proposed methodology requires elicitation of a maturity metric, it is more appropriate than existing Bayesian methods for inference of component or subsystem failure probability because it provides analysts with the capability to model their epistemic uncertainty as a design or technology matures, all at the cost of few implementation steps.

CHAPTER VII

CONCLUSIONS

To fulfill the future aviation needs of the public and military, there are efforts in industry and government to integrate aircraft with enabling technologies to achieve aggressive goals and requirements for performance and capabilities. However, many enabling technologies are immature, and system integrators incur the associated risk when they integrate these technologies. This risk can be reduced through technology development programs, but these programs often require over ten years and significant resources before the technology can be transitioned to the vehicle. Ideally, the process could be accelerated and the required resources reduced by creating the development activities, such as physical experiments and tests, such that they maximize performance improvement, maturation, and risk reduction during the development program. The research in this dissertation comprises contributions toward this vision, and these contributions are summarized here. Additionally, future research opportunities are discussed, and an overarching thesis statement is presented.

7.1 A Novel Framework for Designing Technology Development Activities

The framework presented in Chapter 3 was formulated to address the motivating question:

Motivating Question: How should technology development activities be designed?

It was argued that to meet the technology development goals of uncertainty reduction, performance improvement, and maturation, there are three primary questions that

must be answered to design technology development activities:

1. *Which types of activities should be selected?*
2. *What is the best setup of the physical or computational environment for each activity?*
3. *How should each activity be executed to maximize the value of information that is generated?*

The framework comprises three phases that correspond with answering each question: (1) thought experimentation, (2) detailed definition of the activities, and (3) statistical design of experiments. These phases were applied to a case study for an AFC technology to derive new insights for how the actual technology development program should have been conducted. Also, opportunities for adding rigor to the framework were discussed, and three contributions were presented in Chapters 4, 5, and 6 toward this end.

7.1.1 Limitations and Future Research Opportunities

Future improvements for phase one of the framework are discussed in Sec. 7.2.1. For phase two, the development of quantitative methods for evaluating alternative equipment and variables for each activity is an opportunity for future research. In phase three, there are decision criteria that are difficult to estimate for experimental designs before the activity has been executed. An approach for estimating uncertainty reduction was formulated in Chapter 5, but the need for estimating other decision criteria remains. For instance, how should one estimate the performance improvement potential of an activity with a specific experimental design a priori? To add further rigor to the selection of an experimental design, this type of question must be answered for the decision criteria that are pertinent to each activity.

Even if the framework is improved through further research, there are fundamental limitations that should be considered when implementing it. The framework is not intended to be used for selecting which technologies enter a development program; there are methodologies in the literature that can be used for this purpose. The framework does not include explicit steps for determining the best plan for conducting the portfolio of development activities. There is a tradeoff between minimizing the time to complete a set of development activities and leveraging learning between activities. A processes should be followed for planning execution of the activities to minimize the risks of schedule slips and cost growth while maximizing the uncertainty, performance, and maturation benefits of the activities. Also, the framework was formulated to be generic, and there are many ways which the decision processes can be interpreted and implemented for a given set of activities. Technologists with disciplinary backgrounds must exercise their knowledge and understanding of each technology in phase two to select the most appropriate equipment and variables for each activity. In phase three, it is impossible to list specific steps from DoE theory that are foolproof for any technology and any type of activity. An applied statistician or at least a professional with substantial knowledge of DoE should be consulted for applying the decision process to select a statistical design.

7.2 Multiattribute Utility Analysis for Evaluating Technology Development Activities

Chapter 4 explored the problem of how to inform decisions regarding the selection of technology development activity classes before details of the activities have been defined. The corresponding RQ follows.

Research Question 1.0: Given alternatives defined by combinations of technology development activity classes and technologies, what is an appropriate way for decision makers to evaluate the alternatives for downselection?

It was argued that the primary drawback of the state of the art is the lack of a capability to explicity evaluate alternatives. Thus, it does not address RQ 1.0. A generic decision process provided the foundation for a novel methodology, and ideas from multiattribute utility theory were incorporated to address RQ 1.0. The normative decision support methodology entails establishing objectives and attributes, constructing a utility model to represent decision makers' values, modeling the impacts of the alternatives, and evaluating the alternatives with expected utility. As demonstrated in the illustrative example, the product of the methodology is not simply a single expected utility for each alternative but rather a capability that enables quantitative tradeoffs and sensitivity analyses to provide insights and stimulate deeper thinking about the problem on the part of the decision makers. Compared with the state of the art, the proposed methodology is an improvement because it was shown to enable explicit evaluation of alternatives rather than only providing measures of potential for each technology. The answer to RQ 1.0 is summarized by the following argument.

Argument 1: The proposed methodology improves upon the state of the art and is an appropriate way to evaluate technology development activity alternatives because

1. It aggregates decision makers' preferences, risk attitude, and system-level performance goals in the analysis
2. It quantitatively represents uncertainty surrounding the impacts of the alternatives
3. It enables the quantitative evaluation of alternatives under conditions of risk and uncertainty with a theoretically valid measure of value

7.2.1 Limitations and Future Research Opportunities

Although key attributes were proposed in the methodology, an opportunity for future work is to identify an exhaustive list of attributes that can be used. In practice, some or all of the attributes may not be mutually utility independent in the minds of some decision makers. Another research opportunity is to formulate one or multiple attributes in such a way that mutual utility independence is satisfied and the attributes are still easily interpreted by decision makers. Finally, there may be a feasible way to use the mapping between the technology development activity impacts and multiattribute utility to perform inverse design of activities. In other words, there may be a way to establish a distribution on utility and back out distributions on the activity impacts. Then, a set of technology development activities could be identified that map to those activity impact distributions. Probabilistic inversion is a potential enabler for this inverse design approach.

7.3 Uncertainty Quantification with Multitask Gaussian Processes for Technology Development Experiments

Chapter 5 explored the problems of how to quantify technology integration impact uncertainty in light of data from multiple, heterogeneous experiments and how to quantitatively estimate the uncertainty reduction that a planned experiment will achieve. The RQs are as follows.

Research Question 2.0: What is an appropriate way to quantify the uncertainty surrounding technology integration impacts in light of data from multiple, heterogeneous technology development experiments?

Research Question 3.0: What is an appropriate way to quantitatively estimate expected uncertainty reduction for a planned technology experiment?

It was argued that these problems have not been fully addressed in the technology development context, but the ingredients for a solution were identified in the statistics

and machine learning literature. These ingredients were synthesized and adapted for the technology development context to formulate a methodology that addresses the research gaps. The first three steps of the methodology were borrowed from the data analysis literature. These steps comprise the traditional pipeline of cleaning a data set, identifying a set of predictive models, and evaluating and selecting from the set of models. The fourth step is a novel contribution because it provides an approach for incorporating epistemic technology maturity uncertainty in Gaussian process model predictions. The fifth step is also a novel contribution because it fuses a rigorous information theoretic framework for quantifying uncertainty reduction with predictive models that incorporate the additional layer of epistemic uncertainty associated with technology maturity. The key capabilities provided by the methodology were demonstrated with a simple one-dimensional illustrative example. The following arguments answer RQs 2.0 and 3.0.

Argument 2: The proposed methodology provides an appropriate way to quantify the uncertainty surrounding technology integration impacts in light of data from multiple, heterogeneous technology development experiments because

1. It is anchored in proven machine learning methods for making predictions under uncertainty
2. It provides a flexible, quantitative approach to model the epistemic uncertainty associated with extrapolating technology impacts to the future

Argument 3: The proposed methodology provides an appropriate way to quantitatively estimate uncertainty reduction for a planned experiment because

1. It implements a rigorous information theoretic framework that is the state of the art in experiment design
2. It aggregates prediction uncertainty from a probabilistic regression model and the additional layer of epistemic uncertainty associated with technology maturity in the estimation process

The Gaussian process comparison experiment is also a contribution not only to technology development but to the engineering design, statistics, and machine learning communities as well. New empirical evidence was presented to support claims concerning when multitask Gaussian processes will outperform single-task Gaussian processes.

7.3.1 Limitations and Future Research Opportunities

The methodology formulation is limited to Gaussian process regression models. Thus, an important research opportunity is to extend the ideas to other classes of regression models. Also, the posterior entropy metric can be a component of a composite objective function for selecting experimental designs in phase three of the framework.

Other measures, such as expected performance improvement, could be incorporated to balance exploration and exploitation in the independent variable space. Another future research opportunity is to apply the methodology to real technology development programs that have been conducted and publish the results as a case study that can inform the implementation for future programs. In particular, a valuable exercise would be to calibrate the subjective parameters that govern the maturity uncertainty inflation using actual data from technologies that have been successfully developed. The calibrated parameters would provide an example for the evolution of these parameters as a technology matures. Finally, Gaussian process prediction uncertainty is represented by symmetric distributions. In some situations, analysts may wish to skew the prediction uncertainty distributions. To do this, an approach is needed to parametrically skew the distributions at any desired locations in the independent variable space. The multivariate skew-normal distribution [125] is a potential enabler for this purpose.

7.4 Maturity-Weighted Bayesian Inference for Reliability Analysis of Success/Failure Data

Chapter 6 investigated the problem of how to incorporate multiple data sources in the Bayesian reliability analysis of success/failure data during a design process or a technology development process. The RQ follows.

Research Question 4.0: What is an appropriate way to modify the Bayesian inference process to enable the proper representation of epistemic uncertainty when the success/failure reliability data are potentially nonexchangeable?

It was argued that the current state of the art produces overly-optimistic estimates of the epistemic uncertainty surrounding failure probability and does not provide the flexibility to incorporate the maturity dimension in the Bayesian reliability analysis. An adaptation of the traditional beta-binomial probability model was formulated to

address the research gap. The novel Bayesian reliability analysis methodology begins with traditional Bayesian data analysis steps. Then, a maturity weight is introduced in the posterior beta distribution to enable discounting of the reliability data at a given point in the development process. An illustrative example was presented to compare the proposed methodology with the state of the art. The flexibility provided by the infusion of a maturity weight was shown to enable an analyst to inject additional subjective uncertainty into the inference process, thereby enabling more conservative estimates of failure probabilities, if so desired. The overarching claim answers RQ 4.0.

Argument 4: The proposed methodology improves upon the state of the art and is an appropriate way to modify the Bayesian inference process because it provides analysts with the flexibility to incorporate epistemic uncertainty associated with technology or design maturity in the Bayesian reliability analysis process.

7.4.1 Limitations and Future Research Opportunities

Since this work was limited to success/failure reliability data, there is a research opportunity to extend the idea of discounting data with maturity weights to other types of reliability data. The weighted likelihood approach [126] is one option for injecting a maturity weight into the Bayesian inference procedure in a more generic way. Also, the methodology should be applied to real reliability data for systems that have been developed previously. The maturity weights should be calibrated using the data to provide an example of how they actually vary with maturation.

7.5 Thesis Statement

In Sec. 2.4, the current practices for designing technology development activities were described. The primary problem with the current practices that was identified is the reliance on TRL definitions for design decisions about future development activities.

The issue with using TRL scales as the main driver in these decisions is that they do not characterize the state of uncertainty surrounding the integration impacts of a technology. A result of this is that decisions can be misinformed because they hinge on a maturation criterion and less significantly or not at all on uncertainty reduction criteria. Also, there is not a clear-cut procedure for designing technology development activities in the literature; many important decisions are delegated to decision makers and technologists. The proposed framework addresses these gaps, and the other contributions in this dissertation add rigor to the framework. The overarching thesis statement is as follows.

Thesis Statement: The proposed framework for designing technology development activities improves upon the current practices because

1. It incorporates a set of generic decision-making steps in three phases to provide a systematic process for determining the types of activities that should be pursued, the best setup for each activity, and how each activity should be executed to maximize the value of information that is generated

2. It integrates multiple decision criteria for evaluating alternatives during the activity design process

3. It provides a foundation for the use of quantitative methods to improve the state of decision-support capabilities

The first reason stated in the thesis highlights that the framework is the first—to the best of the author's knowledge—to decompose the activity selection process into explicit decision-making steps. The second reason in the thesis statement is important because the framework fuses multiple decision criteria to qualitatively or quantitatively capture the value of the alternatives, which is a key improvement on the current practices. The third reason in the thesis points out a key characteristic of

the framework: it lays out a generic process with steps that quantitative methods can be "plugged into" to improve decision support capabilities. Three such contributions have been presented in this dissertation, but there are more opportunities that have been identified. Researchers are encouraged to continue this work of populating the framework in the future and to apply the framework in technology development programs. By doing so, the author's hope is that development efficiency will be increased for promising advanced technologies, thereby accelerating the delivery of the technology benefits to society.

APPENDIX A

EDS SURROGATE MODEL ASSESSMENT

In the illustrative example described in Sec. 4.4, artificial neural network surrogate models were used to expedite the uncertainty propagation process for mapping the technology impacts to the system-level metrics with the EDS M&S environment. In this appendix, an analysis of the EDS surrogate models is presented.

To improve the fits of all system-level metric surrogates, the aircraft design takeoff gross weight (TOGW) computed by EDS was used as an input to the models. Thus, a surrogate model of design TOGW was needed as well, and the fit statistics are shown in Fig. 97. The model fit error (MFE) plot shows a histogram of the relative error (%) of the predictions for the training data. The MFE distribution appears to be symmetric with mean near zero and a standard deviation of much less than one, which is a preferred result. Similarly, the model representation error (MRE) histogram is the relative error (%) of the predictions for validation data that was not used in the training process. The MRE mean is also small with a low standard deviation. Any skewness in the MFE or MRE distributions would be a cause for concern because it would indicate a surrogate model that either under-predicts or over-predicts the true values from EDS. The actual by predicted plot shows the actual EDS values plotted against the values predicted by the surrogate model. The ideal case is where the values all lie on a straight line, as indicated in the plot by the solid black line. The validation and training points all lie close to this perfect fit line, which is a good result. Finally, the residual by predicted plot shows the residual values (absolute error between the true EDS values and the predictions) on the vertical axis plotted against the predicted values on the horizontal axis. Ideally, this plot looks random like a

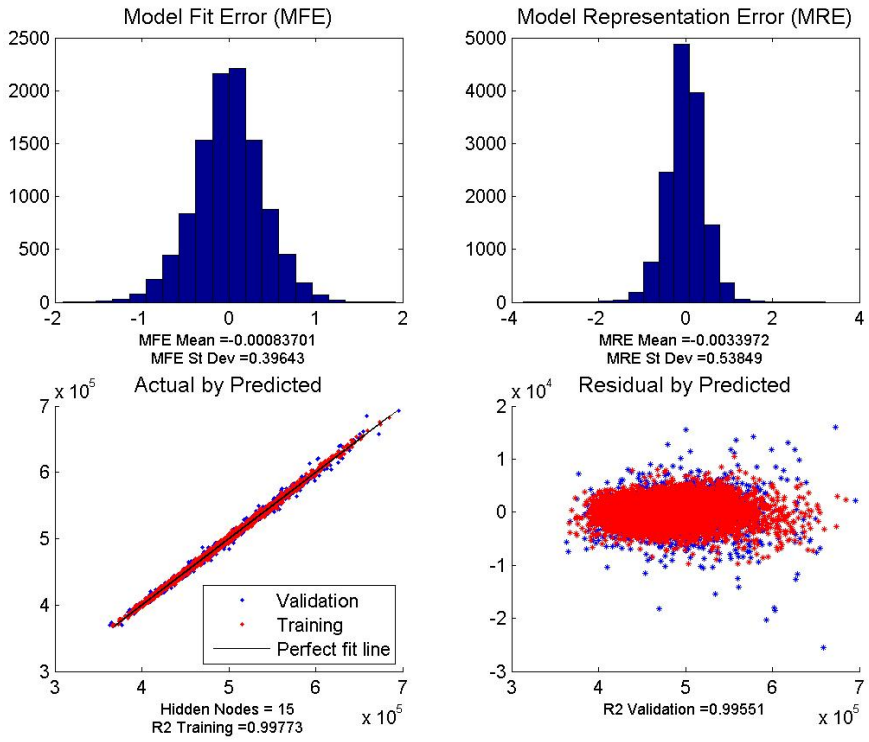


Figure 97: Fit statistics for the design TOGW EDS surrogate model.

shotgun blast, with not obvious patterns. Any pattern in this plot would indicate that a more complex model may be required to fit the data well. A conservative estimate of the worst case prediction error can be identified in this plot by finding the maximum residual value and dividing it by the smallest predicted value. The worst case in the plot is a validation case with a residual of approximately -25,000 divided by the smallest predicted value of approximately 360,000, which results in an error of just under 7% (absolute value). A typical rule of thumb when fitting regression models is to aim for this error to be less than 10%. The coefficient of determination R^2 is a less important evaluation criterion. Nevertheless, a value as close to one as possible is another indication of a good predictive model. This model fit within the evaluation criteria and was judged to be appropriate for use in the example problem.

The fit statistics for the aircraft design block fuel surrogate model are shown in Fig. 98. The MFE and MRE distributions both appear to be symmetric with

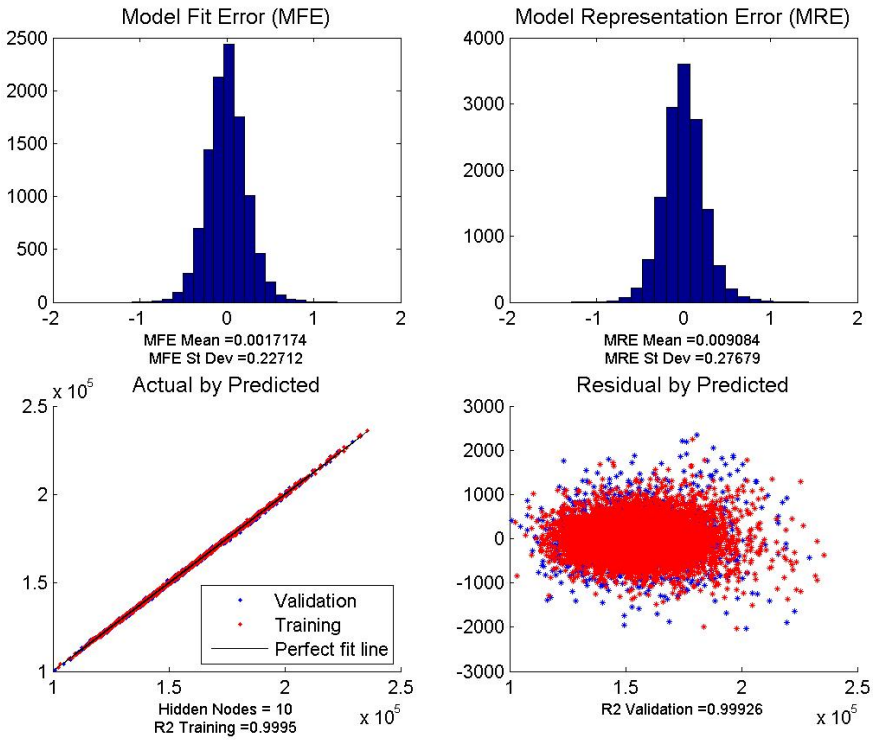


Figure 98: Fit statistics for the design block fuel EDS surrogate model.

low means and standard deviations. The actual by prediction plot exhibits samples that were tightly grouped around the perfect fit line. The residual by predicted plot indicates no obvious patterns, and the worst case prediction error was approximately 2%. These results were satisfactory for use of the surrogate model in the example problem.

The fit statistics for the aircraft sideline noise surrogate model are shown in Fig. 99. The MFE and MRE distributions both appear to have a small degree of skewness with low means and standard deviations. The actual by prediction plot exhibits samples that were grouped around the perfect fit line but with a few points that were separate from the high-density region. The residual by predicted plot indicates no obvious patterns, and the worst case prediction error was approximately 2%. These results were also judged to be satisfactory for use of the surrogate model in the example problem.

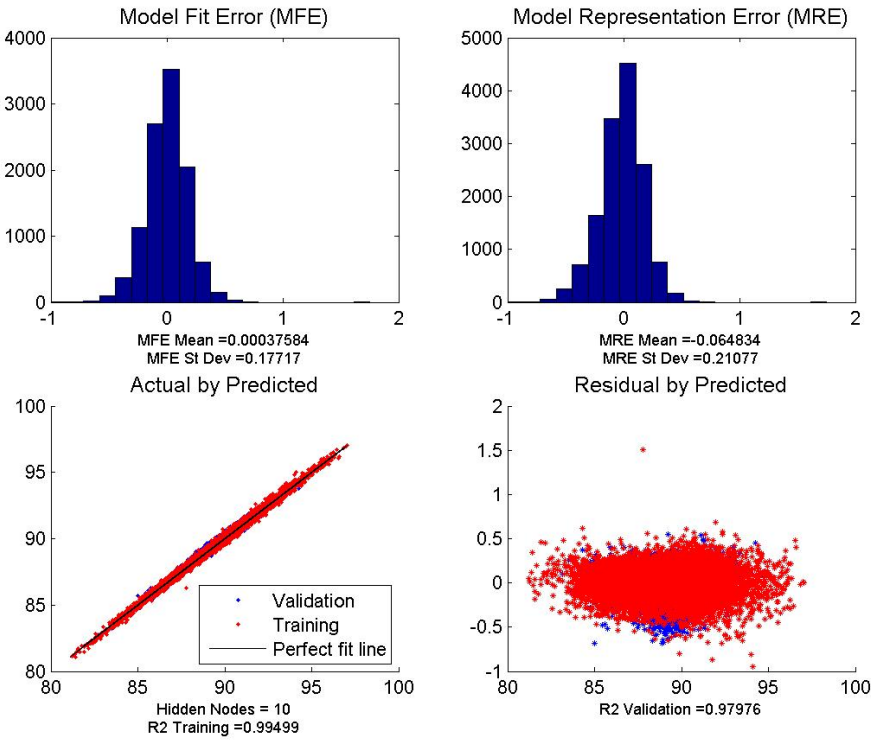


Figure 99: Fit statistics for the sideline noise EDS surrogate model.

The fit statistics for the aircraft TOFL with one engine inoperative surrogate model are shown in Fig. 100. The MFE and MRE distributions both appear to be symmetric with means and standard deviations that are larger than the other surrogates. The actual by prediction plot exhibits many samples that were not tightly grouped around the perfect fit line. The residual by predicted plot indicates that a pattern may be present, and the worst case prediction error was approximately 58%. TOFL has been a notoriously difficult EDS output to regress, and although many of the evaluation criteria were violated, this was the best surrogate model that the analyst was able to produce. It was decided to use this surrogate model in the illustrative example despite the poor predictive performance because the results are notional and will not be used for high-consequence decision making.

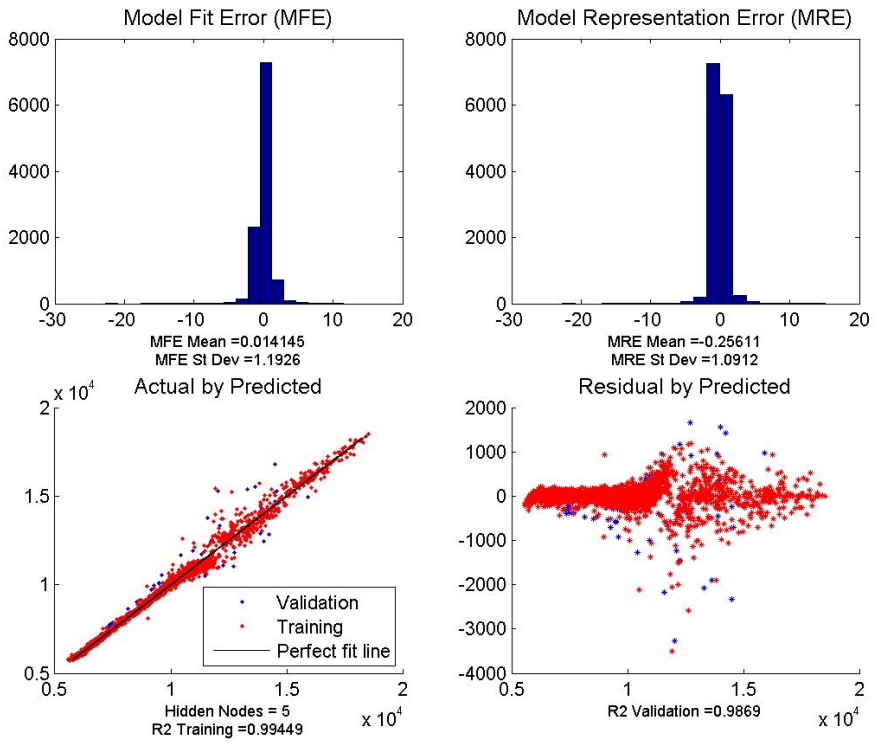


Figure 100: Fit statistics for the TOFL EDS surrogate model.

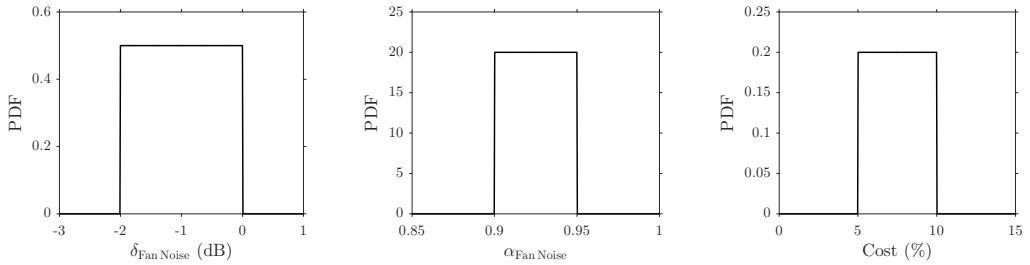
APPENDIX B

EXPECTED UTILITY COMPUTATION EXAMPLE

In the description of step five of the methodology for evaluating technology development activities, Fig. 20 is presented to illustrate the flow of uncertainty from the effects of each alternative to multiattribute utility. The details of this propagation process are not shown for the illustrative example problem described in Sec. 4.4. In this appendix, results from each level of the hierarchy are shown for the alternative defined as a computer experiment for the fan vertical acoustic splitter technology (A_1) in the example problem.

The effects of the activities on the baseline technology impact distributions are at the lowest level of the propagation hierarchy. For A_1, the performance effects were modeled as a translation of the takeoff noise k-factor and a scaling of the variance with the parameters $\delta_{\text{Fan Noise}}$ and $\alpha_{\text{Fan Noise}}$, respectively. A cost impact was modeled as well. The uniform distributions for these effects are shown in Fig. 101. The bounds of the distributions correspond with Table 9. The interpretation of the distribution on $\delta_{\text{Fan Noise}}$ is that the computer experiment is expected to result in performance improvement between 0 and -2 dB. The distribution on $\alpha_{\text{Fan Noise}}$ represents anticipated variance reduction between $100(1 - 0.95^2) = 9.75\%$ and $100(1 - 0.90^2) = 19\%$ from the baseline. The cost distribution indicates that the activity is expected to require between 5% and 10% of the total budget.

Each random sample drawn from the distributions on $\delta_{\text{Fan Noise}}$ and $\alpha_{\text{Fan Noise}}$ was used in Eq. (15) to effect a change of the fan noise technology impact k-factor. The baseline distribution on $k_{\text{Fan Noise}}$ is plotted in Fig. 102a, along with two transformed versions that correspond with two random samples of mean translation and variance

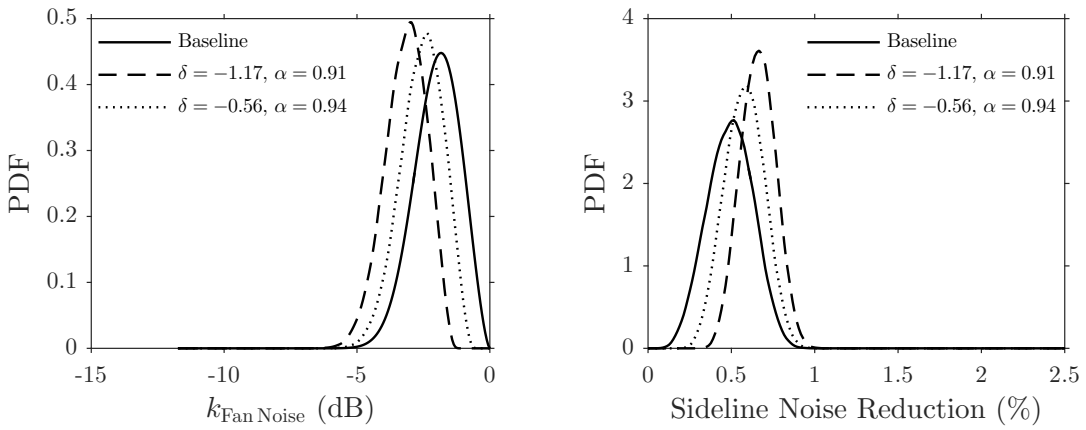


(a) Mean Translation (b) Variance Scaling (c) Cost of the experiment

Figure 101: Uniform distributions used to represent the effects of conducting a computer experiment for the fan vertical acoustic splitter technology.

scaling. The dashed line PDF was created with a mean shift of -1.17 dB and a variance reduction of $100(1 - 0.91^2) = 17.19\%$, whereas the dotted line had a smaller mean shift of -0.56 dB and a smaller variance reduction of $100(1 - 0.94^2) = 11.64\%$. A sample of 10,000 such $k_{\text{Fan Noise}}$ distributions were generated in a similar way. For each $k_{\text{Fan Noise}}$ distribution and the baseline distributions shown in Fig. 22a, Fig. 22b, and Fig. 22c, the EDS M&S environment surrogate models were used to propagate uncertainty to the system-level metrics. As expected, the primary effect of the various realizations of $k_{\text{Fan Noise}}$ distributions was observed for the marginal distribution characteristics of the sideline noise metric. The baseline sideline noise distribution is plotted in Fig. 102b, along with the distributions associated with two versions of the transformed distributions on $k_{\text{Fan Noise}}$. By comparing the PDF shapes of $k_{\text{Fan Noise}}$ and sideline noise reduction, one will notice an intuitive trend: as $\delta_{\text{Fan Noise}}$ and $\alpha_{\text{Fan Noise}}$ simultaneously decreased to shift $k_{\text{Fan Noise}}$ toward lower noise with lower variance, the impact propagated to better sideline noise performance with lower variance.

At the next higher level of the propagation hierarchy were the attributes. For each of the 10,000 distributions of the system-level metrics, the performance attribute $\text{P}(D \geq D_{\text{Target}})$ and average variance reduction attribute were quantified. The histograms of these attributes are shown in Fig. 103. Cost is also an attribute, and the distribution on this attribute is shown in Fig. 101c. The performance attribute values lying between 0.16 and 0.24 were small, considering that the minimum and

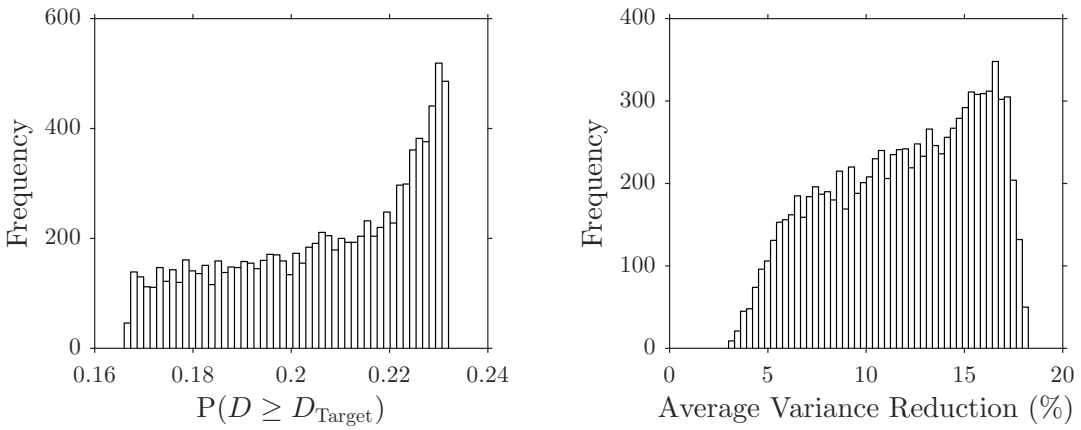


(a) Noise technology impact distributions (b) Noise system-level metric distributions

Figure 102: Baseline and modified noise technology impact distributions and the corresponding system-level noise marginal distributions.

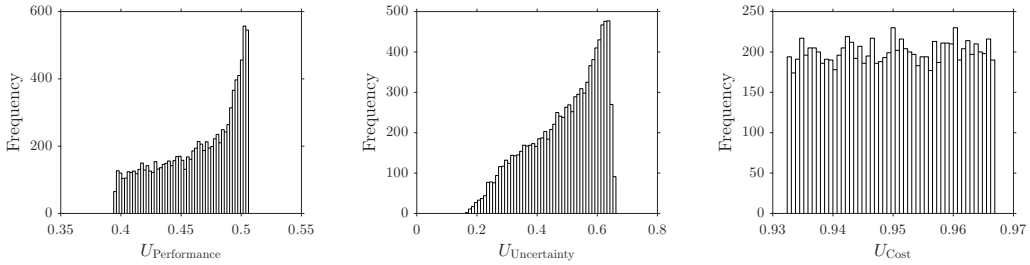
maximum possible values for the attribute were 0 and 1, respectively. Considerable average variance reduction values were quantified given that the uncertainty in the marginal distributions on fuel burn reduction and TOFL reduction were unaffected by A_1 and that the uncertainty reduction attribute was defined as an average variance reduction of all three system-level metrics.

With the attribute distributions characterized, the next step was to propagate these distributions through the single-attribute utility functions show in Fig. 26. The results of propagation were the three single-attribute utility distributions shown in Fig. 104. Comparing the three utility distributions, it is apparent that the cost utility was highest of the three for A_1 due to the low cost of the computer experiment that was modeled. The skewed shape of the performance and uncertainty reduction utility distributions was due to a combination of the skew of the attribute distributions and the high slope of the single-attribute utility functions over the range of the samples from the attribute distributions. The final step in the propagation process was to push the three single-attribute utility distributions through Eq. (10) to compute a distribution on multiattribute utility. The resulting distribution is plotted in Fig. 105, and the vertical dashed line marks the expected utility for A_1, which was estimated



(a) Performance attribute (b) Uncertainty reduction attribute

Figure 103: Histograms of the performance and uncertainty reduction attributes that summarize changes in the system-level metric distributions due to the effects of A_1.



(a) Performance (b) Uncertainty reduction (c) Cost

Figure 104: Histograms of the single-attribute utilities.

by computed the sample mean. Comparing the multiattribute utility distribution with the three single-attribute utility distributions, it is clear that the magnitude of the multiattribute utility samples would not have been nearly as high without the contribution of the high cost utilities. The performance and uncertainty reduction utilities had the effect of discounting the high cost utilities. The relatively large scaling constant value for cost from Table 8 also had a role due to the high weighting of the cost utility.

The process illustrated here for quantifying expected utility was repeated for the three other alternatives to produce the results shown in Fig. 29. For the sensitivity analyses described in Sec. 4.4.4, this process was followed to compute the expected

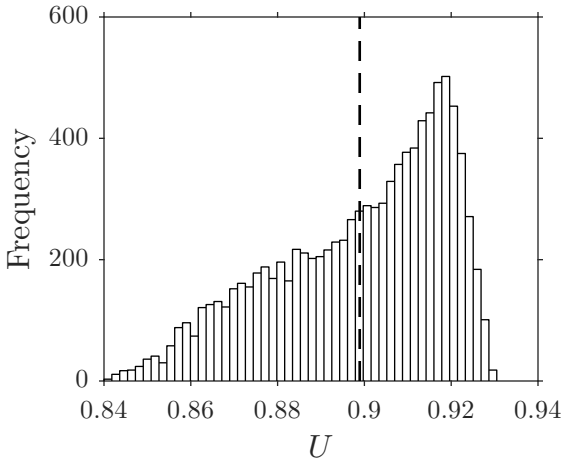


Figure 105: Histogram of multiattribute utility with the expected utility for A_1 indicated by the vertical dashed line.

utilities of all four alternatives at 20 different levels of the indifference probabilities (Scenarios 1 and 2) and the uncertainty reduction scaling constant (Scenario 3).

REFERENCES

[1] Whittenbury, J., "Configuration Design Development of the Navy UCAS-D X-47B," *AIAA Centennial of Naval Aviation Forum "100 Years of Achievement and Progress"*, AIAA Paper 2011-7041, 2011.

[2] Guynn, M. D., Berton, J. J., Tong, M. J., and Haller, W. J., "Advanced Single-Aisle Transport Propulsion Design Options Revisited," *2013 Aviation Technology, Integration, and Operations Conference*, AIAA Paper 2013-4330, 2013.

[3] Gad-El-Hak, M., *Flow Control: Passive, Active, and Reactive Flow Management*, Cambridge University Press, New York, 2000.

[4] Prandtl, L., "Über Flüssigkeitsbewegung bei sehr kleiner Reibung," *Proceedings of the Third International Mathematical Congress*, 1904, pp. 484–491.

[5] Williams, D. R. and MacMynowski, D. G., "Brief History of Flow Control," *Fundamentals and Applications of Modern Flow Control*, edited by R. D. Joslin and D. N. Miller, chap. 1, Progress in Astronautics and Aeronautics, AIAA, Reston, VA, 2009, pp. 1–20.

[6] Liddle, S. C., Jabbal, M., and Crowther, W. J., "Systems and certification issues for civil transport aircraft flow control systems," *The Aeronautical Journal*, Vol. 113, No. 1147, 2009, pp. 575–586.

[7] Moorhouse, D. J., "Detailed Definitions and Guidance for Application of Technology Readiness Levels," *Journal of Aircraft*, Vol. 39, No. 1, 2002, pp. 190–192.

[8] Smaling, R. and Weck, O. D., "Assessing Risks and Opportunities of Technology Infusion in System Design," *Systems Engineering*, Vol. 10, No. 1, 2007, pp. 1–25.

[9] Mankins, J. C., "Technology readiness and risk assessments: A new approach," *Acta Astronautica*, Vol. 65, No. 9-10, 2009, pp. 1208–1215.

[10] Aven, T., "A unified framework for risk and vulnerability analysis covering both safety and security," *Reliability Engineering & System Safety*, Vol. 92, No. 6, 2007, pp. 745–754.

[11] Nikolaidis, E., "Types of Uncertainty in Design Decision Making," *Engineering Design Reliability Handbook*, edited by E. Nikolaidis, D. M. Ghiocel, and S. Singhal, CRC Press, Boca Raton, FL, 2005.

[12] Oberkampf, W. and Roy, C., *Verification and Validation in Scientific Computing*, Cambridge University Press, New York, 2010.

[13] Zang, T. A., "On the expression of uncertainty intervals in engineering," *Theoretical and Computational Fluid Dynamics*, Vol. 26, No. 5, 2012, pp. 403–414.

[14] Coleman, H. W. and Steele, W. G., "Engineering Application of Experimental Uncertainty Analysis," *AIAA Journal*, Vol. 33, No. 10, 1995, pp. 1888–1896.

[15] Roy, C. J. and Oberkampf, W. L., "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing," *Computer Methods in Applied Mechanics and Engineering*, Vol. 200, No. 25-28, 2011, pp. 2131–2144.

[16] Kirby, M. R. and Mavris, D. N., "Forecasting Technology Uncertainty in Preliminary Aircraft Design," *1999 World Aviation Conference*, SAE Paper 1999-01-5631, 1999.

[17] GAO, "Stronger Practices Needed to Improve DOD Technology Transition Processes," GAO-06-883, 2006.

[18] GAO, "Technology Transition Programs Support Military Users, but Opportunities Exist to Improve Measurement of Outcomes," GAO-13-286, 2013.

[19] Dieter, G. E. and Schmidt, L. C., *Engineering Design*, McGraw-Hill, New York, 4th ed., 2009.

[20] Lin, J. C., Whalen, E. A., Eppink, J. L., Siochi, E. J., Alexander, M. G., and Andino, M. Y., "Innovative Flow Control Concepts for Drag Reduction," *54th AIAA Aerospace Sciences Meeting*, AIAA Paper 2016-0864, 2016.

[21] Largent, M. C., "A Probabilistic Risk Management Based Process For Planning And Management Of Technology Development," Ph.D. Dissertation, School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 2003.

[22] Mankins, J. C., "Technology readiness assessments: A retrospective," *Acta Astronautica*, Vol. 65, No. 9-10, 2009, pp. 1216–1223.

[23] Mankins, J. C., "Technology Readiness Levels, A White Paper," NASA, 1995.

[24] "NASA Systems Engineering Handbook," NASA/SP-2007-6105 Rev 1, 2007.

[25] Mankins, J. C., "Research & Development Degree of Difficulty (R&D^3), A White Paper," NASA, 1998.

[26] Gatian, K. N., "A Quantitative, Model-Driven Approach to Technology Selection and Development Through Epistemic Uncertainty Reduction," Ph.D. Dissertation, School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 2015.

[27] Wong, P., "Application of Decision Theory to the Testing of Large Systems," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-7, No. 2, 1971, pp. 379–384.

[28] Thomke, S. H., "Managing Experimentation in the Design of New Products," *Management Science*, Vol. 44, No. 6, 1998, pp. 743–762.

[29] Thomke, S. and Bell, D. E., "Sequential Testing in Product Development," *Management Science*, Vol. 47, No. 2, 2001, pp. 308–323.

[30] Loch, C. H., Terwiesch, C., and Thomke, S., "Parallel and Sequential Testing of Design Alternatives," *Management Science*, Vol. 47, No. 5, 2001, pp. 663–678.

[31] Shannon, C. E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, No. 3, 1948, pp. 379–423 and 623–656.

[32] Urbina, A., Mahadevan, S., and Paez, T., "Resource Allocation Using Quantification of Margins and Uncertainty," *51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, AIAA Paper 2010-2510, 2010.

[33] Hess, J. T. and Valerdi, R., "Test and Evaluation of a SoS using a Prescriptive and Adaptive Testing Framework," *2010 5th International Conference on System of Systems Engineering*, IEEE, 2010, pp. 1–6.

[34] Sankararaman, S., Mahadevan, S., McLemore, K., Bradford, S., Liang, C., and Peterson, L., "Test Resource Allocation for Uncertainty Quantification of Multilevel and Coupled Systems," *52nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, AIAA Paper 2011-1904, 2011.

[35] Sankararaman, S., "Uncertainty Quantification and Integration in Engineering Systems," Ph.D. Dissertation, Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, 2012.

[36] Sankararaman, S., McLemore, K., Mahadevan, S., Bradford, S. C., and Peterson, L. D., "Test Resource Allocation in Hierarchical Systems Using Bayesian Networks," *AIAA Journal*, Vol. 51, No. 3, 2013, pp. 537–550.

[37] Bjorkman, E. A., "Test and Evaluation Resource Allocation Using Uncertainty Reduction as a Measure of Test Value," Ph.D. Dissertation, School of Engineering and Applied Science, George Washington University, Washington, DC, 2012.

[38] Bjorkman, E. A., Sarkani, S., and Mazzuchi, T. A., "Test and Evaluation Resource Allocation Using Uncertainty Reduction," *IEEE Transactions on Engineering Management*, Vol. 60, No. 3, 2013, pp. 541–551.

[39] GAO, "Defense Acquisitions: Assessments of Selected Weapon Programs," GAO-16-329SP, 2016.

[40] Lin, J. C., Andino, M. Y., Alexander, M. G., Whalen, E. A., Spoor, M. A., Tran, J. T., and Wygnanski, I. J., "An Overview of Active Flow Control Enhanced Vertical Tail Technology Development," *54th AIAA Aerospace Sciences Meeting*, AIAA Paper 2016-0056, 2016.

[41] Sauser, B., Ramirez-Marquez, J., Verma, D., and Gove, R., "From TRL to SRL: The Concept of Systems Readiness Levels," *Conference on Systems Engineering Research*, Paper #126, 2006.

[42] Sorensen, R., *Thought Experiments*, Oxford University Press, New York, 1998.

[43] Gendler, T. S., "Thought Experiments Rethoughtand Reperceived," *Philosophy of Science*, Vol. 71, No. 5, 2004, pp. 1152–1163.

[44] Shepard, R., "The Step to Rationality: The Efficacy of Thought Experiments in Science, Ethics, and Free Will," *Cognitive Science*, Vol. 32, No. 1, 2008, pp. 3–35.

[45] Montgomery, D. C., "Experimental Design for Product and Process Design and Development," *Journal of the Royal Statistical Society: Series D (The Statistician)*, Vol. 48, No. 2, 1999, pp. 159–177.

[46] Mavris, D. N., Baker, A. P., and Schrage, D. P., "IPPD Through Robust Design Simulation for an Affordable Short Haul Civil Tiltrotor," *American Helicopter Society 53rd Annual Forum*, 1997.

[47] Mooney, H. P., Brandt, J. B., Lacy, D. S., and Whalen, E. A., "AFC-Enabled Vertical Tail System Integration Study," NASA/CR-2014-218168, 2014.

[48] Tversky, A. and Kahneman, D., "Judgment under Uncertainty: Heuristics and Biases," *Science*, Vol. 185, No. 4157, 1974, pp. 1124–1131.

[49] Hazelrigg, G. A., "A Framework for Decision-Based Engineering Design," *Journal of Mechanical Design*, Vol. 120, No. 4, 1998, pp. 653–658.

[50] Dyer, J. S., Fishburn, P. C., Steuer, R. E., Wallenius, J., and Zionts, S., "Multiple Criteria Decision Making, Multiattribute Utility Theory: The Next Ten Years," *Management Science*, Vol. 38, No. 5, 1992, pp. 645–654.

[51] Saaty, T. L., *Decision Making for Leaders*, RWS Publications, Pittsburgh, 1999.

[52] Keeney, R. L. and Raiffa, H., *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, Cambridge University Press, Cambridge, England, 1993.

[53] Lafleur, J. M., "Probabilistic AHP and TOPSIS for Multi-Attribute Decision-Making under Uncertainty," *IEEE Aerospace Conference*, 2011.

[54] Loken, E., Botterud, A., and Holen, A. T., "Decision Analysis and Uncertainties in Planning Local Energy Systems," *9th International Conference on Probabilistic Methods Applied to Power Systems*, 2006.

[55] Thurston, D. L., "Real and Misconceived Limitations to Decision Based Design With Utility Analysis," *Journal of Mechanical Design*, Vol. 123, No. 2, 2001, pp. 176–182.

[56] Gass, S. I., "Model World: The Great DebateMAUT Versus AHP," *Interfaces*, Vol. 35, No. 4, 2005, pp. 308–312.

[57] Yoon, K., "Systems Selection by Multiple Attribute Decision Making," Ph.D. Dissertation, Department of Industrial Engineering, Kansas State University, Manhattan, KS, 1980.

[58] von Neumann, J. and Morgenstern, O., *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, 2nd ed., 1947.

[59] Keeney, R. L., "Decision Analysis: An Overview," *Operations Research*, Vol. 30, No. 5, 1982, pp. 803–838.

[60] Bigwood, M. P., "Managing the New Technology Exploitation Process," *Research-Technology Management*, Vol. 47, No. 6, 2004, pp. 38–42.

[61] Derringer, G. C., "A Balancing Act: Optimizing a Product's Properties," *Quality Progress*, Vol. 27, No. 6, 1994, pp. 51–58.

[62] Derringer, G. and Suich, R., "Simultaneous Optimization of Several Response Variables," *Journal of Quality Technology*, Vol. 12, No. 4, 1980, pp. 214–219.

[63] Kruithof, J., "Telefoonverkeersrekening," *De Ingenieur*, Vol. 52, No. 8, 1937, pp. E15–E25.

[64] Cooke, R. M., "Parameter fitting for uncertain models: modelling uncertainty in small models," *Reliability Engineering & System Safety*, Vol. 44, No. 1, 1994, pp. 89–102.

[65] Du, C., Kurowicka, D., and Cooke, R. M., "Techniques for generic probabilistic inversion," *Computational Statistics & Data Analysis*, Vol. 50, No. 5, 2006, pp. 1164–1187.

[66] Lee, S. H. and Chen, W., "A comparative study of uncertainty propagation methods for black-box-type problems," *Structural and Multidisciplinary Optimization*, Vol. 37, No. 3, 2009, pp. 239–253.

[67] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*, Wiley, 2nd ed., 2006.

[68] Csiszar, I., "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, Vol. 3, No. 1, 1975, pp. 146–158.

[69] Cooke, R. M., Nauta, M., Havelaar, A. H., and van der Fels, I., "Probabilistic inversion for chicken processing lines," *Reliability Engineering & System Safety*, Vol. 91, No. 10-11, 2006, pp. 1364–1372.

[70] Pratt, J. W., "Risk Aversion in the Small and in the Large," *Econometrica*, Vol. 32, No. 1/2, 1964, pp. 122–136.

[71] Harsanyi, J. C., "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, Vol. 63, No. 4, 1955, pp. 309–321.

[72] Fishburn, P. C., *The Theory of Social Choice*, Princeton University Press, Princeton, NJ, 1973.

[73] Gerchak, Y. and Mossman, D., "On the Effect of Demand Randomness on Inventories and Costs," *Operations Research*, Vol. 40, No. 4, 1992, pp. 804–807.

[74] Garthwaite, P. H., Kadane, J. B., and O'Hagan, A., "Statistical Methods for Eliciting Probability Distributions," *Journal of the American Statistical Association*, Vol. 100, No. 470, 2005, pp. 680–700.

[75] Genest, C. and Zidek, J. V., "Combining Probability Distributions: A Critique and an Annotated Bibliography," 1986.

[76] Kirby, M. R. and Mavris, D. N., "The Environmental Design Space," *26th International Congress of the Aeronautical Sciences*, ICAS Paper 2008-4.7.3, 2008.

[77] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S., *Global Sensitivity Analysis: The Primer*, John Wiley, Chichester, England, 2008.

[78] Anderson-Cook, C. M., "Opportunities and Issues in Multiple Data Type Meta-Analyses," *Quality Engineering*, Vol. 21, No. 3, 2009, pp. 241–253.

[79] Pan, S. J. and Yang, Q., "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 10, 2010, pp. 1345–1359.

[80] Lindley, D. V., "On a Measure of the Information Provided by an Experiment," *The Annals of Mathematical Statistics*, Vol. 27, No. 4, 1956, pp. 986–1005.

[81] DeGroot, M. H., "Uncertainty, Information, and Sequential Experiments," *The Annals of Mathematical Statistics*, Vol. 33, No. 2, 1962, pp. 404–419.

[82] Bernardo, J. M., "Expected Information as Expected Utility," *The Annals of Statistics*, Vol. 7, No. 3, 1979, pp. 686–690.

[83] Sebastiani, P. and Wynn, H. P., "Maximum entropy sampling and optimal Bayesian experimental design," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 62, No. 1, 2000, pp. 145–157.

[84] Wolpert, D. H., "The Lack of A Priori Distinctions Between Learning Algorithms," *Neural Computation*, Vol. 8, No. 7, 1996, pp. 1341–1390.

[85] Wickham, H., "Tidy Data," *Journal of Statistical Software*, Vol. 59, No. 10, 2014, pp. 1–23.

[86] Rasmussen, C. E. and Williams, C. K. I., *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006.

[87] Murphy, K. P., *Machine Learning: A Probabilistic Perspective*, The MIT Press, Cambridge, MA, 2012.

[88] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., "Design and Analysis of Computer Experiments," *Statistical Science*, Vol. 4, No. 4, 1989, pp. 409–423.

[89] Santner, T. J., Williams, B. J., and Notz, W. I., *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.

[90] Gramacy, R. B. and Lee, H. K. H., "Cases for the nugget in modeling computer experiments," *Statistics and Computing*, Vol. 22, No. 3, 2012, pp. 713–722.

[91] Caruana, R., "Multitask Learning," *Machine Learning*, Vol. 28, No. 1, 1997, pp. 41–75.

[92] Lawrence, N. D. and Platt, J. C., "Learning to Learn with the Informative Vector Machine," *21st International Conference on Machine Learning*, 2004.

[93] Lawrence, N., Seeger, M., and Herbrich, R., "Fast Sparse Gaussian Process Methods: The Informative Vector Machine," *Advances in Neural Information Processing Systems 15*, 2003, pp. 625–632.

[94] Menzefricke, U., "Hierarchical Modeling with Gaussian Processes," *Communications in Statistics—Simulation and Computation*, Vol. 29, No. 4, 2000, pp. 1089–1108.

[95] Schwaighofer, A., Tresp, V., and Yu, K., "Learning Gaussian Process Kernels via Hierarchical Bayes," *Advances in Neural Information Processing Systems 17*, 2005, pp. 1209–1216.

[96] Kennedy, M. C. and O'Hagan, A., "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, Vol. 87, No. 1, 2000, pp. 1–13.

[97] Bonilla, E. V., Chai, K. M. A., and Williams, C. K. I., "Multi-task Gaussian Process Prediction," *Advances in Neural Information Processing Systems 20*, 2008, pp. 153–160.

[98] Álvarez, M. and Lawrence, N. D., "Sparse Convolved Gaussian Processes for Multi-output Regression," *Advances in Neural Information Processing Systems 21*, 2009, pp. 57–64.

[99] Álvarez, M. A. and Lawrence, N. D., "Computationally Efficient Convolved Multiple Output Gaussian Processes," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 1459–1500.

[100] Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G., "To Transfer or Not To Transfer," *Advances in Neural Information Processing Systems, NIPS '05 Workshop, Inductive Transfer: 10 Years Later*, 2005.

[101] Toal, D. J. J., "Some considerations regarding the use of multi-fidelity Kriging in the construction of surrogate models," *Structural and Multidisciplinary Optimization*, Vol. 51, No. 6, 2015, pp. 1223–1245.

[102] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2nd ed., 2009.

[103] Conrow, E. H., "Estimating Technology Readiness Level Coefficients," *Journal of Spacecraft and Rockets*, Vol. 48, No. 1, 2011, pp. 146–152.

[104] MATLAB version 8.6.0.267246 (R2015b), The MathWorks Inc., Natick, Massachusetts, 2015.

[105] MT-IVM version 0.142, `https://github.com/lawrennd/mtivm`, retrieved 2-18-2016.

[106] MTGP, `https://github.com/ebonilla/mtgp`, retrieved 2-22-2016.

[107] Forrester, A. I. J., Sóbester, A., and Keane, A. J., "Multi-fidelity optimization via surrogate modelling," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 463, No. 2088, 2007, pp. 3251–3269.

[108] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015.

[109] Gratiet, L. L., *MuFiCokriging: Multi-Fidelity Cokriging models*, 2012, R package version 1.2.

[110] Roustant, O., Ginsbourger, D., and Deville, Y., "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization," *Journal of Statistical Software*, Vol. 51, No. 1, 2012, pp. 1–55.

[111] MULTIGP version 0.13, `https://github.com/SheffieldML/multigp`, retrieved 2-22-2016.

[112] Branin, F. H., "Widely Convergent Method for Finding Multiple Solutions of Simultaneous Nonlinear Equations," *IBM Journal of Research and Development*, Vol. 16, No. 5, 1972, pp. 504–522.

[113] Hedar, A.-R., "Global Optimization Test Problems,", `http://www-optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar_files/TestGO_files/Page2904.htm`, retrieved 3-10-2016.

[114] Loeppky, J. L., Sacks, J., and Welch, W. J., "Choosing the Sample Size of a Computer Experiment: A Practical Guide," *Technometrics*, Vol. 51, No. 4, 2009, pp. 366–376.

[115] Chai, K. M. A., *Multi-task Learning with Gaussian Processes*, Ph.D. thesis, University of Edinburgh, 2010.

[116] Rasmussen, C. E. and Nickisch, H., "Gaussian Processes for Machine Learning (GPML) Toolbox," *Journal of Machine Learning Research*, Vol. 11, 2010, pp. 3011–3015.

[117] ISO, *ISO 8402 International Standard: Quality Vocabulary*, Geneva, Switzerland, 1986.

[118] Fabrycky, W. J. and Blanchard, B. S., *Life-Cycle Cost and Economic Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1991.

[119] Hamada, M. S., Wilson, A. G., Reese, C. S., and Martz, H. F., *Bayesian Reliability*, Springer, New York, NY, 2008.

[120] Whitmore, G. A., Young, K. D. S., and Kimber, A. C., "Two-Stage Reliability Tests with Technological Evolution: A Bayesian Analysis," *Applied Statistics*, Vol. 43, No. 2, 1994, pp. 295–307.

[121] Young, K. D. S., "A Bayesian Analysis of Updated Component Data," *The Statistician*, Vol. 43, No. 1, 1994, pp. 129–137.

[122] Huang, Z. and Jin, Y., "Validation and Adjustment of Prior and Data for Bayesian Reliability Analysis in Engineering Design," *Journal of Mechanical Design*, Vol. 133, No. 5, 2011, pp. 051003–1–051003–12.

[123] Peng, W., Huang, H.-Z., Li, Y., Zuo, M. J., and Xie, M., "Life cycle reliability assessment of new products—A Bayesian model updating approach," *Reliability Engineering & System Safety*, Vol. 112, 2013, pp. 109–119.

[124] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B., *Bayesian Data Analysis*, CRC Press, Boca Raton, FL, 3rd ed., 2014.

[125] Azzalini, A. and Dalla Valle, A., "The multivariate skew-normal distribution," *Biometrika*, Vol. 83, No. 4, 1996, pp. 715–726.

[126] Hu, F. and Zidek, J., "The weighted likelihood," *The Canadian Journal of Statistics*, Vol. 30, No. 3, 2002, pp. 347–371.

VITA

Ryan B. Jacobs was born in Plantation, Florida on November 14, 1985. He received a bachelor's degree in aerospace engineering from Embry-Riddle Aeronautical University in Daytona Beach, Florida in 2008. Thereafter, he worked for QuEST Global as a project engineer conducting high-fidelity analysis and design for gas turbine components. As his interest in engineering design research grew, he decided to enter the graduate program in the Aerospace Systems Design Laboratory at Georgia Tech. He received his master's degree in aerospace engineering in 2012, and he continued in the Ph.D. program. During his time at Georgia Tech, he conducted research with several sponsors including NASA, Raytheon, and Boeing. He graduated with a Ph.D. in aerospace engineering, with a minor in mathematics, from Georgia Tech in 2016.