



**VALIDATION OF DISCRETE AND FUNCTIONAL SIMULATION RESPONSES
OVER EXPERIMENTAL REGIONS USING RESPONSE SURFACES**

DISSERTATION

Scott M. Storm, Major, USAF

AFIT-ENS-DS-17-S-044

**DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY**

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-DS-17-S-044

VALIDATION OF DISCRETE AND FUNCTIONAL SIMULATION RESPONSES
OVER EXPERIMENTAL REGIONS USING RESPONSE SURFACES

DISSERTATION

Presented to the Faculty
Department of Operational Sciences
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Operations Research

Scott M. Storm, B.S., M.S.

Major, USAF

September 2017

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

AFIT-ENS-DS-17-S-044

VALIDATION OF DISCRETE AND FUNCTIONAL SIMULATION RESPONSES
OVER EXPERIMENTAL REGIONS USING RESPONSE SURFACES

Scott M. Storm, B.S., M.S.
Major, USAF

Committee Membership:

Raymond R. Hill, Ph.D
Chair

Joseph J. Pignatiello Jr., Ph.D
Member

G. Geoffrey Vining, Ph.D.
Member

Edward D. White III, Ph.D.
Member

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

Abstract

The Director of Test and Evaluation has placed great emphasis on the use of experimental design to properly collect and analyze data from military assets. This emphasis on statistical based data collection plays a large role in the validation of simulation models. It is prudent to validate simulation models across these experimental regions to ensure accuracy at all desirable configurations, to include points within the region where no experimental data has been collected. This research proposes the use of response surfaces to capture the behavior of the response across the experimental region. The first paper develops a methodology for point-wise validation of model response output at unique points within an experimental region using regression based tolerance intervals. The second paper develops an F-statistic for validation of the simulation model over the entire region based on a ratio of mean square errors constructed from the model and system data with regards to a common response surface. The third paper proposes an extension to the Correlation Analysis (CORA) objective rating system that uses functional regression techniques to allow for assessments of functional response data when no system data is available. The fourth paper applies each methodology to the Sandia thermal challenge problem.

To my wife and children for their unconditional love and support

Acknowledgments

I would like to thank Dr. Raymond Hill for providing guidance and mentor-ship throughout this arduous but rewarding process. I extend my thanks to the other members of my research committee, Dr. Joseph Pignatiello, Dr. Geoffrey Vining, and Dr. Edward White for their unwavering belief in my abilities as a researcher. Lastly, I would like to thank my fellow members of the ENS PhD cohort for always allowing me to vocalize new ideas and even more so for always giving brutally honest feedback.

Scott M. Storm

Table of Contents

	Page
Abstract	iv
Dedication	v
Acknowledgments	vi
Table of Contents	vii
List of Figures	xi
List of Tables	xv
 I. Introduction	 1
1.1 Background	1
1.2 Research Focus	2
1.3 Overview	4
 II. Review of Literature	 7
2.1 Definitions	8
2.1.1 Modeling and Simulation	8
2.1.2 Verification and Validation	8
2.1.3 Assessment, Credibility, and Accreditation	10
2.1.4 Validation Metric	10
2.1.5 Epistemic versus Aleatory Uncertainty	12
2.2 The Modeling Process and How It Relates To V&V	12
2.2.1 The Simple Modeling Process	12
2.2.2 VV&A Principles	14
2.2.3 Common Questions When Validating a Model	19
2.3 Validation Taxonomy	21
2.3.1 Davis Validation Taxonomy	21
2.3.2 Balci VV&T Taxonomy	21
2.3.3 Sargent’s Operational Validity Classification Taxonomy	23
2.4 Statistical Validation Techniques	25
2.4.1 Analysis of Variance	25
2.4.2 Two-Sample <i>t</i> -Test	26
2.4.3 Confidence Intervals/Bounds	26

	Page
2.4.4 Analysis of Paired Observations	27
2.4.5 Factor Analysis	28
2.4.6 Hotelling's T^2 Tests	29
2.4.7 Multivariate Analysis of Variance	30
2.4.8 Nonparametric Goodness-of-fit Tests	30
2.4.8.1 Kolmogorov-Smirnov Test	30
2.4.8.2 Cramer-Von Mises Test	31
2.4.8.3 Chi-square Test	32
2.4.9 Nonparametric Tests of Means	33
2.4.9.1 Mann-Whitney-Wilcoxon Test	33
2.4.10 Regression Analysis	34
2.4.11 Theil's Inequality Coefficient	36
2.4.12 Time Series Analysis	37
2.4.12.1 Correlation/Spectral Analysis	37
2.4.12.2 Error Analysis	40
2.5 Current Research Efforts	45
2.5.1 Wavelets	45
2.5.2 EARTH Method	49
2.5.3 Hypothesis Testing on Intervals	53
2.5.4 Validation in Engineering Applications	58
2.5.4.1 Classical Hypothesis Testing	59
2.5.4.2 Bayesian Validation Metrics	73
2.5.4.3 Frequentest Metrics	84
2.5.4.4 Area Metrics	87
2.5.5 Correlation and Analysis (CORA) Method	94
2.6 Summary and Finding of Literature Review	94
 III. Point-Wise Model Validation Over a Design Space Using Regression Tolerance Intervals and Bayesian Hypothesis Testing	 96
3.1 Introduction	96
3.2 Validation Literature	97
3.3 Methodology	100
3.3.1 Hypothesis Under Test and Corresponding Assumptions	100
3.3.2 Fitting a Response Surface to Experimental Data	101
3.3.3 Regression Tolerance Intervals	102
3.3.4 Model Assessment at Tested/Untested Design Points	104
3.3.5 Using Bayesian Hypothesis Testing to Establish Minimum P	106
3.4 Example	109
3.4.1 Method Sensitivity	111
3.5 Considerations	114
3.6 Conclusion	117

	Page
IV. F-Statistic for Model Validation Over Experimental Regions Using Least Squares Response Surfaces	118
4.1 Introduction	118
4.2 Literature Review	119
4.3 Methodology	121
4.4 Example	127
4.5 Considerations	132
4.6 Conclusion	135
V. Model Validation of Discretized Functional Response Data Across Experimental Regions Using Functional Regression Extensions to the CORA Objective Rating System	136
5.1 Introduction	136
5.2 Literature Review	137
5.3 Functional Regression and the CORA Objective Rating System	139
5.3.1 Functional Regression	140
5.3.2 CORA Objective Rating System	141
5.4 Extensions to CORA Objective Rating System	146
5.5 Improvement to CORA Size Rating	152
5.6 Considerations	157
5.7 Conclusion	158
VI. Application of Response Surface Based Validation Methodologies Using the Sandia Thermal Challenge Problem	160
6.1 Introduction	160
6.2 Problem Development	160
6.2.1 Mathematical Model and Material Characterization	160
6.2.2 Ensemble Validation	161
6.2.3 Generation of Model Data	161
6.2.4 Generation and Evaluation of Response Surface	164
6.3 Point-Wise Model Validation Using Regression Tolerance Intervals	167
6.3.1 Combined Statistical Interval Approach to Model Validation	171
6.4 F-Statistic for Overall Validity Across the Experimental Region	174
6.5 Functional Regression Extension to the CORA Objective Rating System	176
6.6 Conclusion	185
VII. Conclusion	186
7.1 Contributions	186

	Page
7.2 Future Research	188
Bibliography	191

List of Figures

Figure	Page
2.1 Simplified Modeling Process [105]	13
2.2 7-Step Approach for Conducting a Successful Simulation Study [75]	14
2.3 Model Confidence versus Cost [107]	16
2.4 Logic Paths to Type I, II, III Errors [6]	18
2.5 Validation portion of Davis VV&A Taxonomy [27]	22
2.6 Balci VV&T Taxonomy [6]	23
2.7 Sargent's Operational Validity Classification [108]	24
2.8 Example of Deviations with Corresponding Envelope of Acceptable Precision [82]	37
2.9 Plot of Cross Correlation as a Function of Lag [9]	39
2.10 Dynamic Time Warping Example [110]	44
2.11 Illustrative Example of Wavelet Decomposition with Approximations and Details [22]	46
2.12 Procedure for Automatic Mult-Level Comparison [22]	49
2.13 Wavelet Decomposition in Crash Test Data [22]	50
2.14 Time Series Data for Head Acceleration [110]	53
2.15 Graphical Comparison Between EARTH Metric and SME Ranking [110]	55
2.16 Graphical Comparison Between Various Error Metrics [110]	56
2.17 Risk Curve for Two β Options [109]	57
2.18 Risk Curve for Two Sample Size Options [109]	58
2.19 Single Factor Comparison of Experimental Results Against Prediction with Error Bounds for Damped Spring-Mass System [52]	61
2.20 Validation of Single Measurement with $\omega = 1.5$ [52]	62

Figure	Page
2.21 Validation of Two-Dimension Problem with $\omega = 1.0$ and $\omega = 1.5$ [52]	63
2.22 95% Confidence Ellipse and Predictions with Varying Parameter Values (Straight Line) [52]	65
2.23 95% Confidence Ellipse on Predictions with Rectangular Region Around Measurement for Additional Acceptable Error [52]	66
2.24 Procedure for Model Validation [15]	68
2.25 PDFs for Each of Five Design Points [15]	70
2.26 Response PDF Without(Left) and With(Right) Transformation [15]	70
2.27 Response PDF With RSM Errors and Acceptable Level of Error [15]	71
2.28 PDF for Probability of Failure [98]	74
2.29 Ratio of Posterior to Prior Densities [98]	75
2.30 Component Energies for Eight Coefficients [98]	80
2.31 Proposed Bayesian Prediction Model Validation Framework [16]	82
2.32 Prediction of $\delta(x)$ with Corresponding Prediction Intervals and Differenced Data [120]	85
2.33 Three-Dimensional Confidence Region on Regression Coefficients [88]	88
2.34 Example Results with 90% CIs on Regression Curve [88]	89
2.35 Example Distribution Curves and associated shaded differences with $n = 1$ and $n = 4$ [37]	90
2.36 Comparison of U-Pooled Values Against Standard Uniform [37]	91
2.37 Area Metric Using U-Pooling and Back-transformation [37]	92
2.38 Four Examples for Interval (P-Box) Extension to Area Metric [36]	94
3.1 Tolerance Interval Assessment of Model Response Distribution	106
3.2 Validation Assessment of Model Replications ($x_3 = 0$)	113
3.3 Method Sensitivity to Model Bias and Error Variance	114

Figure	Page
3.4 Notional System and Model Response Distributions with Tolerance Intervals (Dashed Lines) Set to $\mu_s \pm 2\sigma_s$	116
4.1 Observed and Hypothesized Probability Density Functions for Doornik- Hansen Multivariate Test Statistic	124
4.2 Evaluation of Model Data Using MSE from Response Surface	127
4.3 Visual Representation of Model and System Data With Respect to the Fitted Surface	130
4.4 F-Statistic Sensitivity to Model Bias(a) and Model Error(b)	131
4.5 Percent Valid Assessments as a Function of Model Bias and Error Variance . .	132
4.6 Validity Contours with Differing σ_s	133
4.7 Confidence of Validation Method with Increasing Model Observations	134
5.1 Structure for CORA Objective Rating System [43]	142
5.2 Constant and Non-Constant Corridor Widths [43]	142
5.3 Three Measures of Signal Disagreement [113]	144
5.4 Extended CORA Methodology for Model Validation Over Experimental Regions	150
5.5 Notional Example of Using CORA Ratings to Isolate Poor Model Performance Within Experimental Regions	151
5.6 Different Model and System Responses with Equal Area	153
5.7 Sensitivity of Shape Rating to Shifts in Magnitude	153
5.8 Model and System Responses with Calculated Size and Area Scores	156
6.1 Material Characterization of k and ρC_p	163
6.2 Ensemble Validation Region [33]	164
6.3 Graphical Comparison of Model and Experimental Data	166
6.4 First-Order Linear Model	167
6.5 Method Sensitivities to Model Bias and Error Variance	172

Figure	Page
6.6 Model and System Responses with Corridors	177
6.7 Influence of Interpolation Factor on Overall CORA Rating	181
6.8 Model and Estimated System Response with Corridors at x_{new}	184

List of Tables

Table	Page
2.1 Balci 15 VV&A Principles [4]	15
2.2 7 Questions When Conducting Model Validation [92]	20
2.3 SME Ranking for 15 Computational Models [110]	54
2.4 Statistical Methods for Model Validation [97]	79
3.1 System Replications and Tolerance Intervals($P = 0.9, \alpha = 0.05$)	110
3.2 First Ten Replications of Model Data	110
3.3 Results From Validation Assessments	112
4.1 Notional System/Model Response Across Design Space	128
6.1 Material Characterization Data	162
6.2 Mean and Variance for Material Characterization	164
6.3 Experimental Data for Ensemble Validation	165
6.4 ANOVA for First-Order Linear Model	167
6.5 Tolerance Intervals	169
6.6 Point-wise Validation Results	169
6.7 Validation Results with Reduced Observations	170
6.8 Tolerance and Confidence Interval Results Compared to Two Sample t-Test . .	173
6.9 Model Validation at New Parameter Configuration	174
6.10 Mean System Response for Each Parameter Configuration Over Time	176
6.11 Weightings Used to Calculate CORA Ratings for Each Parameter Configuration	179
6.12 CORA Ratings for Thermal Problem	180
6.13 CORA Ratings Using Different Corridor Width Specifications	181
6.14 r^2 Values for Ensemble Validation Data [51]	182
6.15 p-Values from Transformed Hotelling T^2 Test [78]	183

VALIDATION OF DISCRETE AND FUNCTIONAL SIMULATION RESPONSES OVER EXPERIMENTAL REGIONS USING RESPONSE SURFACES

I. Introduction

The use of modeling and simulation (M&S) to emulate the behavior of real world systems is becoming commonplace as the cost of running these systems drastically increases. This research considers problems dealing with the validation of simulation models. The catalog of validation techniques currently used to assess simulation models largely focuses on single response output at a single set of input parameters. These existing methods are of great value in situations where the data fits the form in which the methods were intended; however, they are generally insufficient when validating models with complex response behavior. This proposal contributes to the model validation field by considering new techniques for point-wise and simultaneous validation of models throughout their operational domain when the model response is either discrete or in functional form. The significance of these types of validation efforts grows rapidly as we increasingly lean on model reproduction of complex real-world systems.

1.1 Background

Model validation is one component of the model verification, validation, and accreditation (VV&A) process. Validation is concerned with the assessment of model output in regards to the model's ability to accurately represent the real-world system under the conditions for which it is intended to be used [30]. The importance of proper model validation cannot be overstated. Models and simulations are utilized heavily in the science and engineering communities to assess system performance and to aid in the

decision making process. Confidence in the output these models generate is critical. The necessity for adequate model validation dates back to the 1960s, and a recent interest in improving the validation process has led to an explosion of techniques over the last few decades. Methods that handle validation over an array of parameter configurations are primarily found within the engineering community. As part of the engineering design process, competing system designs are assessed according to the performance of their model representations. This places an increased emphasis on the model's ability to accurately predict system output throughout its intended operational domain. To build confidence in a design, the predictive capability of the model must extend to tested and untested configurations of system parameters. While the competing design problem may not translate perfectly to the validation typical of the operations research and computer science communities, a methodology that allows for validation of the entire design space has great utility, particularly in the validation of complex system simulators built from experimental test data collected according to an experimental design.

1.2 Research Focus

The Department of Defense (DoD) guidance on proper data collection and model validation is in its infancy compared to our civilian counterparts. It was not until a National Research Council study in 1998 that the use of statistical methods for the construction and validation of T&E models was formally considered. The recommendations from this study included the use of confidence intervals to measure uncertainty, the use of experimental design, and the critical specification of statistical models [40]. While a portion of the validation procedures within the current V&V literature are applicable to DoD M&S problem, the library of techniques uniquely suited to address the validation of military models remains lacking. This research provides techniques that address the military model validation problem by focusing on two key questions, "How do we collect our data?" and

“What form does our data take?”. Answering these questions allows us to fine-tune our techniques to address the types of validation most prominent in military applications.

First, how do we collect data? Initiatives from the Director of Test and Evaluation (DOT&E) set basic guidance for the use of Design of Experiments (DOE) when planning and executing tests across both developmental and operational test cycles [46]. One key purpose of using experimental design was that it allows for a critical assessment of the response across the entirety of the operational envelope of the system. Furthermore, the use of DOE ensures that test resources are optimally allocated in order to provide a quality assessment across the system’s operational envelope while minimizing the number of test runs [40].

The next question that must be addressed is, “What form does our data take?”. This depends on the system being modeled. Some systems produce responses that are either discrete or their behavior is so benign over time that the response can be treated as discrete. Combat simulations often produce discrete response output. These discrete outputs can include measures of battle damage assessments (BDA), aircraft lost, sorties flown, etc. Other military based simulations, such as flight simulators, record response data over an interval of time. This is known as functional response output. Examples of functional response outputs would be aircraft pitch attitude, lift, or airspeed over the course of a preformed maneuver. Both forms of system response are modeled frequently. It is crucial that we have established methods to validate simulation models under both conditions.

This research provides the M&S community with validation techniques that uniquely address the modeling considerations experienced in the DoD; these techniques are also relevant to non-DoD validation efforts. The developed techniques must address model validation across the model’s operational domain, when the data are collected according to an experimental design, and it must handle response data in either discrete or functional form. To approach these problems, we considered three cases:

1. Model validation of discrete data at individual parameter configurations within the model's operational domain, to include configurations where no experimental data was collected
2. Overall model validation across the entire experimental region with a desired level of confidence
3. Overall and point-wise model validation of function response data over an experimental region

Each of these three cases addresses a specific validation consideration and all three address the DOT&E desire to conduct statistical based analysis of models across the span of the system's operational domain. To characterize the system's operational domain, the developed methodologies utilize Least Squares response surfaces constructed from experimental data. These response surfaces estimate the system response behavior at any point within the experimental region. The response surfaces are used in each of the three cases to establish validation criteria under the assumption that the fitted surface is an adequate representation of the true system response.

1.3 Overview

Chapter II provides an extensive literature review of the work done in model validation to date. The literature review highlights key V&V terminology, compares validation taxonomies, summarizes established methodologies, and elaborates on the techniques currently being implemented in both the scientific and engineering communities. Each section provides an appreciation for the wide ranging scope of the model validation field and helps establish an understanding for the utility of the proposed research.

Chapter III develops a multi-step approach to point-wise model validation across design spaces. The first step is the consideration for experimental designs based on user-desired variance properties. Next, system data is collected and used to fit an Ordinary Least

Squares (OLS) response surface across the design space. Regression tolerance intervals are constructed at each configuration within the design space using modeler defined interval parameters. Finally, multiple replications of the model are conducted at each design configuration and validation at each design point is assessed based on the proportion of the model response distribution within the predetermined tolerance intervals. For small model sample sizes, Bayesian Hypothesis Testing (BHT) is used to establish a lower validation bound on the required proportion of model responses to fall inside each tolerance interval.

Chapter IV formulates an F-statistic, based on a ratio of mean square errors, to assess overall model validity across the model's operational domain. System data is fit to an OLS response surface and the mean square errors between both system and model data and the fitted surface are calculated. The hypothesis under test is that system and model data that possess the same distributional properties across all input configurations will produce similar measures of disagreement with regard to the fitted surface. If the F-statistic indicates the mean square errors are statistically different for a specified α level, it is determined that the system and model data do not share distributional properties and the model is assessed as invalid. A notional example is provided to demonstrate the methodology and considerations to the method's confidence with respect to the model sample size is discussed.

The previous two methods are for discrete response data. The proper evaluation of functional data over an array of desired input parameters, to include points within the experimental region where no corresponding system data is available, is critical to achieving a robust validation assessment of a simulation model. Chapter V extends the Correlation Analysis (CORA) objective rating system so that it can validate functional data across an experimental region. Functional regression analysis is used to generate surrogate estimations of the system response functions at points within the region where experimental observations are absent. These CORA scores provide a measure of disagreement at each

desired parameter configuration. An overall score for model validity is achieved using a weighted linear combination of the individual CORA scores. Lastly, an improved CORA size scoring metric is introduced.

Chapter VI uses data from the Sandia National Laboratories thermal challenge problem as a case study to demonstrate each of the three validation methodologies. Chapter VII discusses the overall contributions of the dissertation as well as future research areas that would extend the methods to accommodate a more robust set of validation problems.

II. Review of Literature

The validation of models and simulations is not a new concept. Articles by Conway [25] and Naylor and Finger [87] discussed the necessity of adequate assessment of simulations through various techniques. These concepts were expanded in the 1980's by scholars such as Osmin Balci and Robert Sargent. Advances in software technology during the late 1980's and into the 1990's resulted in the increased use of models and simulations in the scientific and defense related domains [126]. In 1988, the General Accounting Office reported to congress that "while Department of Defense (DoD) officials agree[d] that credibility [was] important, the DOD general [had] not established the credibility of its simulations systematically and uniformly" [70]. In the mid 1990's, as the DoD increased its emphasis on M&S management, publications such as DoD Directive 5000.59 and subsequently DoD Instruction (DODI) 5000.61 were released to address issues associated with M&S efforts. These documents issued policies, procedures, and assigned responsibilities in the domain of M&S verification, validation, and accreditation [126]. As M&S capabilities grow, it becomes increasingly important that models are validated rigorously. This increased awareness is represented well within the V&V literature as authors continue to research new and improved ways to conduct M&S validation. This literature review provides a foundational understanding of M&S VV&A that will allow for easy comprehension of the research efforts discussed in this dissertation. This chapter is structured in the following manner: Definitions, The Modeling Process, Validation Taxonomies, General V&V Techniques, Statistical Validation Techniques and Current Research Efforts.

2.1 Definitions

The following definitions aid in understanding the basic terminology associated with M&S VV&A, and provide an appreciation for frequently overlooked discrepancies among commonly used terms. For example, the terms verification and validations are often used interchangeably when in fact the two are distinct processes. To fully understand the methodologies proposed in the body of this work, one must understand the different objectives associated with these processes. The literature is full of definitions for commonly used M&S VV&A terms. The definitions provided in this section capture the commonalities found among these various definitions.

2.1.1 Modeling and Simulation.

Models and simulations are closely related but the two entities should not be misinterpreted as being the same. Using the definitions provided by Balci [6], Law [74], Davis [27], and Sargent [106], a model can be defined as a conceptual representation of a real-world system or process. The representation logic is often mathematical in nature and requires relevant data in order to be utilized. This definition is closely aligned with the DoD definition which states that a model is “a physical, mathematical, or otherwise logical representation of a system, entity, or process” [30]. Simulation, as defined by Law [74], Banks et al. [8], and DODI 5000.61, is the implementation of a model with the intent of studying the system behavior over a designated interval of time. Models and simulations are employed together to help analysts assess system behaviors; however, despite their close relationship, modeling and simulation are not interchangeable terms.

2.1.2 Verification and Validation.

Until the late 1960’s, there was little distinction between verification and validation. Naylor and Finger [87] for example, use the terms interchangeably to mean the process by which one tests whether or not a model is “true”. Law [74] credits Fishman and Kiviat [39] as the first to give verification and validation definitions similar to the ones provided in

his text. Another early example of the literature distinguishing between verification and validation is Mihram [80]. Mihram develops a five step modeling processes in which verification and validation are separated as the third and fourth steps of the modeling process respectively. While most literature since has adequately defined verification and validation as separate processes, it is still imperative to clarify the differences so as to not confuse the processes. While many definitions exist within the literature, it is appropriate, given the nature of research addressed in this dissertation, to use the definitions adopted by the DoD. Verification is “the process of determining that a model or simulation implementation, and its associated data, accurately represent the developer’s conceptual description and specifications” [30]. Validation is “the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real-world from the perspective of the intended uses of the model” [30]. In simplified terms, verification is determining whether the model was built right, and validation is determining whether the right model was built [5]. One should note that the verification and validation of a model is only accurate in relation to the intended use of that model. A model that is valid for one purpose may not be valid for another purpose.

The remainder of this literature review will focus on the validation process as it relates more closely to the research being presented. Not all validation efforts are the same. Sargent [106] defines three types of validity which focus on different aspects of the modeling process. Conceptual model validity focuses on determining the correctness of the underlying theories and assumptions used to construct the conceptual model. Operational validity is concerned with determining the accuracy of a model’s behavior over the domain of the intended application. This accuracy is established through some comparison, statistical or visual, of the model output with real or expected output. Finally, data validity is confirming the adequacy and correctness of the data being used to build and evaluate the model.

Much like the different types of validation, there also exists various levels of validation that can be accomplished during the assessment of a model. Godley et al. [47] defines three levels of validation used for their simulator validation effort. Absolute validity is when the difference between output of two systems is determined to be statistically insignificant. Relative validity is when the outputs for the systems differ; however, as experimental conditions change, the behavior of the outputs are similar in terms of direction and magnitude. The authors additionally define interactive relative (dynamic) validity as how actors in a simulation environment dynamically react to stimuli between the experimental conditions. This dissertation focuses on absolute and relative validity.

2.1.3 Assessment, Credibility, and Accreditation.

The intended purpose of the model validation effort is to establish a suitable level of confidence in M&S results. Throughout the literature, this level of confidence is referred to as credibility. The process by which the parties involved establish this credibility, with respect to the models used in decision making, is referred to as model assessment. The V&V process is part of the overall model assessment process. The assessment of a model often leads to accreditation, which is the official determination regarding whether a model is suitable for its intended purpose [30]. Accreditation is completed on the back end of the validation effort and is often the final step in the model development process. Davis [27] notes that application-specific accreditation includes an assessment of the users (people or entire organizations) of the model as well as the model itself. He states that “one can argue that it is more important to accredit people and organizations than the tools they use” [27]. This is a novel perspective that is often lost in the literature, which too often focuses solely on the assessment of the tools rather than the operator of the tool.

2.1.4 Validation Metric.

Before discussing the desired properties of a validation metric, it is imperative that we discuss the requirements for a distance measure to be considered a mathematical metric. A

sound mathematical metric should satisfy the following conditions for all x , y and z in the set X

1. $d(x, y) \geq 0$ (non-negativity)
2. $d(x, y) = 0 \iff x = y$ (identity of indiscernibles)
3. $d(x, y) = d(y, x)$ (symmetry)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

Liu et al [76] provides a comprehensive list of desired features for a validation metric that incorporates their ideas with those presented by Oberkampff and Trucano [89] [90]. The follow list is not exclusive, but rather a guideline for desired properties of validation metrics as defined by those who work in the area of model validation.

1. A metric should be a quantitative measure of agreement between the model and system output. Furthermore, the metric should be objective and thus independent of the analyst's preferences.
2. Criteria used to accept or reject a model should be separate from the validation metric itself. The metric should only provide a quantitative measurement and not a determination of model acceptability.
3. Model and system uncertainties, as well as the correlation among multivariate responses need to be considered.
4. The validation metric should provide a level of statistical confidence associated with the system data.
5. A validation metric should have the capability to differentiate between various models containing various amounts of uncertainty.

6. A metric should be flexible enough to assess model and system agreement at multiple settings of inputs over an intended prediction region so as to assess the global predictive capability of the model.

While these characteristics are highly desirable for a validation metric, the similarly measures used in most validation techniques tend to not possess all of the above properties. The necessity of meeting the desired characteristics is situationally dependent on the unique requirements of each validation effort.

2.1.5 Epistemic versus Aleatory Uncertainty.

When characterizing the uncertainty within a model, the sources of uncertainty are typically categorized as either epistemic or aleatory. Epistemic uncertainty is that in which the modeler feels there is a potential to reduce if the model is refined using additional data. Often this uncertainty can be attributed to insufficient measurement accuracy or the negation of significant effects in the model. Aleatory uncertainty is that in which the modeler does not believe there is a potential to reduce. These uncertainties are associated with the statistical variance that occurs with each measurement (or run) at a particular experimental point. Uncertainties are typically placed into one of these two categories so that the modeler can identify which causes for uncertainty can be controlled [29].

2.2 The Modeling Process and How It Relates To V&V

2.2.1 The Simple Modeling Process.

The modeling process introduced by Sargent [105] illustrates how verification and validation take place throughout the modeling process. The simplified process, as provided in Figure 2.1, addresses the growth of a model from the problem entity (system to be modeled), through the development of the conceptual model, and finally to implementation of the computerized model. The cycle is formed when the computerized model is tested against the problem entity. Between each of these stages of development, verification and

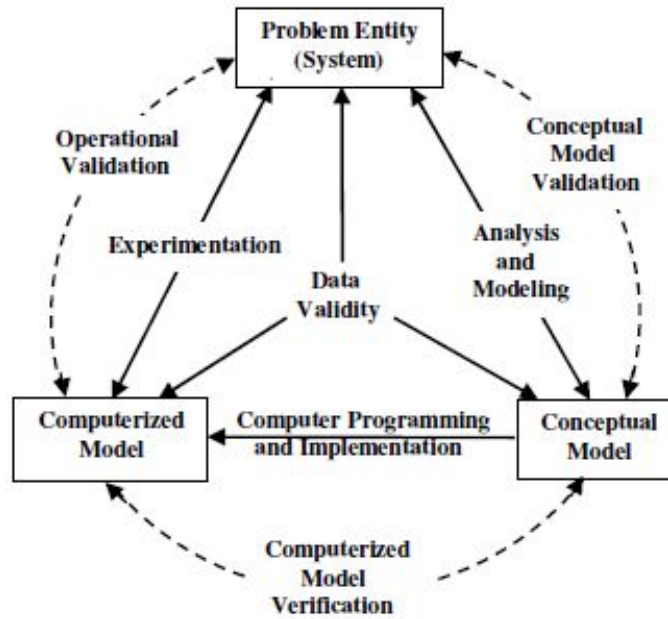


Figure 2.1: Simplified Modeling Process [105]

validation are performed before the model can progress to its next stage of development. The conceptual model is validated against the system being modeled and this conceptual validation is required prior to beginning the development of the computerized model. The computerized model is verified against the conceptual model prior to experimentation. The results of this experimentation are used to conduct operational validation of the model. Centered in the modeling process is data validity since the data used in all stages of the modeling process must be checked for validity throughout model development in order to assure the model is being constructed and tested adequately.

Law [75] presents a 7-step approach to conducting a successful simulation study, depicted in Figure 2.2, in which the steps echo the simple modeling process presented by Sargent. The basic elements of problem formulation, construction of the conceptual model and the computerized model (referred to as the Programmed Model by Law), and the

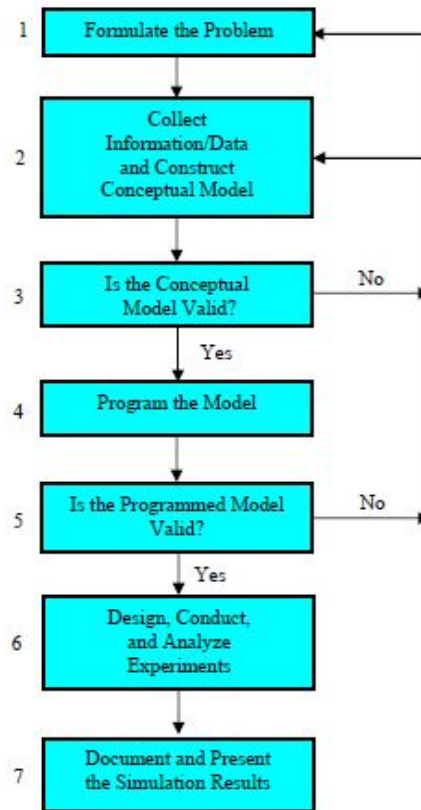


Figure 2.2: 7-Step Approach for Conducting a Successful Simulation Study [75]

experimentation of that model are present in Law's 7-step approach. The Sargent and Law processes are also similar in that they require validation of the conceptual and computerized (programmed) model before experimentation. Law's 7-step process lacks data validation throughout the model development, but addresses the need for documentation of simulation results following experimentation which is absent from Sargent's process.

2.2.2 VV&A Principles.

Along with a basic understanding of the M&S life-cycle, it is important to understand the fundamental principles associated with the VV&A process. Balci [4] presents 15 VV&A principles, provided in Table 2.1, which he claims aid in the comprehension of VV&A and are critical in the success of the M&S effort.

Table 2.1: Balci 15 VV&A Principles [4]

1	V&V must be conducted throughout the entire M&S life cycle.
2	The outcome of VV&A should not be considered as a binary variable where the model or simulation is absolutely correct or absolutely incorrect.
3	A simulation model is built with respect to the M&S objectives and its credibility is judged with respect to those objectives.
4	V&V requires independence to prevent developer's bias.
5	VV&A is difficult and requires creativity and insight.
6	Credibility can be claimed only for the prescribed conditions for which the model or simulation is verified, validated and accredited.
7	Complete simulation model testing is not possible.
8	VV&A must be planned and documented.
9	Type I, II and III errors must be prevented.
10	Errors should be detected as early as possible in the M&S life cycle.
11	Multiple response problem must be recognized and resolved properly.
12	Successfully testing each submodel (module) does not imply overall model credibility.
13	Double validation problem must be recognized and resolved properly.
14	Simulation model validity does not guarantee the credibility and acceptability of simulation results.
15	A well-formulated problem is essential to the acceptability and accreditation of M&S results.

Many of these principles are intuitive. Principle 1 states that the V&V effort must be carried out throughout the M&S process. This fact was highlighted in the discussion of Sargent's simple modeling process. Each stage of the modeling processes was paired with a verification or validation process necessary for advancement within the model development process. Principle 2 states that models are never a perfect representation of a system and thus one should not validate a model as either absolutely correct or incorrect. Instead, the model's validity should be a measure of a degree of credibility. This credibility is a function of the development cost for that model as illustrated by Sargent [107], shown in Figure 2.3, where it is clear that a model's credibility increases with budget; however, the model's value exhibits diminishing returns after a certain confidence is achieved. Principle 3 discusses the

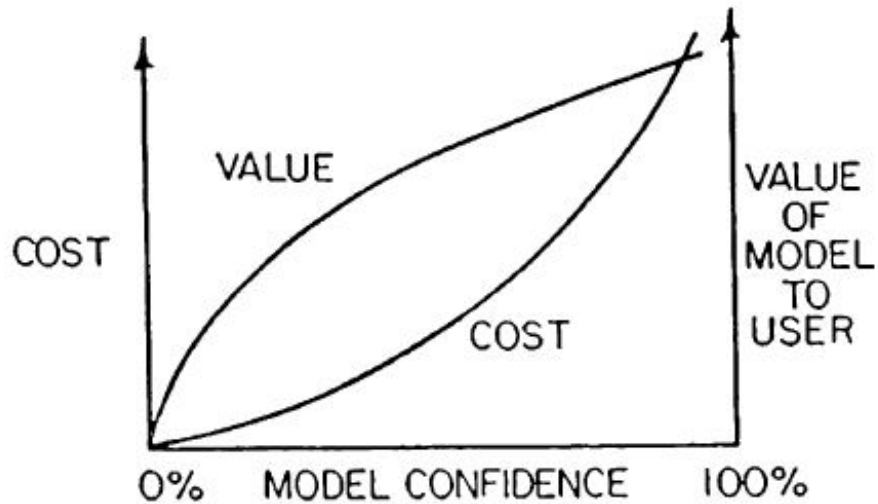


Figure 2.3: Model Confidence versus Cost [107]

fact that the credibility of a model can only be judged based on the objectives of that model. It is appropriate to determine a model's sufficient validity in relation to these objectives, rather than assessing an overall validity. Principle 4 is a concept that is often brought up in V&V literature which is that the V&V effort must be carried out by an independent actor so as to avoid biases in the assessment of model credibility. This independence can be established by either assigning a group within the organization, who are not involved in the model development process, to conduct the V&V, or by hiring an independence third party to conduct the assessment. Principle 5 states that adequate V&V requires individuals who are intimately familiar with the model and respective system they are testing. Often the most qualified individuals are the model developers which, as stated in Principle 4, can create bias in the assessment process. This makes the V&V inherently difficult [6].

Principle 6 states that model credibility can only be established for the conditions under which the model was tested; you cannot validate a model outside of its domain of applicability. It is possible that a model will behave appropriately within the limits of

the conditions in which it was built; however, the same model may behave erratically to conditions outside of this domain. Principle 7 concerns the idea that exhaustive testing of a model or simulation is impossible to do the cost and time considerations associated with the testing of potentially millions of logical paths. The amount of testing required for adequate V&V is determined by the size of the domain of applicability, which in turn is a function of the study objectives. Principle 8 sets the requirement that all V&V must be planned and documented. This requirement stems from the fact that V&V is conducted throughout the modeling processes. Documentation is critical in cataloging testing has been accomplished as well as future testing requirements. This documentation is critical in mitigating the risk of duplication of effort as well as erroneous testing efforts. Principle 9 discusses the need to prevent Type I, II, and III errors. Type I error, also referred to as Model Builder's Risk, corresponds to rejecting simulation results when they are in fact credible. Conversely, Type II error, referred to as Model User's Risk, is concerned with accepting invalid results as sufficiently valid. Type III error deals with solving the wrong problem. Figure 2.4 presents a flow chart which illustrates the potential logic paths that lead to each type of error. Principle 10 states that errors need to be detected as early as possible in the M&S life cycle due to the fact that remedying these errors later in the life cycle can be costly in terms of time, complexity, and monetarily [6].

Principle 11 deals with how to properly assess models with multiple response variables. It is cautioned that the validity of a model with two or more responses cannot be determined using a univariate statistical procedure, but rather requires a multivariate procedure that incorporates the correlations between responses. A more in depth discussion of these procedures can be found later in this chapter. Principle 12 states that overall model credibility cannot be established by achieving credibility for each model subcomponent. The errors for each submodel may be individually acceptable; however, the accumulation of error over all submodels may be unacceptable for the model as a whole. Overall system

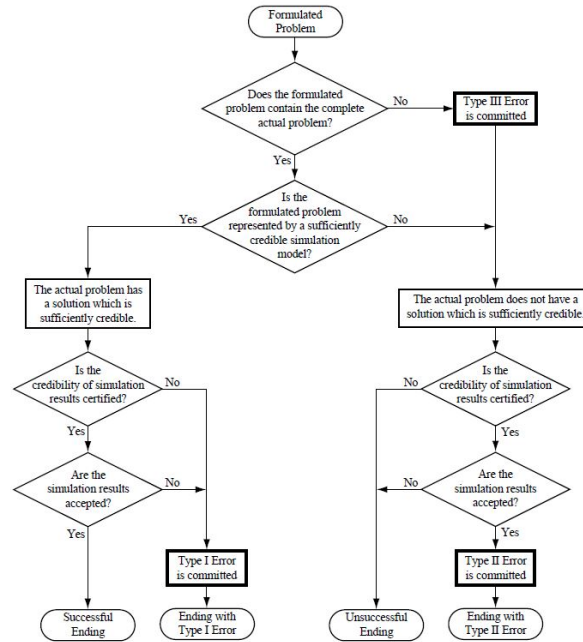


Figure 2.4: Logic Paths to Type I, II, III Errors [6]

validity must be established. Principle 13 addresses double validation. Double validation deals with validating the input data models as well as the model output. The use of invalid input models can lead to incorrectly establishing sufficient validity for a model that would have been found to be insufficient had the model ingested the appropriate input data. Principle 14 discusses the fundamental differences between model credibility and the credibility of the simulation results. Model credibility is a subject of the credibility of the simulation results and deals entirely with the judged credibility of the model with respect to the system definition and study objectives. The credibility of the simulation results is concerned with the problem definition and assessment of the system definition and specified study objectives. Principle 15 states that the formulation of the problem being modeled can greatly affect the credibility of simulation results [6].

2.2.3 Common Questions When Validating a Model.

In line with the fundamental principles discussed above, O’Keefe et al. [92] mentions seven typical problems, listed in Table 2.2, that need to be addressed in order to adequately validate any computer-based model.

What to Validate. The question of what to validate depends on where the model is within the development process. It is possible to validate any combination of the following: final system results, intermediate results and model reasoning. The overall idea is that performance of the model should be assessed as early in development as possible so as to address errors when they are easier and less costly to fix [92].

What to Validate Against. The question of what to validate the model against pertains to determining what information will be used to assess model validity. This can be empirical data, expert performance, or subject matter expertise [92].

What to Validate With. The question of what to validate the model with is concerned with selecting which test cases should be used to conduct the validation effort. Using data from test cases that were used in the construction of the model can bias the validation process because the model has been built to successfully handle this data. If no historical test cases are available, it may be possible to have experts create test cases for validation. This however, should only be considered when no historical cases are available, as it is unlikely that test cases generated by experts will be rigorous enough for satisfactory validation [92].

When to Validate. The question of when to validate the model can only be answered based on the context of the model being developed. Typically, some form of validation should take place in the early stages of the model development process. Later, additional forms of validation are necessary if the model is changed or extended in any way. The amount of validation and the frequency of its occurrence will be heavily dependent on the

cost of conducting the assessment. In some instances, it is possible to validate the model while it is in use [92].

How to Control the Cost of Validation. Determining the amount of money and effort needed in a validation effort is a difficult task. The cost of validation can be controlled during the design phase by integrating validation into the model development process [92].

How to Control Bias. Bias in the validation process is a serious issue that can lead to unjustly validating a poor model. To control expert bias, blind evaluations and cross checking should be utilized. Bias between experts can also be checked with numerous statistical tests. Developer bias, due to selecting test cases that guarantee good model performance, can only be controlled by honest professionalism [92].

How to Cope With Multiple Results. When dealing with a model which incorporates multiple response variables, it is inappropriate to test for overall model validity by testing the validity of each response variable separately. A multivariate validation approach must be used in order to capture the correlations between the response variables and to assess overall validity [92].

Table 2.2: 7 Questions When Conducting Model Validation [92]

1	What to Validate
2	What to Validate Against
3	What to Validate With
4	When to Validate
5	How to Control the Cost of Validation
6	How to Control Bias
7	How to Cope With Multiple Results

2.3 Validation Taxonomy

2.3.1 Davis Validation Taxonomy.

Davis [27] provides a VV&A taxonomy, as provided in Figure 2.5, which separates validation into three evaluation components. These components are empirical evaluation, theoretical evaluation, and evaluations by other comparisons. Empirical evaluation encompasses testing most commonly associated with scientific methods. This includes testing the model against historical data, field-test data, or other relevant and readily available data such as data gathered during operations or exercises. These methods can fall short of complete scientific rigor since that would imply full experimental testing; however, the ability to statistically or visually compare model data to that of empirical data is of great value in regards to the validation effort. Theoretical evaluation consists of assessing a model against “calculations based on rigorous or otherwise well established theories” from science and engineering [27]. The last category of evaluation encompasses all methods not covered under empirical or theoretical evaluation. These methods include assessment of the model through the use of subject matter expertise as well as written documentation such as doctrine and intelligence. Davis stresses that all three categories of evaluation holistically feed into an overarching validation process. No validation approach has been deemed better than the others and it is often necessary to conduct techniques from all three evaluation categories in order to rigorously assess a model for validity.

2.3.2 Balci VV&T Taxonomy.

Balci [3] provides a different taxonomy on validation, which is incorporated into his larger VV&T taxonomy. Balci’s approach is more expansive than that of Davis. Balci divides techniques into six categories. Informal techniques are those that, while remaining structured, lack a mathematical foundation. These techniques often rely on reasoning and subjective measures. The static techniques align themselves more closely to the verification process than the validation process. These techniques are concerned with the assessment

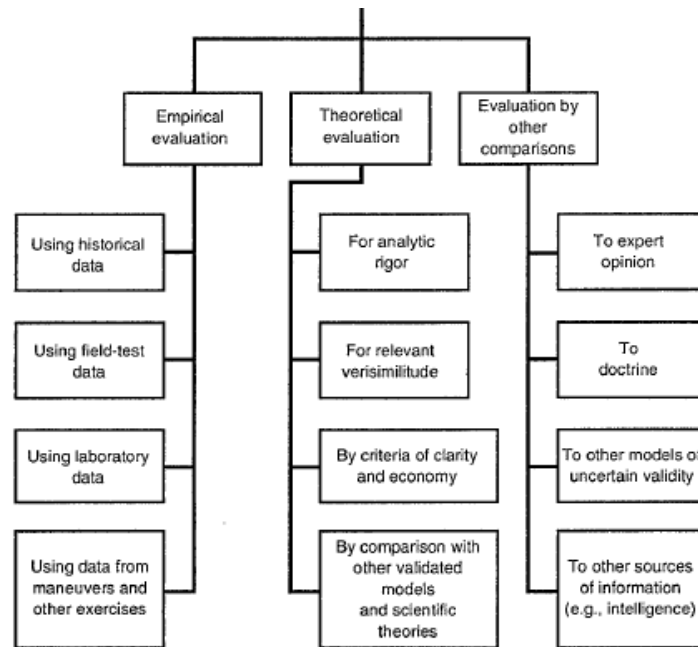


Figure 2.5: Validation portion of Davis VV&A Taxonomy [27]

of the model source code structure and implementation. Dynamic techniques are intended for assessment of model behavior through analysis of model output. This assessment of validity is carried out using an array of statistical or multivariate statistical techniques. Balci notes that “whenever possible, a multivariate statistical technique should be used to conduct model validation”[3]. Symbolic techniques are also concerned with evaluation the dynamic model output behavior. When evaluating a model using symbolic techniques, “symbolic inputs are provided to a simulation model as inputs and expressions are produced as outputs”[3]. These outputs are created from “the transformation of the symbolic data along model execution paths”[3]. Constraint techniques are concerned with assessing model correctness. Finally, formal techniques focus on model correctness through the use of formal mathematical proofs. These techniques are unfortunately not attainable for reasonably complex models and simulations; however, they form the foundation for the other categories of VV&T techniques. Balci separates this taxonomy into two

groups: techniques for conventional simulation models and techniques for object-oriented simulation models. Bacli [6] consolidates the original taxonomy into four categories: Informal, Static, Dynamic, and Formal. This version of the taxonomy, as depicted in Figure 2.6 includes a more expansive list of V&V techniques.



Figure 2.6: Balci VV&T Taxonomy [6]

2.3.3 Sargent's Operational Validity Classification Taxonomy.

Sargent [108] provides a simple taxonomy concerning the operational validation of models. His classification system, presented in Figure 2.7, places the validation techniques

into one of four quadrants based on whether the approach is objective or subjective and whether the system is observable or non-observable. Comparison using graphical displays uses graphs of model and system behavior to visually assess the accuracy of the model's output with regards to the model's intended purpose. In this method, box plots, histograms, and scatter plots are analyzed by system experts to determine differences between model and system behavioral dynamics. Often these graphs incorporate visual representations of standard statistical measures. Exploring model behavior includes assessing the model output behavior in terms of directional and magnitude changes in model responses. These behaviors are typically compared with one of three methods: graphical, construction of confidence intervals, and through hypothesis testing. Sargent notes that in order to achieve a high degree of model confidence, it is necessary to conduct the validation effort using various sets of experimental conditions. This indicates that it is unlikely that one can obtain high confidence in models in which the system they are intended to model is not observable.

	Observable System	Non-observable System
Subjective Approach	• Comparison Using Graphical Displays • Explore Model Behavior	• Explore Model Behavior • Comparison to Other Models
Objective Approach	• Comparison Using Statistical Tests and Procedures	• Comparison to Other Models Using Statistical Tests

Figure 2.7: Sargent's Operational Validity Classification [108]

2.4 Statistical Validation Techniques

This section discusses statistical V&V techniques as structured in the Balci [6] taxonomy. The decision to elaborate on Balci's taxonomy is based on its comprehensiveness in regards to the techniques covered. Balci's taxonomy categorizes statistical techniques as a subset of Dynamic model V&V. The other three classifications for V&V techniques are: Informal, Static, and Formal. Balci [6] discusses these methods at length. The remainder of this section focuses on validation techniques that implement statistical rigor in the assessment of model validity.

2.4.1 Analysis of Variance.

Analysis of Variance (ANOVA) can take on similar yet unique definitions depending on the context in which it is being used. In general linear models, ANOVA attempts to explain dependent variable scores in terms of a model and error component in order to assess which mean scores for the dependent variables differ. The ANOVA can also provide insight into which experimental groups or conditions contribute to the variation in the dependent variable. In a single-factor ANOVA, differences in the means are determined from a statistical test of the null hypothesis which states that the means are equal. ANOVA uses the F statistic to determine statistical significance at a prescribed alpha value. The F statistic is calculated from the ratio of the mean squares between groups and mean squares within groups or commonly called, mean square error [103]

$$F = \frac{MS_{Between}}{MS_{Within}}. \quad (2.1)$$

This ratio of mean square errors, under a true null hypothesis, follows an F distribution and can be used to calculate a probability value (p-value) associated with the likelihood of observing the calculated F ratio or one that is higher, given that the null hypothesis is true.

2.4.2 Two-Sample *t*-Test.

The two-sample *t*-test tests the null hypothesis that the means of two populations ($\bar{Y}_1$ and $\bar{Y}_2$) of data are equal. In the case of validation, the two-sample *t*-test is most relevant and uses the alternative hypothesis that the two means are not equal. The calculation of the test statistic depends on whether or not equal variance between the samples is assumed. The general form for the *T* statistic, when equal variances are assumed, is [117]

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \quad (2.2)$$

where S_p is a pooled estimator of the population variance and N_1 and N_2 are the sizes of each sample, respectively. The *T* statistic is compared to a critical *t* value from the *t*-distribution, which is based on the desired confidence level and the degrees of freedom. If the absolute value of the calculated *t* statistic is greater than the critical value, the null hypothesis of equal means is rejected.

2.4.3 Confidence Intervals/Bounds.

Confidence intervals are another approach to determine whether or not the difference in means is significant. In confidence interval analysis a $100(1 - \alpha)$ percent confidence interval for a parameter, which in validation is often the difference in the means of the response for the system and the model, is calculated and evaluated. In the case of evaluating the difference in the response means between systems, the inclusion of $\hat{\mu}_1 - \hat{\mu}_2 = 0$ within the confidence interval supports the null hypothesis at the specified level of significance [83]. The confidence interval for the differences in the mean is calculated using the following formula [74]

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{\alpha/2, n-1} S \sqrt{\frac{1}{n}} \quad (2.3)$$

where S is the sample standard deviation and n is the number of samples. In the case where the number of observations differs between the two systems and when it is inappropriate to make the assumption that the variances are equal, the Welch confidence intervals, as described by Law [74], can be used to perform the comparison on the difference between the mean response for the system and the model. The Welch confidence interval is defined as

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{\alpha/2, \hat{f}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (2.4)$$

where $\hat{f}$ represents the computed estimated degrees of freedom. Both cases above deal with the comparison of two systems. If a comparison needs to be made on more than two systems simultaneously, the confidence intervals for each two way comparison must be adjusted so that the overall confidence level achieved is at least $1 - \alpha$. The Bonferroni inequality is often used to assure the overall confidence level is achieved. The Bonferroni inequality implies that each of the c confidence intervals should be constructed at a level of $(1 - \alpha)/c$ so that the overall confidence level remains at $(1 - \alpha)$ [74]. The Bonferroni approach is reasonable when there are few comparisons. When there are more than a few comparisons, other multiple comparison methods are used. These methods are not discussed here and the reader is referred to Montgomery [83] for additional details.

2.4.4 Analysis of Paired Observations.

An analysis of paired observations utilizes the paired-sample t-test to test for the difference of means between two populations. The difference of each paired observation, where the pairs represent identical “subjects” under different conditions, are determined and the mean and standard deviation of these differences across all observed pairs are calculated [117].

$$\bar{d} = \frac{1}{N} \sum_{i=1}^N d_i \quad (2.5)$$

$$s_d = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (d_i - \bar{d})^2} \quad (2.6)$$

where d_i represents the differences between the i^{th} observed pair and N is the number of observed pairs. The test statistic for the paired-sample t test is calculated as [117]

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{N}} \quad (2.7)$$

The null hypothesis, that the two populations have the same mean, is rejected (in the two-tailed case) if $|t| \geq t_{1-\alpha/2, \nu}$ where $\nu = N - 1$ degrees of freedom. In a validation effort, the two samples would represent the model output behavior and the system output behavior. The paired observations would be the response for the model and system under the same experimental conditions. In this case, the two-tailed paired-sample t test would be appropriate because the goal of the validation is merely to test for significant differences in the mean response of the two systems. The direction of the difference is typically not of interest.

2.4.5 Factor Analysis.

Factor analysis is a method for simplifying complex sets of data through the construction of factors. These factors are a condensed representation of relationships between sets of dependent variables. The correlations between these constructed factors and the defined variables are known as factor loadings. These loadings are used to deduce the meaning of each factor based on similarities between variables that are highly correlated to a particular factor [69].

Cohen and Cyert [24] use factor analysis to validate a process model's ability to accurately generate time paths. The concept behind the proposed methodology was that a factor analysis could be performed on both the generated and observed time paths. The factor loadings for both the model and the observed system could then be assessed in order to determine statistical differences.

2.4.6 Hotelling's T^2 Tests.

The Hotelling two-sample T^2 test is a multivariate statistical test used to test the hypothesis that the means of two multivariate populations are equivalent. Balci and Sargent [7] propose the use of the Hotelling two-sample T^2 when validating multivariate response models. The use of the Hotelling's two-sample T^2 test assumes that the populations are multivariate normal and that the variance-covariance matrices for the model and system are equal. If the data are not multivariate normal, transformations should be performed on the data to obtain multivariate normal behavior. Furthermore, the two-sample T^2 test requires that the data matrices for the model and system be independent and the observations vectors within the data matrices are also independent. The two-sample T^2 test statistic is calculated as [7]

$$T^2 = \frac{nN}{n + N}(\bar{x} - \bar{y})'S^{-1}(\bar{x} - \bar{y}) \quad (2.8)$$

where n is the number of replications collected on the model, N is the number of replications collected on the system, $\bar{x}$ is a vector of the model response variables averaged over n replications, $\bar{y}$ is a vector of the system response variables averaged over N replications, and S is the sample common variance-covariance matrix[7].

To express T^2 as being from an F distribution, the T^2 must be weighed by the factor $(n + N - k - 1)/k(n + N - 2)$. Therefore, in order to accept the validity of the model being tested (fail to reject the null hypothesis), equation 2.9 must be satisfied [7]

$$T^2 \leq \frac{k(n + N - 2)}{n + N - k - 1} F_{\alpha, k, n+N-k-1} \quad (2.9)$$

otherwise the null hypothesis is rejected and the model is not validated against the real-world system.

2.4.7 Multivariate Analysis of Variance.

A multivariate analysis of variance (MANOVA) is the multivariate response form of ANOVA. The difference lies in that an ANOVA tests the null hypothesis that the population means of k groups are equal whereas a MANOVA tests the null hypothesis that the population means of k groups are equal for all p response variables. While an ANOVA requires the use of the F-test, there are several approaches to conducting a MANOVA [12]. The null hypothesis can be tested using a number of multivariate tests statistics which are summarized in [48].

2.4.8 Nonparametric Goodness-of-fit Tests.

2.4.8.1 Kolmogorov-Smirnov Test.

The Kolmogorov-Smirnov (K-S) test is a distribution-free goodness-of-fit test that determines the degree of agreement between the distributions of two sets of data. Naylor and Finger [87] proposed that the K-S test be used to aid in validating simulation models by comparing the distribution of the model data with the distribution of the system data. If it is assumed that the data from the model and system are samples of their overall population of data, the two sample K-S test can be used to test the hypothesis that both samples come from a population with the same distribution. The test statistic for the two-sample K-S test is simply the largest difference between the two fitted cumulative distribution functions for all values of x and is can be defined as [125]

$$D_{n_1, n_2} = \sup_x |F_1(x) - F_2(x)| \quad (2.10)$$

where n_1 is the size of the first sample, n_2 is the size of the second sample and $F_1(x)$ and $F_2(x)$ are the cumulative distribution functions for the two samples, respectively. The critical value for the two-sample K-S test is [125]

$$D_{n_1, n_2, \alpha} = c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \quad (2.11)$$

where $c(\alpha)$ is the inverse of the Komogorov distribution at α . The null hypothesis is rejected when $D_{n_1, n_2} > D_{n_1, n_2, \alpha}$. When comparing model data to real-world data, this rejection of the null hypothesis implies that the model data does not pass a goodness-of-fit with the system data and therefore cannot be validated.

2.4.8.2 Cramer-Von Mises Test.

The Cramer-Von Mises Test assesses whether or not a sample was drawn from a specific continuous distribution $F(x)$. The criterion for this test is known as the ω^2 criterion and is of the form [2]

$$\omega^2 = \int_a^b [F_N(x) - F(x)]^2 dF(x) \quad (2.12)$$

where $F_N(x)$ is the empirical distribution of the sample. The test statistic for the Cramer-Von Mises test of two samples is [2]

$$\begin{aligned} T = N\omega^2 &= \left[\frac{NM}{(N+M)} \right] \int_a^b [F_N(x) - G_M(x)]^2 dH_{N+M}(x) \\ &= \frac{U}{NM(N+M)} - \frac{4MN-1}{6(M+N)} \end{aligned} \quad (2.13)$$

where N and M are the sizes of the observations from the first and second samples

respectively, $G_N(x)$ is the empirical distribution of the second sample, $H_{N+M}(x)$ is the empirical distribution for the two samples together and U is defined as [2]

$$U = N \sum_{i=1}^N (r_i - i)^2 + M \sum_{j=1}^M (s_j - j)^2 \quad (2.14)$$

where r_i and s_j are the ranked observations for the first and second sets of observations respectively. In the two sample case, the null hypothesis that the two samples are from the same distribution can be rejected if the value for U is too large given a specified significance level. Anderson [2] provides tables to determine when U is large enough to reject the null hypothesis.

2.4.8.3 *Chi-square Test.*

The Chi-square test (Pearson's Chi-squared) is the oldest of the goodness-of-fit tests, dating back to Karl Pearson's paper in 1900. The test determines whether the frequency of binned outcomes follow a specific distribution. In a chi-squared test, the observed outcomes are placed into k adjacent intervals (bins) which create a histogram representing the distribution of the observed data. Next, the expected proportions for each bin are calculated based on a sampling from the fitted distribution [74]. The chi-squared test statistic is of the form [8]

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (2.15)$$

where O_i and E_i are the observed and expected counts for the i^{th} bin, respectively. When all parameters are known, the test statistic can be compared to a chi-squared distribution with $k-1$ degrees of freedom ($\chi_{k-1,1-\alpha}^2$). The null hypothesis that the bin frequencies from the observed follow the expected distribution is rejected if the test statistic is found to be larger than $\chi_{k-1,1-\alpha}^2$, otherwise the test fails to reject the null. When using a chi-squared test for

model validation, a model would be validated if its binned model data (the observed) was found to have the same distribution as the binned system data (expected). The chi-squared approach is deemed less powerful than the K-S test for goodness-of-fit due the chi-squared test's tendency to be highly biased when the expected bins are not of equal proportion. This leads to an undesirably high Type I error [74]. Furthermore, the K-S test can be applied to small data sets, whereas the chi-squared test is only valid for data sets of 50 or more samples [8]. Naylor and Finger [87] also mention that the chi-squared test is relatively sensitive to non-normality.

2.4.9 Nonparametric Tests of Means.

2.4.9.1 Mann-Whitney-Wilcoxon Test.

The Mann-Whitney-Wilcoxon Test (Mann-Whitney U Test) is a nonparametric statistical test with a null hypothesis that the distributions for two samples of data are identical. Observations from two samples are rank ordered and compared in a pairwise manner to establish a count on the number of times an observation from the first sample “bests” the corresponding ranked value from the other sample [77]. The statistic used takes the form [117]

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R \quad (2.16)$$

where n_1 and n_2 are the number of observations for sample 1 and sample 2, respectively and R is the rank sum for sample 1. If the observed U statistic is less than U critical at a specified significance level, then the null hypothesis is rejected and the two samples are deemed not identically distributed. The values for U critical are found in the Mann-Whitney Tables [77].

2.4.10 Regression Analysis.

The basis of using linear ordinary least squares (OLS) regression analysis to conduct model validation is found in Reckhow et al. [99], Kleijnen and Van Groenendaal [62] and Kleijnen [65]. In an assessment of relative validity, the model is assessed based on a positive linear correlation between the model and prediction data. In simple OLS regression, the sum of squares error is minimized to calculate estimates for the intercept and slope parameters in the model which takes the form

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (2.17)$$

where β_0 and β_1 are the estimated intercept and slope parameters and ϵ is the error in the predicted response. Assuming model and system data are bivariate normally distributed, testing for relative model validity is accomplished by fitting an OLS regression of the simulated data based on the system data and testing the null hypothesis that $H_0 : \beta_1 \leq 1$. Absolute validity can also be assessed under a more strict set of null hypotheses. The null hypothesis under test is that $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 1$. To test these hypotheses simultaneously, the Sum of Squared Errors (SSE) with and without the hypothesis are calculated. These SSEs represent the full and reduced regression model respectively.

$$SSE_{full} = \sum_1^n (Y_i - \hat{Y}_i)^2 \quad (2.18)$$

$$SSE_{reduced} = \sum_1^n (Y_i - X_i)^2 \quad (2.19)$$

These SSEs can be used to form an F-statistic for model evaluation. The F-statistic can be written as

$$F_{2,n-2} = \frac{(SSE_{reduced} - SSE_{full})/2}{SSE_{full}/(n-2)}. \quad (2.20)$$

If the F-statistic is high, the null hypothesis is rejected and we can conclude that the model is invalid [62] [65]. Additionally, examination of residual and normal probability plots is conducted to ensure normality and lack of autocorrelation in the residuals [99].

Despite its frequent use as a validation technique, Mitchell [81] discusses five reasons why regression analysis may be inappropriate for validation.

1. Regression is not being used for its intended purpose since the purpose of regression is to estimate y with regards to x whereas there is no desire, in regression validation, to estimate the model output using the observed system data.
2. Larger scatter amongst the data results in greater standard error in the estimate for the slope parameter, which leads to a smaller test statistic. In other words, regression applied to highly scattered data are more likely to have slopes that are not statistically different than the null hypothesis of $\beta_1 = 1$. This can result in falsely validated models (i.e. Type II Error). A similar phenomena regarding large scatters occurs when using confidence limits to assess statistical differences in the estimated regression parameters.
3. Regression is not sensitive enough to quantify how good a line is once past the conventional thresholds for p-values due to the fact that any grouping of points which tend to avoid two opposite corners will form a significant regression line.
4. The fitted line generated from regression analysis is of no direct relevance to model performance since the line is not part of the model or model output. Mitchell notes that the deviations between model and system output do give direct information regarding a model's ability to simulate the intended system. He uses this notion to develop an alternative validation method which will be discussed momentarily.

5. Data used for empirical validation often violate the principle assumptions of regression such as the regressors being known without error and the assumption of independent observations.

Mitchell and Sheehy [82] propose an alternative methodology which focuses on the assessment of the deviations between predicted and observed data. In their method, the deviations between the model data and system data are assessed against a predetermined precision requirement, which is based on 95% confidence limits with an additional buffer to relax the criterion while maintaining practical significance. The deviations are graphed, along with the the established absolute deviation lines (Figure 2.8) and validation is assessed based on the number of deviations that lie outside of the envelope of acceptable precision. If the number of deviations outside of the precision bounds is less than five percent (assuming 95% confidence), the model is considered valid.

2.4.11 Theil's Inequality Coefficient.

Theil's Inequality Coefficient (TIC) is a commonly used validation technique in economics and in particular for determining the accuracy of forecasting models. The TIC assesses the differences between observed and predicted time-series data. This is easily translated into validation of simulation models in which the response is in the form of time-series data. In this case, the error between the model and system at each point in time is measured [100]. Let $A = A_1, A_2, \dots, A_n$ and $P = P_1, P_2, \dots, P_n$ be the time-series data from the system and model, respectively, where the i_{th} observation corresponds to the response at a particular point in time. The expression for the TIC can be written as [10]

$$TIC = \frac{\sqrt{\frac{1}{n} \sum_i^n (A_i - P_i)^2}}{\sqrt{\frac{1}{n} \sum_i^n A_i^2} + \sqrt{\frac{1}{n} \sum_i^n P_i^2}}. \quad (2.21)$$

The numerator for this expression represents the standard deviation of the error between

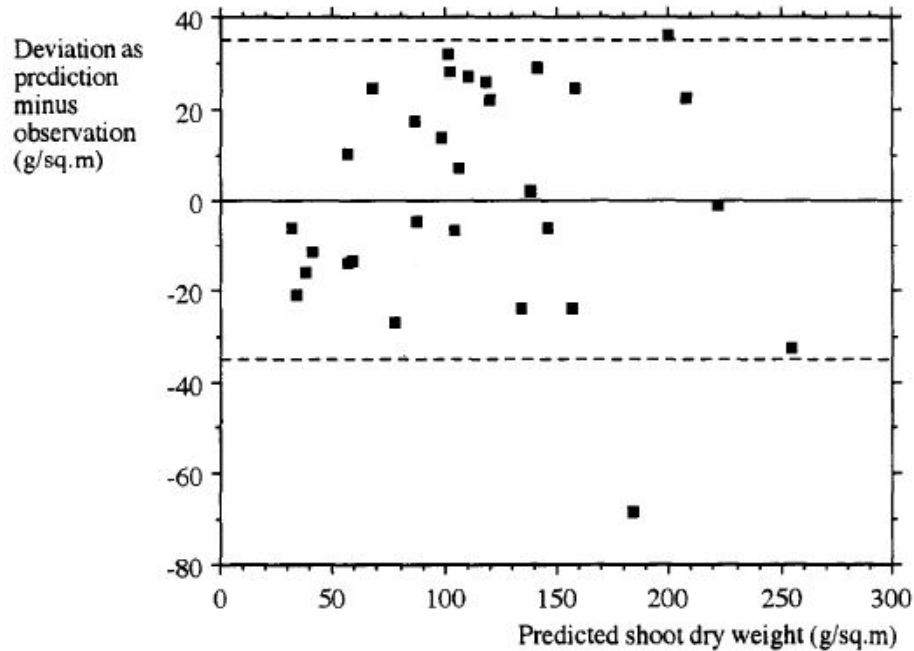


Figure 2.8: Example of Deviations with Corresponding Envelope of Acceptable Precision [82]

the two data sets and the denominator is based on the standard deviations of each set of data. This indicates that the TIC is a measure of percent differences rather than absolute differences. Classification for relating TIC values to the quality of a model depends on the desired level of performance for a given application. Arbitrarily, TIC values above 0.7 correspond to poor models. Values ranging between 0.4 and 0.7 correspond to average to good models, and TIC values below 0.4 correspond to very good or excellent models [100].

2.4.12 Time Series Analysis.

2.4.12.1 Correlation/Spectral Analysis.

Many simulations provide response output in the form of time-series data. This data is often auto-correlated and correlation and spectral analysis capture these auto-correlation characteristics in a convenient form for assessment [116]. Spectral analysis differs from correlation analysis in that it examines the time series data in the frequency domain rather

than the time domain and cross-spectral analysis between two sets of time-series data assesses the interactions or correlations between the pair of variables at each frequency [54]. Procedures exist to test the equivalence of a spectra derived from model data and spectra derived from system data under the same conditions. Equivalent spectra can increase model confidence; however, it is harder to interpret the negative effect of deviations in spectra on model confidence [116].

In general, correlation and spectral analysis are concerned with determining and analyzing autocorrelations in continuous stochastic stationary time-series data. The autocorrelation function for a particular time-series, $x(t)$, is defined as [9]

$$R_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)x(t + \tau)dt \quad (2.22)$$

where τ is the corresponding lag between observations. This function expresses autocorrelation within a time-series dataset but does not address autocorrelation between two separate sets of time-series data. This is known as the cross-correlation function and is expressed as [9]

$$R_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)y(t + \tau)dt. \quad (2.23)$$

A graph of the cross-correlation function against lag can illustrate which lag values are the most vulnerable to high cross-correlation between two sets of time series data. An example of this graph can be seen in Figure 2.9. In this example, it is evident that two regions in the lag domain correspond to high levels of cross correlation between the two sets of data. High levels of cross-correlation could indicate significant dependence between the two sets of data which is a critical insight when validating the model time-series against system time-series data. The correlation coefficient function, which measures the linear correlation

between the two data sets, can be generated from these autocorrelation functions. The correlation coefficient function is defined as [9]

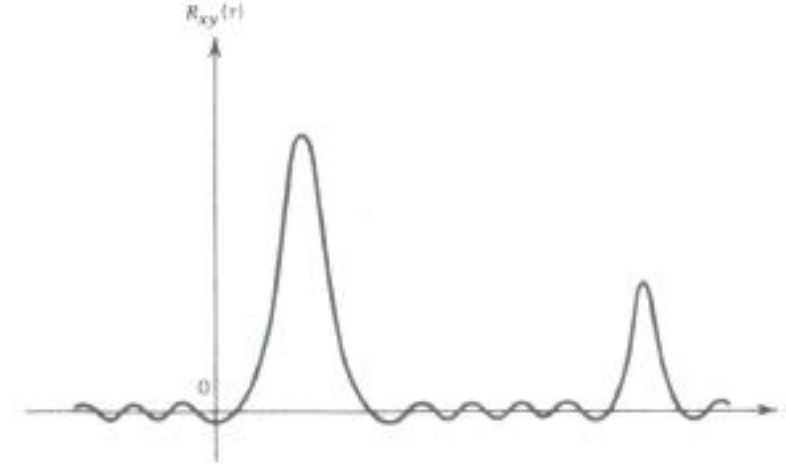


Figure 2.9: Plot of Cross Correlation as a Function of Lag [9]

$$\rho_{xy}(\tau) = \frac{R_{xy}(\tau) - \mu_x \mu_y}{\sqrt{[R_{xx}(0) - \mu_x^2][R_{yy}(0) - \mu_y^2]}} \quad (2.24)$$

where μ_x and μ_y are the average values for the two sets of time-series data. These ρ_{xy} functions express autocorrelation between two sets of time-series data in terms of time and the corresponding lag. In some instances, analysis of the time-series data may benefit from looking at the autocorrelations in terms of the function frequencies rather than as a function of time. This transformation of the autocorrelation function into the frequency domain creates the spectral density function. This function, referred to as the cross-spectral density function, is the Fourier transform of the correlation function between the two sets of time-series data [9].

$$S_{xy}(f) = \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-i2\pi f\tau} d\tau \quad (2.25)$$

where i is imaginary and τ is the lag component. These spectral density functions deal with both the positive and negative frequency ranges. In some instances it is more practical to work only in the positive frequency domain and thus the one-tailed spectral density function can be defined as [9]

$$G_{xy}(f) = 2S_{xy}(f). \quad (2.26)$$

Using this relationship, the coherence function, which is analogous to the squared correlation coefficient function shown in Equation 2.24, can be defined as [9]

$$\gamma_{xy}^2(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)} \quad (2.27)$$

where $0 \leq \gamma_{xy}^2(f) \leq 1$. In many applications, the coherence function proves to be a much more powerful analytical tool than the correlation coefficient function. Ultimately, the appropriate function for analysis will be application dependent [9].

2.4.12.2 Error Analysis.

There are many error measures and metrics used to quantify the differences between two sets of time series data. Sarin et al. [110] discusses some of the more popular error measures as a introduction to their EARTH metric, which is discussed later in this chapter. This section highlights the utility of the more commonly used error measures.

Vector Norms. When dealing with finite-dimensional time-series data, the most common error measure used is vector norms. The L_p norm of two time history vectors A and B of equal size N is calculated as [110]

$$\|A - B\|_p = \left(\sum_{i=1}^N |a_i - b_i|^p \right)^{1/p}. \quad (2.28)$$

The three most commonly used vector norms are the L_1 , L_2 and L_∞ norms. The L_∞ norm, also known as the sup-norm, takes the largest difference between the two sets of data. Vector norms provide a solid foundation for quantifying the discrepancies between time series data; however, they have two major deficiencies. First, using different vector norms can lead to different conclusions regarding the discrepancies between the two sets of data. Secondly, the vector norm measure does not distinguish between the errors due to phase and magnitude differences [110].

Average Residual and Residual Standard Deviation. Another popular error measure is the average residual and the residual standard deviation. The average residual measures the average difference between two sets of time series data and is calculated as [110]

$$\bar{R} = \frac{\sum_{i=1}^N (a_i - b_i)}{N}. \quad (2.29)$$

The residual standard deviation provides a measure for the differences in the individual residuals from the mean residual and is calculated as [110]

$$S_{N-1} = \sqrt{\frac{\sum_{i=1}^N (R_i - \bar{R})^2}{N - 1}}. \quad (2.30)$$

These measures, much like the vector norm measure, provide an aggregated measure of error between two sets of time series data; however, the average residual has the distinct disadvantage of differences being canceled out due to the averaging of positive and negative residuals [110].

Sprague and Geers (S&G) Metric. The S&G metric captures the errors due to phase and magnitude, combining these errors into a single metric. The calculations for these errors are as follows [110]

$$M_{S\&G} = \sqrt{\frac{\psi_{AA}}{\psi_{BB}}} - 1 \quad (2.31)$$

$$P_{S\&G} = \frac{1}{\pi} \cos^{-1}\left(\frac{\psi_{AB}}{\sqrt{\psi_{AA}\psi_{BB}}}\right) \quad (2.32)$$

$$C_{S\&G} = \sqrt{M_{S\&G}^2 + P_{S\&G}^2} \quad (2.33)$$

where

$$\psi_{AA} = \frac{\sum_{i=1}^N a_i^2}{N}, \psi_{BB} = \frac{\sum_{i=1}^N b_i^2}{N}, \psi_{AB} = \frac{\sum_{i=1}^N a_i b_i}{N}. \quad (2.34)$$

The advantage to the S&G metric over vector norms and average residual measures is that it allows for the a more detailed investigation into the error contributions from phase and magnitude differences. A major limitation of the S&G method is that it cannot account for the error due to the shape of the data because of the aggregation of the time history information into the ψ 's [110].

Russell's Error Measure. Russell's error metric is similar to the S&G metric except for an adjustment in the magnitude error factor. This adjustment scales the magnitude error factor to the phase error factor when there exists an order-of-magnitude difference in the amplitude between the sets of time-series data. The new magnitude error factor is calculated as [110]

$$M_R = \text{sign}(\psi_{AA} - \psi_{BB}) \log^{10} \left(1 + \left| \frac{\psi_{AA} - \psi_{BB}}{\sqrt{\psi_{AA}\psi_{BB}}} \right| \right) \quad (2.35)$$

While Russell's error measure removes the asymmetric limitation found in the S&G metric, it still cannot identify and quantify the error due to magnitude differences [110].

Normalized Integral Square Error (NISE). The NISE is a measure of error that considers errors due to phase shift, magnitude difference, and shape difference. NISE uses the principle of cross-correlation ($\rho(n)$) to compute the phase difference between two sets of time-series data denoted as n_* . One of the two time histories is then shifted by n_* to compensate for the error in phase. After this adjustment, $\psi_{AB}(n_*)$ is calculated and used in the construction of the three error measure contributors. The equations for phase, magnitude and shape error are [110]

$$P_{NISE} = \frac{2\psi_{AB}(n_*) - 2\psi_{AB}}{\psi_{AA} + \psi_{BB}} \quad (2.36)$$

$$M_{NISE} = \rho(n_*) - \frac{2\psi_{AB}(n_*)}{\psi_{AA} + \psi_{BB}} \quad (2.37)$$

$$S_{NISE} = 1 - \rho(n_*) \quad (2.38)$$

The overall NISE metric is the sum of the three error components which mathematically works out to [110]

$$C_{NISE} = 1 - \frac{2\psi_{AB}(n_*)}{\psi_{AA} + \psi_{BB}} \quad (2.39)$$

The major drawback to the NISE metric is that while it attempts to address the shape error, the $\rho(n_*)$ component cancels during the summation of the three errors and thus the overall NISE metric is independent of $\rho(n_*)$ and does not account for shape error [110].

Dynamic Time Warping (DTW). DTW is an algorithm for measuring differences in time history data sets commonly used in matching and categorization applications such as speech recognition and signature matching. In DTW the peaks and valleys of the two

sets of time-series data are aligned by expanding or compressing the time axis according to a given cost function. The DTW algorithm finds a monotonically increasing sequence of ordered adjacent pairs from the two sets of data in order to minimize the cumulative cost function. Figure 2.10 illustrates the DTW of two sets of time series data (time histories 2 and 3) against an empirical set of time series data (time history 1). The DTW distance is the square root of the cumulative cost function for conducting the DTW on the data. The smaller DTW distance between time histories 1 and 2 indicates that time history 2 is better than time history 3 at representing time history 1 [110].

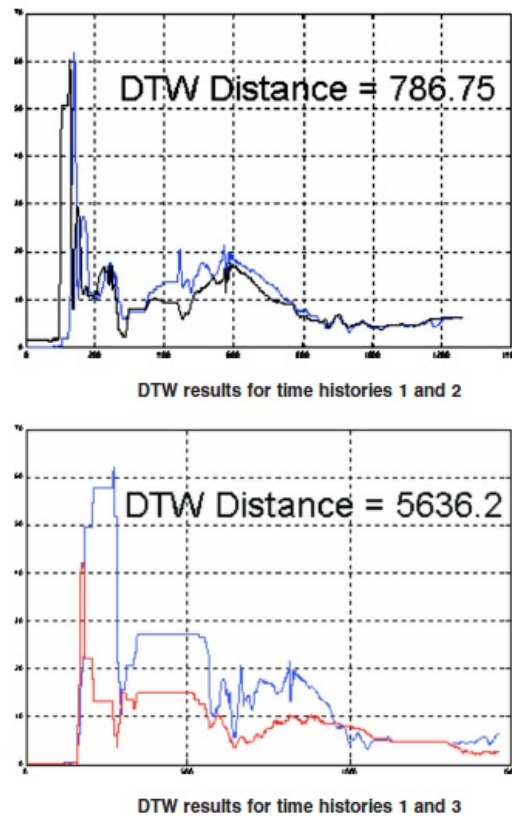


Figure 2.10: Dynamic Time Warping Example [110]

2.5 Current Research Efforts

2.5.1 Wavelets.

The use of wavelets to model and characterize time-series responses has garnered much attention in recent years for a variety of reasons. In signals analysis, the use of wavelets has gained popularity over the more traditional Fourier transforms due to the fact that wavelets provide valuable information in both the time and frequency domains, whereas Fourier transforms only provide information in the frequency domain. The literature on wavelet utilization for time-series analysis and characterization is robust, not only in terms of the volume of literature, but also in the diverse implementation of wavelets within a multitude of real world applications. A quick look through the wavelet literature highlights its usefulness in areas such as data compression, audio and image processing, data smoothing/de-noising, classification, and analysis of dynamic time-series data [114] [14] [26] [17] [1]. The literature in regards to the use of wavelets for time-series data validation is less robust but recent efforts have been made to implement wavelets into the model validation domain.

Let us define a basic mother wavelet as $\psi(t)$. This mother wavelet is manipulated, in terms of a binary dilation and dyadic translation in order to represent a time-series function, $f(t)$. Mother wavelets are unique in their selection in that they must be able to represent any function in $L^2(R)$ solely on their dilation and translation [23].

A continuous wavelet transform can be defined as [91]

$$(W_\psi f)(a, b) = \langle f, \psi_{a,b} \rangle = a^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-a}{a}\right) dt \quad (2.40)$$

where a and b represent the dilation and translation factors respectively. For the discrete case where $a = 2^{-j}$, $b = 2^{-j}k$, the discrete wavelet coefficient, $c_{j,k}$, is equal to [91]

$$c_{j,k} = (W_\psi f)(2^{-j}, 2^{-j}k) \quad (2.41)$$

in which the wavelet coefficients correspond to the details of the original time-series signal.

In some instances, the general behavior of a time-series response is of more importance than the details of the same response. In these instances, it may be best to use wavelet decomposition to separate details from the signal to get a clearer picture of the general behavior of the data over time. In wavelet decomposition, the original signal is filtered using mother (ψ) and father (φ) wavelets to separate the data into an approximate function and its details (Figure 2.11). The father wavelet is also known as a scaling function and is directly related to the mother wavelet chosen. At each level of decomposition, the time-series data is represented by wavelet and scaling coefficients representing the details and approximations of the function respectively, which are calculated as [91]

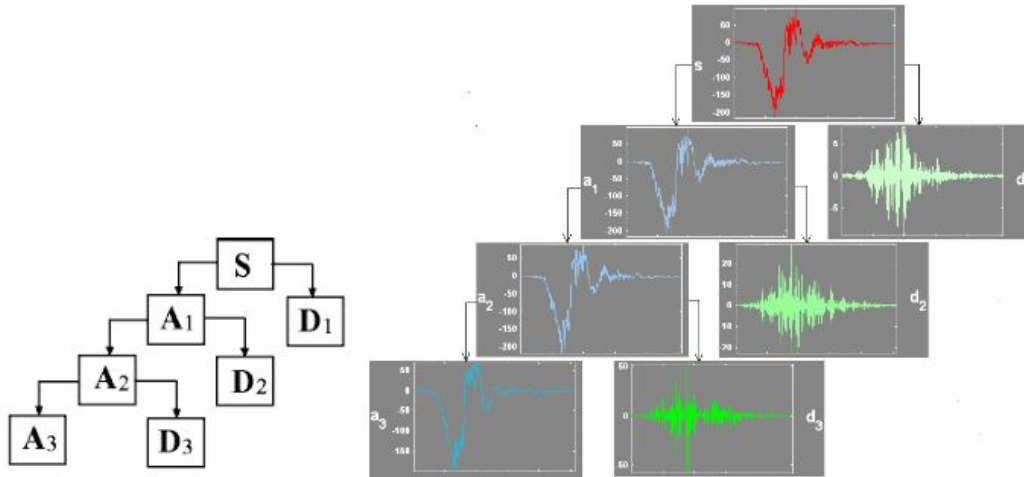


Figure 2.11: Illustrative Example of Wavelet Decomposition with Approximations and Details [22]

$$c_{j,k} = \int f(t)\psi_{j,k}(t)dt \quad (2.42)$$

$$a_{j,k} = \int f(t)\varphi_{j,k}(t)dt. \quad (2.43)$$

The coefficients can then be run through an inverse transform function that will generate the time-series representation of the approximate and detailed data. Multiple levels of decomposition are conducted to sequentially remove more detail from the approximated signal.

It is possible to use selective reconstruction to increase the detail incorporated into the approximated function being analyzed. In selective reconstruction, only those wavelet coefficients with significantly large values are used in the reconstruction process for the signal [91]. This allows the analyst to perform modified wavelet decomposition on a time-series data of interest. The analyst can look at the approximated function and the corresponding wavelet coefficients, which represent the removed detail at that decomposition level, and assess whether or not any of the wavelet coefficients warrant being added back into the approximated function. The process of determining whether a coefficient is significant is typically referred to as thresholding.

Jiang and Mahadevan [58] use a continuous wavelet transform (CWT) on model and system time-series data to calculate a wavelet cross-spectrum, which is then used to construct a time-frequency phase difference map of the two sets of time-series data. Box-plots are used to interpret the significance of the phase differences for validation purposes. This methodology is useful when attempting to validate non-stationary time-series behavior in which Fourier methods are not appropriate due to their inability to capture features in the time domain. The quantitative measures used to assess model validity are the time-frequency-based coherence and the phase relationships between the predicted and observed responses. The coherence function measures the correlation between the two sets of time-series data as a function of their frequency. The authors perform a continuous wavelet transform (CWT) on the response data to capture the total energy at each scale length.

The wavelet chosen as the filter for the CWT is the Morlet wavelet, which the authors justify based on its ability to provide a balance between time and frequency localizations and its scale being closely related to the Fourier period. The authors analyze the wavelet phase spectrum, which estimates the phase difference between two sets of time-series data with respect to scale and time, in order to assess model validity. A coherence map is constructed based on a smoothed cross-wavelet spectrum, which is normalized by the squared absolute value of the smoothed wavelet power spectrum. This map provides a visualization of localized correlation between time-series. If the cross-correlation is consistently zero throughout, the two sets of data are completely unrelated. If the cross-correlation is consistently equal to one, a perfect linear relationship is implied at the time and scale. Lastly, hypothesis tests are conducted to determine whether the time-frequency coherences between two sets of time-series data are statistically significant [58] .

Cheng et al. [22] use wavelet decomposition to generate a smooth approximation of the time-series data prior to the validation effort. In their methodology, Figure 2.12, a wavelet filter with desired features is selected and a threshold value is set for the correlation based metric [22]

$$R' = [\alpha_1(1 - \rho_{am}^j) + \alpha_2 \left| \frac{\tau_m^j}{T} \right| + \alpha_3 \left| \frac{A_t^j - A_s^j}{A_t^j} \right|] \times 100\% \quad (2.44)$$

where $\alpha_1, \alpha_2, \alpha_3$ are weighted coefficients, ρ_{am}^j is an adjustment to the classical correlation coefficient, T is the pulse duration, and A is the amplitude of the time-series approximation. Weighting factors are selected and the maximum decomposition level is determined. An iterative process of decompositions and approximations is conducted until a best level representation of the time-series data is achieved [22].

The methodology is demonstrated on a crash test model validation effort. The crash responses are decomposed at six levels using a fourth order Daubechies wavelet. Figure 2.13 shows one response, from both test and simulation data, at each decomposition level.

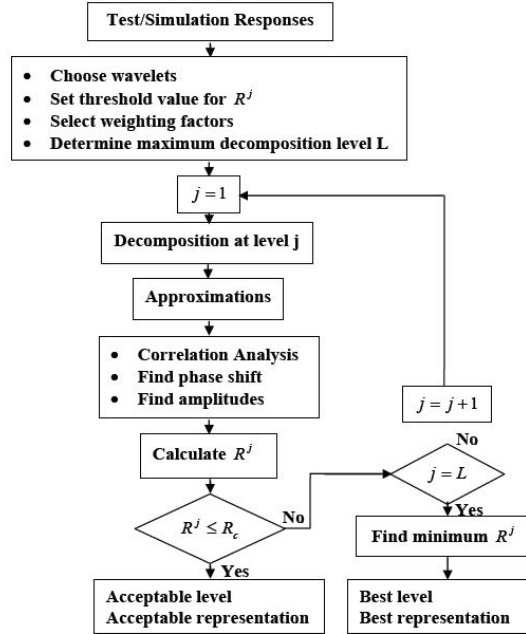


Figure 2.12: Procedure for Automatic Mult-Level Comparison [22]

The approximated signals become more similar at high decomposition levels due to the reduction of details, which are attributed to white noise. The basis behind the methodology is that wavelet decomposition will smooth out the noise in the signal so that only the general behavior of the signal is used for validation [22]. Cheng, in conjunction with various coauthors, has discussed the used of utilizing wavelet decomposition for model validation in the same or similar form as discussed above on numerous occasions [20] [21] [19] [18].

2.5.2 *EARTH Method.*

The error assessment of response time histories (EARTH) method decomposes time-series error into three components: phase, magnitude and frequency. These errors are calculated and then aggregated using a linear regression model based on subject matter expert ratings. This aggregation creates a regression-based validation metric for comparing model and system time-series data sets [110].

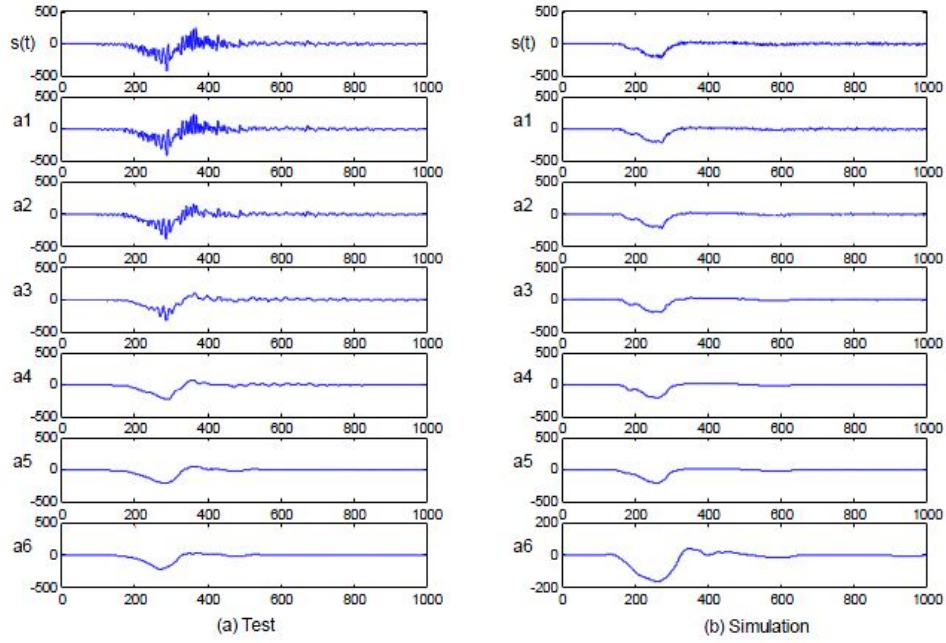


Figure 2.13: Wavelet Decomposition in Crash Test Data [22]

Phase Error. The measure used by the EARTH method to quantify phase error is based on cross-correlation. This measure was selected over the S&G and Russell metrics due to its greater sensitivity to phase differences. A tunable penalty function is assigned to the phase error measure to properly weigh the effect of local versus global phase errors. The established measure for phase error is defined as

$$Error_{phase} = e^{\frac{n_s - c}{r}} \quad (2.45)$$

where c represents the critical size phase shift at which its contribution is no longer negligible, and r changes the rate of phase error increase above the critical value of c . The values for c and r are application specific and determined by subject matter expertise.

Magnitude Error. To isolate the magnitude error, the errors due to phase and frequency are minimized. The global phase error is minimized through a time shift of

length n_* . To minimize the local phase and frequency errors, a DTW is applied to the time-shifted data using a specified cost function. Finally, to isolate the magnitude error, the L_1 norm on the warped time-shifted data is calculated

$$Error_{magnitude} = \frac{\|A^{[ts+w]} - B^{[ts+w]}\|_1}{\|B^{[ts+w]}\|_1} \quad (2.46)$$

where ts represents the time shift and w represents the DTW on the time-series data.

Slope Error. Slope error is calculated from the derivative of the time-series data. To minimize the effect of global phase error, the slope is calculated on the time-shifted data. Then the derivative of the slope is taken at each point. To minimize the effect of the local time shifts, a DTW is applied to the derivatives. Finally, to isolate the error due to the slope, the L_1 norm is calculated

$$Error_{slope} = \frac{\|A^{[ts+d+w]} - B^{[ts+d+w]}\|_1}{\|B^{[ts+d+w]}\|_1} \quad (2.47)$$

where the superscript $[ts + d + w]$ indicates the order of operations taken on the time-series data to get the resulting error contribution. Thus, the ts value represents the time shift in the data, d represents the computation of the derivative of the slope and w represents the warping of the resulting computation.

Regression Model. To aggregate the three measures of error into a single metric, a first degree polynomial regression model is constructed and tested using SME evaluations and ranking of empirical data sets. To build the regression model, SMEs are asked to evaluate N sets of time-series data. M of those data sets are used to construct the regression model and $N-M$ are used to test the predictive capability of the constructed model. The final SME ranking metric is of the form

$$R_p = 10 - (c_1 Error_{phase} + c_2 Error_{magnitude} + c_3 Error_{slope}) \quad (2.48)$$

where R_p is the predicted ranking of the model in which 10 implies a perfect match between the model and system.

Sarin et al. [110] use error measurements for three crash test head impact responses in order to illustrate their EARTH metric. These responses are head impactor displacement, head acceleration in the x-direction and neck force in the x-direction. The authors begin by visually inspecting the differences between the computational and test time histories for each response in terms of the three error components. In each case, three models were tested against three sets of test data. For example, Figure 2.14 provides the time-series plots for three sets of test and model data in regards to head acceleration in the x-direction. From the plot, it can be seen that while the computational models have acceptable phase errors, the errors in magnitude and slope are consistently larger than their corresponding test data set.

To construct the regression model, six SMEs were asked to provide rankings for 15 computational models related to the deceleration crash test data. The ratings ranged between 1 (worst match) and 10 (perfect match) and each SME provided one ranking per computational model and the average rating for each model was tabulated (Table 2.3). The rankings for the first ten models were used to construct the regression model and the last five were held back for testing of the regression model. The three components of error were calculated for all 15 models and used as coefficients in the regression metric equation. The EARTH rankings were compared graphically to the SME ranking for both the fitted and test data sets. The results of the test set comparison can be seen in Figure 2.15. For all five models, the EARTH metric fell within the range of SME rankings. Finally, the authors conducted the same graphical analysis on other error metrics. Figure 2.16 indicates that the

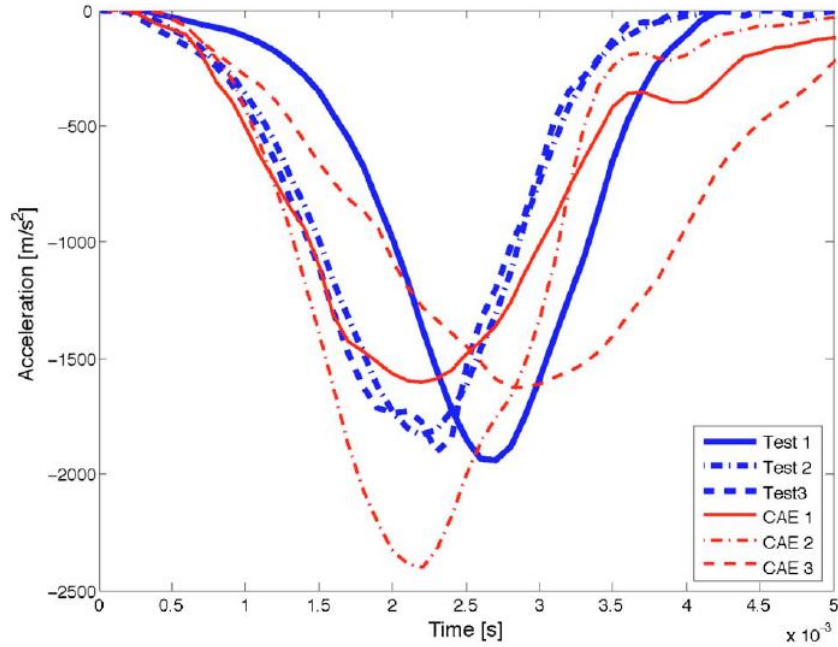


Figure 2.14: Time Series Data for Head Acceleration [110]

EARTH metric outperformed all other metric and that only metric to fall within the range of SME rankings in the majority of cases was the wavelet decomposition metric.

2.5.3 Hypothesis Testing on Intervals.

Sargent [104] proposes the use of a new statistical procedure for M&S validation that encompasses hypothesis testing where the required model accuracy is specified as an interval rather than a single value. The procedure uses Operating Characteristic (OC) curves and risk curves to assess and adjust Type I and Type II error as well as the desired number of observations for the model and the system. The method assumes that the data are Normal Independent and Identically Distributed (NIID) and that the variances for the populations represented by the two data sets are equal but unknown. There are statistical methods that can be used to obtain approximately NIID data such as the method of replications and the method of batched means [109].

Table 2.3: SME Ranking for 15 Computational Models [110]

SME ratings for the fifteen CAE models (first ten models used to build the regression-based validation metric; last five used to test it)

CAEID	SME ranking						Average ranking
	1	2	3	4	5	6	
1188	5	3	1	2	3	4	3.00
1189	3	4	4	3	3	3	3.33
1130	4	4	4	3	4	5	4.00
1047	5	4	5	4	5	5	4.67
1020	6	4	5	6	7	4	5.33
1041	6	5	5	6	6	5	5.50
1028	7	6	5	5	7	6	6.00
1005	8	7	7	6	8	6	7.00
1083	7	7	7	9	9	7	7.67
1052	8	7	8	10	10	8	8.50
1042	4	3	4	3	4	4	3.67
1100	5	4	3	3	4	6	4.17
1009	7	6	6	7	7	6	6.50
1016	7	7	7	8	9	5	7.17
1022	8	9	8	10	10	8	8.83

The statistical procedure focuses on determining whether the difference between the means of the model and system response is within a specified acceptable range of accuracy. The difference in the means, $D = \mu_M - \mu_S$, is considered insignificant if contained within the specified accuracy interval. The statistical form of the hypothesis test can be expressed as [104]

$$H_0 : L \leq D \leq U$$

$$H_1 : D < L \text{ or } D > U$$

where U and L are the upper and lower bounds of the specified model accuracy interval. In many cases $U = -L$ and in all cases the interval will contain the value zero [104].

Given the underlying assumptions, the test statistic used is based on the t-distribution. The α level associated with the test statistic is the maximum of α 's found at the bounds of

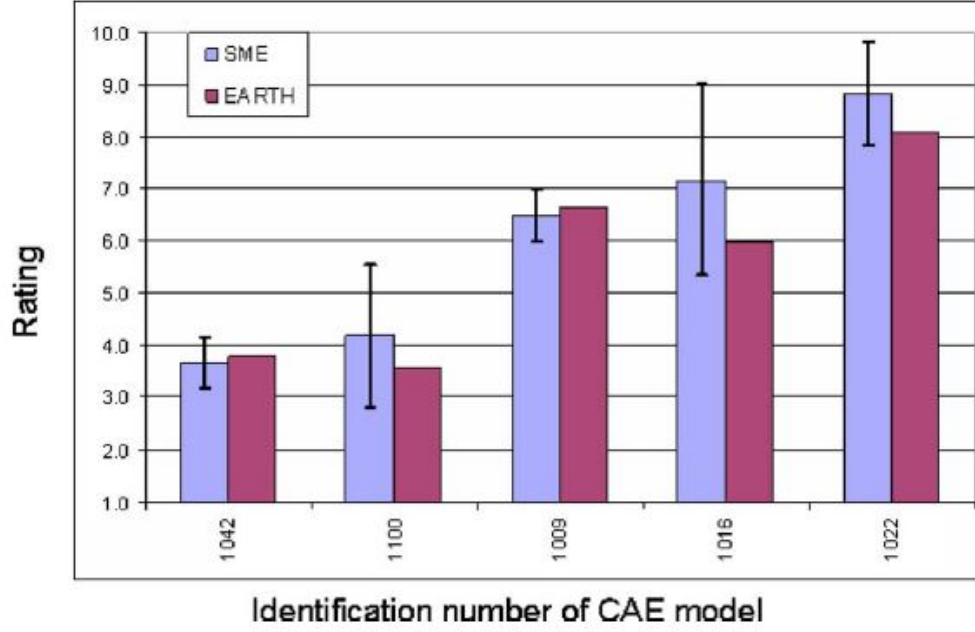


Figure 2.15: Graphical Comparison Between EARTH Metric and SME Ranking [110]

the accuracy interval, $\max(\alpha_L, \alpha_U)$. The acceptance region for the hypothesis test is given by $T \in A$, where A can be expressed as[109]

$$A = \left[t \left(\alpha_L, n_m + n_s - 2, \frac{\frac{L}{\sigma}}{\sqrt{\frac{1}{n_m} + \frac{1}{n_s}}} \right), t \left(1 - \alpha_U, n_m + n_s - 2, \frac{\frac{U}{\sigma}}{\sqrt{\frac{1}{n_m} + \frac{1}{n_s}}} \right) \right] \quad (2.49)$$

where n_m and n_s are the number of observations for the model and system, respectively.

The test statistic can be expressed as [109]

$$T \equiv \frac{\bar{M} - \bar{S}}{S_P \sqrt{\frac{1}{n_m} + \frac{1}{n_s}}} \quad (2.50)$$

where $\bar{M}$ and $\bar{S}$ are the sample means for the model and system respectively and S_P is calculated from the pooled variance estimator denoted by [109]

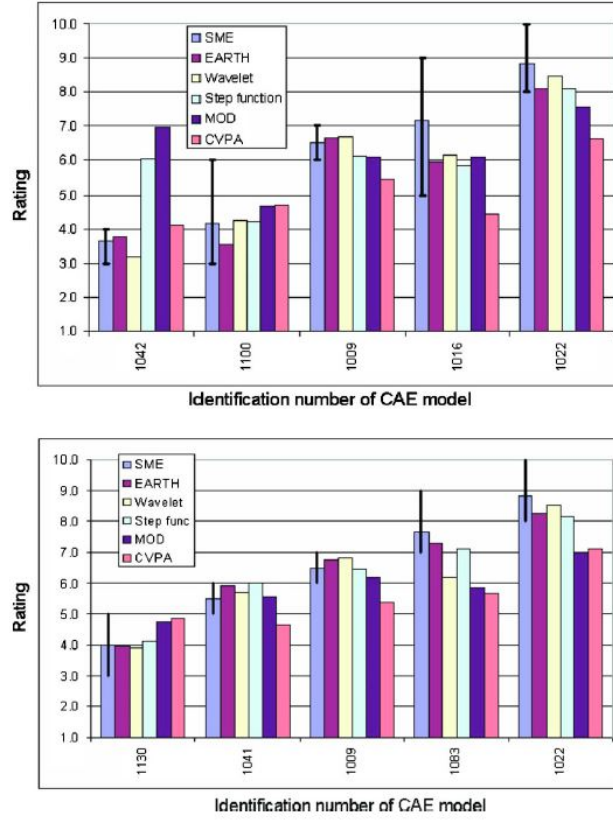


Figure 2.16: Graphical Comparison Between Various Error Metrics [110]

$$S_p^2 = \frac{(n_m - 1)S_m^2 + (n_s - 1)S_s^2}{n_m + n_s - 2}. \quad (2.51)$$

The first step of the interval statistical procedure is to determine the acceptable range of accuracy for the performance measure being tested. This interval depends on the context of the model application and is often determined by subject matter expertise. Next, an initial sample size for the model and system data are selected and collected for analysis of the mean and variances estimates. Next, different levels of β are selected. These β values are used, along with the mean and variance statistics, to construct risk curves which are then used to evaluate trade-offs between model builder's risk and model user's risk for different

β values. This trade-off analysis results in the selection of a single β value for the hypothesis test. Figure 2.17 illustrates the assessment of risk curves for two distinct β values. In this figure, the x-axis represents the difference between the difference in means and the lower bound of the acceptance region. The y-axis represents the probability of Type II error. The graph to the left side of zero represents the model user's risk and is the primary concern when validating. The appropriate choice for β is indicated by the risk curve that represents a lower probability of Type II error throughout the range of differences. In this example, that would correspond to $\beta = 0.4$ because the risk curve associated with that value falls underneath the risk curve associated with $\beta = 0.6$ [109].

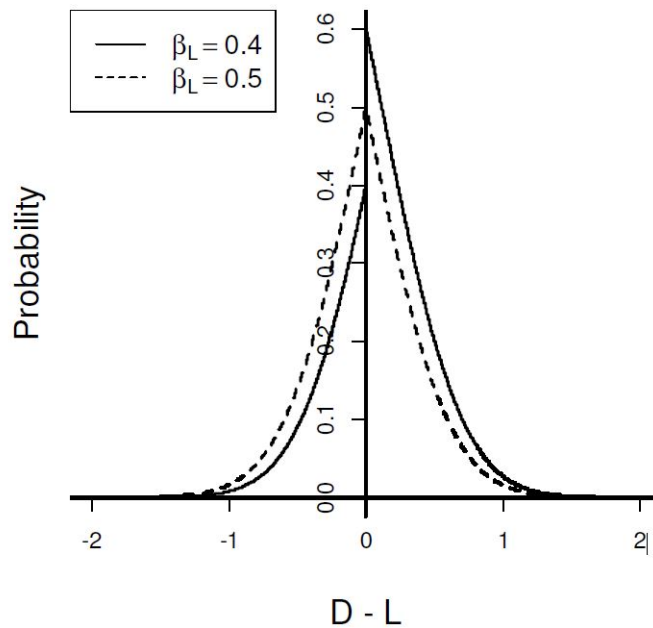


Figure 2.17: Risk Curve for Two β Options [109]

After a β value has been selected, more risk curves are generated by adjusting the model and system sample sizes while holding the β level constant. In the example illustrated by Figure 2.18, if the desire is to minimize model user risk, it would be best

to use $n_m = 40$ and $n_s = 15$ due to this risk curve falling below the risk curve generated using the other sampling option. Once the desired sample sizes has been established, it is necessary to collect the required observations and to recalculate the mean and variance statistics. These statistics, along with the specified sample sizes and β level are used to generate the acceptance region, A, and the T test statistic. If the test statistic falls outside of the acceptance region, the null hypothesis is rejected and the model is not validated. Conversely, if the test statistic falls within the acceptance region, the null cannot be rejected and the model is validated [109].

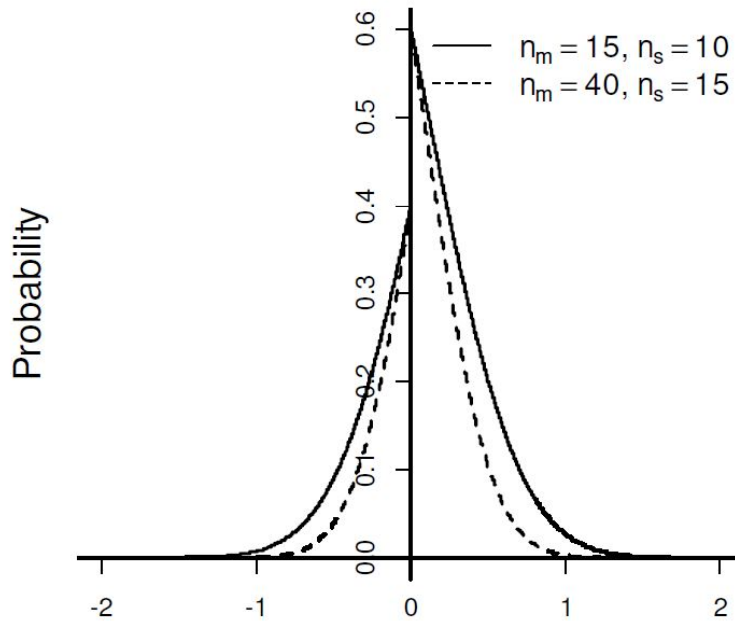


Figure 2.18: Risk Curve for Two Sample Size Options [109]

2.5.4 Validation in Engineering Applications.

The operations research (OR) and applied statistics communities have a robust library of research in regards to model validation. Likewise, the engineering community has their own set of model validation techniques. The majority of validation techniques used in the

assessment of engineering models, like those used in the OR and statistics communities, are rooted in computational science where validation is a process of measuring the agreement between model and experimental response data [16]. In many instances, the methodologies share common statistical procedures. These methods translate easily to use outside of the engineering community. A perspective which is addressed predominantly in the engineering model validation literature, and is mostly absent in the OR literature, is the necessity of validation at system configurations for which there is no experimental data. This necessity in the engineering community is a natural response to the engineering design process. In engineering design, the response characteristics for competing system designs are assessed using model representations of each design under consideration. The analysis of these models is an integral part of the selection for new engineering systems; therefore, it is imperative that the system designer have confidence in the model's predictive capabilities throughout the intended operational domain of the system. This necessitates validation of the model at more than just the tested configurations to aid in determination of the optimal engineering design.

Methodologies pertinent to engineering model validation have expanded greatly over the last fifteen years. A good starting point for reference is Liu et al. [76] which provides a sufficient summary of the four most common forms of validation metrics used in engineering applications. These categories are: classical hypothesis testing, Bayesian approaches, frequentest metrics, and area metrics.

2.5.4.1 Classical Hypothesis Testing.

The work presented in Hills and Trucano [52] has been used in several engineering design applications and is therefore, a suitable starting point for the classical hypothesis testing approach to engineering design model validation. Their methodology focuses on propagation of total uncertainty, either through Monte Carlo simulation or sensitivity analysis, followed subsequently by comparisons of experimental data against uncertainty

bounds around the model predictions. Hills and Trucano [52] begin the development of their methodology assuming a single factor as well as normally distributed and uncorrelated uncertainty in the parameters due to measurement error. Uncertainty bounds at a particular confidence level, generated via uncertainty propagation, are attached to the predicted response values where total uncertainty is considered the sum of the uncertainty from the parameters and the measurement error.

$$\sigma_{total}^2 = \sigma_{meas}^2 + \sigma_{pred}^2. \quad (2.52)$$

The 95% confidence bounds are constructed as $1.96\sigma_{total}$ from the mean response value. Figure 2.19 illustrates the response across the range of a parameter with the observed experimental response represented by the dotted points and the predicted model response, with associated confidence bounds, represented as the solid line with error bars. In the example, it is clear that the majority of the experimental points fall within their respective confidence bounds.

The authors note that, in their example, 7 percent of the experimental data falls outside of the prediction confidence bounds. They claim that this does not conclude that the model is invalid, even though more data fell outside the confidence bounds than anticipated. To develop an understanding for how one should appropriately test models using multiple measurements, Hills and Trucano first look at the comparison of one measurement and then extend to multiple measurements.

In the single measurement problem, a response measurement for the damped spring-mass problem associated with a particular factor value of $\omega = 1.5$ is chosen. To propagate the uncertainty, a Monte Carlo generates a population of 5,000 possible predictions and normally distributed random measurement error is added. Figure 2.20 provides the resulting histogram, where the vertical dashed lines represent the 95% confidence bounds on the model predictions. The mean predicted response is represented by the thick solid

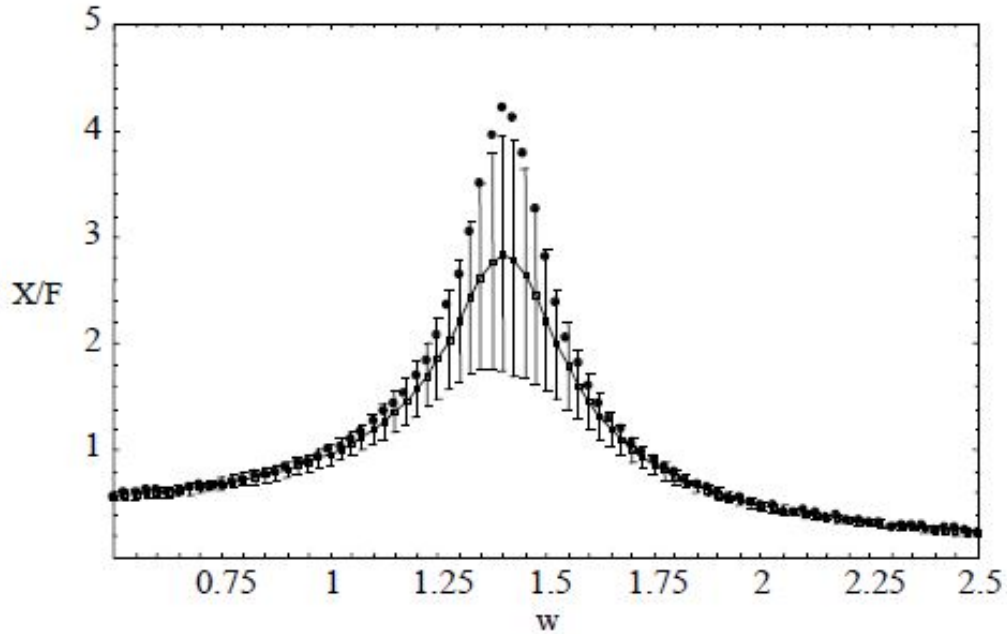


Figure 2.19: Single Factor Comparison of Experimental Results Against Prediction with Error Bounds for Damped Spring-Mass System [52]

vertical line, and the experimental data point is the thin solid vertical line. The experimental data point falls within the confidence bounds. In a statistical sense, this implies that we would expect only a 5 percent chance of rejecting a valid model at the specified confidence level.

Hills and Trucano extend this procedure to the more rigorous case in which experimental data exists at multiple values for the input variable of interest. For the two measurement case, the authors look at $\omega = 1.0$ and $\omega = 1.5$. A Monte Carlo simulation is used to generate a two-dimensional histogram (shown as a contour plot) using 50,000 samples. Figure 2.21 provides this contour with the dashed contour representing the 95% confidence region and the solid contours corresponding to various counts of a particular response prediction. The authors note the crudeness of the contours due to low number

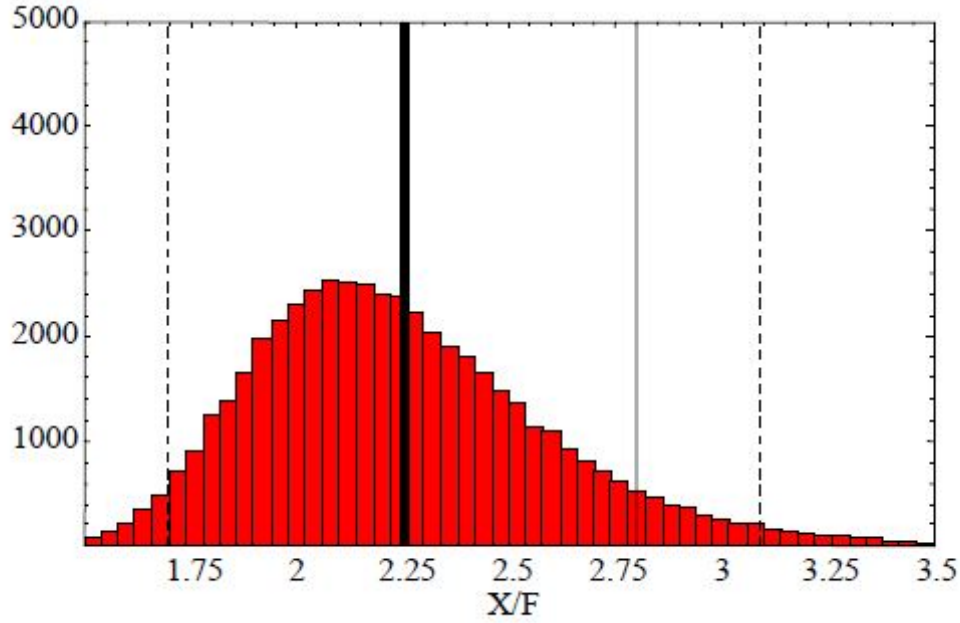


Figure 2.20: Validation of Single Measurement with $\omega = 1.5$ [52]

of Monte Carlo simulations run for the two-dimensional case. To adequately resolve the multi-dimensional contours, an increase in the number of simulation runs is required with an increase in the number of prediction variables. The 50,000 simulation runs were not enough to adequately smooth the contours for the two dimensional problem. The dark point in Figure 2.21 represents the measurement pair for the two values of ω . The paired experimental point lies within the contour bounds and therefore there is not sufficient statistical evidence, at the 95% confidence level, to invalidate the model. This method, unlike the one previously discussed, considers the interrelations between the acceptance ranges for the two separate measurements, which is necessary to accurately evaluate the confidence levels for multiple measurements.

The extension to this method is conceptually straightforward. If there are m measurement locations (factor levels), a $m - 1$ dimensional contour surface, including confidence bounds, can be generate to evaluate model validity. This method becomes

computationally inefficient as m becomes large due to the number of required Monte Carlo simulation runs necessary to characterize the contour surfaces. Hills and Trucano present two alternatives to the above methodology.

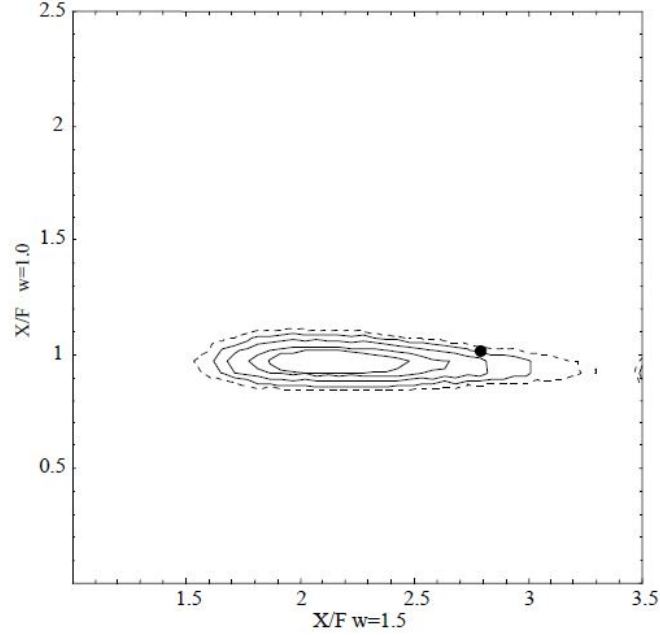


Figure 2.21: Validation of Two-Dimension Problem with $\omega = 1.0$ and $\omega = 1.5$ [52]

The first alternative uses the assumption that the total model uncertainty for the response is adequately modeled by jointly distributed normal probability density functions. The correlation between the uncertainty in different measurements can be estimated with either a sensitivity analysis or through Monte Carlo simulation, which provides the resulting means and covariance matrix, $\mathbf{V}$. Multivariate distributions symmetric about their means produce constant probability contours given by the following ellipses [52]

$$r^2 = \begin{bmatrix} p_1 - p_{mean1} & p_2 - p_{mean2} & \dots & p_n - p_{mean\ n} \end{bmatrix} \mathbf{V}^{-1} \begin{bmatrix} p_1 - p_{mean1} \\ p_2 - p_{mean2} \\ \vdots \\ p_n - p_{mean\ n} \end{bmatrix} \quad (2.53)$$

where each p_i represents a prediction of the desired response for the i^{th} factor level. Since each distribution is normal, r^2 can be related to the probability through the χ^2 distribution and the critical value for statistical test can be calculated as [52]

$$r_{critical}^2 = l_{1-\alpha}^2(n) \quad (2.54)$$

where l is the value associated with the $100(1 - \alpha)\%$ confidence region for the n parameters in the model. If the calculated r^2 is less than the critical value, then there is not enough evidence to declare the model invalid.

Hills and Trucano then propose a nonparametric approach to model validation that does not require the use of Monte Carlo simulation to propagate the uncertainty. The uncertainty of the model parameter is characterized by its mean and standard deviation. This requires that well defined statistical models exist for the input parameters of interest. As in the previous section, the 95% confidence region is an ellipse defined by Eq. 2.53. Furthermore, the model prediction as a function of the model parameter is plotted along with the confidence bound. The model is then considered valid if the line formed by the model predictions passes through the ellipse at any point Fig 2.22. The advantage of this method is that instead of performing thousands of Monte Carlo simulations, optimization techniques can be used to evaluate whether or not there exists values for the model input parameter that lie within the confidence ellipse. The authors demonstrate the extension of this methodology to multiple measurements over multiple input parameters. In this case, r^2 represents a higher dimensional hyper ellipse of the form [52]

$$r^2 = \mathbf{d}^T \mathbf{V}^{-1} \mathbf{d} \quad (2.55)$$

where the vector $\mathbf{d}$ represents the differences between the measured and mean parameter and response values and $\mathbf{V}^{-1}$ is a concatenated matrix of the form [52]

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{\text{param}} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{\text{meas}} \end{bmatrix} \quad (2.56)$$

Optimization techniques are used to find the values of the input parameters that minimize r^2 . If the minimized value for r^2 is less than the established critical value, then there is insufficient statistical evidence to claim that the model is invalid.

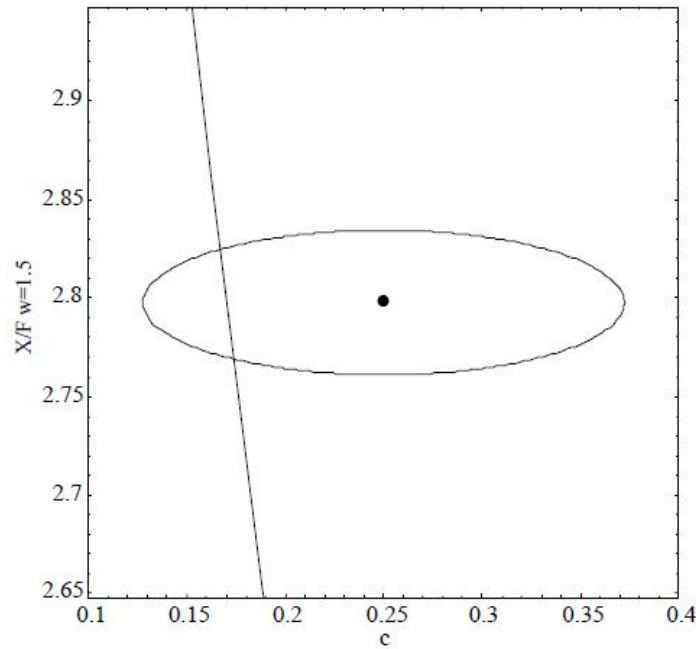


Figure 2.22: 95% Confidence Ellipse and Predictions with Varying Parameter Values (Straight Line) [52]

A last extension to the methods of Hills and Trucano deals with cases in which the model in question is designed to approximate a particular behavioral response of a system within an acceptable level of error rather than to reproduce the behavior exactly. This additional acceptable error is represented as a rectangular region around the observation measurement (Figure 2.23). If there is at least one point within the rectangular region that lies within the 95% confidence ellipse, the model cannot be declared invalid. This methodology can be extended to the multi-measurement space where the acceptable error region becomes a multi-dimensional hyper prism [52].

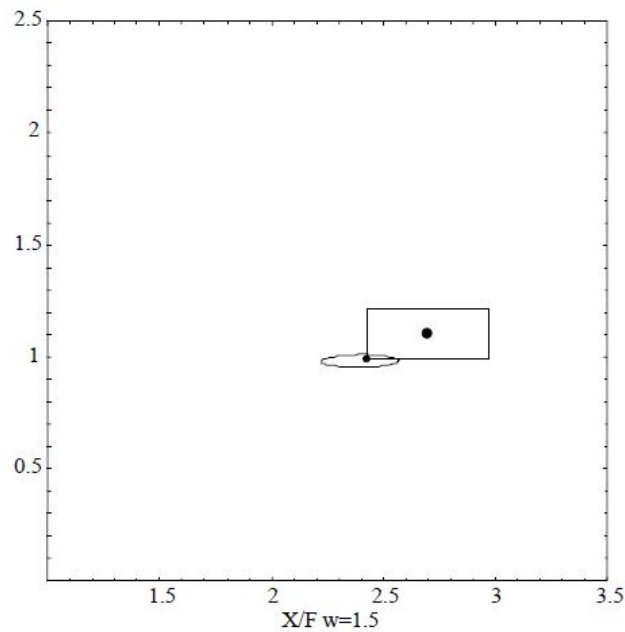


Figure 2.23: 95% Confidence Ellipse on Predictions with Rectangular Region Around Measurement for Additional Acceptable Error [52]

Many authors have used the methodologies proposed in Hill and Trucano [52] as a foundation for their own model validation applications. Ghanem et al. [45] use polynomial chaos expansions (PCE) to represent the uncertainty of model parameters in

a virtual dynamical system. Uncertainty is then propagated using Monte Carlo simulation and a hypothesis test metric which incorporates the observed response's location within a generated probability distribution function with corresponding confidence intervals, is utilized to assess model validity. Hills [50] illustrates the application of the weighted least squares metric (r^2) in three models validation efforts: Burger's equation, convective dispersion equation, and a contact resistance model. Buranathiti et al. [13] illustrates the merits of three validation metrics using a sheet metal flanging process application. The three validation metrics demonstrated are a graphical approach, confidence intervals, and the weighted least squares (r^2) metric. The latter two approaches were discussed in Hills and Trucano [52].

Another validation approach that uses the r^2 metric is presented by Chen et al. [15]. Due to its use of many of the same analytical tools as the methodologies proposed in this research effort, albeit in a different manner, an extensive review of the method is provided. Chen et al. [15] make use of Response Surface Methodologies (RSM) to create meta-models, which act as less costly approximations of the models under test, for the purpose of uncertainty propagation. This approximation of the design space through the construction of response surfaces is a concept that lends itself greatly to model evaluations at points within the design space for which there is no observed data. This is a key consideration for the research discussed later in this document.

Chen et al. [15] propose a four phase approach to model validation, diagrammed in Figure 2.24. The first phase is the problem setup, where models for all known uncertainty types are established. In addition to the investigation of uncertainty, phase 1 also incorporates the creating of the design space using design of experiments (DOE) or other statistical sampling techniques such as Latin hypercube designs. Phase 2 is the experiment stage. In this phase, only one physical experiment is required at each design point. This is critical to the methodology's applicability due to the costs associated

with conducting test for data collection. In Phase 3, simulations of the model are run in order to construct a response surface based on samples of simulation results. Uncertainty propagation on the model is then conducted using Monte Carlo simulations of the meta-model. Once a sufficient number of simulation runs are obtained, estimates of uncertainty in the parameter estimates are established. The Monte Carlo simulation provides estimates of the PDFs for the data which are then analyzed for normality. If normality is rejected, the data are transformed to achieve normality due to the underlying assumptions of the r^2 validation method. Phase 4 is the model validation phase, where the r^2 metric is used to assess model validity. The validation methodology is demonstrated for both the single and multiple design point problems. For brevity, and relevance to the proposed research, only the multiple design space problem is examined further.

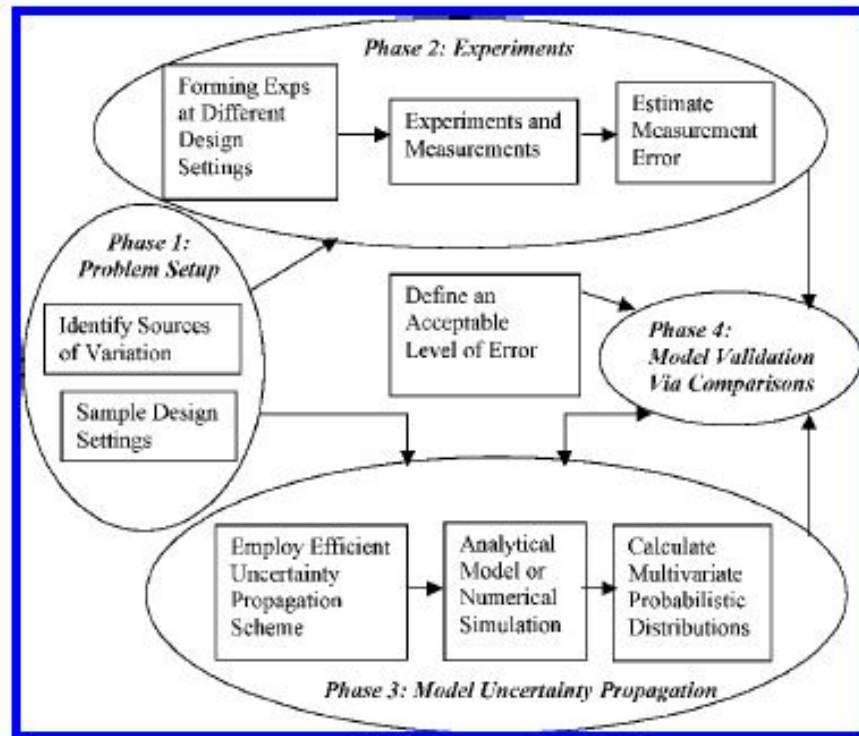


Figure 2.24: Procedure for Model Validation [15]

The multiple design space problem is illustrated using a sheet metal flanging process. Two design variables, flange length L and gap space g are used to create a sample design space. The space chosen consisted of low and high configurations for each design variable, as well as a center point. The variation in gap space was assumed normally distributed with a standard deviation of 0.6 mm. The variation in the flange length was ignored due to the flanging accuracy tolerance being insignificant compared to other factors. In Phase 3, 243 simulations were conducted to create the second-order response surface meta-model which represented the final flange angle as a function of the two design variables as well as five uncontrollable design parameters. Once it was deemed that the SSE from the model was adequate, the surface was used, along with 200,000 random samples from MCS, to propagate the model uncertainty in the form of PDFs (Figure 2.25). A Kolmogorov-Smirnov (K-S) test was conducted to test for normality and based on the results of that test, the distribution of the PDFs were deemed jointly normal. Finally, the r^2 metric was calculated and was found to be smaller than the critical value and therefore there was not enough statistical evidence to conclude that the polynomial model was invalid.

Another model was validated using the same process to show the necessity of data transformations when the PDFs are not jointly normal. In this case, the response distributions are non normal at each design point, and a K-S test rejected the null hypothesis at $\alpha = 0.05$. The left side of Figure 2.26 shows the response distribution at one design point compared against a smooth normal distribution curve. A natural log transformation is applied to both the response and model inputs and a polynomial response surface using the transformed data is constructed. The right side of Figure 2.26 shows the same design point with the response distribution from the transformed data. Visually, it is clear that the transformation was successful in achieving normality.

Lastly, errors associated with the response surface are estimated as the mean and variance associated with the differences between the simulation and the response surface

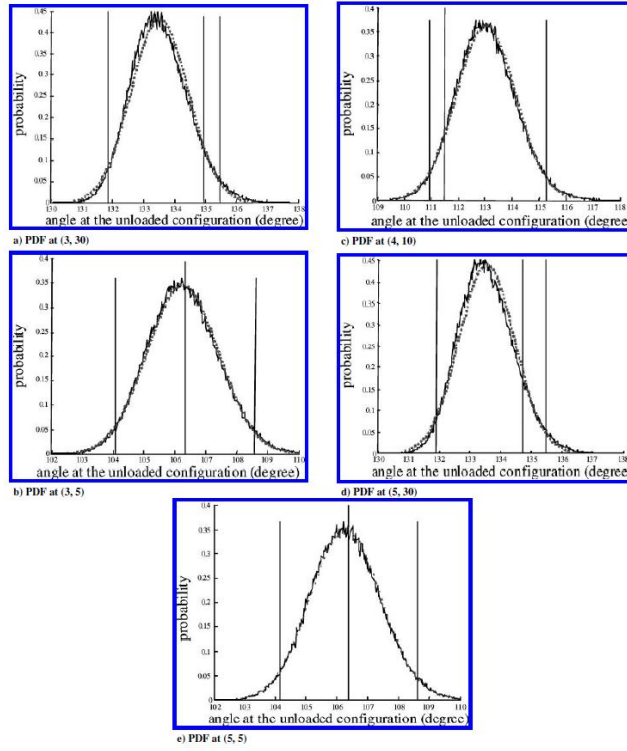


Figure 2.25: PDFs for Each of Five Design Points [15]

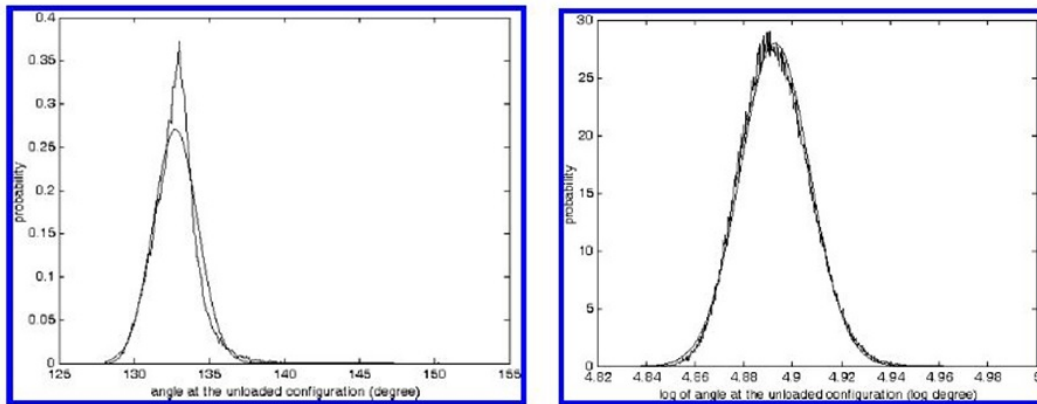


Figure 2.26: Response PDF Without(Left) and With(Right) Transformation [15]

constructed using Latin hypercube sampling. These simulated errors are added back to distributions from the Monte Carlo simulation. In addition, an acceptable level error is established and accounted for by adjusting the physical experiment results. These modifications are depicted in Figure 2.27. When considered, the modified confidence limit for rejecting is lowered due to the relaxed rejection criteria from the additional error considerations. In the multiple design point example, the 95% confidence limit reduced to about 40% due to a reduction in the r^2 metric.

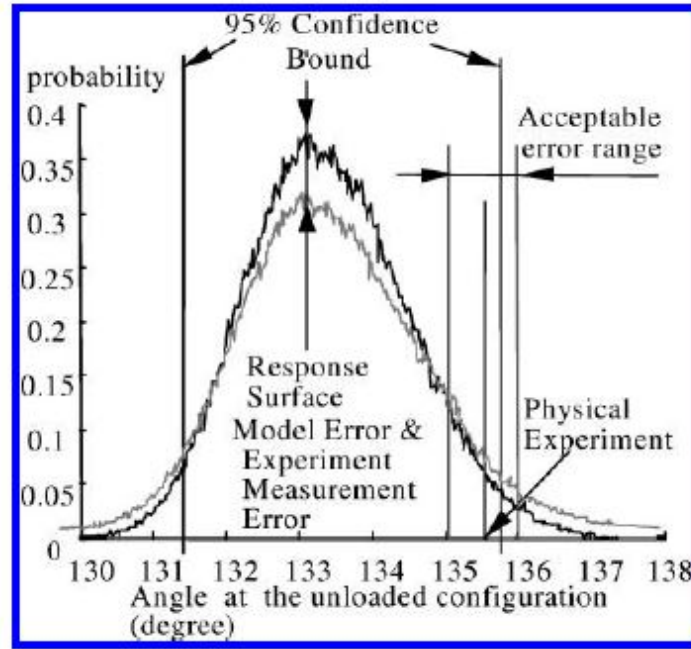


Figure 2.27: Response PDF With RSM Errors and Acceptable Level of Error [15]

Rebba and Mahadevan [96] demonstrate a multivariate extension to the classical hypothesis testing approach. They combine two similarity measures, Mahalanobis distance and covariance similarity into an aggregate metric for measuring the overall discrepancy among multiple response variables. Let matrix $\mathbf{X}_{n \times p}$ be the multivariate output where n is the number of realizations and p is the number of different response variables. Let μ_0 be

the vector of means values for each column in $\mathbf{X}$. Similarly, let $\mathbf{Y}$ be a matrix of the same size, representing the observation data. Let $\bar{\mathbf{Y}}$ be the mean vector and $\mathbf{S}$ be the covariance matrix of $\mathbf{Y}$. The Mahalanobis distance is computed as [96]

$$d^2 = n(\bar{\mathbf{Y}} - \mu_0)^T \mathbf{S}^{-1} (\bar{\mathbf{Y}} - \mu_0). \quad (2.57)$$

The observed data is deemed to be from the population of the model data if the calculated d^2 is less than a critical value, which is calculated as [96]

$$d_{critical}^2 = (fp/(f - p + 1))F_{p, f-p+1, \alpha} \quad (2.58)$$

where $f = n - 1$ and $F_{p, f-p+1, \alpha}$ comes from the F distribution.

The coviance similarity measure is based on the log-likelihood ratio test and is calculated as [96]

$$R = -(2m/f)\log(\lambda) \quad (2.59)$$

where $m = f - 2d$, $f = n - 1$, and $d = 2p^2 + 3p + 1/12(p + 1)$. The likelihood ratio is estimated as [96]

$$\lambda = e^{\frac{fp}{2}} |S|^{\frac{f}{2}} |\Sigma_0|^{-\frac{f}{2}} [etr(\frac{-1}{2}f \sum_0^{-1} S)] \quad (2.60)$$

where Σ_0 and S are the population and sample covariance respectively and etr represents the exponential trace of a matrix ($e^{tr(A)}$). The two metrics are aggregated and take the form of Equation 2.59 with $f = n$ and $d = 2p^2 + 9p = 11/12(p + 3)$. In additon, the likelihood ratio is modified with an extra term [96]

$$\lambda = e^{\frac{fp}{2}} |S|^{\frac{f}{2}} |\Sigma_0|^{-\frac{f}{2}} [etr(\frac{-1}{2}f \sum_0^{-1} S + (\bar{\mathbf{Y}} - \mu_0)(\bar{\mathbf{Y}} - \mu_0)^T)] \quad (2.61)$$

The hypothesis that $\mathbf{X}$ and $\mathbf{Y}$ are similar is rejected, at a particular significance level, if R exceeds a threshold determined from the χ^2 distribution with $p(p + 3)/2$ degrees of freedom.

2.5.4.2 *Bayesian Validation Metrics.*

Validation methodologies based on the principles of Bayes Theorem have become a popular alternative to the classical hypothesis testing methods previously discussed. These methods, centered around the idea of Bayesian Hypothesis Testing (BHT) and the use of the Bayes factor as a validation metric, have been used in numerous applications of model validation to include univariate and multivariate response systems, evaluations of model predictive capabilities outside the tested domain, and for the evaluation of dynamical systems [98] [94] [95] [97] [96] [57] [56].

The basis for BHT centers around the relative posterior probabilities of two competing models or hypothesis, denoted as M_i and M_j , which can be expressed as [98]

$$\frac{P(M_i|observation)}{P(M_j|observation)} = \left[\frac{P(observation|M_i)}{P(observation|M_j)} \right] \left[\frac{P(M_i)}{P(M_j)} \right] \quad (2.62)$$

where the first set of square brackets on the right side of the equation is known as the Bayes factor. The Bayes factor provides an estimate of the ratio of relative likelihoods between the null and alternative hypotheses. If the Bayes factor is greater than one, then an inference can be made that the null hypothesis is favored over the alternative [98] [94] [97].

Rebba et al. [98] introduces the easiest form of BHT used on univariate systems with pass/fail test data. In these instances, the null hypothesis is that the predicted and true failure probabilities are the same. The alternative hypothesis is that they are not the same. Let x_0 and x be the predicted and true failure probabilities respectively. If k failures are observed out of n experiments, the probability of observing the data given that the true failure probability is equal to x follows a binomial distribution where [98]

$$P(k|x, n) = C_k^n x^k (1 - x)^{n-k} \quad (2.63)$$

Under the null hypothesis that the predicted and true failure rates are equal, this probability, $P(data|H_0)$ is estimated by substituting x_0 into Equation 2.63. Assuming no prior knowledge regarding the alternative, a uniform distribution in $[0, 1]$ is assumed for the prior density under the alternative, $f(x|H_1)$. The Bayes factor is calculated as [98]

$$\begin{aligned} B(x_0) &= \frac{P(data|H_0 : x = x_0)}{P(data|H_1 : x \neq x_0)} \\ &= \frac{C_k^n x_0^k (1 - x_0)^{n-k}}{\int_0^1 C_k^n x^k (1 - x)^{n-k} f(x|H_1) dx} \\ &= (n + 1) C_k^n x_0^k (1 - x_0)^{n-k}. \end{aligned} \quad (2.64)$$

For all cases in which a uniform prior is assumed and the data is pass/fail, the Bayes factor, $B(x_0)$ is the posterior PDF of x evaluated at the predicted value x_0 (Figure 2.28).

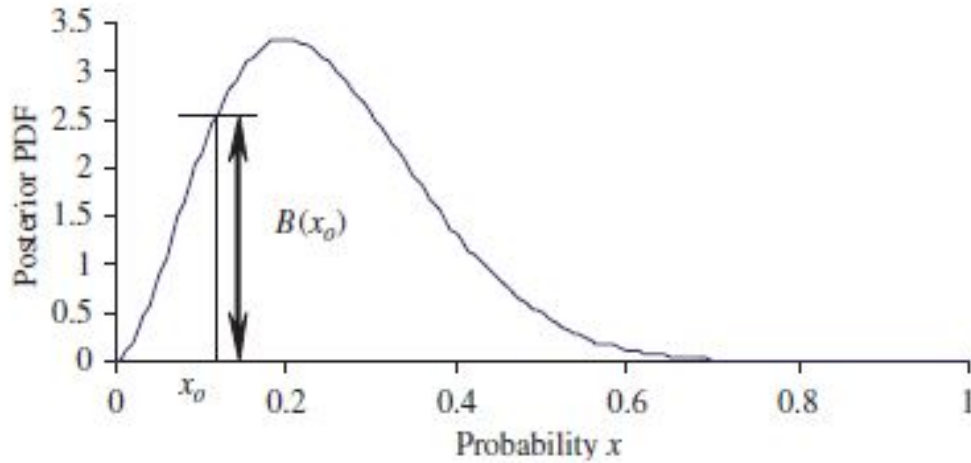


Figure 2.28: PDF for Probability of Failure [98]

When the response data is not pass/fail, the likelihood function is proportional to the probability density of the data y given the parameter x . Let x be a response quantity with

density function $f(x)$ and x_0 remains the predicted quantity from the model. The probability of observing the data given the null hypothesis is $L(x_0) = \epsilon f(y|x_0)$. The probability of observing the data under the alternative is then $\int L(x)g(x)dx$ where $g(x)$ is the prior density under the alternative hypothesis. Similar to the pass/fail scenario, equal priors are assumed ($g(x) = f(x)$) and the Bayes factor is computed as [98]

$$\begin{aligned}
 B(x_0) &= \frac{P(data|H_0 : x = x_0)}{P(data|H_1 : x \neq x_0)} \\
 &= \frac{L(x_0)}{\int L(x)f(x)dx} \\
 &= \frac{f(y|x_0)}{\int f(y|x)f(x)dx} \\
 &= \frac{f(x|y)}{f(x)} \Big|_{x=x_0}
 \end{aligned} \tag{2.65}$$

and therefore the Bayes factor becomes the ratio of posterior to prior PDFs when $g(x) = f(x)$. This ratio of PDFs is illustrated in Figure 2.29.

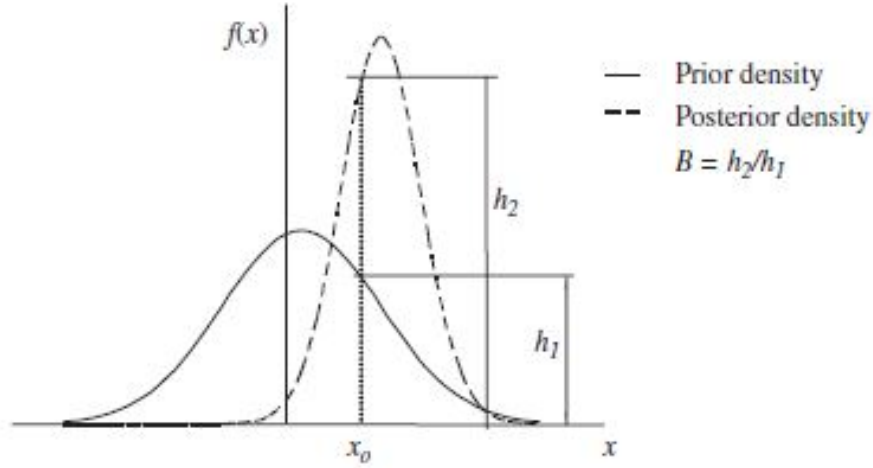


Figure 2.29: Ratio of Posterior to Prior Densities [98]

Rebba et al. [98] estimate the error in prediction using a nonlinear response surface constructed from the input data error, discretization error, and uncertainty quantification error, to estimate the statistics for the model form error. Sensitivity analysis was conducted to show which parameters and error components significantly contributed to variations in the validation metric.

Rebba and Mahadevan [98] extend the Bayes factor approach to the multivariate case such that the metric is defined as the ratio of posterior joint probability density to the prior joint probability density. The overall likelihood function is then proportional to the multi-dimensional Gaussian distribution. The Bayes factor is adjusted to [98]

$$B = \frac{\exp(-\frac{1}{2}(\mathbf{y} - \mathbf{x})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{x}))}{\int \int \dots \int \exp(-\frac{1}{2}(\mathbf{y} - \mathbf{x})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{x})) f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p) d\mathbf{x}_1 d\mathbf{x}_2 \dots d\mathbf{x}_p} \quad (2.66)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the sets of predicted and observed responses respectively and $\mathbf{V}$ is the covariance matrix of the observed data. The authors note that this methodology can be used for time-series data by simply replacing a response quantity x_i with $x(t)$. In order to assure compliance with the normality assumptions of standard statistical tests, the authors suggest the use of data transformations on non-normal data. The proposed transformation methods discussed in Rebba and Mahadevan [98] include the Rosenblatt, Nataf, Power and Modulus transformations, as well as Pericchi's Bayesian method. The suitability of each method is discussed and ultimately the Box-Cox transformation is used to normalize the data.

Jiang and Madadevan [56] propose an alternate form to the Bayes factor for multivariate model validation in which the data is Gaussian. In this form, the likelihood function of the multiple observations, $L(Y)$, is calculated as [56]

$$L(Y) \propto \frac{|\Sigma|^{-1/2}}{(2\pi)^{m/2}} \exp \left[-\frac{1}{2} (Y - \mu)^T \Sigma^{-1} (Y - \mu) \right] \quad (2.67)$$

where Y is an $m \times n$ matrix representing m variables each with n observations, Σ is an $m \times m$ covariance matrix of all variables and μ is a vector of the m corresponding mean values. The multivariate validation problem becomes a hypothesis test where $H_0 : \mu = E_0$ and $H_1 : \mu \neq E_0$ with $\mu|H_1 \sim N(\rho, \Lambda)$. The Bayes factor, converted into the logarithmic scale, is calculated as [56]

$$b_M = \frac{1}{2} \ln \left(\frac{n|\Lambda| + |\Sigma|}{|\Sigma|} \right) + \frac{n}{2} \left[(\bar{Y} - \rho)^T (n\Lambda + \Sigma)^{-1} (\bar{Y} - \rho) - \bar{Y}^T \Sigma^{-1} \bar{Y} \right] \quad (2.68)$$

Rebba and Mahadevan [95] proposes a Bayesian based methodology for quantifying the confidence of the model in the application domain through extrapolation of the confidence in the validation domain. Two cases are considered. The first extrapolates validation inferences for a response quantity in which data are absent. The second extrapolates at inputs conditions which were not tested. Both cases make use of Bayesian networks. These networks are directed acyclic graphical representations with nodes representing random variables and arcs representing the dependencies between variables. Each node possesses a probability density function and each arc carries a conditional density function based on the parent/child nodal relationship. The entire network is then represented as a joint probability density function. The marginal PDF for a particular variable is obtained through the use of Markov Chain Monte Carlo techniques which integrate the joint PDF over all remaining variables [95].

Rebba and Mahadevan [97] propose an interval based Bayesian approach to model validation when dealing with uncertainties. In the interval methodology, the null hypothesis is that the model prediction θ is near the observed θ_0 based on a predefined accuracy requirement denoted as ϵ . It is assumed that the null and alternative hypotheses have

equal priors. If prior information on the model prediction is known and in the form of a probability density $f(\theta)$, then the Bayes factor can be expressed as [97]

$$B = \frac{\int_{\theta_0-\epsilon}^{\theta_0+\epsilon} f(\bar{x}|\theta)f(\theta)d\theta}{\int_{-\infty}^{\theta_0-\epsilon} f(\bar{x}|\theta)f(\theta)d\theta + \int_{\theta_0+\epsilon}^{\infty} f(\bar{x}|\theta)f(\theta)d\theta}. \quad (2.69)$$

If one assumes that $f(\bar{x}|\theta)$ is $N(\theta, \sigma^2/2)$ and the null hypothesis has a density that is $N(\theta_0, \beta^2)$, where β is some constant, then the posterior null probability, also referred to as the confidence in the assessment, can be written as [97]

$$\begin{aligned} C = P(H_0|data) &= \frac{B}{B+1} = \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} f(\theta|\bar{x})d\theta \\ &= \Phi\left(\frac{n\lambda[\bar{x} - \theta_0]}{\sigma^2} + \frac{\epsilon}{\lambda}\right) - \Phi\left(\frac{n\lambda[\bar{x} - \theta_0]}{\sigma^2} - \frac{\epsilon}{\lambda}\right) \end{aligned} \quad (2.70)$$

where $\lambda = \beta\sigma / \sqrt{n\beta^2 + \epsilon^2}$. The posterior null probability can be simplified by assuming that θ has a very flat density and letting $\beta \rightarrow \infty$. In this case the expression for C reduces to [97]

$$C = 2 - \left[\Phi\left(\frac{\sqrt{n}|\bar{x} - \theta_0|}{\sigma^2} - \frac{\epsilon\sqrt{n}}{\sigma}\right) + \Phi\left(\frac{\sqrt{n}|\bar{x} - \theta_0|}{\sigma^2} + \frac{\epsilon\sqrt{n}}{\sigma}\right) \right]. \quad (2.71)$$

Rebba and Mahadevan [97] use this C value as the basis for a reliability metric $r = P(-\epsilon < D < \epsilon)$ where D is the difference between the model and system outputs. The model prediction is only accepted if r is greater than or equal to some adequacy requirement c .

Let $X = \{x_1, x_2, \dots, x_n\}$ be replicated experimental measurements for a particular response at the same input conditions. Suppose that $\bar{X}$ follows a normal distribution $N(\bar{x}, s/\sqrt{n})$ where $\bar{x}$ and s are the sample mean and sample standard deviation, then the model reliability metric can be written as [97]

$$P(H_0) = r = \Phi \left[\frac{\sqrt{n}(\epsilon - |\bar{x} - \theta_0|)}{s} \right] - \Phi \left[\frac{\sqrt{n}(-\epsilon - |\bar{x} - \theta_0|)}{s} \right] \quad (2.72)$$

Table 2.4 provides a brief discussion regarding the advantages and disadvantages of classical and Bayesian point null and interval validation metrics [97].

Table 2.4: Statistical Methods for Model Validation [97]

Method	Advantages	Disadvantages
Classical point null hypothesis testing (p -values)	Simple to estimate, well-established mathematical methods, objective and frequentists, addresses adequacy and very stringent accuracy requirement	Confusing and often misused, ignores Type II error, cannot incorporate prior knowledge, no direct assessment of the model, cannot be used for extrapolation purposes
Bayesian point null hypothesis testing (Bayes factors)	Includes both Type I and II errors, incorporates prior knowledge, direct assessment of the model, addresses adequacy and very stringent accuracy requirement, can be used for extrapolation to application domain	Relatively difficult to compute, likelihood is assumed sometimes, needs prior distributions (based on model, however)
Probability intervals (classical)	Simple to use, no prior assumptions on the distributions, addresses adequacy and no accuracy requirement	Not clear how to extrapolate to the application domain, possibility of misuse in interpreting the confidence level
Interval hypothesis testing (classical)	Better method than point null tests, easy to estimate and interpret the result, addresses accuracy and adequacy	No direct assessment of hypothesis, ignores Type II error, cannot incorporate prior knowledge, cannot be used for extrapolation purposes
Interval hypothesis testing (Bayesian)	Better than point null tests, includes both types of errors, incorporates prior knowledge (model prediction), direct inference on the null hypothesis, addresses accuracy and adequacy, can be used for extrapolation	Difficult to estimate, likelihood is assumed sometimes
Model reliability formulation	No priors required, direct assessment of model prediction quality, addresses accuracy and adequacy, relatively easy to compute	Not clear how to extrapolate to the application domain
Decision theoretic approach	Incorporates prior knowledge, easy to interpret the results, can include accuracy and adequacy	Subjective definition of utility functions, use for extrapolation purposes is questionable
Effect size indicators	Simple to compute, several available, addresses adequacy but do not address accuracy	Qualitative measures, no direct assessment of confidence in model, cannot be used for extrapolation purposes

Jiang and Mahadevan [57] propose an energy-based Bayesian wavelet methodology for the validation of multivariate models consisting of dynamic time-series data responses. The time-series responses for both model and system are first decomposed using wavelet packets, which are linear combinations of wavelet functions. The signals are decomposed, to a desired decomposition level, into their respective wavelet and scaling coefficients and then reconstructed into a discrete wavelet packet transform coefficient, which is used to calculate the energy contribution from each component. The energy contribution is calculated as [57]

$$E^f = \sum_{i=1}^{2^j} E_i^f = \sum_{i=1}^{2^j} \int_{-\infty}^{\infty} [f_j^i(t)]^2 dt \quad (2.73)$$

where $f_j^i(t)$ is the discrete wavelet packet transform coefficient. The component energies for each set of decomposed coefficients are then sorted in descending order to determine the level of contribution each coefficient has to the overall energy of the signal. Figure 2.30 provides an example where a three-level discrete wavelet packet decomposition has resulted in eight component energies, which correspond to the different approximation and detail coefficients. In this example, it can be seen that the vast majority (99.14%) of the component energy is captured in the first four coefficients. A cut off is established and the signal is reconstructed using only the coefficients corresponding to the majority of the component energy.

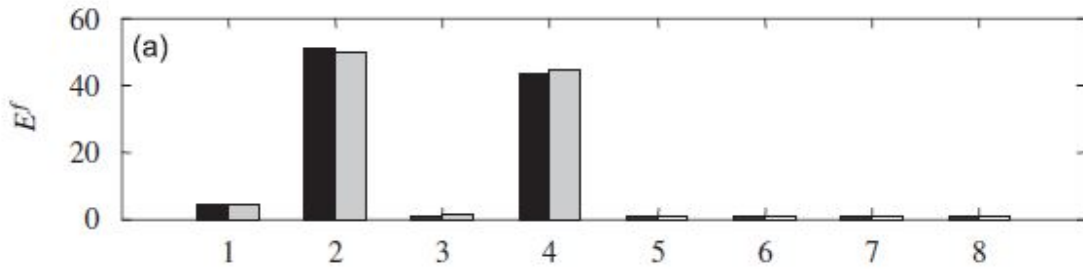


Figure 2.30: Component Energies for Eight Coefficients [98]

The appropriateness of the reconstructed signal is evaluated using cross-correlation and cross-coherence. The similarity between the original signal and the reconstructed signal within the time domain is measure with cross-correlation computed as [57]

$$r(k) = E[\tilde{f}(t)f(t - k)]/E[f(t)f(t - k)], \quad k = 0, 1, \dots, N - 1, \quad (2.74)$$

where k is the time delay, the numerator is the cross-covariance, and the denominator is only the variance of $f(t)$. The cross-correlation ranges between zero and one with a zero indicating no similarities between the signals and a one representing a perfect

match between the two signals. Clearly, a higher cross-correlation is desirable. The similarity between the original signal and the reconstructed signal in the frequency domain is measured using cross-coherence, which is written as [57]

$$c(\omega) = \frac{S_{\bar{f}f}(\omega)}{S_f(\omega)} \quad (2.75)$$

where $S_{\bar{f}f}(\omega)$ is the cross-power-spectrum between the original and reconstructed signal and $S_f(\omega)$ is the power spectrum of the original signal. Similarly to cross-correlation, a cross-coherence close to one is desirable.

Once the assessment of the reconstructed signal is complete, the Bayes factor B_M is calculated as the validation metric. The authors use both the point-null multivariate [56] and interval [97] hypotheses forms of the Bayes factor to demonstrate their methodology. Both methods were previously discussed.

Chen et al. [16] propose a design-driven validation methodology that utilizes Bayesian prediction models to compare competing models during the design selection phase of the engineering design process. Data from computer models and from physical experiments are combined to form a Bayesian based predictive model for the purpose of drawing inferences about the model's ability to generate accurate predictions in the intended but untested design domain. The methodology, outlined in Figure 2.31, is used to select the optimal design from a pool of design candidates using a predetermined validation metric, $M_D(x^*)$.

Let $\mathbf{x}$ be a point in the p -dimensional design variable space. Let $\mathbf{y}^e$ and $\mathbf{y}^m$ be the corresponding experimental and model response, respectively. The prediction bias, $\delta(\mathbf{x})$, is treated as a Gaussian process with parameters $\theta = (\sigma_\delta^2, \beta_\delta, \psi_\delta, \text{ and } \sigma_\epsilon)$ which represent the variance, location, correlation, and variance with regard to experiment error parameters, respectively. The priors for the variance and location parameters take the form [16]

$$\sigma_\delta^2 \sim IG(\alpha_\delta, \gamma_\delta), \beta_\delta | \sigma_\delta^2 \sim N(b_\delta, \sigma_\delta^2 V_\delta)$$

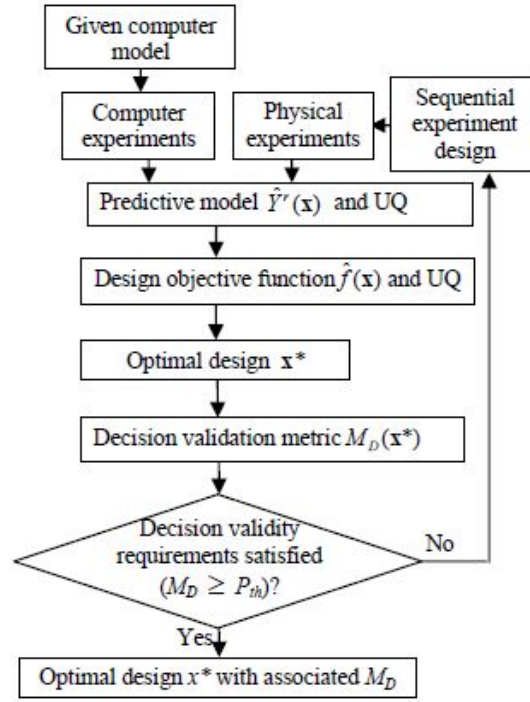


Figure 2.31: Proposed Bayesian Prediction Model Validation Framework [16]

No priors are specified for ψ_δ and σ_ϵ and these values are treated as fixed and estimated from the data. The posterior of the prediction bias can be obtained by integrating out θ as [16]

$$p(\delta(\mathbf{x})|\mathbf{y}^e, \mathbf{y}^m) = \int_{\theta} p(\delta(\mathbf{x})|\mathbf{y}^e, \mathbf{y}^m, \theta) p(\theta|\mathbf{y}^e, \mathbf{y}^m) d\theta \quad (2.76)$$

where it can be shown that the posterior for $\delta(\mathbf{x})$ follows a t-distribution [16]

$$\delta(\mathbf{x})|\mathbf{y}^e, \mathbf{y}^m, \psi_\delta, \tau \sim T(n_{\delta|e,m}, \mu_{\delta|e,m}(\mathbf{x}), \sigma_{\delta|e,m}^2(\mathbf{x})) \quad (2.77)$$

where $\tau = \sigma_\epsilon^2/\sigma_\delta^2$. Formulation of the degrees of freedom, non-centrality, and scale parameters for the t-distribution are omitted for brevity but can be found in [16].

The Bayesian prediction model is calculated using the predictor of the true behavior and its uncertainty quantification which can be calculated as

$$\hat{Y}^r(\mathbf{x}) = Y^m(\mathbf{x}) + \hat{\delta}(\mathbf{x}), \quad (2.78)$$

$$\text{Var}[Y^r(\mathbf{x})] = \text{Var}[\delta(\mathbf{x})] = \sigma_{\delta|e,m}^2(\mathbf{x}). \quad (2.79)$$

Let $f(\mathbf{x})$ and $\hat{f}(\mathbf{x})$ represent the design objective function and its prediction, respectively. A candidate optimal design, $\mathbf{x}^*$ can be found by optimizing $\hat{f}(\mathbf{x})$. Chen et al. [16] propose three decision validation metrics, presented as follows

(1) The Multiplicative Metric

$$M_D^M(\mathbf{x}^*) = \left\{ \prod_{\mathbf{x} \in \Omega_d} P(f(\mathbf{x}^*) < f(\mathbf{x}_i)) \right\}^{1/K} \quad (2.80)$$

(2) The Average Metric

$$M_D^A(\mathbf{x}^*) = \frac{1}{K} \sum_{\mathbf{x} \in \Omega_d} P(f(\mathbf{x}^*) < f(\mathbf{x}_i)) \quad (2.81)$$

(3) The Worst-Case Metric

$$M_D^W(\mathbf{x}^*) = \min_{\mathbf{x} \in \Omega_d} P(f(\mathbf{x}^*) < f(\mathbf{x}_i)) \quad (2.82)$$

where Ω_d is the set of all K feasible and differentiable design alternatives. M_D provides a probabilistic measure regarding modeler confidence in taking $\mathbf{x}^*$ as the optimal design over other design choices, with respect to the design objective [16].

Wang et al. [120] extends the methodology proposed in Chen et al. [16] to the validation of computer models. This methodology is based on the construction of prediction intervals for $\delta(\mathbf{x})$ and $Y^r(\mathbf{x})$. The $100(1 - \alpha)\%$ confidence interval for the model bias at any point $\mathbf{x}$ in the input domain is calculated as [120]

$$\mu_{\delta|e,m}(\mathbf{x}) \pm \sigma_{\delta|e,m}(\mathbf{x}) * t_{n_{\delta|e,m}, \alpha/2} \quad (2.83)$$

where $\mu_{\delta|e,m}(\mathbf{x})$ and $n_{\delta|e,m}$ are the non-centrality parameter and degrees of freedom, respectively. Refer to [120] for the formulation of these parameters. To establish validity criteria, define[120]

$$\Delta_{min}(\mathbf{x}) = \begin{cases} 0 & \text{if } 0 \in (l_{\delta}(\mathbf{x}), u_{\delta}(\mathbf{x})) \\ \min\{|l_{\delta}(\mathbf{x})|, |u_{\delta}(\mathbf{x})|\} & \text{otherwise} \end{cases} \quad (2.84)$$

and

$$\Delta_{max}(\mathbf{x}) = \max\{|l_{\delta}(\mathbf{x})|, |u_{\delta}(\mathbf{x})|\} \quad (2.85)$$

where $l_{\delta}(\mathbf{x})$ and $u_{\delta}(\mathbf{x})$ are the lower and upper bounds of the prediction interval, respectively.

The validation rules are established as define[120]

- If $\Delta_{min}(\mathbf{x}) > \Delta_0$, then the computer model is rejected at $\mathbf{x}$.
- If $\Delta_{max}(\mathbf{x}) < \Delta_0$, then the computer model is accepted at $\mathbf{x}$.
- If $\Delta_{min}(\mathbf{x}) < \Delta_0 < \Delta_{max}(\mathbf{x})$, then no conclusion can be reached, and more physical observations are needed.

Δ_0 represents a predetermined tolerance threshold assigned by the engineer based on subject matter expertise. Figure 2.32 graphically depicts the evaluation of differenced data against the prediction intervals on $\delta(\mathbf{x})$ [120].

2.5.4.3 Frequentest Metrics.

Oberkamp and Barone [88] develop a validation metric based on the statistical concept of confidence intervals. The proposed method multiple distinct scenarios. The first is the case where one wishes to estimate the error in the computational model based on the differences between the model data and the estimated mean of the population, which is derived from the experimental samples. Let the estimated error be denoted as [88]

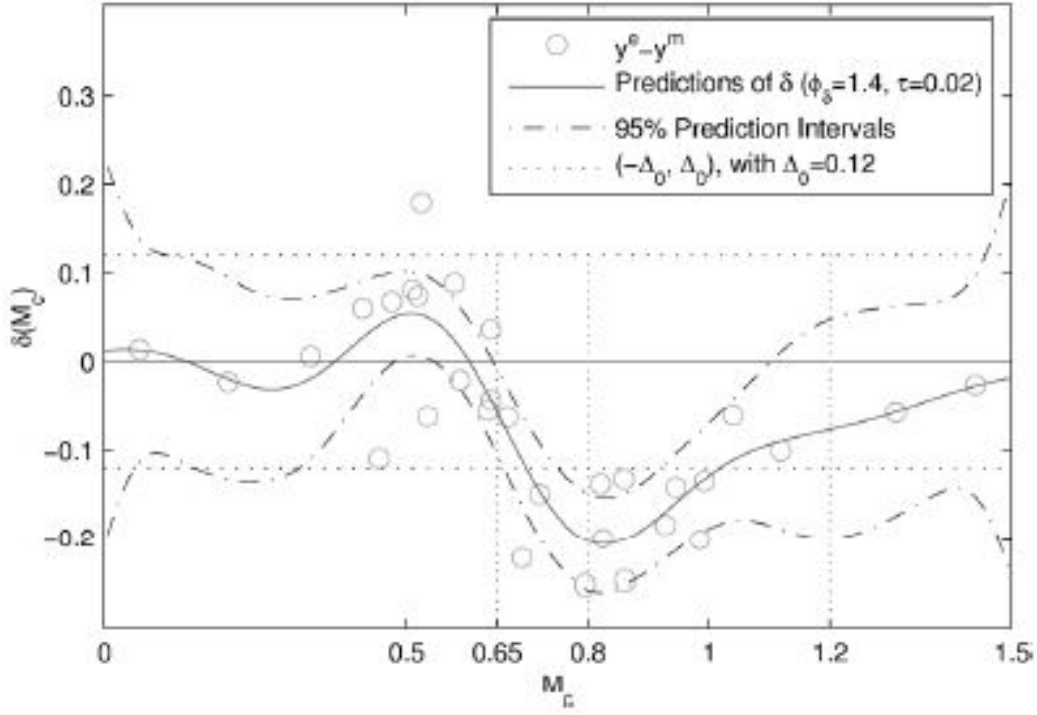


Figure 2.32: Prediction of $\delta(x)$ with Corresponding Prediction Intervals and Differenced Data [120]

$$\hat{E} = y_m - \bar{y}_e \quad (2.86)$$

where $\bar{y}_e$ is the sample mean based on n experiments. Then, the interval containing the true error associated with a particular confidence level, can be found as [88]

$$\left(\hat{E} - t_{\alpha/2, v} * \frac{s}{\sqrt{n}}, \hat{E} + t_{\alpha/2, v} * \frac{s}{\sqrt{n}} \right) \quad (2.87)$$

where s is the sample standard deviation from the experimental data. This method requires the assumption of independent, identically, normally distributed $\hat{E}$ estimates in cases when

n is small; however, when n is large, the central limit theorem provides the normality assumption. The second step in the validation methodology is to assess whether the estimated accuracy is adequate for the model's intended use. This assessment usually entails comparing the magnitude of the estimated error against the experimental uncertainty [88].

The second scenario of interest is when the computational and experimental results are functions of an input variable x . Equation 2.87 can be modified to address the true error as a function of x being in the interval [88]

$$\left(\hat{E} - t_{\alpha/2,v} * \frac{s(x)}{\sqrt{n}}, \hat{E} + t_{\alpha/2,v} * \frac{s(x)}{\sqrt{n}} \right) \quad (2.88)$$

where

$$s(x) \sim \left[\frac{1}{n-1} \sum_{i=1}^n (y_e^i(x) - \bar{y}_e(x))^2 \right]^{1/2} \quad (2.89)$$

is the standard deviation as a function of x and $y_e^i(x)$ is the interpolated value from the i th experimental replication [88].

Oberkampf and Barone [88] fit the data to a nonlinear regression when there is insufficient experimental data to use the interpolation method. The regression function acts as an estimated mean over the range of the data. Simultaneous confidence intervals, referred to as confidence regions in higher-dimensional space, are constructed over the range of input parameters. Let the estimated mean of the data be represented as [88]

$$\bar{y}_e(x) = f(x; \vec{\theta}) + \epsilon \quad (2.90)$$

where $f(x; \vec{\theta})$ is the regression function over the range of input parameter x with the corresponding vector of the regression function coefficients $\vec{\theta}$ and ϵ is the random

measurement error. The set of $\vec{\theta}$ that minimizes the error sum of squares, $S(\vec{\theta})$ is the solution vector and is written as $\vec{\hat{\theta}}$. The confidence regions around $S(\vec{\theta})$ are computed by solving the inequality [88]

$$\vec{\theta} \text{ such that } S(\vec{\theta}) \leq S(\vec{\hat{\theta}}) \left[1 + \frac{p}{n-p} F(p, n-p, 1-\alpha) \right] \quad (2.91)$$

where p is the number of estimated coefficients, n is the number of experimental measures, and F is the F probability distribution with p and $n-p$ degrees of freedom and $1-\alpha$ is the confidence interval quantile. To determine the confidence interval widths, all regression functions with $\vec{\theta}$ s within the desired confidence contour are evaluated and the maximum ($y_{CI}^+(x)$) and minimum ($y_{CI}^-(x)$) values of the regression function over the range of x are recorded. $y_{CI}^+(x)$ and $y_{CI}^-(x)$ form the upper and lower confidence intervals for the regression function [88]. Figure 2.33 illustrates a three-dimensional 90% confidence region on the regression coefficients and Figure 2.34 shows the resulting confidence intervals on the regression function along with the computational and experimental data points.

2.5.4.4 Area Metrics.

Ferson et al. [37] introduce a validation metric that uses the area between cumulative probability curves to quantify the disagreement between predictive results and empirical data from the real-world system. The proposed methodology can be utilized when the predictive and/or observed data is expressed as probability distributions. The authors claim that the utility of the area metric is that, unlike many of the other common approaches, it can detect differences in distributions beyond the standard measures of central tendency found in the lower order moments of mean and variance. Furthermore, the area metric is noted for satisfying all four properties of a mathematical metric.

The area metric assumes that the probabilistic prediction is characterized by a cumulative distribution function $F(x)$ which is specified by the model and given to the

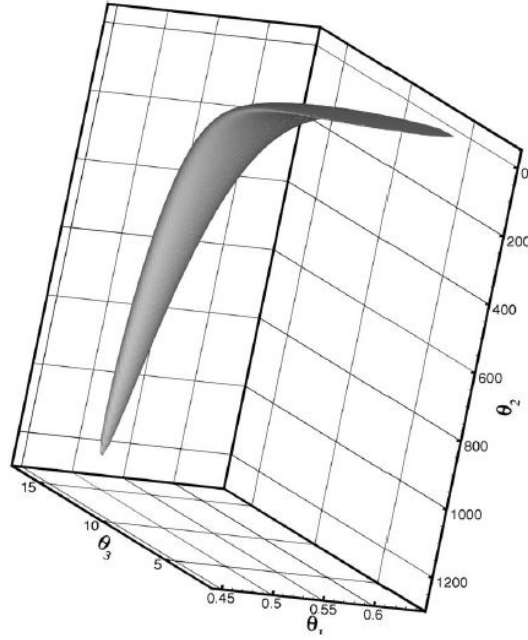


Figure 2.33: Three-Dimensional Confidence Region on Regression Coefficients [88]

analyst in this format. The system observations are provided as a collection of points and the empirical distribution (EDF) is constructed as a non-decreasing step function with a constant vertical step size of $1/n$, where n is number of data points in the sample. The distribution for the data x_i is represented as [37]

$$S_n(x) = \frac{\sum_{i=1}^n I(x_i, x)}{n} \quad (2.92)$$

where

$$I(x_i, x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

and therefore $S_n(x)$ is the fraction of data points within the sample that are at or below each magnitude x . The area metric uses the area between the prediction distribution F and

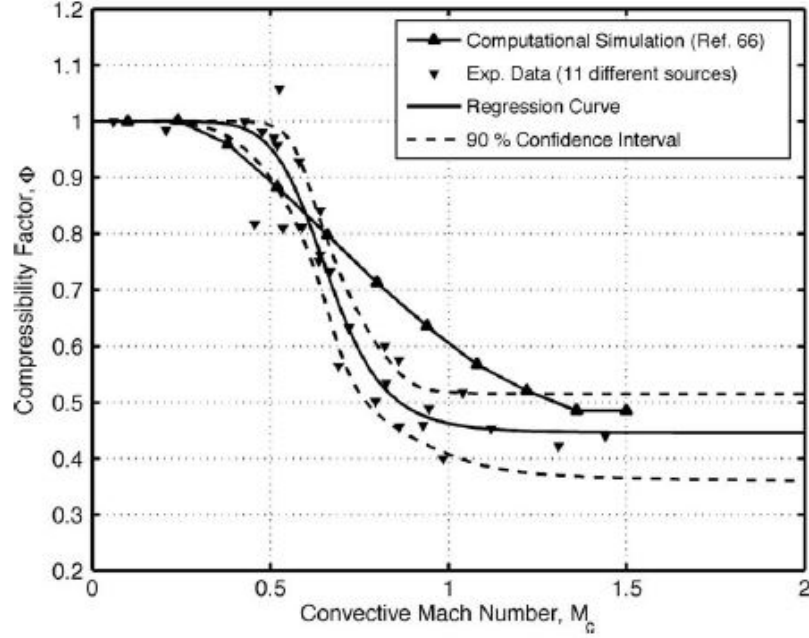


Figure 2.34: Example Results with 90% CIs on Regression Curve [88]

the empirical distribution S_n as the measure of disagreement between model and system. Mathematically, this difference between the two distribution functions can be written as [37]

$$d(F, S_n) = \int_{-\infty}^{\infty} |F(x) - S_n(x)| dx \quad (2.93)$$

which equates to the average horizontal distance between the two functions. Figure 2.35 illustrates the area metric and how it's measure of difference can differ depending on the number of data points used to build the empirical distribution. The smooth gray line represents the prediction distribution curve and the solid black line represents the empirical distribution curve. The plot on the left uses one data point to construct the empirical distribution and the plot on the right depicts an empirical distribution constructed using four data points. The smaller the shaded area, the more closely the model represents the system data.

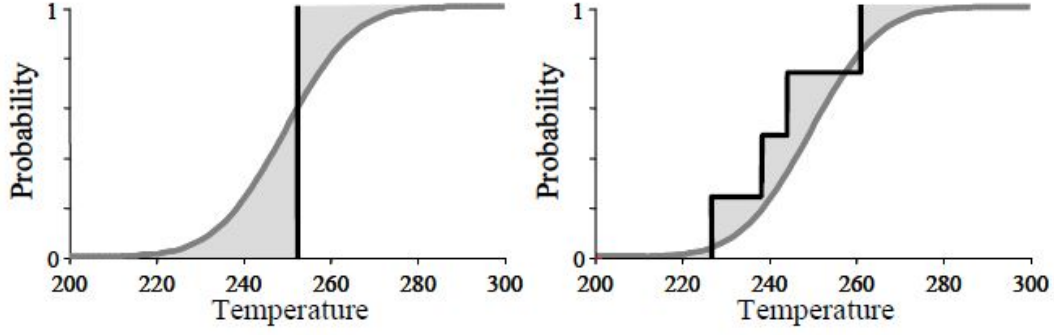


Figure 2.35: Example Distribution Curves and associated shaded differences with $n = 1$ and $n = 4$ [37]

The area metric can be extended to cases in which the data for predictions is sparse. In these cases, the available prediction data is considered to be a sample of the smoothed prediction distribution. Therefore, empirical distributions are constructed for both the prediction and system data and the previously discussed methodology is applied. This method provides a crude measure of the true differences between the distributions and thus the amount of epistemic error due to sampling error in the prediction is expected to be high.

The area metric approach is further extended to integrate evidence from all relevant data over the validation domain into a single measure of disagreement. This method is referred to as the u-pooling method. In this method, every data point x_i is transformed according to its prediction distribution F_i such that $u_i = F_i(x_i)$, with values uniformly distributed on the unit interval $[0, 1]$. These u-values are then pooled together to yield a set of values that are collectively uniformly distributed and form the basis for an overall summary metric to assess the accuracy of the model predictions. The u-values can be assessed directly against the standard uniform distribution (Figure 2.36) or the pooled u-values can be transformed back into their original unit of scale through the use of a suitable distribution function G [37]

$$y_i = G^{-1}(u_i) = G^{-1}(F_i(x_i)) \quad (2.94)$$

where the back-transforming distribution G is dependent on the context of the data and validation effort. Figure 2.37 illustrates the area metric using u-pooling and back-transformation.

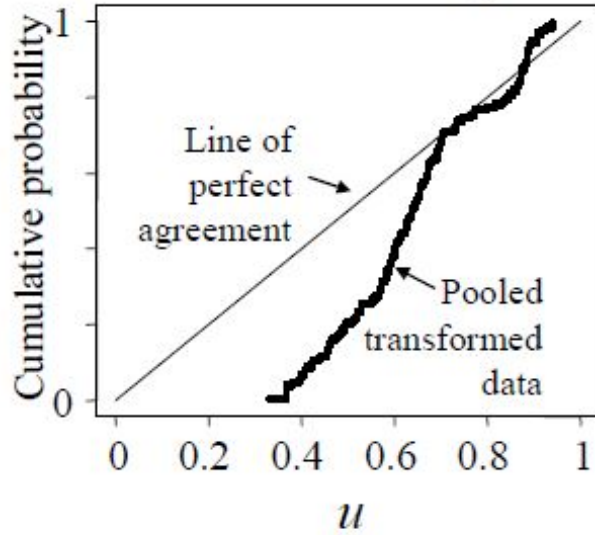


Figure 2.36: Comparison of U-Pooled Values Against Standard Uniform [37]

The above methodologies provide a measure of disagreement between model and system data but in themselves do not provide an assessment of validation. The authors suggest applying standard statistical tests to the u-pooling or y-values in order to detect significant mismatches between the model and system performance. They provide the example of using a Kolmogorov-Smirnov test, applied to the u-pooled values to assess statistical departures from uniformity. These departures would indicate poor model predictive capability.

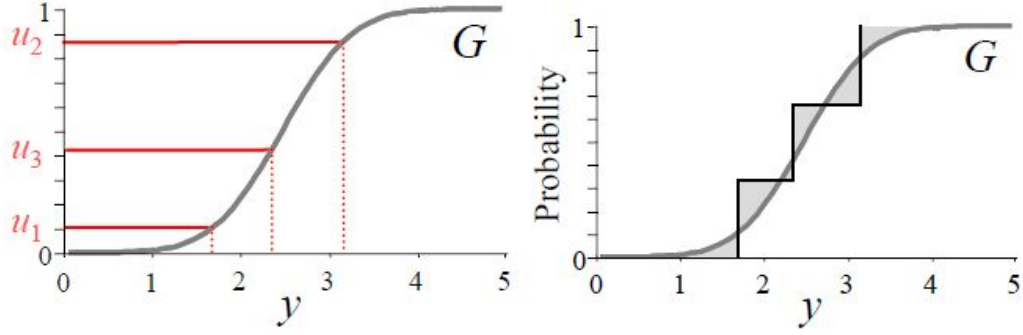


Figure 2.37: Area Metric Using U-Pooling and Back-transformation [37]

Ferson et al. [36] extend the area metric to stochastic models in which the epistemic uncertainty is captured via intervals. The data are represented by a collection of intervals, the breath of the intervals are viewed as epistemic uncertainty and the scatter of the intervals as aleatory uncertainty. The authors contest that in cases where scalar values are replaced with interval ranges, that correctness of predictions should be based on overlap between prediction and observation. In cases where both the observed and predicted data are of interval form, the validation measure should be the shortest distance between the two intervals. To compare intervals A and B, the measure of disagreement would be [36]

$$\inf |X - Y| \quad (2.95)$$

where $X \in A$ and $Y \in B$. This is referred to as a realization based comparison and is not actually based on the shapes of the intervals and is not a valid mathematical metric due to violating the property of identity of indiscernibles. The authors sync the area/shape based approach with the realization based approach by relating them as special cases of the Wasserstein distance [36]

$$\inf E|X - Y| \quad (2.96)$$

where $X \sim F$ and $Y \sim S_n$. Probability boxes, or p-boxes, to demonstrate the calculation of the mismatch distance. The area measuring mismatch for p-boxes is calculated as [36]

$$\int_{-\infty}^{\infty} \Delta([F_R(x), F_L(x)], [S_{nR}(x), S_{nL}(x)]) dx \quad (2.97)$$

where F and S_n represent the prediction and observed data distribution, respectively, and the L and R subscripts denote the left and right bounds for each distribution. Furthermore, [36]

$$\Delta(A, B) = \min|a - b| \quad (2.98)$$

where $a \in A$ and $b \in B$ is the shortest distance between the two intervals. Figure 2.38 illustrates the p-box interval method for four distinct cases. The gray p-box represents the prediction distribution and the black p-box represents the observed empirical distribution. In the first two examples, the p-boxes do not overlap and the distance measure is the area between the rightmost distribution in the gray p-box and the leftmost distribution in the black p-box. In the third example, overlap occurs and the distance measure is the area in which the p-boxes do not overlap. In the fourth example, the distance measure is equal to zero because there exists distributions that lie entirely within both the predicted and observed p-boxes.

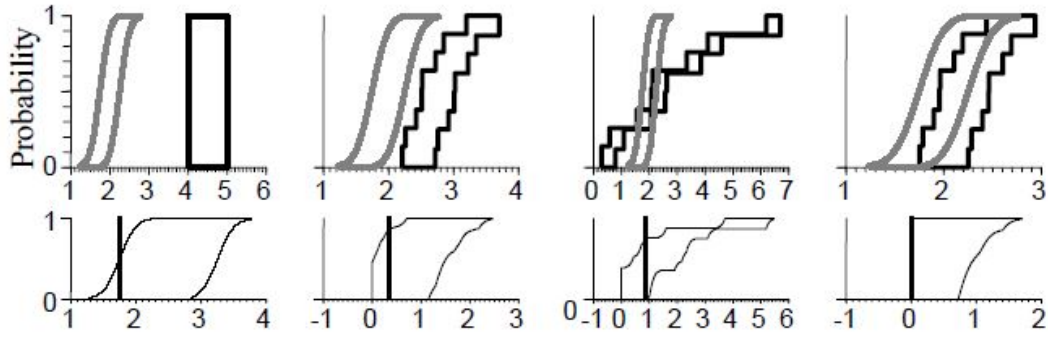


Figure 2.38: Four Examples for Interval (P-Box) Extension to Area Metric [36]

2.5.5 Correlation and Analysis (CORA) Method.

The CORA method presented by Gehre et al. [43], and demonstrated by Gehre and Stahlschmidt [44], is an objective evaluation tool used to assess similarities between model and system response curves. The method has become popular in crash test model validation due the implementation of a user-friendly CORA toolbox. The method combines two independent sub-methods, whose limitations are countered by one another, to provide an overall measure of signal similarity. A detailed overview of the CORA rating system is presented in Chapter 5.

2.6 Summary and Finding of Literature Review

This literature review has provided a broad overview on key elements related to model validation. Terminology was discussed, varying validation taxonomies were provided, validation techniques were provided with an elaboration on statistical techniques, and finally, current validation research efforts were discussed. This literature review has shown the vastness of the model validation effort. This vastness leads to discrepancies between and within the various scientific communities in regards to how model validation should be conducted. There exists a multitude of model validation procedures, each with advantages and disadvantages, that are used in various modeling contexts. While

the different communities use different validation techniques, both validation efforts share commonalities in terms of the desired validation result.

While model validation is still too often a visual assessment, an academic push for statistically rigorous validation has been progressing over the last three decades. As the use of highly complicated models to emulate complex system behavior grows, the statistical techniques used to validate these modes must adapt to these inherent complexities. Two of the larger complexities under consideration are systems tested at multiple input configurations and systems with time-dependent response data. This focus is evident in the current research efforts discussed in this literature review. A familiarity with these efforts is critical in that the research presented in subsequent chapters of this document addresses these considerations.

III. Point-Wise Model Validation Over a Design Space Using Regression Tolerance Intervals and Bayesian Hypothesis Testing

3.1 Introduction

As computer technology continues to advance, so does our propensity to use computerized models to emulate the behavior of real-world systems for the purpose of gaining insight into those real-world systems. This is particularly true in test and evaluation (T&E) where conducting tests on an actual system, such as an aircraft, can be costly in terms of budget, manpower and time resources. This resource drain limits the availability of these systems for various other purposes. In the case of aircraft, efforts such as pilot training can be negatively affected. Simulations, from desktop analytical simulations through human-in-the-loop simulators, are a potentially cost effective alternative to conducting full-scale, real-world test and evaluation efforts. To be an effective alternative, simulation models must be assessed for their ability to accurately represent the system for which they are built; an assessment referred to as model validation. Many complex simulation models, such as an aircraft training simulator, are designed to emulate the true-system behavior over a multidimensional space of operating conditions, commonly referred to as the model design space. For the model to achieve overall validity, validation must consider the entire model design space.

Despite growing dependencies on simulation, the rigor of the model validation process can often be crude. Too often, the preferred validation technique is simply a graphical comparison of experimental system results with the corresponding model results. These techniques are mostly visual in nature and too dependent on subjective assessments. Statistical validation methodologies exist; however, the vast majority of classical statistical validation methodologies focus on validation of a single response variable at a single design configuration. Recently, as part of the engineering design process, the engineering

community has begun to focus its validation efforts on a model's predictive capability over its intended operational domain. These validation methodologies are essential when assessing competing system designs. While the competing design problem may not translate directly to the validation problem typical of the operations research and science communities, a methodology that allows for validation over the entire design space has great utility, particularly in the validation of complex system simulators or constructive simulations built on experimental test data.

The methodology proposed in this article bridges the classical validation approaches commonly used to validate simulations in the operations research and simulation communities with the engineering design validation approaches implemented by engineering communities to assess the predictive capability of system models. Specifically, the methodology targets the use of system-level tolerance intervals to assess the adequacy of a stochastic simulation model representing that system. Section 2 provides a discussion of the prominent validation techniques currently in use by both communities. Section 3 discusses regression tolerance intervals. Section 4 formulates the proposed validation methodology. Section 5 uses a numerical example to illustrate the utility of the methodology. Section 6 discusses key considerations when using the proposed methodology.

3.2 Validation Literature

The literature on simulation model validation is extensive. Validation methodology taxonomies by Davis [27], Balci [3], and Sargent [108] help place each technique into suitable categories based on similar procedural characteristics. One such category is statistical validation. Statistical validation techniques utilize statistical rigor to assess the level of agreement between model and system generated data for both the univariate and multivariate cases. In the Operations Research (OR) community, statistical techniques appear to be the most frequently utilized methods for simulation model validation. Methods include, but are not limited to, confidence intervals [74], t-tests [111], Hotelling's T^2 tests

[7], Chi-Squared and Kolmogorov-Smirnov goodness-of-fit tests [87], non-parametric tests of means [111], univariate and multivariate analysis of variance (ANOVA/MANOVA) [87] [42], and time-series analysis [38] [41] [116] [110]. Balci [6] provides a comprehensive list of available statistical techniques used for validation. While many of these techniques are still utilized, most have origins in model validation dating back to the 1960's and 1970's. This is a testament to their effectiveness, but also highlights a critical need, as a scientific community, to reassess how simulation model validation should be conducted and what techniques should be used on increasingly complex simulation models.

The engineering community has their own set of model validation techniques. Some draw from the statistical methods mentioned above, while others provide entirely new approaches to the validation problem. Liu et al. [76] segregates these techniques into one of four categories dependent upon the methods utilized. These categories are classical, Bayesian, frequentist, and area based methods.

Classical hypothesis testing focuses on the propagation of total uncertainty and a subsequent comparison of experimental data against the uncertainty bounds around a model prediction. This methodology was first introduced by Hills and Trucano [52] and utilizes a weighted least squares test statistic, r^2 , to test the null hypothesis that the experimental data come from the prediction populations. This method forms the basis for other hypothesis testing methods used in engineering design validation [45][50] [13][15].

Validation methodologies based on the principles of Bayes' Theorem have become a popular alternative to the classical hypothesis testing methods. These methods, centered around Bayesian Hypothesis Testing and the use of the Bayes factor as a validation metric, have been used in numerous applications of model validation to include univariate and multivariate response systems, evaluations of model predictive capabilities outside the tested domain, and for the evaluation of dynamical systems [98] [94] [95] [97] [96] [57] [56]. These Bayesian-based approaches provide an assessment of the null hypothesis in

the form of a relative likelihood of the null against the alternative hypothesis. Frequentist approaches quantify the difference between model and system performance by measuring the distance between the output data means. The validation metric is an interval around the estimated error in the model, at an associated confidence level, in which the true error is believed to reside [88]. Ferson et al. [37] introduce a validation methodology that uses the area between cumulative probability curves to quantify the disagreement between empirical data from the real-world system and predictive results. This methodology is useful when the predictive and/or observed data is expressed as probability distributions and when the uncertainty is captured via intervals [36].

Validation of models at untested parameter settings is discussed in the engineering literature but largely overlooked in the OR literature. Oberkampf and Barone [88] fit system data to a nonlinear regression curve, which provides an estimated mean over the range of data. Confidence intervals are constructed for some response over the range of input parameters. The upper and lower bounds of these intervals are used to assess model validity. Chen et al. [16] utilize Bayesian prediction models to draw inferences about a model's ability to generate accurate predictions in the intended but untested design domain. Wang et al. [120] extends the methodology proposed in Chen et al. [16] to the validation of computer models by constructing confidence intervals around the Bayesian prediction model and its associated bias.

The OR literature discusses validation across the experimental region of regression based metamodels against simulation models data rather than validating the simulation model against the true system data[68]. These assessments typically involve a subjective evaluation of the signs of the metamodel's $\hat{\beta}$ estimates or by cross-validation of the simulation and metamodel by constructing the metamodel using only a portion of the simulation data and then comparing the remaining simulation data against the predicted output from the metamodel at those input parameter configurations[65] [64] [63]. The use

of bootstrapped regression statistics for validation of metamodels has also been considered [67] [66].

The term validation is typically associated with a measure of agreement between a model and the system for which it is intended to represent. Past efforts use the term validation in this manner and our focus is similar to that of past efforts; however, we wish to avoid the philosophical debate regarding whether any model can be deemed fully valid. In this paper, the term validity refers to an assessment of model adequacy and we are therefore evaluating the model's ability to adequately emulate a real-world system in the context of its intended use.

3.3 Methodology

3.3.1 *Hypothesis Under Test and Corresponding Assumptions.*

Rigorous validation of models involves assessing whether or not the simulation model response output at a design point is statistically different than the corresponding system response output at that same point. This assessment is represented as the hypothesis test

$$\begin{aligned} H_0 : Y^s &\equiv Y^m \\ H_A : Y^s &<> Y^m \end{aligned} \tag{3.1}$$

where Y^s and Y^m represent system and model data, respectively. H_0 implies the data are equivalent, and H_A implies rejection of H_0 , the data are not equivalent. If the model and system are statistically equivalent, then it is reasonable to believe that the two data sets share distributional properties. Our methodology uses statistical regression tolerance intervals, constructed around a fitted response surface of the system data, to evaluate the distributional spread of the model response data. Under the null hypothesis, the tolerance interval about the predicted system response contains a user specified proportion of the model response distribution with a user-specified confidence level. If the proportion of the model's distribution falling outside of the tolerance interval exceeds a specified amount,

the null hypothesis is rejected, the data sets are viewed as not equivalent, and the model is considered invalid at that design point. The proposed tolerance interval methodology requires the following assumptions.

1. The model and system response errors are i.i.d normally distributed.
2. Constant variance across the design space.
3. The fitted response surface provides tolerance interval widths that represent a true measure of system response uncertainty at each design point.
4. System behavior is adequately modeled using linear regression based response surface fitting techniques.

3.3.2 Fitting a Response Surface to Experimental Data.

This methodology uses a response surface of the system experimental data and its associated tolerance intervals to generate a multi-dimensional region within the design space which should capture simulation model output for the response of interest. The modeler determines if the generated region is appropriate for validation purposes within the context of the current model validation effort. This means it is imperative the modeler consider which design configurations are tested to ensure that the response behavior within the design space is properly characterized and that the associated tolerance intervals are of a size appropriate for model validation purposes. The extensiveness of the chosen design depends on the cost of running experiments at a single design configuration; therefore, the best response surface design for one validation test may not be suitable for another.

In response surface methodology (RSM), there are several design considerations to achieve quality prediction capability. Standard designs for nonlinear surfaces, such as central composite designs (CCDs) and face-centered cube designs (FCDs), are flexible and effective at characterizing second-order models. Alphabetic optimal designs, or variance-optimal designs, use various design point criteria to achieve a desired variance property

throughout the design space. D-Optimal designs maximize the determinant of the moment matrix, $\mathbf{M} = \mathbf{X}'\mathbf{X}/N$, which contains the variances and covariances of the regression coefficients, to maximize the precision of parameter estimates in the model. I-optimal designs minimize the average scaled prediction variance over the design region to achieve a single measure of predictive performance. A-, V-, and G-optimal achieve similar desired properties within the design space and are discussed in Myers et al. [86].

Optimal designs are also useful for finding designs tailored to hypothesized response surface forms. In the case of systems where interesting but unknown response behavior is expected throughout the design region, space-filling designs such as Latin hypercube [79] or sphere packing designs [61], may be appropriate; however, these designs are more commonly utilized for deterministic, not stochastic, models. The choice of experimental design is ultimately a function of both the complexity of the system being modeled and the modeler's criteria for the model validation effort.

The example in this paper uses a FCD for modeling second-order response behavior. These are useful when resource limitations are not a significant factor in design consideration because they provide low scaled prediction variance over the design space. If significant limitations to experimental runs exist, I-optimal designs provide a low average scaled prediction variance which may be suitable for obtaining tolerance interval widths appropriate for simulation model validation.

3.3.3 Regression Tolerance Intervals.

Tolerance intervals are too often overlooked in statistical analysis in favor of confidence or prediction intervals. The tolerance interval provides the analyst with a level of confidence associated with a proportion of future observations falling within the interval

$$Pr[F(U) - F(L) > P] = 1 - \alpha \quad (3.2)$$

where $F(\cdot)$ is the cumulative distribution function (CDF) for the future observations, U and L are the upper and lower tolerance bounds respectively, P is proportion of the CDF to fall within the tolerance interval, and $1 - \alpha$ is the user-defined confidence level.

The construction of tolerance intervals on normally distributed data can be traced back to Wilks [122]. Others using tolerance intervals under the normality assumption include [121] [53] and more recently [55]. Research into the use of tolerance intervals when the univariate normality assumption does not apply includes [123] [118] [60] [49] [72]. The extension of tolerance intervals to regression was proposed by Wallis [119]. Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad (3.3)$$

where $\mathbf{Y}$ is the response vector, $\mathbf{X}$ is the design matrix, β is a vector of unknown parameters, and ϵ is a vector of normally distributed error. The Ordinary Least Squares (OLS) predictions for the responses are defined as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \quad (3.4)$$

where $\hat{\beta}$ corresponds to the OLS estimates for the regression parameters. $\hat{\mathbf{Y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n]$ is a vector of predicted response values for the n input configurations. The two sided regression tolerance interval is then given by [124]

$$TI_{P,\alpha,h} = \hat{y}_h \pm k_{2,h}s \quad (3.5)$$

where h denotes the input configuration, s denotes the root mean square error of the fitted surface, and $k_{2,h}$ is the 2 sided k -factor for the h^{th} configuration. The k -factor is chosen such that [119]

$$Pr(A \geq P) = 1 - \alpha \quad (3.6)$$

where [119]

$$A = Pr(Y + ks > y > Y - ks). \quad (3.7)$$

An exact solution to the 2-sided k-factor can be found in Krishnamoorthy and Matthew [71]; however, the solution is computationally complex and time consuming. Several authors propose construction of tolerance intervals using an approximation of the k-factor [53] [129] [28]. Rucker [101] uses Monte-Carlo simulations to show that these approximations are comparable to the exact solution. The k-factor approximation used for analysis of individual design points is defined as [124]

$$k_{2,h} = \sqrt{\frac{f\chi_{1;p}^2(1/n_h^*)}{\chi_{f;\alpha}^2}} \quad (3.8)$$

where $f = n - p$, $n_h^* = (X_h^T(\mathbf{X}^T\mathbf{X})^{-1}X_h)^{-1}$, and $\chi_{d;\alpha}^2(\delta)$ is the α th quantile of a noncentral chi-squared distribution with d degrees of freedom and δ noncentrality parameter [124]. For model validation purposes, under H_0 , the tolerance interval for the system should also capture a similar proportion of the model observations.

3.3.4 Model Assessment at Tested/Untested Design Points.

To motivate this validation approach, consider any configuration or design point of interest, X_h . We have some response surface characterization of the system based on the results from the experimental design captured in the design matrix X . We generate a regression tolerance interval for this surface at the desired design point. For a model adequately representing the system, the observed model output, Y_h^m , at any X_h should fall within these system tolerance intervals. Thus, under the predetermined test assumptions, the model is considered valid at design point, X_h , if 100P percent of the model response distribution is captured within the corresponding system response tolerance interval. Let r_h^s be the number of system replications at X_h and r_h^m be a sufficiently large number of model replications at X_h . Each replication of the model response is assessed against the

prescribed tolerance interval with a single replication of the model response falling within the prescribed tolerance interval defined as a success.

Define $y_{h,i}^s$ as the system output at a design point X_h for replicate i and $y_{h,i}^m$ as the model output at design point X_h for replicate i . Clearly,

$$Y_h^s = \{y_{h,i}^s | i = 1, \dots, r_h^s; X_h\} \quad (3.9)$$

$$Y_h^m = \{y_{h,i}^m | i = 1, \dots, r_h^m; X_h\} \quad (3.10)$$

Define

$$Z_{h,i} = \begin{cases} 0 & \text{if } y_{h,i}^m \notin TI_{P,\alpha,h} \\ 1 & \text{if } y_{h,i}^m \in TI_{P,\alpha,h} \end{cases} \quad (3.11)$$

and

$$S_h = \sum_{i=1}^{r_{h,m}} Z_{h,i} \quad \forall X_h \quad (3.12)$$

Then the model is assessed as valid at the design point X_h if

$$S_h \geq P \times r_h^m. \quad (3.13)$$

This approach holds regardless of whether X_h represents a design point sampled using the system (i.e. a test point) or is an untested point. Figure 3.1 depicts a notional assessment of r_h^m replications of model output at an untested configuration X_h . The success region is bounded between the two dashed lines representing the system tolerance interval at X_h , which is based on parameters $P = 0.9$ and $\alpha = 0.05$. These tolerance intervals are centered around the predicted system response for the untested configuration. The histogram depicts the distribution of the model replications and the red line represents the fitted normal PDF for the model data. Model validity at X_h is determined by the proportion of the data

replications that fall within the tolerance interval lines. In this case, the required proportion would be 90% as specified by the P parameter of the tolerance interval. An evaluation of the model replications shows that more than 90% of distribution lies within the prescribed tolerance intervals, indicating that the model is valid at X_h .

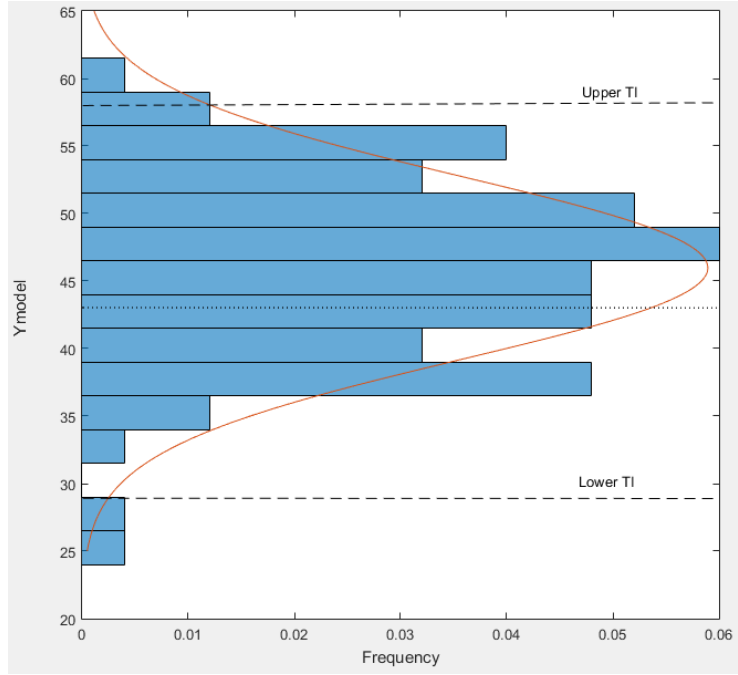


Figure 3.1: Tolerance Interval Assessment of Model Response Distribution

3.3.5 Using Bayesian Hypothesis Testing to Establish Minimum P .

This regression tolerance interval methodology can be overly restrictive when an observed proportion of model output $\hat{P}$ falls just below a prescribed proportion P , yielding an invalid assessment of the model. This is particularly true when the number of model replications is limited. It is possible that the true distribution of the model response is not sufficiently captured with limited model replications. Bayesian Hypothesis Testing (BHT) allows us to relax the criteria by evaluating our belief that $\hat{P}$ is equal to P given the number of replications used to generate the distribution. Consider each model assessment

at a particular design point as a Bernoulli trial where the outcome is either that the model response falls within the tolerance interval (success) or it does not (failure). Under the null hypothesis that $Y^s \equiv Y^m$, we assume that the variability between the predicted system response and each model response replication at X_h is due to the error term ϵ which is i.i.d. with $\epsilon \sim N(0, \sigma^2)$. This aggregation of multiple, independent Bernoulli trials has a binomial distribution with success parameter equal to the prescribe proportion P used to establish the tolerance interval widths.

The binomial pass/fail BHT validation criterion established by Rebba et al. [98] calculates a Bayes factor, which is used to assess $\hat{P}$ for model validation. Let P_o be the theoretical probability of success for a binomial distribution. In our case, P_o represents the P parameter from the tolerance interval. The null hypothesis under test is that the observed probability of success, $\hat{P}$, is equal to the theoretical probability of success, P_o . The alternative hypothesis is that the observed and theoretical probability of success are not the same.

$$\begin{aligned} H_0 : \hat{P} &= P_o \\ H_A : \hat{P} &\neq P_o \end{aligned} \tag{3.14}$$

If no prior information is known about $\hat{P}$, under the alternative hypothesis we can assume a uniform distribution in $[0,1]$ for the PDF of the observed probability of success given the alternative hypothesis, $f(\hat{P}|H_0)$. The Bayes factor is then computed as [98]

$$B(P_o) = \frac{P(data|H_0 : \hat{P} = P_o)}{P(data|H_1 : \hat{P} \neq P_o)} \tag{3.15}$$

The Bayes factor provides a measure for the ratio of likelihoods for the two hypotheses. A Bayes factor larger than one implies that the null hypothesis is more likely than the alternative. This is the basis for assessing estimated parameters that fall below the theoretical value. To determine the minimum allowable $\hat{P}$ value necessary for a model

validity assessment, a minimum allowable Bayes factor is determined. In the most relaxed case, any Bayes factor greater than one is considered associated with a valid model. A more strict validation criteria is assigned by adjusting the required Bayes factor to some number larger than one. For our purposes, the smallest $\hat{P}$ for which the resulting Bayes factor is greater than one is used as the validation criteria.

Rebba et al. [98] manipulate $B(P_0)$ into a computable form

$$\begin{aligned}
 B(P_0) &= \frac{P(data|H_0 : \hat{P} = P_0)}{P(data|H_1 : \hat{P} \neq P_0)} \\
 &= \frac{\binom{n}{y} P_0^y (1 - P_0)^{n-y}}{\int_0^1 \binom{n}{y} \hat{P}^y (1 - \hat{P})^{n-y} f(\hat{P}|H_1) d\hat{P}} \\
 &= (n+1) \binom{n}{y} P_0^y (1 - P_0)^{n-y}
 \end{aligned} \tag{3.16}$$

where n and y are the number of assessments and number of observed successes, respectively. In the case where $\hat{P} < P_0$, the model is assessed as valid if $B(P_0) > 1$. The Bayes factor is not used in cases where $\hat{P} \geq P_0$ because we are only interested in establishing a lower validity bound.

The BHT method can be used to establish a lower bound on the observed proportion $\hat{P}$ required for model validation. The required minimum number of successes y is determined using an integer programming formulation

Min y

Subject To:

$$(n+1) \binom{n}{y} P_0^y (1 - P_0)^{n-y} > 1$$

$$0 \leq y \leq n$$

y: Integer

where n and P_0 are constants unique to the validation effort under consideration. The minimized y value corresponds to the minimum $\hat{P}$ estimate required for validation at X_h . Note that the closeness between $\hat{P}$ and P_0 necessary for BHT to determine the difference between the two statistically insignificant is affected by the number of replications at X_h . As n increases, the $\hat{P}$ required to achieve a Bayes factor greater than one approaches the theoretical P_0 .

3.4 Example

Consider a notional response of the form $Y = 10 + 2.5x_1^2 + 2x_2^2 + 1.75x_3^2 - 1.25x_1x_2 - 1.5x_1x_3 - x_2x_3 + 3x_1 + 2.75x_2 + 2x_3$. Let the stochastic system response be of the form $Y^s = Y + \epsilon$ where $\epsilon \sim N(0, 1)$. The experimental data is gathered from three replications of a face-centered design with three parameters and three levels. This data is subsequently fit to an OLS response surface, $\hat{Y}^s$ with a RMSE equal to 0.869. The regression tolerance intervals for $P = 0.9$ and $\alpha = 0.05$ are calculated at each design point X_h . Table 3.1 provides the system replications, fitted response values, and the upper and lower tolerance interval bounds for each point within the FCD.

Let us also consider the point $X_{15} = (0.5, 0.5, 0.5)$. The system response for this configuration of input parameters was not included in our FCD. As a result, let's assume that we do not have experimental data at this design point. Using the regression equation for $\hat{Y}^s$, we can calculate the predicted response value at X_{15} to be $\hat{Y}_{15} = 14.411$. Using (3.5) and (3.8), the tolerance interval for the new point is $TI_{15} = (12.526, 16.296)$.

Let the additive model bias be denoted as δ , where $\delta = 0$ indicates no model bias, $\delta > 0$ indicates a model bias above the true mean, and $\delta < 0$ indicates a model bias below the true mean. Define the model response data as $Y^m = (Y + \delta) + \epsilon$ where $\epsilon \sim N(0, \sigma_m^2)$. Twenty replications of the model response at each desired design point were generated with

Table 3.1: System Replications and Tolerance Intervals($P = 0.9, \alpha = 0.05$)

Factor Levels			System Replications			$\hat{Y}$	Tolerance Intervals	
x1	x2	x3	1	2	3		Lower	Upper
1	-1	-1	16.194	14.793	15.560	15.411	13.368	17.454
-1	1	-1	13.909	16.072	15.421	15.297	13.254	17.340
-1	-1	1	13.595	14.237	13.901	13.697	11.654	15.740
1	1	-1	22.439	19.683	20.259	20.938	18.893	22.982
1	-1	1	18.835	20.307	20.087	19.510	17.466	21.555
-1	1	1	19.091	19.162	20.225	19.528	17.484	21.573
1	1	1	21.234	20.552	21.502	21.026	18.946	23.107
-1	0	0	9.601	9.719	8.607	9.325	7.342	11.307
1	0	0	15.137	14.622	14.605	15.051	13.075	17.028
0	-1	0	8.511	6.389	9.563	8.706	6.723	10.688
0	1	0	13.270	15.286	15.417	14.385	12.408	16.361
0	0	-1	9.903	9.408	10.578	9.762	7.779	11.744
0	0	1	13.433	13.152	13.757	13.927	11.951	15.904
0	0	0	10.319	10.419	9.764	9.610	7.717	11.503

$\delta = 0.2$ and $\sigma_m^2 = 1.4$. The first ten replications of the model data can be seen in Table 3.2. To validate the model at each design point, we calculate how many of the 20 model replications fall within the specified tolerance intervals.

Table 3.2: First Ten Replications of Model Data

Factor Levels			Model Replications									
x1	x2	x3	1	2	3	4	5	6	7	8	9	10
1	-1	-1	17.233	15.343	16.721	16.442	19.168	14.893	15.755	16.847	15.452	15.661
-1	1	-1	17.307	12.552	14.695	14.840	14.893	13.124	15.527	16.471	16.025	12.890
-1	-1	1	15.012	13.521	13.897	12.845	12.088	14.695	13.486	15.259	11.774	15.715
1	1	-1	21.551	21.046	22.000	20.338	21.983	19.641	21.956	20.314	23.330	22.283
1	-1	1	18.761	19.751	18.920	18.297	18.356	17.435	20.755	18.924	20.466	18.837
-1	1	1	20.550	20.293	18.684	22.092	20.009	20.897	17.251	20.362	20.082	20.086
1	1	1	19.869	19.627	20.263	20.210	21.744	19.665	22.615	20.380	19.057	20.174
-1	0	0	9.986	8.724	9.855	10.828	10.070	11.194	8.942	12.742	9.223	7.946
1	0	0	16.634	16.652	16.311	16.119	16.060	14.149	17.280	16.226	16.005	12.241
0	-1	0	10.559	8.992	9.414	11.681	9.213	8.713	11.312	9.975	10.899	9.393
0	1	0	15.656	13.305	14.401	16.062	14.774	15.240	13.187	13.945	13.710	14.395
0	0	-1	11.604	9.962	10.666	8.478	9.826	10.609	10.162	8.381	11.504	9.258
0	0	1	16.171	14.195	13.814	13.557	10.591	14.511	18.110	14.903	17.272	12.935
0	0	0	7.923	9.707	11.168	11.270	10.672	10.430	9.942	9.963	9.606	8.019
0.5	0.5	0.5	12.566	15.496	14.592	15.046	15.888	14.424	15.284	15.479	15.263	14.074

Due to the low number of replications, the validation criteria of $P = 0.9$ is relaxed using BHT. To find the minimum proportion of model replications necessary for validation, we solve the follow integer program,

$$\begin{aligned}
& \text{Min } y \\
& \text{Subject To:} \\
& (21) \binom{20}{y} 0.9^y (0.1)^{20-y} > 1 \\
& 0 \leq y \leq 20 \\
& y: \text{ Integer}
\end{aligned}$$

which results in $y = 16$, indicating an observed success rate of 0.8 (16/20) warrants a valid assessment when evaluating 20 model replications. Table 3.3 shows the results of these assessments at each X_h . Figure 3.2 provides a graphical depiction of the assessments. The marginal plot (factor x_3 set constant) shows the OLS response surface fit from the system data, the model replications at points within the FCD (dots), and the corresponding upper and lower tolerance bounds (triangles). The plot shows that the necessary proportion of the model replications is observed at each design setting, indicating validity of the model at those settings.

3.4.1 Method Sensitivity.

Simulations were used to investigate the proposed method's sensitivity to changing levels of additive model bias and error variance. Similar to the example above, the sensitivity analysis was conducted on a quadratic response with interactions. The baseline response value was

$$Y^b = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_3^2 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 + \beta_8 x_2 + \beta_9 x_3 \quad (3.17)$$

Table 3.3: Results From Validation Assessments

Factor Levels			Num. Reps Within TIs	Observed Proportion	Model Assessment
x1	x2	x3			
1	-1	-1	18	90	Valid
-1	1	-1	17	85	Valid
-1	-1	1	18	90	Valid
1	1	-1	18	90	Valid
1	-1	1	17	85	Valid
-1	1	1	18	90	Valid
1	1	1	20	100	Valid
-1	0	0	19	95	Valid
1	0	0	17	85	Valid
0	-1	0	16	80	Valid
0	1	0	19	95	Valid
0	0	-1	18	90	Valid
0	0	1	16	80	Valid
0	0	0	17	85	Valid
0.5	0.5	0.5	20	100	Valid

where each β_i was randomly drawn from a uniform distribution between -10 and 10. One thousand samples of Y^b , for each combination of model bias and error variance, were simulated with unique β coefficients. In each instance our system and model data were defined as

$$Y^s = Y^b + \epsilon^s \quad (3.18)$$

$$Y^m = (Y^b + \delta) + \epsilon^m$$

where $\epsilon^s \sim N(0, 1)$ and $\epsilon^m \sim N(0, \sigma_m^2)$. Model validity was assessed using 20 model replications for each of the 1,000 iterations of the simulation. The validation criteria was set using the BHT approach and any iteration with 16 or more replications within the prescribed tolerance intervals was deemed valid. The model bias ranged from 0-3 and the model error variance ranged from 0-9. Figure 3.3a provides a contour map of the resulting simulation. This contour highlights the fraction of instances, out of 1,000, that the proposed method declares a model valid given a specified bias and error variance. It

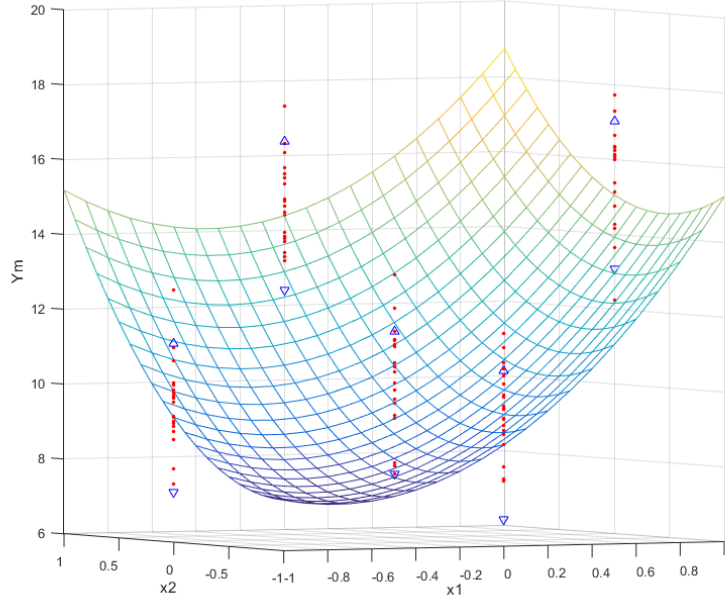


Figure 3.2: Validation Assessment of Model Replications ($x_3 = 0$)

is apparent that the tolerance interval method is sensitive to both sources of model-system deviation. An observed trade-off is present, where smaller error variance results in more allowable model bias and vice versa. Despite this trade-off, the number of valid assessments quickly deteriorates as model bias and error variance become larger.

A more commonly used statistical interval is the confidence interval. The regression confidence interval bounds the mean response rather than a proportion of future observations and is calculated as [83]

$$\hat{Y}(X_h) - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 X_h' (X'X)^{-1} X_h} \leq \mu_{y|X_h} \leq \hat{Y}(X_h) + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 X_h' (X'X)^{-1} X_h} \quad (3.19)$$

where $\hat{\sigma}^2$ is the root mean square error. To show the confidence interval's sensitivity to model bias and error variance, the same procedure as above was completed with the

exception that the mean model response was calculated rather than assessing the model response for each replication. Figure 3.3b depicts this sensitivity.

If we compare the sensitivities of the tolerance interval and confidence interval approaches, we can see that the confidence interval approach is more sensitive to model bias due to the model bias being defined as a shift in the mean response. Conversely, the confidence interval approach is nearly insensitive to increased error variance due the confidence interval only being an evaluation on the mean response. The tolerance interval approach demonstrates sensitivity to both forms of model-system disagreement.

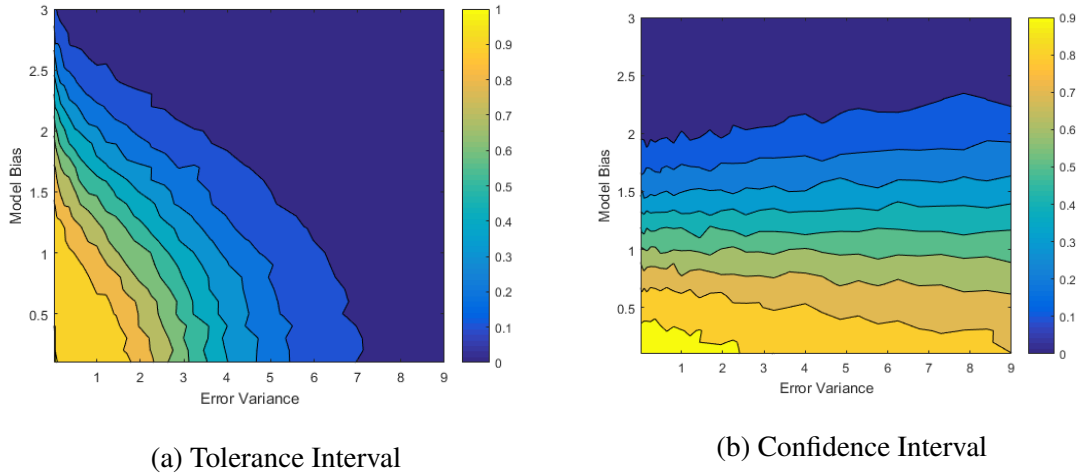


Figure 3.3: Method Sensitivity to Model Bias and Error Variance

3.5 Considerations

Tolerance intervals are an appropriate statistical method for validation of model data due to their inherent characterization of the distributional spread of all future observations with a user defined level of confidence. This evaluation of the response distribution, combined with a separate level of confidence, which can be conducted simultaneously over a design region, makes tolerance intervals more appropriate than the more commonly

used confidence and prediction intervals. The downside to using tolerance intervals for model validation is that their widths are inherently larger than the widths of confidence and prediction intervals. Particular attention must be paid to these widths as their size, if too large, may limit their utility for validation purposes.

The utility of regression tolerance intervals in model validation is limited by the quality of the fitted surface for which the intervals are constructed. The methodology proposed assumes an adequate fit for all models; there is no inflation of tolerance intervals as a result of increased error estimates due to lack-of-fit. Poorly fit response surfaces will result in unacceptable levels of Type II error (validating invalid models). This can be avoided in part by performing a lack-of-fit assessment of the fitted surface.

The current methodology provides upper and lower bounds for the model response data; however, the method does not consider cases where significant model bias is present but low error variance results in a valid assessment. Figure 3.4 illustrates this situation, where the average model response is significantly different than the average system response, but a small variance in the model output results in the model distribution falling within the validation bounds created by the system tolerance intervals (dashed lines). A statistical test on the mean response may be necessary to ensure the model bias is not being compensated by a small error variance. Under the assumption that the OLS fit is adequate, we can perform a one-sample t -test to determine if the sample mean from the model replications is statistically equivalent to the fitted system response value. The test statistic for each X_h can be calculated as

$$t_h = \frac{\bar{Y}_h^m - \hat{Y}_h^s}{S_h^m / \sqrt{r_h^m}} \quad (3.20)$$

where $\bar{Y}_h^m$ and S_h^m are the sample mean and standard deviation of the model replications, respectively. The test statistic would then be evaluated against the critical value based on a desired confidence and $r_h^m - 1$ degrees of freedom.

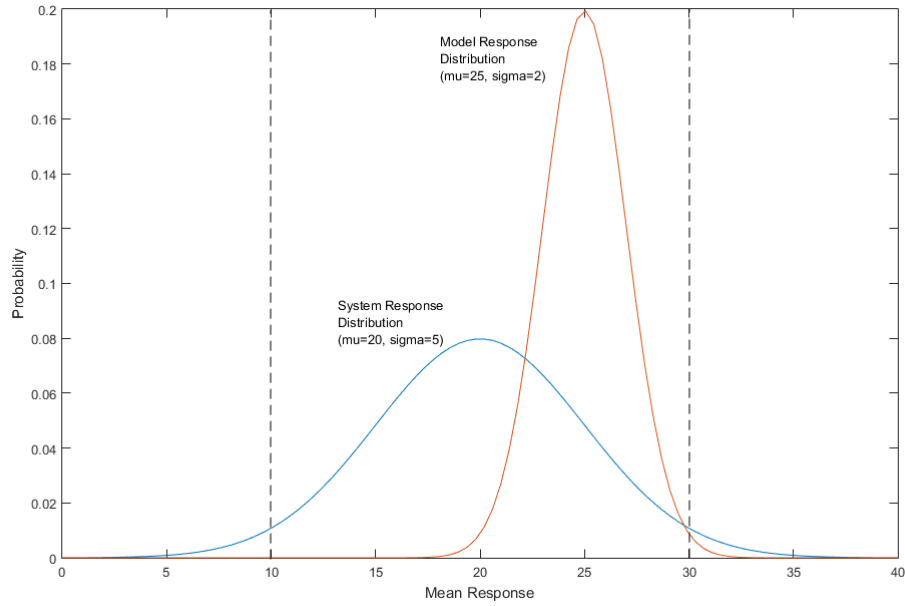


Figure 3.4: Notional System and Model Response Distributions with Tolerance Intervals (Dashed Lines) Set to $\mu_s \pm 2\sigma_s$

The one sample t -test helps to identify significant model bias and reduces the tolerance interval evaluation to a test of the variance; however, under the proposed validation criteria, the model is not penalized for being overly precise ($\sigma_m^2 < \sigma_s^2$). This is an acceptable approach for some models, but other models may require the model output possess the same level of variability as the system. In these cases, the tolerance interval method will need to be extended to penalize the model when the variation among replications is too small.

Another consideration is the assessment of overall validity across the design space. The use of simultaneous intervals is appealing because these intervals maintain a desired confidence over all model assessments; however these interval widths are greatly inflated due to the use of Bonferroni's inequality to maintain the desired overall confidence. As a result, the k -factor for the simultaneous intervals is magnitudes larger than the k -factor

from the non-simultaneous case. This results in high Type II errors. Future research is needed to determine if methods other than Bonferroni's inequality can be used to establish simultaneous tolerance intervals with reasonable interval widths. Ultimately, it is up to the modeler to subjectively assess the width of the intervals to determine their utility in validating the model of interest.

Lastly, experimentation has shown that for the fitted response surface to adequately capture the true system variability, some replications of the system are required. It has also been observed that the number of system replications $r_{h,s}$ does not need to be large, and that as $r_{h,s}$ increases, there is a diminishing improvement on the response surface RMSE term.

3.6 Conclusion

The methodology proposed in this paper provides a simple yet effective way to conduct model validation at any input configuration within a design space for which the model is intended to predict the system response. The use of Bayesian Hypothesis Testing allows for relaxation of the validity criteria when the number of model replications is small. The proposed methodology focuses on univariate response systems. A natural extension is to address the multivariate validation problem where models must be simultaneously validated across a design space for more than one response. This limitation can be addressed by replacing the k -factor approximations used in this paper with multivariate approximations found in the literature. The use of alternative k -factors can also allow for extensions of the methodology to non-parametric data distributions and non-linear regression surfaces. Another extension is to incorporate the validation of dynamical system response that are time-dependent. These types of responses introduce additional problems when it comes to validation; however, if a time-series validation methodology with an associated confidence can be established, an extension of this methodology to the time-series problem is possible.

IV. F-Statistic for Model Validation Over Experimental Regions Using Least Squares Response Surfaces

4.1 Introduction

The library of techniques for model validation is robust; however, much of the literature focuses on validation of systems with “simple” behavior. We define simple behavior as that which is only focused on few input settings, or single valued responses at an input setting and not changing over time. While this was at one time adequate, our increased reliance on computers to emulate complex, multivariate systems dictates a need for validation methods which can address a multitude of multi-variable response outputs.

While validation of a stochastic simulation model over its operational domain of inputs is not novel, as both the engineering design and operations research communities have addressed the issue with various techniques, most of these techniques are limited by a final subjective assessment of model validity which requires human judgment for a final indication of validity. The work presented in this paper provides a more objective and statistically robust validation assessment based on a test statistic formed as the ratio of two mean square errors. Evaluation of this test statistic provides, with some defined level of statistical confidence, an assessment of model validity based on the distributional properties of the system and model data under consideration.

The remainder of this paper is structured as follows: Section 2 provides a review of other related methodologies within the validation literature. Section 3 developed the validation hypothesis and test statistic and structures the validation procedure. Section 4 provides a numerical example that illustrates the utility of the proposed work. Section 5 discusses limitation and potential future research.

4.2 Literature Review

The verification and validation (V&V) of models and simulations is a large and rich research field. The existing methodologies can be categorized based on similarities between their procedural characteristics and the subset of validation problems in which they apply. Validation taxonomies by Davis [27], Balci [3], and Sargent [108] separate these techniques into their appropriate categories. We focus on statistical validation techniques as these are most relevant to the work presented in this paper. These techniques focus on the use of statistically rigorous methods to assess the level of agreement between model and system data. The use of statistically rigorous techniques for simulation model validation is essential to work done in both the Operations Research (OR) and engineering communities. Some of the more commonly utilized methods include, confidence intervals [74], t-tests [111], Hotelling's T^2 tests [7], Chi-Squared and Kolmogorov-Smirnov goodness-of-fit tests [87], non-parametric tests of means [111], univariate and multivariate analysis of variance (ANOVA/MANOVA) [87] [42], and time-series analysis [38] [41] [116] [110]. A more comprehensive list of statistically based validation methods used in the OR community can be found in Balci [6]. These methods collectively and individually address a subset of the validation problems associated with validating data of various distributive properties and functional forms. Many of these methods originated in the 1960's and 1970's yet still remain relevant to many of the validation problems we face today; however, it is essential that we expand our library of techniques to address the new challenges associated with validating more complex systems.

One such new challenge is the validation of simulation models which operate over a domain of input parameters. This challenge has been addressed to some degree in the engineering community. Liu et al. [76] provides a discussion on the different validation approaches utilized in engineering design. These methods are typically based on a hypothesis test regarding the distributional properties of the model and system data. The

methods assess the hypothesis under test through the development of a test statistic based on an error measure between model and system output. Hills and Trucano [52] utilize a weighted least squares test statistic, r^2 , to test the null hypothesis that the experimental data come from the prediction populations. Their r^2 statistic represents the weighted distance from the mean to a measured point, and if this distance is less than the weighted distance to a desired confidence surface, then there is insufficient evidence to reject the model as representative of the system. This statistic is used in several other validation methodologies [45][50] [13][15].

Rebba and Mahadevan [97] mention that most classical approaches focus on comparisons of the first two statistical moments (mean and variance) through the use of a t-test of the means and a χ^2 test for variance. They propose a direct comparison of the cumulative distribution functions, with a distance metric based on the Anderson-Darling statistic. Ferson et al. [37] introduce a similar validation methodology that uses the area between cumulative probability curves to quantify the disagreement between empirical data from the real-world system and predictive model results. While this method does not provide a distinct yes/no validation decision like many of the classical hypothesis testing methods, it is useful when the predictive and/or observed data is expressed as probability distributions and when the uncertainty is captured via intervals [36].

Bayesian approaches to model validation have become increasingly popular in recent years and many deal with assessing model and system agreement at multiple input configurations. These methods, centered around utilizing Bayes Theorem, are wide ranging in application. Bayesian Hypothesis Testing uses a Bayes factor as a validation metric, where the Bayes factor is a relative likelihood of a null hypothesis against the alternative hypothesis, and have been used in numerous applications of model validation to include univariate and multivariate response systems, evaluations of model predictive capabilities outside the tested domain, and for the evaluation of dynamical systems [98] [94] [95] [97]

[96] [57] [56]. Chen et al. [16] and Wang et al. [120] use Bayesian prediction models to draw inferences about a model's ability to generate accurate predictions in the intended but untested design domain.

Oberkampf and Barone [88] take a frequentist approach to model validation in which they quantify the difference between data means. The validation metric becomes a confidence interval around the estimated error in the model in which the true error is believed to reside. This allows for a point-wise assessment of model data at tested and untested design parameters.

While less prevalent in the OR literature, the idea of validating a model across experimental regions is discussed in regards to regression based metamodels. In these papers, metamodels are validated against the simulation model from which they were formed [68]. These assessments involve a subjective evaluation of the metamodel's coefficient estimates $\hat{\beta}$ or by cross-validating the metamodel by constructing the model using only a portion of the simulation data and then using the remaining data for validation of the predictive capabilities of the metamodel at the remaining parameter configurations [65] [64] [63]. Bootstrapping is also discussed as a technique for validation of metamodels [67] [66].

4.3 Methodology

The validation methodology proposed here uses a novel cross-residual method for testing differences in the system and model mean and variance. A cross-residual is defined as the response error measure when the model and system responses are each compared to some common baseline expected value function.

Define Y^s and Y^m as random vectors of system and model responses, respectively. Assume each response in Y^s and Y^m is independently distributed with a mean equal to the expectation of the random variable at a design point, X_i . Our goal is to determine whether the model response Y^m sufficiently represents the system response Y^s over all design points

within the design space. This measure of sufficiency forms the null hypothesis under test. Our null hypothesis is that a valid model representation of a system response is one in which the first two moments of the system and model distributions are equivalent. The null hypothesis is

$$\begin{aligned} H_o : E[Y^s] &= E[Y^m] \\ V[Y^s] &= V[Y^m] \end{aligned} \quad (4.1)$$

Under this null hypothesis, Y^s and Y^m possess the same mean and variance respectively, at corresponding design settings; the distribution for the system and model responses agree in the first two moments.

The true system and model response are represented by

$$\begin{aligned} Y^s &= X\beta^s + \epsilon^s \\ Y^m &= X\beta^m + \epsilon^m \end{aligned} \quad (4.2)$$

where the β^s and β^m coefficients are unknown, X is the $n \times p$ matrix of design settings, and ϵ^s and ϵ^m are normally distributed and independent of one another.

Define $\hat{\beta}^s$ as the Ordinary Least Squares (OLS) estimates of the coefficients describing the system response based on the true system data, Y^s . Under the usual assumptions of OLS, we can define the empirical model as

$$\hat{Y}^s = X\hat{\beta}^s \quad (4.3)$$

where $\hat{Y}^s$ is the predicted value of Y^s at each design setting.

Assume that the model Y^m sufficiently predicts the system of interest. Then the response model should be equivalent to the system model if (4.1) is true. Thus, we can define the model predicted values as

$$\hat{Y}^m = X\hat{\beta}^s. \quad (4.4)$$

Clearly, $\hat{Y}^s = \hat{Y}^m$. Next, define the cross-residuals as

$$\begin{aligned} e^s &= Y^s - \hat{Y}^s \\ e^m &= Y^m - \hat{Y}^s \end{aligned} \tag{4.5}$$

where e^s and e^m are normally distributed random error vectors for the system and model residuals at each design setting, respectively.

To show that e^s and e^m are independent we establish that they are bivariate normal. We use the statistical test of multivariate normality developed by Doornik and Hansen [32]. The proposed multivariate test statistic is approximately distributed as [32]

$$E_p \sim \chi^2(2p) \tag{4.6}$$

where $p = 2$ for bivariate normality. To show bivariate normality of our residuals, we need to show that (4.6) holds for (4.5).

Ten thousand samples of Y^s and Y^m were generated under our null hypothesis using the same baseline Y values and normally distributed error component with a mean of zero and standard deviation of one. The residuals were calculated using the system response surface $\hat{Y}^s$ and each instance of e^s and e^m was tested for bivariate normality. Each instance of the residuals consisted of 180 observations. Maximum likelihood estimation (MLE) was used to fit the 10,000 sample Doornik-Hansen test statistics to a χ^2 distribution. The estimated χ^2 parameter was found to be 3.84, which closely follows (4.6). Figure 4.1 illustrates the closeness between the PDFs for the sample and the hypothesized $\chi^2(4)$ distribution.

Theorem 1: *Under the assumption of Y^m sufficiently representing Y^s , the corresponding residuals e^s and e^m are independent.*

Proof. Define H as an $n \times n$ idempotent matrix that projects Y^s onto $\hat{Y}^s$. Without loss of generality, assume that Y^s and Y^m are vectors of equal length. Let $H \neq 0$, $Y^m \neq 0$, and

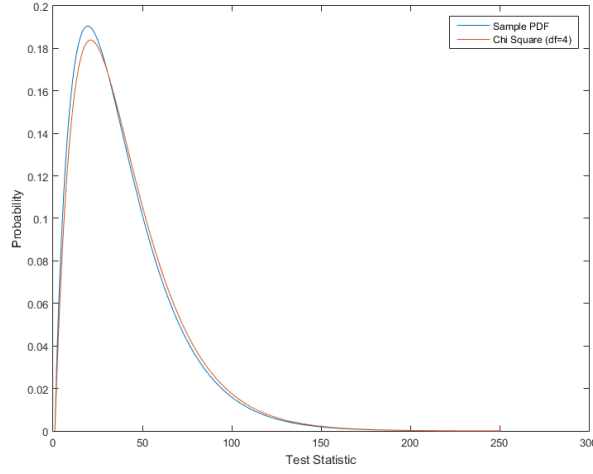


Figure 4.1: Observed and Hypothesized Probability Density Functions for Doornik-Hansen Multivariate Test Statistic

$X = [1, X_c]$ where X_c corresponds to configurations of the independent variables in X . The cross-residuals are defined as

$$\begin{aligned} e^s &= Y^s - HY^s \\ e^m &= Y^m - HY^s \end{aligned} \tag{4.7}$$

Under the null hypothesis, We show that

$$Cov(e^s, e^m) = E[e^s, e^m] - E[e^s]E[e^m] = 0 \tag{4.8}$$

By OLS, we know that $E[e^s] = 0$ and therefore it is sufficient to show

$$Cov(e^s, e^m) = E[e^s, e^m] = 0 \tag{4.9}$$

$$\begin{aligned}
Cov(e^s, e^m) &= E[e^s, e^m] \\
&= E[(Y^s - HY^s)^T (Y^m - HY^s)] \\
&= E[(Y^s)^T Y^m - (Y^s)^T HY^s - (HY^s)^T Y^m + (HY^s)^T HY^s] \\
&= E[(Y^s)^T Y^m] - E[(Y^s)^T HY^s] - E[(HY^s)^T Y^m] + E[(HY^s)^T HY^s] \\
&= E[(Y^s)^T]E[Y^m] - E[(Y^s)^T HY^s] - E[(HY^s)^T Y^m] + E[(HY^s)^T HY^s] \\
&= E[(Y^s)^T]E[Y^m] - E[(Y^s)^T HY^s] - E[(HY^s)^T Y^m] + E[(Y^s)^T H^T HY^s] \\
&= E[(Y^s)^T]E[Y^m] - E[(Y^s)^T HY^s] - E[(HY^s)^T Y^m] + E[(Y^s)^T HY^s] \\
&= E[(Y^s)^T]E[Y^m] - E[(HY^s)^T]E[Y^m] \\
&= E[(Y^s)^T]E[Y^m] - E[(Y^s)^T H^T]E[Y^m] \\
&= E[(Y^s)^T]E[Y^m] - E[(Y^s)^T]E[Y^m] \\
&= 0
\end{aligned}$$

□

The last step is a result of the idempotent properties of H leading to $(Y^s)^T = (Y^s)^T H^T$ as shown by

$$\begin{aligned}
(Y^s)^T &= (Y^s)^T H^T \\
(Y^s)^T H &= (Y^s)^T H^T H \\
(Y^s)^T H &= (Y^s)^T H
\end{aligned}$$

The Mean Square Error (MSE) is a measure of disagreement between a fitted surface (expectation of the response variable) and the data used to generate it. Define the MSE as

$$MSE = \frac{\sum_{i=1}^N e_i^2}{N - p} \quad (4.10)$$

where e_i is the residual at the i^{th} observation and $N = nk$ is the total observations across n design settings and k replications. $N - p$ is the number of error degrees of freedom. The MSE is an unbiased estimate of the variance of the model error. Let MSE_s and MSE_m be the mean square error for the system and model residuals, respectively. Under the assumption that (4.1) holds, both MSE_s and MSE_m provide estimates for the model error variance, $\sigma^2 = \sigma_s^2 = \sigma_m^2$.

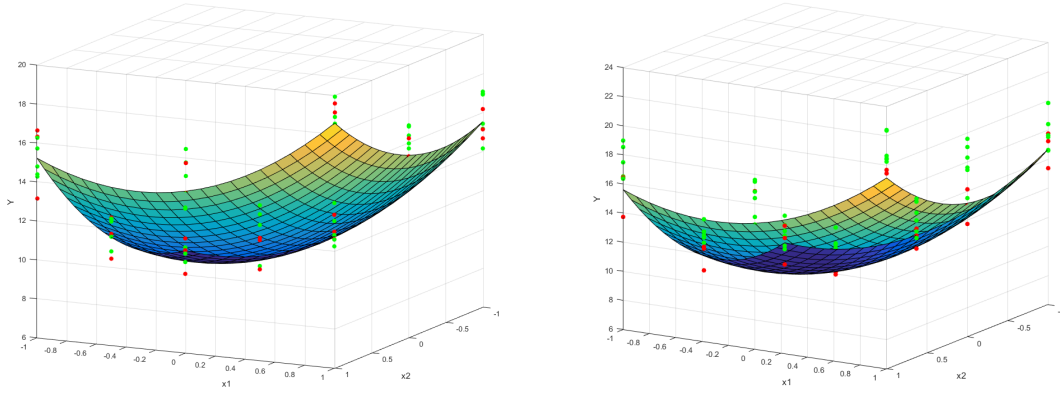
Corollary: Since e_s and e_m are independent, then MSE_s and MSE_m are independent estimates of σ^2 .

Under the assumptions of OLS, since $\epsilon \sim N(0, \sigma^2)$, MSE_s and MSE_m both follow χ^2 with the degree of freedom $N-p$ so their ratio follows an F-distribution [84]. Under the null hypothesis in (4.1), $MSE_s = MSE_m$ so we use the ratio of these values to test for statistically significant differences between σ_s^2 and σ_m^2 . Define the test statistic as

$$F = \frac{MSE_m}{MSE_s} \quad (4.11)$$

for testing $H_0 : \sigma_s^2 = \sigma_m^2$. The null hypothesis is rejected if $F > F_{\alpha/2, DF_m, DF_s}$ or $F < F_{1-\alpha/2, DF_m, DF_s}$ where DF_m and DF_s are the degrees of freedom of the model and system mean square error, respectively. A failure to reject the null hypothesis, means we conclude that Y^s and Y^m vary equivalently around $\hat{Y}^s$ and therefore the model sufficiently represents the system behavior over the entire design space.

A rejection of the null means $E(MSE_m) \neq \sigma^2$, thus $\sigma_m^2 \neq \sigma_s^2$. This disagreement is due to differences in the model mean and/or the model variance. Figure 4.2 shows notional model data(green), system data(red), and a response surface fit to the system data. Figure 4.2a is a valid model where both sets of data share distributional properties (same variance and centered around $\hat{Y}_s$), while Figure 4.2b is an invalid model where the model data is clearly biased above the response surface.



(a) $MS E_m = 1.217$, $MS E_s = 1.264$, $F = 0.963$ (b) $MS E_m = 4.268$, $MS E_s = 1.296$, $F = 3.294$

Figure 4.2: Evaluation of Model Data Using MSE from Response Surface

Under OLS, the residuals have a mean equal to zero when that residual is calculated using expected values based on the empirical model of that data. Rejecting (4.1) means that the distribution of e^m , may not have a mean equal to zero at one or more design settings. Let μ_i be the true mean of e^m at design setting i . Then $MS E_m$ follows a non-central χ^2 with a non-centrality parameter (λ). The ratio of the mean square errors becomes a ratio of a non-central χ^2 and a central χ^2 , which follows a non-central F-distribution [93]. The likelihood of a Type II error is represented by the overlap between the central and non-central F-distributions established above.

The selection of experimental design or the determination of empirical model significance and adequacy are not discussed; however, these concepts are covered at length in Montgomery et al. [84] and Myers et al. [86], respectively.

4.4 Example

Consider a notional response of the form $Y = 10 + 2.5x_1^2 + 2x_2^2 - 1.25x_1x_2 + 3x_1 + 2.75x_2$. Let the stochastic system response be of the form $Y^s = Y + e^s$ where $e^s \sim N(0, 1)$. The

experimental data is gathered from three replications of a full factorial design with two parameters and three levels. This is summarized in Table 4.1. This data are fit to an OLS response surface, yielding $\hat{Y}^s$.

Table 4.1: Notional System/Model Response Across Design Space

Design Settings		System Response (Replication)			Model Response (Replication)					
x1	x2	1	2	3	1	2	3	4	5	6
-1	-1	9.359	6.639	6.678	8.659	7.142	8.399	8.962	7.159	9.514
0	-1	8.645	10.035	9.156	9.134	9.139	9.720	9.925	9.141	9.762
1	-1	16.103	16.309	16.336	16.629	15.630	16.694	17.374	15.103	18.004
-1	0	10.063	9.266	8.595	8.664	8.674	8.217	9.827	9.207	8.977
0	0	10.114	8.943	9.712	12.078	9.291	9.180	9.543	10.019	10.727
1	0	14.595	15.216	15.850	16.641	15.490	15.253	14.219	15.746	15.440
-1	1	15.032	15.413	13.664	16.487	14.001	15.810	15.855	15.636	16.300
0	1	14.625	13.281	15.786	14.074	15.558	16.079	15.769	15.561	15.544
1	1	20.479	19.192	21.424	19.520	19.082	18.910	18.907	19.309	17.014

Define the functional form for the model data as $Y^m = (Y + \delta) + e^m$, where δ represents the model bias and $e^m \sim N(0, \sigma_m^2)$. The model bias is defined as a constant deviation between the model and system response across the experimental region under test. A value of $\delta = 0$ indicates no model bias, $\delta < 0$ indicates a model bias below the system response, and $\delta > 0$ indicates a model bias above the system response. Let the notional model have a bias ($\delta = 0.2$) and an error variance equal to the variance from the system error ($\sigma_m^2 = 1$). Data from six replications of the model data across the experimental region were generated for model validation. This data are also summarized in Table 4.1.

Equations (4.5) and (4.10) were used to calculate the level of disagreement between each set of data and the fitted response surface. The mean square errors were calculated as $MS E_m = 1.155$ and $MS E_s = 0.835$, yielding a test statistic:

$$F = \frac{MS E_m}{MS E_s} = \frac{1.155}{0.835} = 1.383. \quad (4.12)$$

Our critical F values using a significance level of 0.05, $DF_m = 48$, and $DF_s = 21$ establish a model validity range for our F-statistic such that the model is valid if

$$0.503 < F < 2.215. \quad (4.13)$$

Since our statistic falls within the corresponding critical values, we fail to reject the null and declare the model as valid.

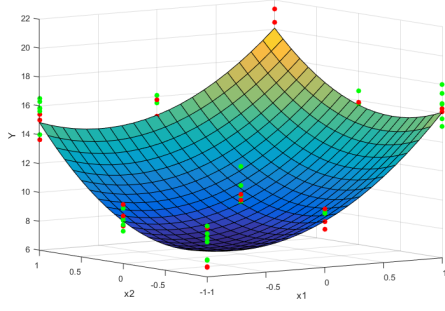
Now consider a model with increased bias and a different error variance. Let the model have a bias ($\delta = 0.6$) and an error variance larger than the system error variance ($\sigma_m^2 = 1.1$). Again we generate six replications of the model data for validation purposes.

This example yields $MSE_m = 1.937$, $MSE_s = 0.788$, and the corresponding F-statistic of $F = 2.459$, which falls outside the critical values in (4.13). This indicates a rejection of the null hypothesis and we declare the model invalid. This assessment makes intuitive sense because the additional model bias and increased variance indicate that Y_s and Y_m are not equivalently distributed.

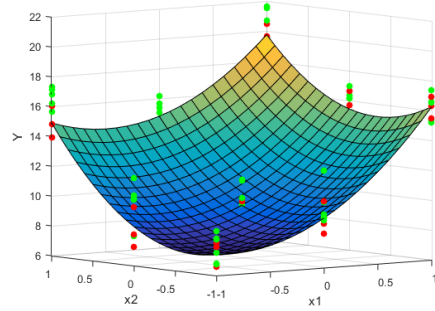
Graphically, the differences between a valid model and an invalid model can be difficult to determine. The data from the two examples above, along with their respective response surface are shown in Figure 4.3. There are some indications that the model in Figure 4.3a is a better fit than the model in Figure 4.3b; however it may not be so clear as to declare one valid and the other invalid based on a visual assessment. However, our F-statistic provides a clear distinction between the validity of the two models.

Next consider the same notional system and model functional forms and vary the two causes for disagreement (bias and error variance) to show the sensitivity of the F-statistic method to changes in the model quality both sequentially and simultaneously.

To assess the method sensitivities over a range of underlying baseline response functions, define the baseline response function as



(a) Valid Model



(b) Invalid Model

Figure 4.3: Visual Representation of Model and System Data With Respect to the Fitted Surface

$$Y^b = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 \quad (4.14)$$

where $\beta_i \sim U(-10, 10)$. This allows us to test a unique function for each iteration of our sensitivity analysis. Define our system and model response as

$$\begin{aligned} Y^s &= Y^b + e^s \\ Y^m &= (Y^b + \delta) + e^m \end{aligned} \quad (4.15)$$

where $e^s \sim N(0, 1)$.

To test the sensitivity of the F-statistic to increases in model bias, we vary model bias from $\delta = 0.02 - 2.0$ while $\sigma_m^2 = 1$. The number of iterations resulting in a valid assessment (out of 1,000 iterations) per bias level is tabulated. Figure 4.4a shows that the F-statistic is highly sensitive to increases in model bias when the system and model error variances are equal. This is evident by the steep drop in valid assessments as model bias increases. The magnitude of the model bias constant with regard to a valid model assessment is directly

proportional to the variance of the system response, σ_s^2 . In other words, a larger σ_s^2 results in a larger allowable model bias for model validity.

To test the sensitivity of the F-statistic to changes in the model error variance, the variance is varied from $\sigma_m^2 = 0.0 - 4.0$ without model bias. Figure 4.4b shows that the F-statistic is highly sensitive to the model's error variance as evident by the steep drop off in valid assessments as the variance deviates from $\sigma_m^2 = \sigma_s^2 = 1.0$.

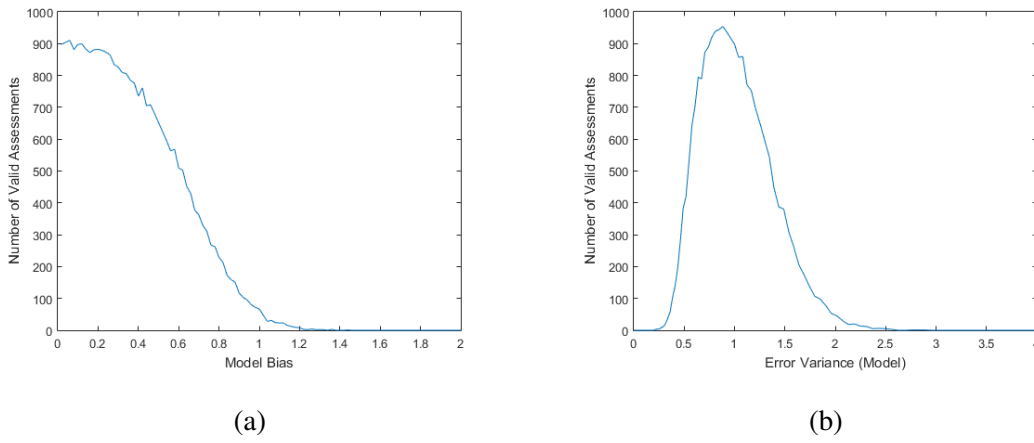


Figure 4.4: F-Statistic Sensitivity to Model Bias(a) and Model Error(b)

Next we examine how the method performs under different levels of model bias and error variance simultaneously. Figure 4.5 provides a contour map of the percentage of valid assessments (over 1,000 iterations) under differing combinations of model bias and error variance. From the contour map, we can see that a trade-off exists between the two competing forms of model discrepancy. As the error variance decreases, the allowable model bias for a valid assessment increases. This is because the F-statistic method cannot distinguish between the two sources of system to model disagreement. The region of high validity appears to be a nonlinear combination of model bias and error variance, with the maximum allowable model bias resulting from a zero error variance. Through

simulation, we can show that the maximum allowable model bias (under no error variance) is approximately equal to the standard deviation of the system response, σ_s . Figure 4.6 gives four examples of validity contours under different σ_s .

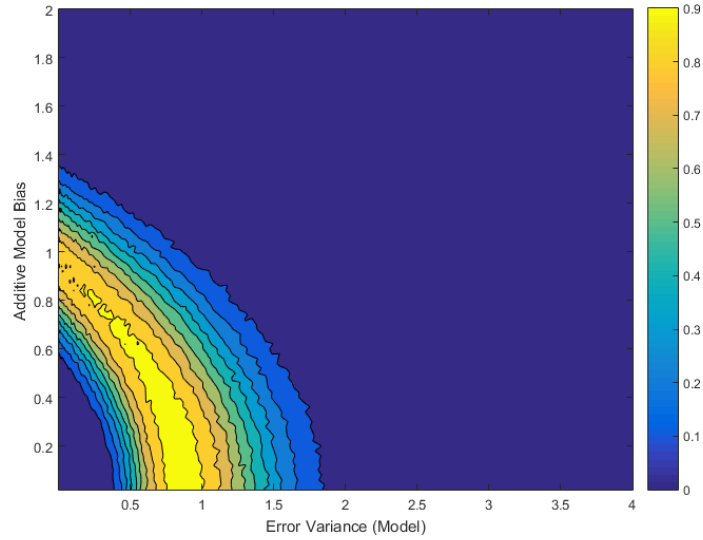


Figure 4.5: Percent Valid Assessments as a Function of Model Bias and Error Variance

4.5 Considerations

The F-statistic developed in this paper tests the similarity of system and model distributions using the null hypothesis that $\sigma_s^2 = \sigma_m^2$. The estimated $\hat{\sigma}_s^2$ and $\hat{\sigma}_m^2$ values are generated based on models derived from the observed samples. We do not directly test either $\hat{\sigma}_s^2 = \sigma_s^2$ or $\hat{\sigma}_m^2 = \sigma_m^2$. There are instances when a χ^2 test of variances will conclude that $\hat{\sigma}_s^2 = \hat{\sigma}_m^2$, but the F-test in (4.11) may show that the differences between estimates is statistically significant. When $\sigma_s^2 = \sigma_m^2$ but $\hat{\sigma}_s^2 \neq \hat{\sigma}_m^2$, this leads to Type I error in which a valid model is assessed as invalid. This occurs when the estimates deviate from the true variance in opposite directions. In a χ^2 test, the allowable deviation between the variance

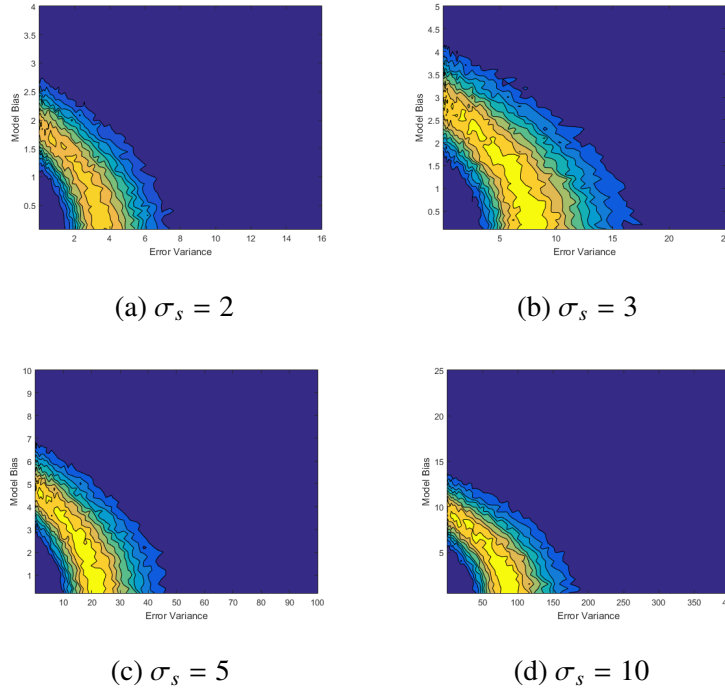


Figure 4.6: Validity Contours with Differing σ_s

estimation to the true population variance (based on the critical χ^2 values) decreases as our sample size, or subsequently our test degrees of freedom, increases.

We can reduce the Type I error by increasing the model and/or system observations. Increasing the model observations is often the more practical option given the high cost of obtaining additional experimental runs. To illustrate this point, we used the F-statistic to evaluate the notional response from the second example, with no model bias and the same error variance ($Y^s = Y^m$). This, under our null hypothesis, will result in the model being evaluated as valid. Conducting the F test with $\alpha = 0.05$ should result in a confidence of test equal to 95%. The number of model observations was incrementally increased and the number of valid assessments out of 10,000 iterations was tabulated, which corresponds to the observed confidence of the test. Figure 4.7a shows that the test statistic performs poorly at low observations, but quickly improves with increased observations. It can be seen that

the confidence asymptotically approaches a confidence above 90% with approximately 100 model observations (approx. 10 replications). The remaining gap between the expected confidence (95%) and the observed confidence (90%) can be attributed to $\hat{Y}^s$ being fit using only three replications of the system data. Figure 4.7b shows that when the system replications are increased to 50, the confidence of the test approaches 95% as the model observations increase. We acknowledge that the F-statistic can be susceptible to Type I error with small sample sizes, but leave it to the modeler to decide an appropriate sample size for their validation effort.

The proposed methodology was developed using OLS regression. We recognize that in some instances, characteristics of the system data will require the use of other data fitting techniques. Our methodology only requires that the technique used produces an adequate model of the true system behavior. Kleijnen et al. [67] mention that correlation among simulation responses exists in the presence of common random numbers. In cases where the errors across the experimental region are correlated and/or non-constant, Generalized or Weighted Least Squares may be appropriate. As general forms of OLS, these fitting techniques should have natural extensions to the cross-residual method.

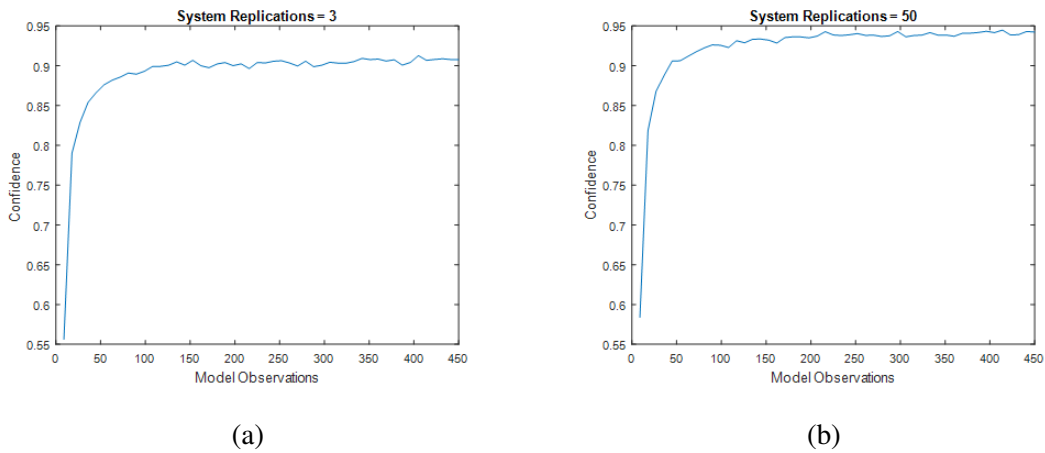


Figure 4.7: Confidence of Validation Method with Increasing Model Observations

4.6 Conclusion

The proposed methodology provides a means for assessing model validation across an experimental design region through the use of cross-residual based measures of disagreement. An F-statistic, calculated from the ratio of model and system mean square errors, is used to statistically detect differences in the distributional properties of the system and model responses. This methodology is a baseline validation technique with plenty of opportunities for future research in regards to relaxations of the Least Squares assumptions. The utility of neural networks or robust regression fitting techniques should be explored as these techniques can be less susceptible to response outliers or non-Gaussian residuals. Future research could also extend the methodology to model validation of multivariate response output.

V. Model Validation of Discretized Functional Response Data Across Experimental Regions Using Functional Regression Extensions to the CORA Objective Rating System

5.1 Introduction

The models engineers and analysts develop to emulate dynamic real-world systems are becoming more complex; our validation methods must be capable of addressing these complexities. This paper examines functional response data that vary over a region of input parameter configurations. The Correlation Analysis (CORA) objective rating system is typically used to evaluate a model response at a single parameter configuration. This paper implements the CORA rating system at multiple parameter configurations to compare model and system data over an experimental region. The CORA rating is then extended to include points within the region in which no experimental data was collected. At these non-sampled points, estimates of the system response are calculated using functional regression techniques. Using the estimated response, the CORA methodology is employed to assess model validity. A weighted linear combination of the CORA scores at each evaluated design point provides an overall validation score for the model across the entire experimental region. Lastly, a new area rating is introduced that overcomes limitations of the current CORA size and shape scores.

The remainder of this paper is structured as follows: Section 2 provides a review of model validation with a focus on methodologies that handle considerations to the experimental region as well as validation methodologies for functional response data. Section 3 outlines general functional response regression and the CORA rating system. Section 4 proposes an extension to the CORA rating system. Section 5 discusses the new CORA area rating. Section 6 discusses limitations and considerations for these extensions.

5.2 Literature Review

Validation within the modeling and simulation (M&S) community is a dense field of research. Balci [3], Davis [27], and Sargent [108] provide extensive taxonomies that categorize the validation methodologies based on similarities between each method's approach and the subset of problems in which they apply. Both the Operations Research (OR) and engineering communities emphasize the desire to use statistically rigorous techniques to validate simulation models. Some of the more commonly utilized statistical methods include confidence intervals [74], t-tests [111], Hotelling's T^2 tests [7], Chi-Squared and Kolmogorov-Smirnov goodness-of-fit tests [87], non-parametric tests of means [111], univariate and multivariate analysis of variance (ANOVA/MANOVA) [87] [42], and time-series analysis [38] [41] [116] [110]. Balci [6] provides a complete list of statistical validation techniques as part of his comprehensive look at the model verification, validation, and testing (VV&T) process. Despite being decades old, several of these techniques are still widely used in industry today; however, many have been modified to address the growing challenges of validating complex response behavior. Two of these challenges are model validation over changing parameter configurations and validation of functional response data.

Literature on validation of models over a domain of varying input parameters is more prevalent in the engineering community and plays a critical role in the engineering design selection process. Various engineering methods validate a model's predictive capability over a design region. Hills and Trucano [52] developed a weighted least squares test statistic, r^2 , that measures the weighted distance between an observed response and the mean response in relation to the weighted distance to a desired confidence region. This statistic tests a null hypothesis that the experimental data comes from the prediction population. This validation statistic has been utilized by other authors in the field [45][50] [13][15]. Rebba and Mahadevan [97] propose a direct comparison of cumulative

distribution functions using a distance metric based on the Anderson-Darling statistic. Oberkampf and Barone [88] quantify the difference between response data means using a metric based on a confidence interval around the estimated model error in which the true error is believed to reside. Ferson et al. [37] quantifies the disagreement between system and model results by measuring the area between cumulative probability curves.

Several authors use Bayesian techniques to conduct model validation, many of which are based on the use of a Bayes factor. These methods evaluate model predictive capabilities in an array of modeling applications to include univariate and multivariate response systems, prediction at untested points within and outside the test domain, and evaluation of dynamical data [98] [94] [95] [97] [96] [57] [56] [16] [120]. The simulation communities most prevalent contribution to model validation across experimental regions is focused on the validation of metamodels built from simulation data [68] [65] [64] [63] [67] [66].

The literature on validation of functional data is fairly recent but expanding rapidly. This is in part due to the V&V communities inability to reach an agreement on a universal approach to validate functional responses. Two common methods for measuring disagreements between functional data sets are correlation and spectral analysis [9] [54]. Both methods measure the cross-correlation between two sets of time-series data; correlation operates in the time domain and spectral analysis operates in the frequency domain. Error measures, or metrics, are another commonly used technique to compare sets of functional data. These error measures include vector norms, assessments of the average residual error, coefficient of correlation and cross-correlation, as well as measures that are based on assessments of errors due to phase, magnitude, and shape differences between functional data sets [31] [102] [112] [110].

Two metrics of interest that combine the error measures with additional information to produce an improved validation measure are the error assessment of response time histories (EARTH) metric [110] and the correlation and analysis (CORA) objective rating system

[43]. The EARTH metric uses dynamic time warping along with a linear regression on the three sources of error (phase, magnitude, slope) to establish a metric of disagreement based on subject matter expert (SME) evaluations of historical data. An enhanced version of the EARTH metric (EEARTH), proposed by Zhan et al. [127], aims to make the scoring system more intuitive to the engineers using it. The CORA rating is a global rating constructed from a weighted linear combination of a corridor score and a cross-correlation score based on measured differences (errors) between the phase, shape, and size properties of the functional data. This objective rating system is predominantly used to evaluate finite element biofidelity models in the field of vehicle impact testing. The details for the CORA rating system are elaborated on in the next section.

Other methodologies for functional data validation focus on the use of Bayesian techniques and wavelet analysis. The Bayesian techniques are founded on the calculation of a Bayes factor and the evaluation of the functional response data using Bayesian interval hypothesis testing and principle components analysis (PCA) to reduce the dimensionality of multivariate data [59] [128] [73]. Wavelet analysis supplements the model validation effort by using the wavelet coefficients to generate component energies [56], calculating time-frequency coherence [58] or as a smoothing procedure prior to the model validation assessment [22].

5.3 Functional Regression and the CORA Objective Rating System

Validation using the CORA rating system is currently limited by requiring both model and system responses at specified points of interest in the design region. Thus, its utility with regard to validation of simulations over their full intended operations domain is limited. The CORA method is extended to allow for validation assessments anywhere within an experimental region. The extension uses functional regression analysis to estimate the time-discretized system response at configurations in which no system response data is available. This estimation of the true system response allows assigning of

a CORA rating for the model at all parameter configurations, and thereby provides a means to better evaluate the simulation's predictive capability throughout the defined experimental region.

5.3.1 *Functional Regression.*

Functional regression analysis involves regression in one of three unique cases: functional predictors with scalar responses, scalar predictors with functional responses, and functional predictor and responses. Morris [85] provides a summary of functional regression techniques for all three cases. This paper focuses on the case in which the predictors are scalar and the response takes a functional form. In particular, we utilize an approach similar to the general functional response regression approach proposed by Faraway [34], which uses Ordinary Least Squares (OLS) to model multidimensional response surface functions over discrete time steps.

Suppose we observe N system responses over m discrete time intervals across an experimental region with n design points and k replications, where $N = nk$. The responses come from a true response function model

$$y(t) = X\beta(t) + \epsilon \quad (5.1)$$

where $\beta(t) = (\beta_1(t)^T, \dots, \beta_p(t)^T)^T$, X is a $N \times p$ design matrix, and $\epsilon = (\epsilon_1(t), \dots, \epsilon_N(t))^T$. Then $\beta = (\beta_1(1), \dots, \beta_p(m))^T$ is a $p \times m$ matrix of the response function coefficients and $Y = (y^T(1), \dots, y^T(m))$ is a $N \times m$ matrix of collected system responses at each discretized time step.

Assume we collect sufficient data to adequately estimate the system response function at each of the discretized time steps, $t = 1, \dots, m$. We can estimate the system response function coefficients as $\hat{\beta} = (X^T X)^{-1} X^T Y$ where $\hat{\beta}$ is a $p \times m$ matrix of β coefficient estimates used to estimate the system response as

$$\hat{Y}_{(N \times m)} = X_{(N \times p)} \hat{\beta}_{(p \times m)}. \quad (5.2)$$

Each row of $\hat{Y}$ gives $\hat{y}(t)$ for the discretized time steps $t = 1, \dots, m$ and each column of $\hat{\beta}$ gives the coefficient estimates for each discretized time step. An estimate of the system response at a new parameter configuration is calculated as [34]

$$\hat{y}_0(t) = x_0^T \hat{\beta}(t). \quad (5.3)$$

where x_0^T is a $1 \times p$ vector of parameter values corresponding to the desired point in the design space and $\hat{y}_0(t)$ is a $1 \times m$ vector of the estimated response at time step $t = 1, \dots, m$.

Faraway [34] [35] presents multiple methods for testing lack of fit for $\hat{y}(t)$ based on a comparison of nested linear models. Model adequacy can be checked either through a visual inspection of the eigenvectors from the estimated residual covariance matrix or by visual inspection of the predicted vs residual plots for the OLS fit at each time step. The former is preferred when the number of discretized points in time is large. For more details regarding lack-of-fit and model adequacy testing, refer to [34] and [35].

5.3.2 CORA Objective Rating System.

The CORA rating system was proposed by Gehre et al. [43] as an objective nonparametric evaluation of time-discretized experimental and model response curves, with an emphasized application in vehicle safety testing. The CORA rating consists of two scoring methods combined using a weighted linear combination to provide a global rating quantifying the level of agreement between experimental and model functional data sets. The first scoring method evaluates the position of the model response within user specified bounds above and below the average system response. These bounds form corridors of evaluation. The second scoring method measures the cross-correlation components between the two responses in terms of phase, size, and shape. Figure 5.1 shows the CORA rating structure in terms of its sub-methods.

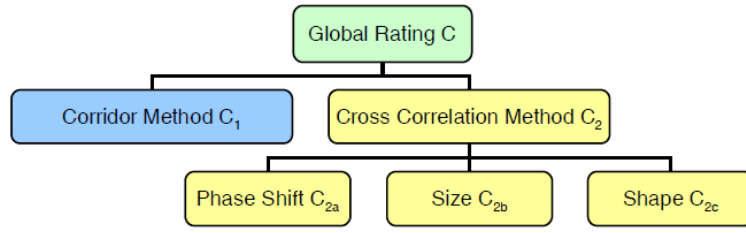


Figure 5.1: Structure for CORA Objective Rating System [43]

The corridor method fits inner and outer corridors around the mean curve of the system signal. If the model curve, at a discretized point in time, falls within the inner corridor, a score of “1” is given. If the model curve falls outside of the outer corridor, a score of “0” is given. If the model curve falls between the two corridors, a score between 0-1 is given based on a user-defined interpolation factor. This interpolation is typically linear but can be higher order. The evaluation is completed at each time step and the average score over all time steps is the model’s corridor score. Figure 5.2a illustrates the corridor scoring scheme at time $t = 1$ with respect to the three zones where the location of the model response within the prescribed corridors (black dots) dictates the corridor rating at that time step.

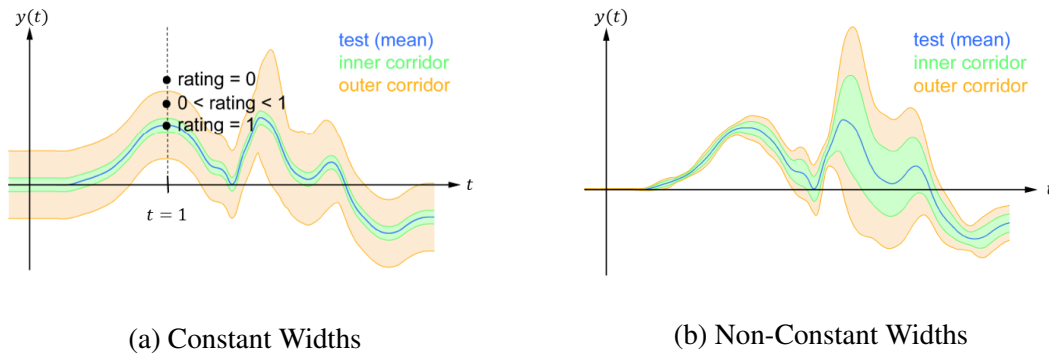


Figure 5.2: Constant and Non-Constant Corridor Widths [43]

Corridor widths are user-defined. A common approach is to construct constant corridor widths by taking a percentage of the global absolute maximum system response within the interval of evaluation. Let y_{max}^s be the maximum value of the system response over all discrete time steps. Untaroiu et al. [115] use $\delta_i = 0.05y_{max}^s$, $\delta_o = 0.25y_{max}^s$ as their inner and outer corridor widths, respectively. Gehre et al. [43] suggest using narrower inner corridor widths and wider outer corridor widths to limit the number of “0” and “1” scores. This allows the corridor portion of the CORA method to better distinguish between responses of differing quality. The use of the global absolute maximum system response to generate corridor width is appropriate, particularly when a lack of system replications restricts the ability to incorporate the variability of the system response. A downside to using y_{max}^s is that response curves with large deviations in response magnitude will result in corridor widths that do no score each time step proportional to the mean response at that point. Calculations, such as the standard deviation of the system response at each time step, can be added to or used in place of the constant corridor widths to provide changing corridor widths at discretized time steps, as shown in Figure 5.2b. The sample standard deviations at each time step should be incorporated into the corridor widths when replications of the system response are available.

The CORA rating system is supported by an open source software that will generate CORA scores for the user. The corridor scoring mechanism is defined in the CORA user’s manual as [113]

$$c_t = \begin{cases} 1 & \text{if } |y^s(t) - y^m(t)| < \delta_i(t) \\ \left(\frac{\delta_o(t) - |y^s(t) - y^m(t)|}{\delta_o(t) - \delta_i(t)} \right)^{k_c} & \text{if } \delta_i(t) < |y^s(t) - y^m(t)| < \delta_o(t) \\ 0 & \text{if } |y^s(t) - y^m(t)| > \delta_o(t) \end{cases} \quad (5.4)$$

where c_t is the corridor score at time step i , $y^s(t_i)$ and $y^m(t_i)$ are the system and model

response at time step t , respectively, δ_i and δ_o are the inner and outer corridor widths, respectively, and k_c is an interpolation factor for the corridor rating. This factor determines the rate at which the score decreases from one to zero. The interpolation factors are typically set to represent a linear or quadratic decrease in the component score. Each of the four component scores receives it's own interpolation factor. The average of the corridor scores over the m discretized time steps becomes the overall corridor rating.

The cross-correlation method calculates a correlation score from a weighted linear combination of three sub scores based on assessing phase, size, and shape (progression) disagreements between system and model response data (Figure 5.3). Each component receives a score between 0-1, resulting in the cross-correlation score also being between 0-1.

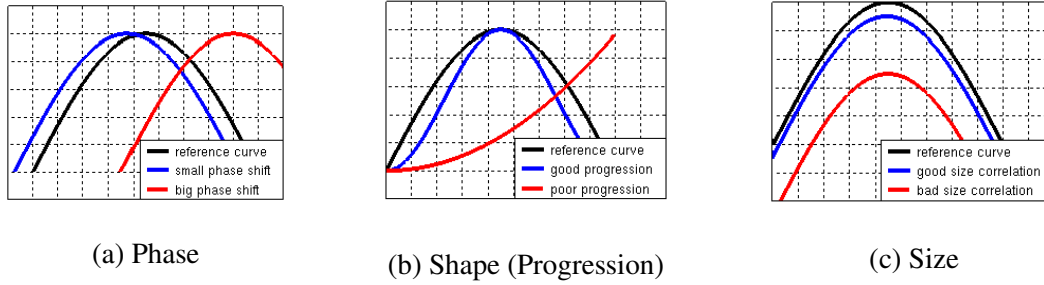


Figure 5.3: Three Measures of Signal Disagreement [113]

Each part of the cross correlation rating is based on the cross correlation between the model and system response curves, which calculates as [113]

$$K_{y^m y^s}(M) = \frac{\sum_{t=0}^{m-1} y^m(t_{min} + (M + t)\Delta t) y^s(t_{min} + t\Delta t)}{\sqrt{\sum_{t=0}^{m-1} (y^m)^2(t_{min} + (M + t)\Delta t) \sum_{t=0}^{m-1} (y^s)^2(t_{min} + t\Delta t)}} \quad (5.5)$$

where M is the number of time shifts, t represents the current time step, t_{min} is the first recorded time within the interval of evaluation, Δt is the length between each discretized time step, and m is the total number of discretized time steps.

The phase shift ranking is calculated similar to the corridor rating. First, minimum and maximum time shifts are calculated as

$$\begin{aligned}\delta_{min} &= D_{min}(t_{max} - t_{min}) \\ \delta_{max} &= D_{max}(t_{max} - t_{min})\end{aligned}\tag{5.6}$$

where D_{min} and D_{max} are user defined percentages of the interval of evaluation between $0 < D_{min}, D_{max} \leq 1$. The time shift at the maximum $K_{y^m y^s}$ is δ . The phase shift score is calculated as [113]

$$C_P = \begin{cases} 1 & \text{if } |\delta| < \delta_{min} \\ \left(\frac{|\delta_{max} - |\delta||}{\delta_{max} - \delta_{min}} \right)^{k_p} & \text{if } \delta_{min} < |\delta| < \delta_{max} \\ 0 & \text{if } |\delta| > \delta_{max} \end{cases}\tag{5.7}$$

where k_p is the phase score interpolation factor.

The shape or progression score is a function of the maximum cross-correlation $K_{y^m y^s}$. Let $K = \max(K_{y^m y^s}(M))$. The shape score is calculated as [113]

$$C_V = \left(\frac{1}{2}(K + 1) \right)^{k_v}\tag{5.8}$$

where k_v is the shape interpolation factor.

The size score is a function of the ratios between the the signals and the time axis. Let A^s and A^m be the areas for the system and model, respectively. These areas are approximated as the sum of the response magnitudes across all discretized time steps, defined by [113] and expressed as

$$A = \sum_t (y(t))^2. \quad (5.9)$$

The size score is then calculated as found in [115] [113] as

$$C_G = \begin{cases} \left(\frac{A^m}{A^s}\right)^{k_s} & \text{if } A^s \geq A^m \\ \left(\frac{A^s}{A^m}\right)^{k_s} & \text{if } A^s < A^m \end{cases} \quad (5.10)$$

where k_s is the size interpolation factor. The total correlation score is calculated as

$$C_2 = w_p C_P + w_v C_V + w_G C_G \quad (5.11)$$

where w_i indicates a user specified weighing for the i component with each $w_i > 0 \forall i$ and $\sum_{i \in p, v, g} w_i = 1$.

The overall CORA rating is

$$C = w_1 C_1 + w_2 C_2 \quad (5.12)$$

where C_1 is the corridor score, C_2 is the cross-correlation score, and w_1 and w_2 where $w_i > 0 \forall i$ and $\sum_{i=1}^2 w_i = 1$ are the weights for the corridor and cross-correlation scores, respectively. More information on the CORA rating system can be found in [113]. A discussion on the selection of the weighting factors is provided later in this paper.

5.4 Extensions to CORA Objective Rating System

The CORA rating is typically used to assess the validity of a functional model response, $y^m(t)$. Using a weighted linear combination of CORA ratings, the CORA method is extended to evaluate a response across multiple design points, to include parameter configurations in which no experimental data exists. The current CORA rating system requires experimental observations at each desired design point to assess model validity. When validating across an experimental region, model validation at points within the

region where no experimental responses were collected is necessary. General functional regression methods provide a prediction of the system response at any point within the experimental region. These predictions are used in lieu of observed data to validate model responses using the CORA method.

The following assumptions are required to extend the CORA method to assess model validity at untested parameter configurations:

- Model and system functional response curves are filtered to reduce noise;
- All responses have been aligned according to the time axis;
- A sufficiently large number of m time steps are available to approximate the behavior of the entire response function; and
- The OLS assumptions of homoscedastic and uncorrelated errors hold

First, consider using the CORA rating system to validate at design points with corresponding experimental data. An overall CORA score across the entire experimental region is calculated by first calculating the CORA scores at each of the n design points and then using a weighted linear combination of the individual scores to establish a CORA score for the region, defined as

$$C_R = \sum_{i=1}^n w_i C_i \quad (5.13)$$

where w_i is the user assigned weighting at each design point. A similar approach is used in the biofidelity modeling literature when trying to establish an overall CORA rating for finite element models which provide functional response data for multiple impact responses; however, in those cases, equal weights are used so that the overall CORA score is an average of the individual scores. The analyst can choose to use equal weights, but unique weights can be utilized to place additional emphasis on portions of the experimental region which are of particular interest to the modeler.

Recall that the CORA method assesses the validity of $y^m(t)$ using the average of the replications of the system response at a design point. This averaged system response curve is denoted as $y^s(t)$. General functional regression discretizes $y^s(t)$ over m time steps and uses multivariate multiple OLS regression to fit $\hat{y}^s(t)$ at all time steps. By the Gauss-Markov Theorem, under the assumptions of OLS, this fit provides the best linear unbiased estimators for the $\beta(t)$ coefficients at each of the m time steps. It can be shown that using these $\hat{\beta}(t)$ results in the expectation of $\hat{y}^s(t)$ equaling the mean of the system response over all replications, thus making $\hat{y}^s(t)$ an appropriate estimation of $y^s(t)$.

Let x_0 be a configuration within the experimental region in which no experimental data is available. To validate $y_0^m(t)$, calculate $\hat{\beta}(t)$ using multivariate OLS regression on the time-discretized system response over all experimental observations. Use Equation 5.3 to estimate $y_0^s(t)$ and use $\hat{y}_0^s(t)$ for $t = 1, \dots, m$ to conduct CORA analysis on the discretized time steps of $y_0^m(t)$, in lieu of having $y_0^s(t)$.

When using $\hat{y}_0^s(t)$ to estimate $y_0^s(t)$, particular attention should be paid to how one defines the corridor widths as part of the CORA method. The CORA User's Manual provides a means for accounting for variability in the system response by adding the standard deviation at each time step to the constant corridor widths. The non-constant corridor widths at time t are defined as [113]

$$\begin{aligned}\delta_i(t) &= a_o y_{max}^s + a_s \sigma(t) \\ \delta_o(t) &= b_o y_{max}^s + b_s \sigma(t)\end{aligned}\tag{5.14}$$

where a_o , a_s , b_o , and b_s are scaling parameters for the inner and outer corridors, respectively. In OLS regression, the root mean square error (RMSE) is an unbiased estimator of the error standard deviation across the experimental region. Since the functional regression presented by Faraway [34] uses OLS fitting at each time step t , estimate $\sigma(t)$ as

$$\hat{\sigma}(t) = \text{RMSE}(t). \quad (5.15)$$

with

$$\text{RMSE}(t) = \sqrt{\frac{\sum_{l=1}^N (y_l^s(t) - \hat{y}_l^s(t))^2}{N - p}} \quad (5.16)$$

where $y_l^s(t)$ denotes the l^{th} entry in the N length vector of system responses and p denotes the number of input parameters being estimated. N and p should be constant at all t time steps. While the RMSE provides an unbiased estimation of the variance of the system error, further consideration for the variability of the fit itself is necessary. The variance of the estimated response at a design point x_0 is a function of the true system error variance and its location within the design space. The variance of the fitted response is calculated as [83]

$$V[\hat{y}_{x_0}] = \sigma^2(x_0(X'X)^{-1}x_0). \quad (5.17)$$

The MSE of the fitted system response at each discretized time step is used as an estimate for $\sigma^2(t)$. This estimate is used in Equation 5.17 to calculate the variance of the fitted system response across all m discretized time steps, $V[\hat{y}_{x_0}^s(t)]$, for all assessed design points within the experimental region. Incorporating this source of variability into the corridor widths addresses the additional uncertainty inherent with using an estimated system response rather than the true response curve.

When using the OLS fit as a surrogate mean system response, one may consider not using the global absolute maximum of the system response to define the corridor widths, but rather a linear combination of the estimated system error variance and the variance of the estimated mean response. The inner and outer corridor widths at design point x_0 for time step t are calculated as

$$\begin{aligned}\delta_i(t) &= a_o RMS E(t) + a_s \sqrt{MSE(t)(x'_0(X'X)^{-1}x_0)} \\ \delta_o(t) &= b_o RMS E(t) + b_s \sqrt{MSE(t)(x'_0(X'X)^{-1}x_0)}\end{aligned}\tag{5.18}$$

where a_o , a_s , b_o , and b_s are user defined scaling parameters for the inner and outer corridors, respectively.

Let x_i be a particular point within the experimental region and let X be the design matrix of points for which response data was collected. For each $x_i \notin X$, calculate its estimated response, $\hat{y}_{x_i}^s(t)$ for $t = 1, \dots, m$ and calculate the CORA rating for $y_{x_i}^m(t)$. The CORA ratings at each of the x_i are aggregated using a weighted linear combination where $\sum_i^q w_i = 1$. The extended CORA rating system is summarized in Figure 5.4.

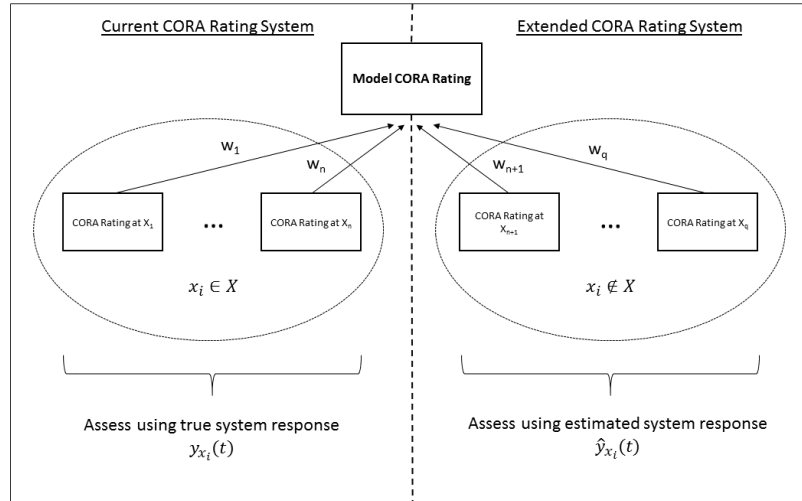


Figure 5.4: Extended CORA Methodology for Model Validation Over Experimental Regions

Lastly, consider the CORA rating for the model. This rating will range between 0-1 with model validity associated with larger scores. The CORA rating system is “objective” in that the score is based on objective measures of discrepancy between two sets of

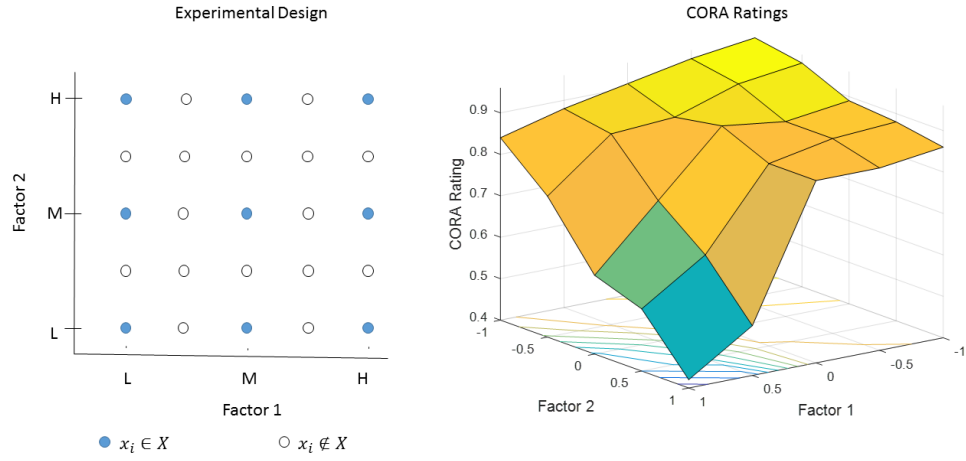


Figure 5.5: Notional Example of Using CORA Ratings to Isolate Poor Model Performance Within Experimental Regions

functional data; however, interpretation of the CORA rating is subjective. The modeler will assess the level of model validity based on the calculated CORA score. Ideally, the modeler could create a library of CORA results for their respective application, which over time, could be used to develop a sliding scale for objective model assessments with respect to that application.

The CORA ratings can be used to compare the adequacy of competing simulation models across the desired experimental region. Similarly, the CORA scores at individual design points can be compared to assess areas of high or low model validity within the experimental region. Figure 5.5 provides a notional depiction of this assessment. A 3^2 design is evaluated at 25 input configurations (nine with true system response and 16 with estimated system response) and the resulting CORA scores indicate poor model performance isolated in one corner of the experimental region.

5.5 Improvement to CORA Size Rating

The CORA size rating in Equation 5.10 approximates the areas of the model and system response curves using the size of the responses at each time step. The ratio of these areas is then used to generate a size score. This approach has a fundamental flaw. Functional responses can exhibit completely different behavior over the interval of evaluation but can result in a similar area calculation. Figure 5.6 illustrates this occurrence. The two responses are perfectly mirrored and yet they produce the same area between the x -axis and the response. While discussed in [43], the metric is retained and compensated by the shape score. Similarly, the shape score, based on the maximum cross-correlation value between the two responses, is insensitive to differences in magnitude between responses. The disagreements due to magnitude are captured in the size score. Figure 5.7 shows two instances where the assessed responses exhibit the same behavior but disagree in magnitude. The model response in Figure 5.7a has been shifted upward by 100 units ($0.12Y_{max}^s$). The corresponding shape score is $C_V = 0.9291$ and is compensated by a size score $C_G = 0.6972$. The model in Figure 5.7b has been shifted upward by 200 units ($0.24Y_{max}^s$). The corresponding shape score is still reasonably high at $C_V = 0.8329$ but is compensated by a low size score of $C_G = 0.4984$.

An improved rating, based on a modification to the current size score, considers both size and shape of the area produced by the functional response which removes the need to have separate size and shape scores. This area score is denoted as C_A . The limitations of C_G result from responses that interchange dominance over the interval of evaluation. In these cases, disagreements in area due to the shape of the response can cancel as they switch dominance. The area score does not have the same limitations as the size rating because it partitions the data into two assessments in which one response is always dominating the other. When one response dominates the other throughout the assessment, the current size rating appropriately accounts for disagreements in shape.

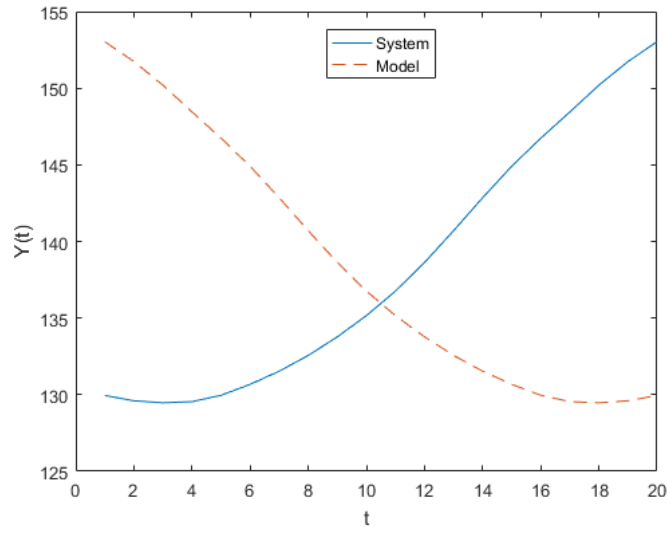
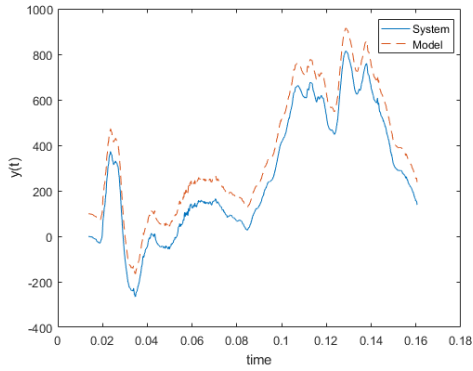
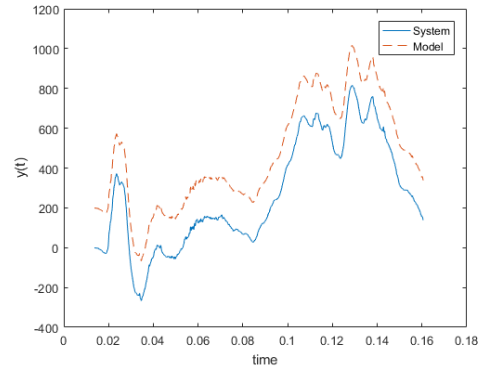


Figure 5.6: Different Model and System Responses with Equal Area

Let $y^s(t)$ and $y_*^m(t)$ be the system response and the time-shifted model response, respectively. Redefine the area of a functional response $y(t)$ as



(a) Model shift by $0.12Y_{max}^s$ ($C_V = 0.9391$)



(b) Model shift by $0.24Y_{max}^s$ ($C_V = 0.8329$)

Figure 5.7: Sensitivity of Shape Rating to Shifts in Magnitude

$$A = \sum_t (y(t) - J)^2 \quad (5.19)$$

where J is calculated as

$$J = \min(0, y^s(t), y_*^m(t)). \quad (5.20)$$

The value J either corresponds to the y-intercept of the time-axis ($y^s(t), y_*^m(t) > 0 \forall t$) or the minimum response value of either the model or system. The CORA manual assumes $J = 0$, which is appropriate for their calculations; however, this assumption does not hold in the area score when response values are negative.

The area rating is a piecewise metric that calculates scores based on the relative values of $y^s(t)$ and $y_*^m(t)$ at each t . Let S_s be the set of all t where $y^s(t) \geq y_*^m(t)$ and let S_m be the the set of all t where $y_*^m(t) > y^s(t)$. Equation 5.19 extends to the following four area calculations

$$\begin{aligned} A_L^m &= \sum_{t \in S_s} (y_*^m(t) - J)^2 \\ A_U^s &= \sum_{t \in S_s} (y^s(t) - J)^2 \\ A_U^m &= \sum_{t \in S_m} (y_*^m(t) - J)^2 \\ A_L^s &= \sum_{t \in S_m} (y^s(t) - J)^2 \end{aligned} \quad (5.21)$$

where the superscript indicates the response being used to calculate the area and the subscript identifies whether the response identified in the superscript was the upper (U) or lower (L) response.

A sub-score is calculated for each condition. These scores are calculated as

$$\begin{aligned} C_A^s &= \left(\frac{A_L^m}{A_U^s} \right)^{k_A} \\ C_A^m &= \left(\frac{A_L^s}{A_U^m} \right)^{k_A} \end{aligned} \quad (5.22)$$

where k_A is the user-specified area interpolation factor. The total area rating is calculated as

$$C_A = w_s C_A^s + w_m C_A^m. \quad (5.23)$$

The weights w_s and w_m are assigned according to the proportion of the interval of evaluation to fall under that respective condition. The weights are calculated as

$$\begin{aligned} w_s &= \frac{|S_s|}{|S_s| + |S_m|} \\ w_m &= \frac{|S_m|}{|S_s| + |S_m|}. \end{aligned} \quad (5.24)$$

where $|\cdot|$ is the cardinality of the set.

To illustrate the differences between the current CORA size (C_G) and shape (C_V) ratings and the proposed area rating (C_A), consider the model and system responses shown in Figure 5.8. The current CORA size and shape ratings are found to be $C_G = 0.8926$ and $C_V = 0.8845$. In this instance, scores are similar. Assuming equal weighting, the two scores average out to a size/shape component equal to 0.8886. The area rating penalizes the model on discrepancies between area sizes as well as discrepancies in the shape of the areas. First, J is calculated using Equation (5.20) and found to be $J = -266$, which corresponds to the minimum value of the model response. The four area components are calculated and the resulting area sub-scores are

$$C_a^s = 0.8017$$

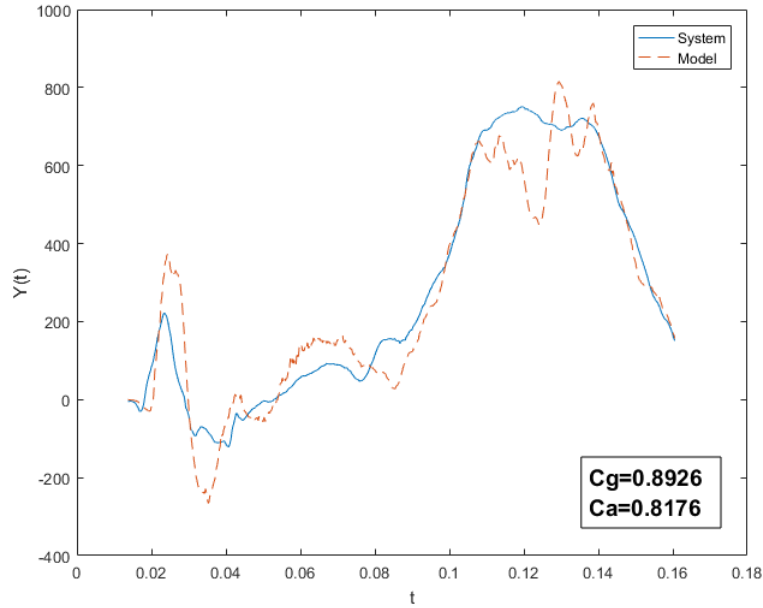


Figure 5.8: Model and System Responses with Calculated Size and Area Scores

$$C_a^m = 0.8363$$

using a $k = 1$ interpolation factor. The weights for the two sub-scores are $w_s = 0.5388$ and $w_m = 0.4612$, respectively. Equation (6.16) is used to calculate the total area score as

$$C_a = 0.5388(0.8017) + 0.4612(0.8363) = \mathbf{0.8176}$$

which is lower than either the current CORA size and shape ratings.

The affect of this change in score on the overall CORA rating depends on the selection of weights used in the analysis. This particular example comes from a packaged data set included with the CORA 3.6 software. In their analysis the weights were $w_1 = 0.5$, $w_2 = 0.5$, $w_P = 0.25$, $w_G = 0.25$, and $w_V = 0.5$. With these weights and using the current size rating, the overall rating for the model is $C = 0.8683$. Replacing the size and shape scores with the area score, and adjusting weights to $w_A = 0.75, w_P = 0.25$, results in an adjusted overall rating of $C_{adj} = 0.8422$. This change may seem insignificant; however,

the use of a larger k interpolation factor would further emphasize the impact of C_A on the overall CORA rating.

5.6 Considerations

Using functional regression to estimate the system response for validation purposes requires that consideration is given to the experimental design chosen when collecting data. Since the estimated response is used as a surrogate in lieu of true experimental data, the OLS response surface's predictive properties are critical to accurate validation assessments. Furthermore, a better understanding of the predictive properties of the chosen design will help to determine the desired weighting for various design points within the experimental region. An extensive discussion on experimental designs and their relationship to the predictive capabilities of a response surface can be found in [86].

The weightings used to construct the CORA rating are subjectively chosen and can have a great effect on the resulting score. Since the CORA rating system has not been widely utilized outside of the biofidelity field of research, no guidelines for the selection of weights exists. It is the responsibility of the modeler to appropriately select these weights and to be ready to justify the weights if their resulting CORA score is brought into question. It is unlikely that a standard for the weights can be established that can encompass all validation efforts, but rather that unique weights for various model applications will need to be established. One approach would be to weight all components equality, which reduces the weighted linear combination of scores to an average score. Prior knowledge of the application may suggest that the cross-correlation score is more critical to model validation than the corridor score, or vice versa. In these instances, the weights should be adjusted to place more emphasis on the preferred measure of disagreement. Weighting used to aggregate individual CORA scores across an experimental region should also be justified. In the literature, these weights are considered equal in value across all design points or responses. In the validation of an entire region, it may be beneficial to set these

weights according to the predictive accuracy of the utilized response surface at each design point. An alternative approach would be to assign weights based on the importance of particular points or sub-regions within the space. For instance, the modeler may be more concerned with the predictive capability of the model near the edges of its operational domain. The modeler would then assign higher weights to edge and corner points within the experimental region to ensure that those points have more emphasis on the global rating.

How the responses are preprocessed is crucial to an accurate assessment of model validity. Both the CORA and functional regression methods require that the data be filtered to reduce noise. An alignment of the data with regard to the discretized points in time is also required. This applies to the alignment of the model and system responses as well as the alignment of the responses across the experimental design. The latter is necessary for an accurate fitting of the OLS response surface at each time step. Lastly, an interval of evaluation must be established. This interval is selected such that the validation assessment is only based on a proportion of the data that is most relevant to the validation effort (i.e. removal of transient phases or portions of data that exhibit initialization bias). A proposed method for selecting the interval of evaluation is in [44] and [113].

5.7 Conclusion

This paper extends the CORA objective rating system to permit model validation of time-discretized functional data at points within an experimental region where no experimental data exists. An OLS based functional regression approach estimates the unknown responses, which is then used as a surrogate responses in lieu of true system data, in order to conduct the CORA procedure. CORA ratings from multiple design points within the experimental region can be aggregated using a weighted linear combination of the individual scores to form an overall CORA score for the model under assessment. Alternatively, the individual scores can be used to assess model adequacy at various points within the design region. This procedure provides a means for model assessment across

an array of desirable input parameters when the response output is time dependent. An area rating, C_A has been introduced, which replaces the need for separate shape and size ratings. Further research will extend the methodology beyond the limiting assumptions of OLS fitting techniques.

VI. Application of Response Surface Based Validation Methodologies Using the Sandia Thermal Challenge Problem

6.1 Introduction

This paper applies three response surface based validation methodologies to model and system data taken from a validation challenge problem presented by Sandia National Laboratories [33]. This challenge required the validation of a thermal model representing the internal temperature in the body of a solid slab based on various thermodynamic properties. The thermal problem consists of multiple challenge criteria. This paper focus on the first two criteria: material characterization and ensemble validation. When applicable, validation results are compared to the results from other validation methods.

6.2 Problem Development

6.2.1 Mathematical Model and Material Characterization.

A mathematical model of the temperature at depth x and time t is provided as

$$T(x, t) = T_i + \frac{qL}{k} \left[\frac{(k/\rho C_p)t}{L^2} + \frac{1}{3} - \frac{x}{L} + \frac{1}{2} \left(\frac{x}{L} \right)^2 - \frac{2}{\pi^2} \sum_{n=1}^6 \frac{1}{n^2} e^{-n^2 \pi^2 \frac{(k/\rho C_p)t}{L^2}} \cos \left(n\pi \frac{x}{L} \right) \right] \quad (6.1)$$

where T_i is the initial temperature condition, k is the thermal conductivity, ρC_p is the volumetric heat capacity, L is the thickness of the slab, and q is the heat flux.

The thermodynamic properties k and ρC_p are stochastic and their distributions need to be established in order to properly utilize the temperature model. Material characterization data is provided as part of the problem and found in Table 6.1. Experimental results for both k and ρC_p at various internal temperatures are provided for three different sample sizes. Characterization is focused on the $N_c = 30$ case because it provides more data from which to estimate the distributional properties of k and ρC_p . The challenge problem assumes that both parameters are independent of temperature. From visual inspection of Figures 6.1a

and 6.1b, it can be seen that this assumption likely does not hold for k , as evident by the linear increase in k as temperature increases, but appears appropriate for ρC_p . Maximum likelihood estimation (MLE) is used to fit both set of data to a normal distribution. For k , only the experimental data from $T = 500^\circ\text{C}$ is used because this condition most closely represents the observed experimental temperatures at $t = 1000\text{s}$. Since ρC_p behaves independently of temperature, all data is used to characterize its distributional properties. Table 6.2 provides the estimates for the mean and standard deviation for both parameters. These values represent the uncertainty in the parameters and are used to propagate that uncertain through to the model response.

6.2.2 Ensemble Validation.

Experimental data are collected over two slab thicknesses (L) and two heat flux magnitudes (q). The response is collected in one hundred second increments up to one thousand seconds. Each parameter configuration consists of four replications. The thermal challenge problem requires validation of the model in Equation 6.1 using the data provided in Table 6.3.

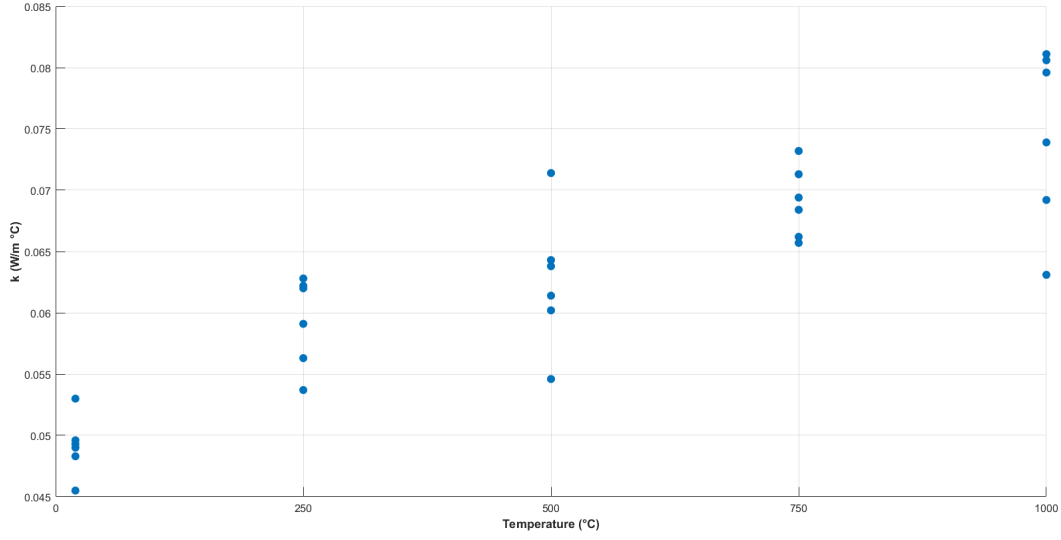
The Regression Tolerance Interval and F-Statistic validation methodologies in Chapters III and IV were developed to validate discrete response data. The experimental data at $t = 1000\text{s}$ is used to demonstrate these methods. This provides 16 data points to characterize the response region depicted in Figure 6.2. To demonstrate the CORA rating extensions in Chapter V, the data across all discretized time steps is treated as a time-discretized functional response.

6.2.3 Generation of Model Data.

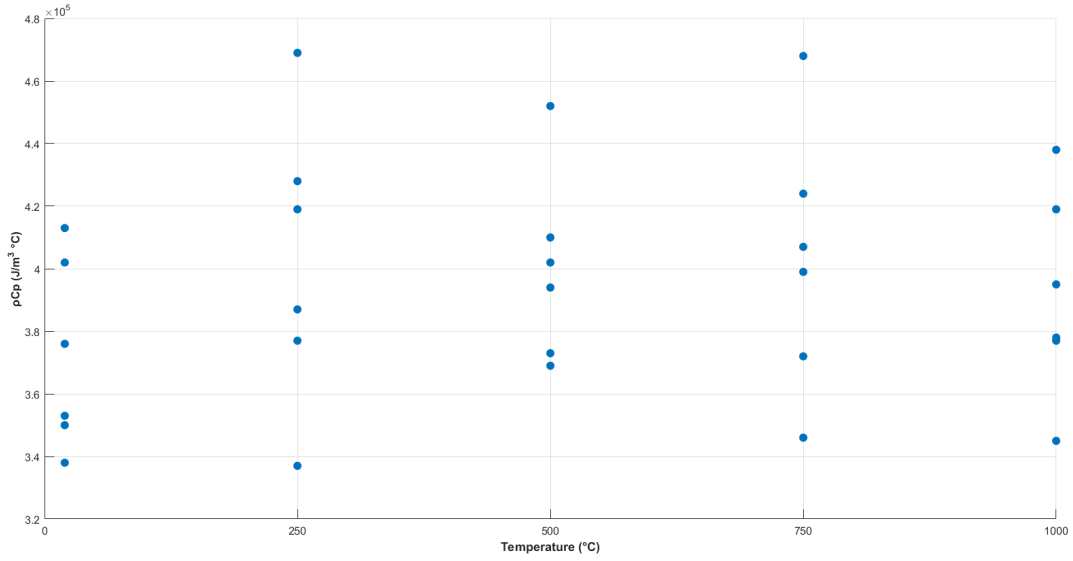
The material characterizations for k and ρC_p are used in conjunction with the mathematical model in Equation 6.1 to generate the desired replications of the model at

Table 6.1: Material Characterization Data

k(20°C)	k(250°C)	k(500°C)	k(750°C)	k(1000°C)
Thermal conductivity, $k(T)$			$(W/m^{\circ}C)$	
$N_c = 30$				
0.0496	0.0628	0.0602	0.0657	0.0631
0.053	0.062	0.0546	0.0713	0.0796
0.0493	0.0537	0.0638	0.0694	0.0692
0.0455	0.0561	0.0614	0.0732	0.0739
0.0483	0.0563	0.0643	0.0684	0.0806
0.049	0.0622	0.0714	0.0662	0.0811
$\rho C_p(20^{\circ}C)$	$\rho C_p(250^{\circ}C)$	$\rho C_p(500^{\circ}C)$	$\rho C_p(750^{\circ}C)$	$\rho C_p(1000^{\circ}C)$
Volumetric heat capacity, $\rho C_p(T)$			$(J/m^3^{\circ}C)$	
$N_c = 30$				
3.76E+05	3.87E+05	4.52E+05	4.68E+05	4.19E+05
3.38E+05	4.69E+05	4.10E+05	4.24E+05	4.38E+05
3.50E+05	4.19E+05	4.02E+05	3.72E+05	3.45E+05
4.13E+05	4.28E+05	3.94E+05	3.46E+05	3.95E+05
4.02E+05	3.37E+05	3.73E+05	4.07E+05	3.78E+05
3.53E+05	3.77E+05	3.69E+05	3.99E+05	3.77E+05



(a) Thermal Conductivity (k)



(b) Volumetric Heat Capacity (ρC_p)

Figure 6.1: Material Characterization of k and ρC_p

each of the four ensemble parameter configurations. The derived distributions for k and ρC_p are used to develop the error component for the model. For the F-Statistic and Regression Tolerance Interval methods, 20 replications of the model response were generated at each

Table 6.2: Mean and Variance for Material Characterization

Parameter	Mean	Variance
$k(W/m^{\circ}C)$	0.063	0.005
$\rho C_p(J/m^3^{\circ}C)$	3.939E+05	35642.063

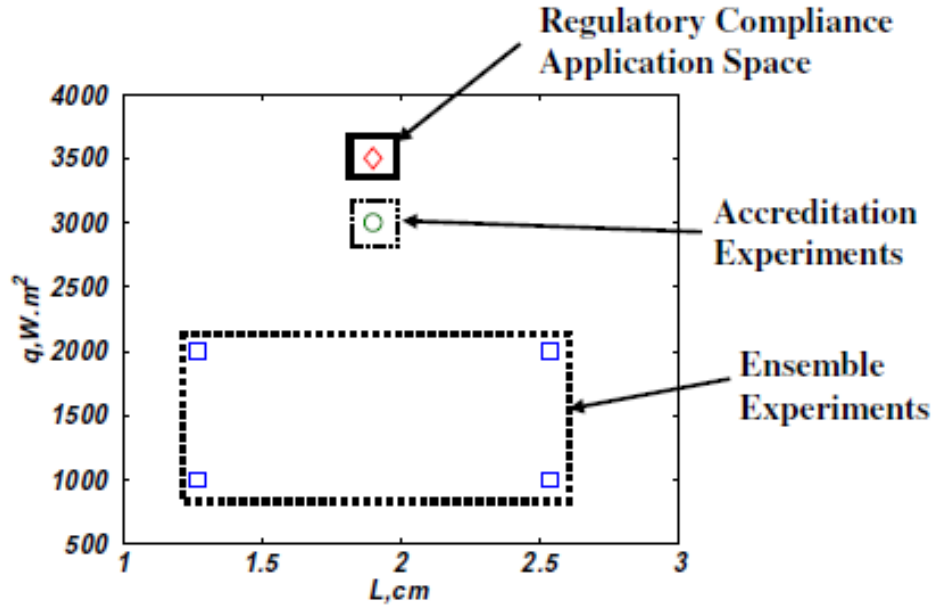


Figure 6.2: Ensemble Validation Region [33]

parameter configuration. A visual comparison between the model and experimental data is shown in Figure 6.3.

6.2.4 Generation and Evaluation of Response Surface.

The experimental data follows a 2^2 factorial design with replications. The first-order empirical model for the 16 experimental observations, using coded independent variables, is

Table 6.3: Experimental Data for Ensemble Validation

	1) $q = 1000$ $L = 0.0127$				2) $q = 1000$ $L = 0.0254$			
	Replication				Replication			
Time(s)	1	2	3	4	1	2	3	4
100	105.5	109	96.3	111.2	99.5	106.6	96.2	101.3
200	139.3	143.9	126	146.9	130.7	140.4	126.1	133.1
300	165.5	170.5	148.7	174.1	154.4	165.9	148.7	157.2
400	188.7	193.5	168.5	197.5	174.3	187.2	167.7	177.2
500	210.6	214.6	186.9	219	191.7	205.8	184.3	194.8
600	231.9	234.8	204.6	239.7	207.3	222.4	199.3	210.6
700	253	254.6	222	259.9	221.7	237.6	213	225
800	273.9	274.2	239.2	279.9	235	251.7	225.7	238.4
900	294.9	293.6	256.4	299.9	247.6	264.9	237.6	251
1000	315.8	313.1	273.5	319.8	259.5	277.4	248.9	262.9
	3) $q = 2000$ $L = 0.0127$				4) $q = 2000$ $L = 0.0254$			
	Replication				Replication			
Time(s)	1	2	3	4	1	2	3	4
100	183.1	177.8	187.2	171.3	173.4	178.9	179.3	188.2
200	247.4	240.2	254.2	231.6	234.2	241.9	242.6	254.6
300	296.3	287.4	306.1	277.6	279.7	289.1	290.1	304.2
400	338.7	327.8	351.9	317.4	317.4	328.4	329.6	345.4
500	378	364.8	395.4	354.2	350.2	362.6	363.9	381.1
600	416	400.2	437.9	389.7	379.5	393.2	394.7	413.1
700	453.5	434.8	480	424.5	406.3	421.1	422.8	442.1
800	490.8	469.1	522.1	459.1	430.9	446.9	448.8	469
900	528.2	503.3	564.1	493.7	454	471.1	473.3	494
1000	565.6	537.6	606.2	528.2	475.6	493.9	496.4	517.6

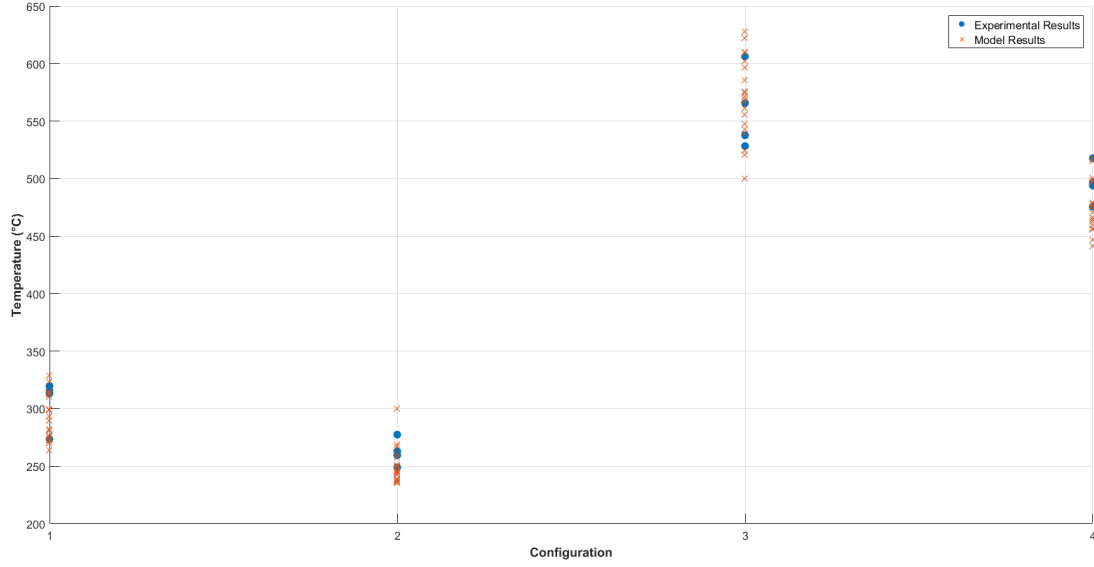


Figure 6.3: Graphical Comparison of Model and Experimental Data

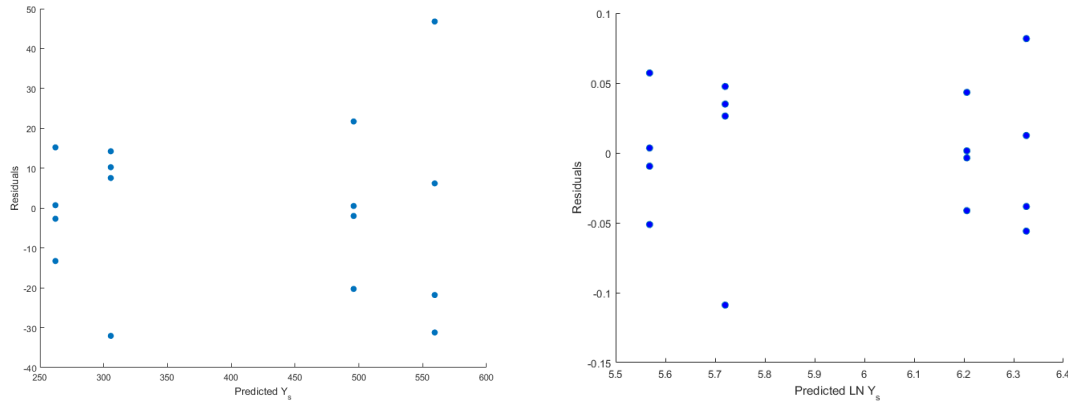
$$\hat{y}_{1000s} = 405.75 + 121.89q - 26.73L - 5.04qL. \quad (6.2)$$

Based on the residual versus predicted plot shown in Figure 6.4a, the experimental response of $T = 606.2^{\circ}\text{C}$ found at $q = 2000$ and $L = 0.0127$ looks like a potential outlier; however, an examination of the standardized residuals supports the inclusion of the data point in the empirical model. The shape of the scatter indicates that a variance stabilizing transformation on the system response data is necessary. A log transform on the response data gives the following empirical model

$$LN\hat{y}_{1000s} = 5.299 + 5.729 \times 10^{-4}q - 14.51L + 0.003qL. \quad (6.3)$$

The residual versus predicted plot for the transformed response, shown in Figure 6.4b indicates improved variance stabilization. The ANOVA in Table 6.4 indicates that the fitted response surface is significant. The response surface of the inverse transformed response values is used to conduct model validation on the model data using the Regression

Tolerance Interval method and the transformed response surface is used to conduct the F-Statistic method.



(a) Residuals vs Predicted (Original Response) (b) Residuals vs Predicted (Transformed)

Figure 6.4: First-Order Linear Model

Table 6.4: ANOVA for First-Order Linear Model

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	1.6194735	0.539825	174.7811
Error	12	0.0370629	0.003089	Prob $\geq$ F
C. Total	15	1.6565364		< .0001

6.3 Point-Wise Model Validation Using Regression Tolerance Intervals

The regression tolerance interval validation methodology creates a bound on a user-specified proportion of the model data at a design point X_p , which establishes model validity. These bounds are formed from regression tolerance intervals around a response

surface of the experimental data. The two sided regression tolerance interval is given by [124]

$$TI_{P,\alpha,h} = \hat{y}_h \pm k_{2,h}s \quad (6.4)$$

where h denotes the design point, s is the root mean square error (RMSE) of the response surface, and $k_{2,h}$ is the 2 sided k -factor for the h^{th} configuration where the k -factor is calculated as

$$k_{2,h} = \sqrt{\frac{f\chi_{1:P}^2(1/n_h^*)}{\chi_{f;\alpha}^2}} \quad (6.5)$$

where α and P are the desired significance and proportion, respectively, f is the error degrees of freedom, $n_h^* = (X_h^T(\mathbf{X}^T\mathbf{X})^{-1}X_h)^{-1}$, and $\chi_{d,\alpha}^2(\delta)$ is the α th quantile of a noncentral chi-squared distribution with d degrees of freedom and δ noncentrality parameter [124].

The choice of desired proportion P is user dependent. Larger P values will focus the validation at the tails of the distributions, whereas smaller P values will force the assessment closer to the mean. If the mean model response is not of great concern, then a larger P is more appropriate. If the mean response is a primary concern for model validity, than focusing on the center of the distribution using a smaller P is appropriate.

Equations 6.4 and 6.5 are used to construct tolerance intervals about the fitted ensemble data using $P = 0.5$ and $P = 0.9$ to compare the validation results for differing proportional requirements. In both cases $\alpha = 0.05$. Table 6.5 provides these tolerance interval bounds.

The 20 model replications are tested against tolerance intervals for both P values. The results can be seen in Table 6.6. When using $P = 0.5$, the model data does not meet the proportional requirements for model validation at configuration 2; however, the Bayes factor associated with the observed proportion is larger than one. Therefore, the model is

Table 6.5: Tolerance Intervals

Configuration			P=0.5		P=0.9	
q	L	$\hat{Y}^s$	TI _L	TI _U	TI _L	TI _U
1000	0.0127	304.953	285.993	332.167	261.206	356.027
1000	0.0254	261.978	245.69	288.792	224.396	305.854
2000	0.0127	558.593	523.863	586.017	478.46	652.147
2000	0.0254	495.652	464.835	522.492	424.548	578.664

valid at configuration 2 if the Bayesian relaxation is used. When $P = 0.9$, the model is valid at all configurations.

Table 6.6: Point-wise Validation Results

Configuration		P=0.5		P=0.9	
q	L	Obs P	Valid	Obs P	Valid
1000	0.0127	0.55	Yes	1	Yes
1000	0.0254	0.45	No*	1	Yes
2000	0.0127	0.6	Yes	1	Yes
2000	0.0254	0.65	Yes	1	Yes

The tolerance interval method is sensitive to changes in the OLS fit. Assume that the modeler decided that rather than transforming the response to accommodate the non-constant variance, they chose to remove the observation corresponding to $T = 606.2^\circ\text{C}$. Evaluating the model replications using tolerance intervals generated from the reduced observation response surface results in the validation assessments found in Table 6.7.

Table 6.7: Validation Results with Reduced Observations

Configuration		P=0.5		P=0.9	
q	L	Obs P	Valid	Obs P	Valid
1000	0.0127	0.55	Yes	1	Yes
1000	0.0254	0.65	Yes	1	Yes
2000	0.0127	0.3	No	0.7	No
2000	0.0254	0.45	No	0.95	Yes

In the new assessment, the model fails to meet the minimum proportional requirements for both P values at configuration 3 and fails the same requirements for $P = 0.5$ at configuration 4. Changes in the validation results for configurations 3 and 4 indicate that changes to the OLS response surface, due to the removal of the observation, have influenced the assessment. Residual analysis of the original OLS empirical model does not justify the removal of the observation and as a result, its erroneous removal has reduced the quality of $\hat{y}_h$ and s , both of which are integral in the tolerance interval calculations. That each configuration is affected differently by the change in response surfaces likely indicates the presence of heteroscedastic errors in the system response. This non-constant variance over the design space can result in the MSE being a poor estimation of the error variability at certain configurations. Further investigation into the effects of non-constant variance with regard to Type I and Type II error rates is left to future research.

The validation assessment of configuration 2, as shown in Table 6.6, highlights a limitation of the tolerance interval methodology. If only large P values are considered, differences at the center of the distribution can be masked by lower model error variance. The model replications have a smaller variance than indicated by the system's tolerance interval and therefore, despite the disagreement at the center of the distribution, the required proportion of replications fall within the tolerance interval bounds set by $P = 0.9$. If

statistical agreement at the center of the distribution, particularly the mean, is necessary for validation, this phenomenon can result in the test exhibiting low statistical power.

6.3.1 Combined Statistical Interval Approach to Model Validation.

To improve the power of the test, consider the strengths and weaknesses of regression confidence and tolerance intervals. Regression confidence intervals, defined in [83] as

$$\hat{Y}(X_h) - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 X_h' (X'X)^{-1} X_h} \leq \mu_{y|X_h} \leq \hat{Y}(X_h) + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 X_h' (X'X)^{-1} X_h} \quad (6.6)$$

only evaluate the mean response at X_h and are insensitive to changes in error variance. Tolerance intervals are sensitive to changes in mean and error variance when the model variance is larger than the system variance; however, they fail to properly assess disagreements in the mean response when the model variance is smaller than the system variance. The use of both statistical intervals, in a combined validation procedure, compensates for the weaknesses in each, while addressing the following adjusted null hypothesis that

$$\begin{aligned} H_o : E[Y^m] &= E[Y^s] \\ V[Y^m] &\leq V[Y^s] \end{aligned} \quad (6.7)$$

Figure 6.5 shows the sensitivity of each approach with regard to additive model bias and error variance as well as the sensitivity for the combined approach. The sensitivity contours clearly show that the combined interval approach maintains sensitivity to increasing error variance but also gains sensitivity to model bias similar to the confidence interval approach. The concentration of valid assessments to the bottom left corner of the sensitivity contour is an indicator of improved power of test.

The new procedure constructs regression confidence intervals to bound the mean of the model replications at X_h . Regression tolerance intervals are used to bound a desired

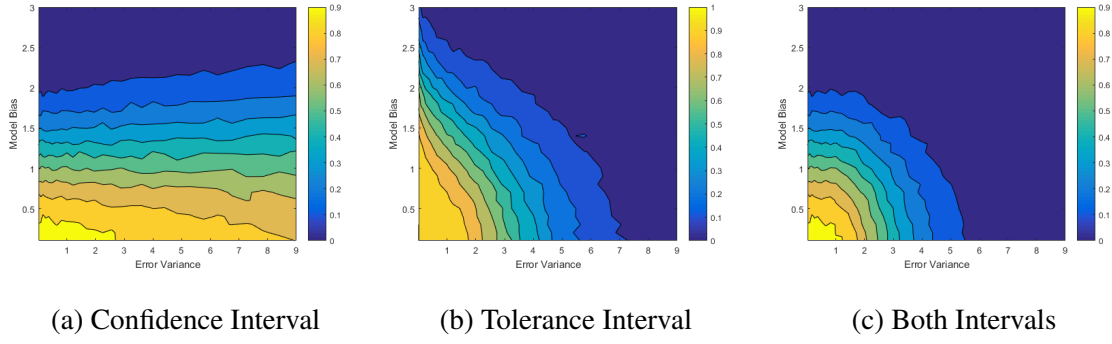


Figure 6.5: Method Sensitivities to Model Bias and Error Variance

proportion of the model response at the tails of the distribution. The validation assessment considers the accuracy (mean) and precision (error variance) of the model response. If the mean of the model replicates falls within the confidence interval and the desired proportion of replicates fall with the tolerance interval, fail to reject the adjusted null hypothesis, and assess the model as valid. If the model falls outside the bounds of either statistical interval, reject the null hypothesis and assess the model as invalid. Bonferroni's inequality is used to aggregate the confidences levels of each interval into an overall confidence level for the validation assessment.

Using the transformed experimental data set, the combined statistical interval approach is applied to the model data at configuration 2. Equation 6.6 is used to construct a 95% confidence interval to bound the mean response of the model replications, and a tolerance interval with $\alpha = 0.05$ and $P = 0.9$ is used to limit disagreement in the tails of the distribution. The resulting confidence intervals is (246.588, 278.329). The mean model response is $\bar{Y}_4^m = 250.402$ which indicates that the difference in means is statistically insignificant at the specified α level. The results of both statistical interval assessments indicate that the model is valid at configuration 2.

Brandyberry [11] addresses the validation of the ensemble data set using a two sample t-test on the experimental and model data to test for a statistical differences in the means. Brandyberry operates under the assumption that the two samples have equal variance. The two sample t-test with equal variance is calculated as

$$t = \frac{\bar{Y}_1^m - \bar{Y}_2^s}{s_p \cdot \sqrt{\frac{1}{n_m} + \frac{1}{n_s}}} \quad (6.8)$$

where

$$s_p = \sqrt{\frac{(n_m - 1)s_{Y_m}^2 + (n_s - 1)s_{Y_s}^2}{n_m + n_s - 2}}. \quad (6.9)$$

To compare methods, model validation at each configuration is conducted using the two sample t-test and the combined statistical interval method with the full 16 observations using $\alpha = 0.05$ and a desired proportion $P = 0.9$. The results of these tests are provided in Table 6.8. Both methods result in a valid model assessment at $\alpha = 0.05$. Unlike the two sample t-test, the combined statistical interval method also gives the modeler insight into the spread of the model replications with regard to the tolerance interval.

Table 6.8: Tolerance and Confidence Interval Results Compared to Two Sample t-Test

Configuration		Tolerance Interval (P=0.9)		Confidence Interval		Model	t-test	t-test	TI/CI
q	L	Obs P	Valid	CI _L	CI _U	Mean	P-value	Result	Result
1000	0.0127	1	Yes	287.038	323.986	295.805	0.4	Valid	Valid
1000	0.0254	1	Yes	246.588	278.329	250.402	0.166	Valid	Valid
2000	0.0127	1	Yes	525.777	593.457	570.016	0.584	Valid	Valid
2000	0.0254	1	Yes	466.534	526.588	476.388	0.105	Valid	Valid

The use of regression based statistical intervals allows for validation assessments at any parameter configuration within the experimental region. Consider the parameter configuration $q = 1500$ and $L = 0.0191$, which represents the center point for the

experimental region defined by the ensemble validation data. The two sample t-test used in [11] is not capable of conducting validation at this design point because no experimental data exists for a comparison of sample means. The regression based statistical intervals utilize the OLS response surface at the desired configuration for an assessment of model validity.

Using a significant level and desired proportion of $\alpha = 0.05$ and $P = 0.9$, respectively, the regression confidence and tolerance intervals at the new design point are $CI = (373.950, 397.290)$ and $TI = (363.507, 408.703)$ with a fitted response value $\hat{y}_{new} = 385.443$. Twenty model replications at the new parameter configuration are generated from Equation 6.1. These replications are assessed for validity using the statistical intervals above. The results in Table 6.9 indicate that the model is valid at this parameter configuration.

Table 6.9: Model Validation at New Parameter Configuration

Configuration		Tolerance Interval		Confidence Interval	
q	L	Obs P	Valid	Mean	Valid
1500	0.0191	1	Yes	380.438	Yes

6.4 F-Statistic for Overall Validity Across the Experimental Region

The F-Statistic validation method establishes an aggregated model validity assessment over the ensemble region shown in Figure 6.2 at a desired confidence level. The validation hypothesis is defined as

$$\begin{aligned}
 H_o : E[Y^s] &= E[Y^m] \\
 V[Y^s] &= V[Y^m].
 \end{aligned}
 \tag{6.10}$$

If the model and experimental data possess statistically equivalent mean and variance, they should produce the same OLS response surface such that

$$\hat{Y}^s = \hat{Y}^m = X\hat{\beta}^s \quad (6.11)$$

and therefore, their cross residuals, defined as

$$e^s = Y^s - \hat{Y}^s \quad (6.12)$$

$$e^m = Y^m - \hat{Y}^s$$

will produce statistically equivalent measures of the MSE.

The MSE for the experimental data is presented in Table 6.4. The MSE for the model replications is calculated using the cross residuals in Equation 6.12 and found to be $MSE_m = 0.0049$. The F-statistic formed from the ratio of these MSEs is calculated as

$$F = \frac{MSE_m}{MSE_s} = \frac{0.0049}{0.0031} = \mathbf{1.594}. \quad (6.13)$$

At $\alpha = 0.05$ and with the model and experimental degrees of freedom equal to $DF_m = 76$ and $DF_s = 12$, respectively, the model is considered valid (failure to reject H_o) over the assessed experimental region if $0.472 \leq F \leq 2.823$. Using these critical values, the model is assessed as valid.

The size of the critical F-values for this particular validation assessment needs discussion. An interpretation of the critical F values tells us that in order to reject our null hypothesis, the MSE_m must be nearly three times the MSE_s . This indicates a low power of the test under small degrees of freedom. Recall that the 2^2 factorial design with four replications only produces 16 experimental observations and only 12 degrees of freedom for the error. If the experimental design resulted in 30 degrees of freedom for the error, the critical F-values would fall between $0.568 \leq F \leq 1.910$, which is a more strict validation

criteria. This illustrates that the F-statistic method has increased power as the number of model and experimental observations increases.

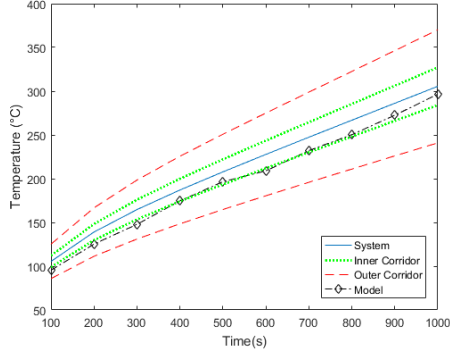
6.5 Functional Regression Extension to the CORA Objective Rating System

While the tolerance interval and F-statistic methods address the validation of discrete model data, the developed extensions to the CORA rating system allow for point-wise and overall validation of time-series response data across an experimental regions. The CORA method uses the averaged system response at each discretized time step over multiple replications as a baseline for the scoring process. Table 6.10 provides the averaged experimental response for the ensemble data at each parameter configuration. For this example, assume that the each response is appropriately preprocessed so that the resulting response at each time step is aligned across all parameter configurations.

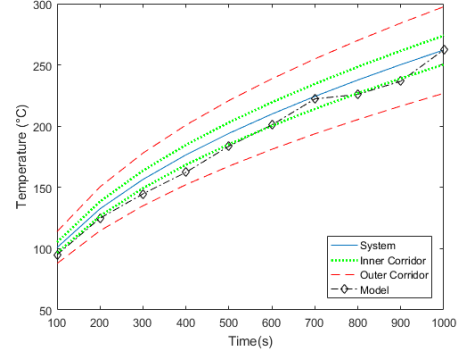
Table 6.10: Mean System Response for Each Parameter Configuration Over Time

	time (seconds)									
Config	100	200	300	400	500	600	700	800	900	1000
1	105.5	139.025	164.7	187.05	207.775	227.75	247.375	266.8	286.2	305.55
2	100.9	132.575	156.55	176.6	194.15	209.9	224.325	237.7	250.275	262.175
3	179.85	243.35	291.85	333.95	373.1	410.95	448.2	485.275	522.325	559.4
4	179.95	243.325	290.775	330.2	364.45	395.125	423.075	448.9	473.1	495.875

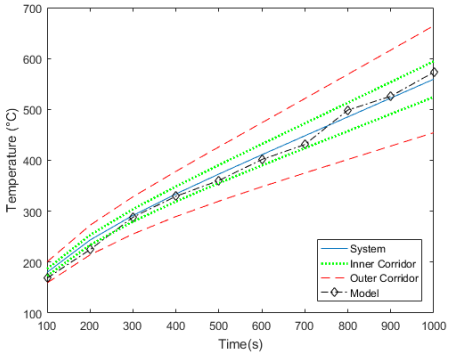
Equation 6.1 is used to generate 10 replications of the response at each discretized time step across all parameter configurations. The averaged response at each time step is used to assess model validity. A non-constant corridor width, based on the sample standard deviation of the system response at each time step, denoted s , is used to calculate corridor rating. The inner and outer corridor width are set to $\delta_i = s$ and $\delta_o = 3s$, respectively. Figure 6.6 shows the mean model and system responses and the inner and outer corridors for each parameter configuration.



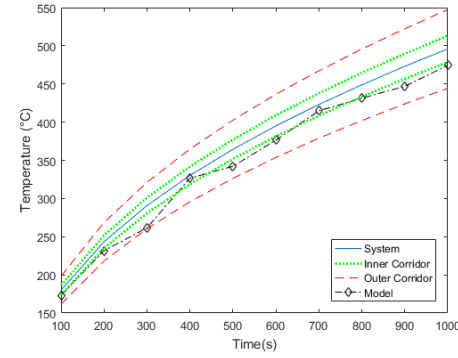
(a) Configuration 1



(b) Configuration 2



(c) Configuration 3



(d) Configuration 4

Figure 6.6: Model and System Responses with Corridors

The equations for calculating the CORA rating can be found in [113]. The global CORA rating in this paper use the area score (C_A) in place of the size (C_G) and shape (C_V) scores. The area score is a piecewise metric that calculates scores based on the relative values of $y^s(t)$ and $y_*^m(t)$ at each t . Let S_s be the set of all t where $y^s(t) \geq y_*^m(t)$ and let S_m be the the set of all t where $y_*^m(t) > y^s(t)$. The resulting areas are calculated as

$$\begin{aligned}
A_L^m &= \sum_{t \in S_s} (y_*^m(t) - J)^2 \\
A_U^s &= \sum_{t \in S_s} (y^s(t) - J)^2 \\
A_U^m &= \sum_{t \in S_m} (y_*^m(t) - J)^2 \\
A_L^s &= \sum_{t \in S_m} (y^s(t) - J)^2
\end{aligned} \tag{6.14}$$

where the superscript indicates the response being used to calculate the area and the subscript identifies whether the response identified in the superscript was the upper (U) or lower (L) response.

A sub-score for the two conditions is calculated as

$$\begin{aligned}
C_A^s &= \left(\frac{A_L^m}{A_U^s} \right)^{k_A} \\
C_A^m &= \left(\frac{A_L^s}{A_U^m} \right)^{k_A}
\end{aligned} \tag{6.15}$$

where k_A is the user-specified area interpolation factor. The total area rating is calculated as

$$C_A = w_s C_A^s + w_m C_A^m. \tag{6.16}$$

The weights w_s and w_m are calculated according to the proportion of m time steps to fall under that respective condition. The weights are calculated as

$$\begin{aligned}
w_s &= \frac{|S_s|}{|S_s| + |S_m|} \\
w_m &= \frac{|S_m|}{|S_s| + |S_m|}.
\end{aligned} \tag{6.17}$$

All interpolation factors (k) are set equal to one to indicate a linearly interpolated decline in the component scores as the quality of the model response decreases. The weightings used to calculate the CORA scores are in Table 6.11.

Table 6.11: Weightings Used to Calculate CORA Ratings for Each Parameter Configuration

	Weightings					
Config	w₁	w₂	w_P	w_A	w_m	w_s
1	0.5	0.5	0.25	0.75	0	1
2	0.5	0.5	0.25	0.75	0.1	0.9
3	0.5	0.5	0.25	0.75	0.3	0.7
4	0.5	0.5	0.25	0.75	0	1

The resulting CORA ratings, shown in Table 6.12, indicate that the model performs reasonably well at all four design points. The CORA rating metric does not assign a valid/invalid assessment to the model; however, it does allow for comparisons at various design points. The CORA ratings indicate that the model had the highest level of agreement with the system response at configuration 3. Conversely, the lowest level of agreement was observed at configuration 4. A overall CORA rating for the experimental region can be calculated using a weighted linear combination of the individual ratings at each design point. Since all design points are equally distances from the center of the experimental region, and no other information is provided to the contrary, the overall CORA rating is calculated using equal weights at each design point. The model's overall CORA rating across the experimental region is

$$C_R = 0.25 \times 0.914 + 0.25 \times 0.905 + 0.25 \times 0.939 + 0.25 \times 0.856 = \mathbf{0.903}. \quad (6.18)$$

While the CORA rating metric cannot prescribe an assessment on validity, a rating of 0.903 is generally favorable on a scale from zero to one.

Table 6.12: CORA Ratings for Thermal Problem

	Component Scores				Total
Config	C1	C2	CP	CA	C
1	0.914	0.914	1	0.885	0.914
2	0.875	0.935	1	0.914	0.905
3	0.920	0.959	1	0.945	0.939
4	0.777	0.935	1	0.913	0.856

The CORA method is sensitive to the user specified inputs. Consider the effects of selecting corridor widths. Table 6.13 shows the CORA rating for each configuration under differing corridor widths. It is clear that the CORA rating is sensitive to the selection of the width sizes. Widths that are too narrow result in low CORA ratings and larger inner corridor widths result in high CORA ratings. Now consider changes to the interpolation factor. The thermal challenge problem was validated with all interpolation factors equal to one. Figure 6.7 shows how increasing the interpolation factors can influence the overall CORA rating. Even just changing the interpolation factor from linear to cubic results in a drop of 0.122 in the overall CORA rating.

Solutions to the thermal challenge problem proposed by Hills and Dowding [51] and McFarland and Mahadevan [78] provide the most appropriate comparisons to the CORA rating metric results above. It should be noted that the model data used by these authors are different than the model data in this paper due to slight differences in material characterization and the stochastic nature of the sampled model response. To compare

Table 6.13: CORA Ratings Using Different Corridor Width Specifications

Corridor Width (s)		Configuration				
Inner	Outer	C1	C2	C3	C4	Overall
1	3	0.914	0.905	0.939	0.856	0.903
1	2	0.871	0.843	0.899	0.789	0.851
0.5	1	0.586	0.575	0.838	0.561	0.640
2	3	0.957	0.968	0.979	0.923	0.957
0.5	3	0.848	0.839	0.919	0.797	0.851
0.25	4	0.852	0.855	0.917	0.823	0.862

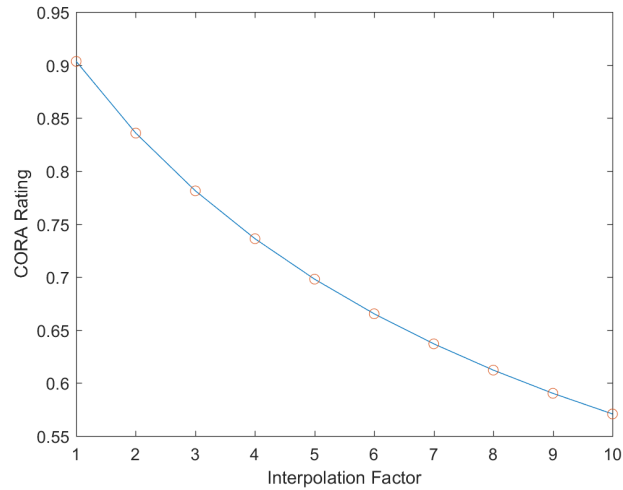


Figure 6.7: Influence of Interpolation Factor on Overall CORA Rating

methods, assume that these differences are inconsequential to the final assessment of model validity.

The opposing solutions to the thermal problem, rather than treating the problem as a single response time-series, consider the model response at each time step to be individual

responses with high correlation. This allows them to assess model validity using standard multivariate statistics. Hills and Dowding use the weighted least squares r^2 metric first developed in [52] and extended in [50]. The statistic is distributed as χ^2 with a degree of freedom equal to the number of measurements considered. The r^2 statistic is calculated as [51]

$$r^2 = (\gamma - T)^T (\text{cov}(\gamma - T))^{-1} (\gamma - T) \quad (6.19)$$

where γ and T are vectors of the system and model data, respectively. The results of their analysis on the ensemble data is in Table 6.14. The results from the r^2 test indicate that the model is valid for configurations 1-3 but invalid at configuration 4.

When comparing these results with the results in Table 6.12, it can be seen that the CORA ratings exhibit similar orders of quality between the configurations. Both methods found the model most agreed at configuration 3 and least agreed (or invalid in the case of the r^2 statistic) at configuration 4. Similarly, both methods indicate similar measures of model/system disagreement between configurations 1 and 3.

Table 6.14: r^2 Values for Ensemble Validation Data [51]

Configuration				
1	2	3	4	r^2_{critical}
18.1	32.8	18	425.4	55.8

McFarland and Mahadevan [78] use a statistic based on a Hotelling T^2 test on variables that have been transformed using Principle Component Analysis (PCA). The statistic is calculated as

$$\frac{f - k + 1}{fk} n \left[A_{(k)}^T (\bar{x} - \mu_0) \right]^T \left[A_{(k)}^T \sum A_{(k)} \right]^{-1} \left[A_{(k)}^T (\bar{x} - \mu_0) \right] \quad (6.20)$$

where n is the number of observations, k is the number of principle components, $f = n - 1$, Σ is the covariance matrix of the observed data, and $A_{(k)}^T$ is the PCA transformation matrix containing the first k eigenvectors of Σ .

The p-values from the test statistic, using $\alpha = 0.05$ are in Table 6.15. The high p-values indicate an inability to reject the null hypothesis, and McFarland and Mahadevan assess the model at each design point as valid. These results agree with the CORA ratings if one were to interpret the high CORA ratings to indicate a valid assessment.

Table 6.15: p-Values from Transformed Hotelling T^2 Test [78]

Configuration			
1	2	3	4
0.56	0.76	0.53	0.76

The CORA rating method produces comparable results to the methods above. The extended CORA method has the ability to calculate CORA ratings at parameter configurations with no corresponding system response data. This is achieved by estimating the system response at each time step using general functional regression techniques.

To demonstrate, estimate the time-series response at a new design point $x_{new} = [1500, 0.0191]$. The parameter coefficient estimates at each time step are calculated as

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (6.21)$$

where X is the $N \times p$ design matrix, Y is a $N \times m$ matrix of observed system response values at each m time step, and $\hat{\beta}$ is a $p \times m$ matrix of the coefficient estimates at each time step. Recall that for this example, a log transformation on Y is required to stabilize the variance. The estimations for the system response for x_{new} at each time step are calculated as

$$\hat{y}_{new}(t) = x_{new}^T \hat{\beta}(t). \quad (6.22)$$

The root mean square error (RMSE) of the fitted surface at each t is used as an unbiased estimate of the system standard deviation at each discretized time step of the system response. The RMSE is used to create the inner and outer corridor widths for the CORA rating, defined as

$$\begin{aligned} \delta_i(t) &= RMSE(t) + \sqrt{MSE(t)(x_0'(X'X)^{-1}x_0)} \\ \delta_o(t) &= 3RMSE(t) + \sqrt{MSE(t)(x_0'(X'X)^{-1}x_0)} \end{aligned} \quad (6.23)$$

where $MSE(t)(x_0'(X'X)^{-1}x_0)$ is an estimation of the variance for the fitted response at x_0 at time t . Figure 6.8 depicts the predicted system response, the corridor widths, and the model response for the new parameter configuration.

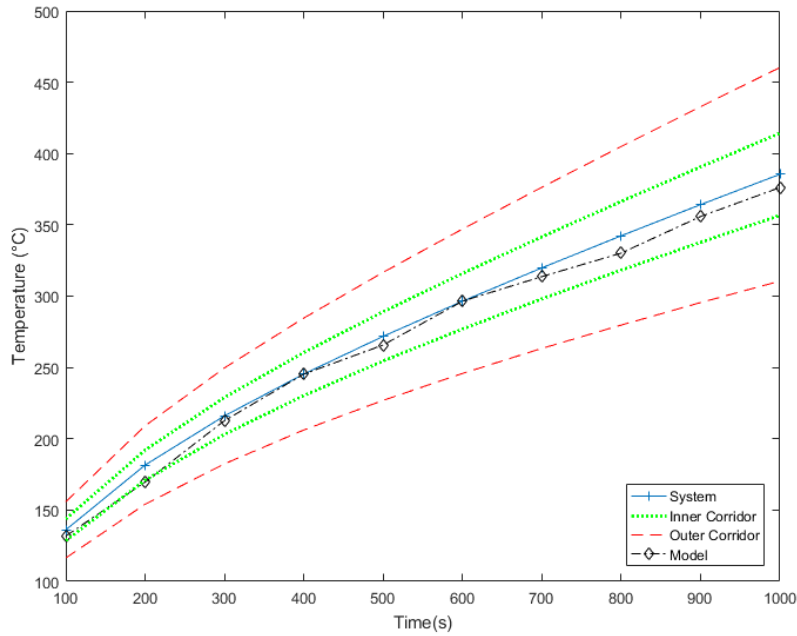


Figure 6.8: Model and Estimated System Response with Corridors at x_{new}

The component CORA scores are calculated using the same weight as the other four design points and the resulting CORA score is $C_{x_{new}} = 0.980$. When adding this CORA rating to the overall rating of the experimental region, adjustments to the weighting are considered. Rather than using a equal weighting scheme, the weight on the new design point is lower to reflect the fact that the CORA rating at that point was based on an estimate of the mean system response rather than an observed mean system response. The new weighting scheme is $w_{new} = 0.1$, $w_1 = w_2 = w_3 = w_4 = 0.225$. The resulting CORA rating for the experimental region with x_{new} included is $C_R = 0.911$.

CORA ratings of $C_{x_{new}} = 0.980$ and $C_R = 0.911$ are strong indications of model validity both at x_{new} and across the experimental region. The CORA rating for x_{new} agrees with the valid model assessment obtained from the combined statistical interval method at $t = 1000s$.

6.6 Conclusion

The thermal challenge problem presented by Sandia National Laboratories was used as a case study to demonstrate the utility of the response surface based validation methods developed and presented in this dissertation. The ensemble validation experimental data was utilized to validate a mathematical model using each of the three developed methods. When applicable, the results of the applied methods were compared to other submitted approaches to the thermal challenge problem found in the literature. In these instances, the developed methods provided comparable results while also being able to validate the model at untested design parameters within the experimental region.

VII. Conclusion

This dissertation addresses the validation of models over experimental regions, using methods from response surface approaches. Validation methods are a necessary consideration in the test and evaluation and the simulation communities. Three methodologies were developed to assess model validity under unique considerations. Chapter III presents a methodology that utilizes regression based tolerance intervals to assess model validity at individual parameter configurations within the experimental region. Chapter IV presents an F-statistic based validation methodology that assesses the entire experimental region simultaneously. Chapter V uses functional regression extensions to the CORA rating system to perform validation assessments over experimental regions when the response data is time-dependent. Chapter VI provides a case study to demonstrate the utility of all three methodologies. This chapter discusses the contributions of each method and proposes future research to address each method's key limitations.

7.1 Contributions

The three methodologies developed in this research provide the analytic communities with response surface-based approaches to model validation that focus on model validity over the entirety of the model's operational or experimental domain. While the evaluation of models over multiple parameter configurations is discussed as part of the engineering design process, this focus is overlooked in the simulation and Operations Research communities. Additionally, due to the implementation of regression based OLS fitting techniques, the methods presented in this dissertation are closely related and highly dependent on experimental design concepts. Therefore, these methods are easy to implement when data has been collected using design of experiments, as is often the case with DoD related test problems.

Chapter III presented a validation methodology that allows for the assessment of models based on a desired proportion of model observations to fall within a defined tolerance interval. Previous statistical interval based approaches either use confidence intervals or prediction intervals. These intervals either focus the validation assessment on the mean response or a single future observation, respectively. In both instances, the variability of the model response is ignored. The use of tolerance intervals is novel in that it assesses model validity by bounding the spread of all future observations given a desired proportion and significance level. In addition, the use of regression based tolerance intervals allows for the validation of model responses at parameter configurations in which no experimental data is available. The method additionally provides a means for relaxing the validation criteria using Bayesian techniques that assess the belief that the observed proportion of model observations to fall within the prescribed statistical bounds is equal to a desired theoretical proportion. Chapter VI combines the use of regression based confidence and tolerance intervals to better assess the mean and to bound the variability of the model response. By combining these statistical intervals, the limitations of each is compensated. This results in a method with higher power than either individual statistical interval approach. The concept of combining statistical intervals to perform model validation is absent in the literature.

Chapter IV presented an F-statistic based validation approach that statistically measures the disagreement between model and system response data using residuals constructed with a common reference, in this case the system response surface. This disagreement is assessed as a ratio test of mean square errors calculated from the cross-residuals. This cross-residual approach is new to the V&V field, and provides a simple yet effective way to assess a model's overall validity across multiple parameter configurations. In addition, the cross-residuals are proven independent, which is also a concept previously undiscussed in the literature.

Chapter V takes the concepts of point-wise and overall model validity over experimental regions and extends them to functional response data, in particular, data that is time-dependent. An established validation metric, known as the CORA objective rating system, is implemented to assess model validity. The CORA method is typically used to assess biofidelity models at a single configuration of input parameters. We extend the use of this method to assessments at multiple input configurations through the use of a weighted linear combination of the calculated rating at each parameter configuration. Furthermore, to conduct CORA assessments at untested points within the experimental region, the CORA rating system is extended using general functional regression techniques to create response surfaces at user defined discretized time steps. This results in OLS based estimations of the functional response at any parameter configuration. These are used to validate the model response in lieu of available experimental data. The CORA rating system has never been utilized in this manner. Lastly, an improved rating is developed that addresses known limitations of the current CORA rating components and therefore reduces the number of calculated scores required and improves the accuracy of the validation metric.

7.2 Future Research

All validation methodologies were developed using OLS response surface fitting. The intent was to develop a baseline approach under the simplifying assumptions of OLS. Most real-world response data fails to conform to these underlying assumptions. The most common violations are heteroscedasticity and correlation in the errors between observations. Studies need to be conducted to determine what effects these violations have on the proposed methodology's validity assessments. Results from these studies will help to prioritize improvements to the methods based on the criticality of these violations. Modifications to the methods that address violations of the OLS must be considered. In the case of heteroscedasticity, the use of WLS or a transformation of the response may be necessary. In the case of correlated errors, a GLS approach may be appropriate.

These two extensions are of particular interest for the tolerance interval method. The k-factor approximation would need to be adjusted accordingly using either WLS or GLS. Such extensions to regression based tolerance intervals are not extensive in the literature particularly with regard to model validation. A look into the use of quantile regression methods may also be beneficial as these methods are designed to handle errors with unequal variance.

While not a strict OLS assumption, non-normal errors are commonly observed in real-world data. The tolerance interval methodology requires normally distributed errors as the k-factor approximation utilized was developed under the assumption of normality. The implementation of nonparametric tolerance intervals needs to be explored. Krishnamoorthy and Mathew [71] discuss nonparametric tolerance intervals based on ordered statistics. Furthermore, the use of bootstrapping to develop the desired tolerance intervals has been discussed in the literature. Conversely, if one can estimate the distribution of the response, there are developed tolerance interval forms for many defined distributions.

Another area of future research is the extension of the methodologies to address multiple responses. The multivariate response problem states that, due to potential dependencies between response variables, it is inadequate to validate each response independently. Multivariate methods must be researched to address the covariance structure between response variables. These considerations should extend to cases in which the response variables are of various forms (i.e. functional, discrete, binary, etc.). Multivariate validation methods exist within the literature. These methods need to be researched to determine applicability to the methodologies presented in this research.

A limitation of the validation methods presented in this dissertation is that their ability to conduct model validation at untested design points is highly dependent on the quality of the response surface. Future research might determine if other fitting techniques can provide a more accurate estimation of the system response. This could include extensions

using non-linear regression methods, or the use of less conventional fitting techniques such as neural networks. Exploration into the utilization of robust regression techniques may result in methods that are less susceptible to response outliers or non-constant variance.

The methodology presented in Chapter V lends itself to a unique set of limitations and subsequent future research. First, the CORA rating system only provides a metric to quantify the level of disagreement between two sets of functional response data. This leaves the assessment of validity up to the subjective interpretation of the analyst. Research is needed to establish an application specific sliding scale for model validity based on the CORA ratings. Similarly, the current literature subjectively selects weights, which can greatly influence the final CORA rating calculation. Research could develop statistically-based objective procedures for assigning component weights. The functional regression technique used to extend the CORA rating system is limited in that it ignores within-function and between-function correlation. Morris [85] provides an extensive survey paper on functional regression techniques. A number of the methodologies described by Morris attempt to incorporate these forms of correlation. Research into these methods should be conducted to determine their applicability to the proposed extension to the CORA rating system. Capturing or estimating the within-function and between-function covariance structure will result in better estimations of the system response and will greatly increase the utility of the proposed method when validating at untested design points.

Bibliography

- [1] Adeli, H. and X. Jiang (2006). Dynamic fuzzy wavelet neural network model for structural system identification. *Journal of Structural Engineering* 132(1), 102–111.
- [2] Anderson, T. W. (1962). On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics* 33, 1148–1159.
- [3] Balci, O. (1994). Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of Operations Research* 53(1), 121–173.
- [4] Balci, O. (1997). Verification validation and accreditation of simulation models. In *Proceedings of the 1997 Winter Simulation Conference*, pp. 135–141. Atlanta, GA: Institute of Electrical and Electronics Engineers Computer Society.
- [5] Balci, O. (1998a). Verification, validation, and accreditation. In *Proceedings of the 1998 Winter Simulation Conference*, pp. 41–48. Washington D.C.: Institute of Electrical and Electronics Engineers Computer Society.
- [6] Balci, O. (1998b). Verification, validation, and testing. *Handbook of simulation* 10, 335–393.
- [7] Balci, O. and R. G. Sargent (1981). A methodology for cost-risk analysis in the statistical validation of simulation models. *Communications of the ACM* 24(4), 190–197.
- [8] Banks, J., J. Carson, and B. Nelson (1996). *Discrete-Event System Simulation* (5 ed.).
- [9] Bendat, J. S. and A. G. Piersol (1980). Engineering applications of correlation and spectral analysis. *New York, Wiley-Interscience, 1980. 315 p. 1.*
- [10] Bliemel, F. (1973). Theil's forecast accuracy coefficient: A clarification. *Journal of Marketing Research* 10(4), 444–446.
- [11] Brandyberry, M. D. (2008). Thermal problem solution using a surrogate model clustering technique. *Computer Methods in Applied Mechanics and Engineering* 197(29), 2390–2407.
- [12] Bray, J. H. and S. E. Maxwell (1985). *Multivariate analysis of variance*. Number 54. Newbury Park: Sage.
- [13] Buranathiti, T., J. Cao, W. Chen, L. Baghdasaryan, and Z. C. Xia (2006). Approaches for model validation: methodology and illustration on a sheet metal flanging process. *Journal of Manufacturing Science and Engineering* 128(2), 588–597.

- [14] Carter, J. A. and J. N. Winn (2009). Parameter estimation from time-series data with correlated errors: a wavelet-based method and its application to transit light curves. *The Astrophysical Journal* 704(1), 51.
- [15] Chen, W., L. Baghdasaryan, T. Buranathiti, and J. Cao (2004). Model validation via uncertainty propagation and data transformations. *AIAA Journal* 42(7), 1406–1415.
- [16] Chen, W., Y. Xiong, K.-L. Tsui, and S. Wang (2008). A design-driven validation approach using bayesian prediction models. *Journal of Mechanical Design* 130(2), 021101.
- [17] Chen, Y., B. Yang, and J. Dong (2006). Time-series prediction using a local linear wavelet neural network. *Neurocomputing* 69(4), 449–465.
- [18] Cheng, Z. and J. A. Pelletiere (2003). Correlation analysis of automobile crash responses based on wavelet decompositions. *Mechanical Systems and Signal Processing* 17(6), 1237–1257.
- [19] Cheng, Z. and J. A. Pelletiere (2005). Validation methods and metrics for biodynamics modeling. Technical report, SAE.
- [20] Cheng, Z., J. A. Pelletiere, and W. D. Pilkey (2006). Energy distribution analysis of impact signal based on wavelet decompositions. Technical report, Air Force Research Laboratory.
- [21] Cheng, Z., J. A. Pelletiere, and N. L. Wright (2006a). Multi-scale validation of automobile impact modeling. In *11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Portsmouth, VA*.
- [22] Cheng, Z., J. A. Pelletiere, and N. L. Wright (2006b). Wavelet-based test-simulation correlation analysis for the validation of biodynamical modeling. In *Conference & Exposition on Structural Dynamics*.
- [23] Chui, C. K. (2014). *An introduction to wavelets*, Volume 1. San Diego: Academic Press.
- [24] Cohen, K. J. and R. M. Cyert (1961). Computer models in dynamic economics. *The Quarterly Journal of Economics* 75(1), 112–127.
- [25] Conway, R. W. (1963). Some tactical problems in digital simulation. *Management Science* 10(1), 47–61.
- [26] Crouse, M. S., R. D. Nowak, and R. G. Baraniuk (1998). Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on Signal Processing* 46(4), 886–902.

- [27] Davis, P. K. (1992). Generalizing concepts and methods of verification, validation, and accreditation (VV&A) for military simulations. No. RAND/R-4249-ACQ. Technical report.
- [28] De Gryze, S., I. Langhans, and M. Vandebroek (2007). Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression. *Chemometrics and Intelligent Laboratory Systems* 87(2), 147–154.
- [29] Der Kiureghian, A. and O. Ditlevsen (2009). Aleatory or epistemic? does it matter? *Structural Safety* 31(2), 105–112.
- [30] DoD Instruction 5000.61 (2003). DoD modeling and simulation (M&S) verification, validation, and accreditation (VV&A).
- [31] Donnelly, B., R. M. Morgan, and R. H. Eppinger (1983). Durability, repeatability and reproducibility of the NHTSA side impact dummy. In *Proceedings of the International Research Council on the Biomechanics of Injury Conference*, Volume 11, pp. 299–310. International Research Council on Biomechanics of Injury.
- [32] Doornik, J. A., H. Hansen, et al. (1994). An omnibus test for univariate and multivariate normality. Technical report, Economics Group, Nuffield College, University of Oxford.
- [33] Dowding, K. J., M. Pilch, and R. G. Hills (2008). Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering* 197(29), 2385–2389.
- [34] Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics* 39(3), 254–261.
- [35] Faraway, J. J. (2000). Modeling reach motions using functional regression analysis. Technical report, SAE Technical Paper.
- [36] Ferson, S. and W. L. Oberkampf (2009). Validation of imprecise probability models. *International Journal of Reliability and Safety* 3(1-3), 3–22.
- [37] Ferson, S., W. L. Oberkampf, and L. Ginzburg (2008). Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering* 197(29), 2408–2430.
- [38] Fishman, G. S. and P. J. Kiviat (1967). The analysis of simulation-generated time series. *Management Science* 13(7), 525–557.
- [39] Fishman, G. S. and P. J. Kiviat (1968). The statistics of discrete-event simulation. *Simulation* 10(4), 185–195.
- [40] Freeman, L. J. Implementing design and analysis of experiments in the U.S. department of defense testing community. https://www.ida.org/idamedia/Corporate/Files/Publications/IDA_Documents/OED/ida-document-ns-d-4684.ashx. Accessed May 17, 2017.

- [41] Gallant, A., T. M. Gerig, and J. Evans (1974). Time series realizations obtained according to an experimental design. *Journal of the American Statistical Association* 69(347), 639–645.
- [42] Garratt, M. (1974). Statistical validation of simulation models. In *Proceedings of the 1974 Summer Computer Simulation Conference, Houston, TX*, Volume 2, pp. 915–926.
- [43] Gehre, C., H. Gades, and P. Wernicke (2009). Objective rating of signals using test and simulation responses. In *21st International Technical Conference on the Enhanced Safety of Vehicles Conference (ESV), Stuttgart, Germany*, pp. 15–18.
- [44] Gehre, C. and S. Stahlschmidt (2011). Assessment of dummy models by using objective rating methods. In *22nd International Technical Conference on the Enhanced Safety of Vehicles, Washington, DC*.
- [45] Ghanem, R. G., A. Doostan, and J. Red-Horse (2008). A probabilistic construction of model validation. *Computer Methods in Applied Mechanics and Engineering* 197(29), 2585–2595.
- [46] Gilmore, J. M. (2009). Test and evaluation initiatives.
- [47] Godley, S. T., T. J. Triggs, and B. N. Fildes (2002). Driving simulator validation for speed research. *Accident Analysis & Prevention* 34(5), 589–600.
- [48] Haase, R. F. and M. V. Ellis (1987). Multivariate analysis of variance. *Journal of Counseling Psychology* 34(4), 404–413.
- [49] Hahn, G. J. and W. Q. Meeker (2011). *Statistical intervals: a guide for practitioners*. John Wiley & Sons, Hoboken.
- [50] Hills, R. G. (2006). Model validation: model parameter and measurement uncertainty. *Journal of Heat Transfer* 128(4), 339–351.
- [51] Hills, R. G. and K. J. Dowding (2008). Multivariate approach to the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering* 197(29), 2442–2456.
- [52] Hills, R. G. and T. G. Trucano (1999). Statistical validation of engineering and scientific models: Background. *Sandia National Laboratories, Albuquerque, NM, Report No. SAND99-1256*.
- [53] Howe, W. (1969). Two-sided tolerance limits for normal populations some improvements. *Journal of the American Statistical Association* 64(326), 610–620.
- [54] Jenkins, G. M. (1961). General considerations in the analysis of spectra. *Technometrics* 3(2), 133–166.

- [55] Jensen, W. A. (2009). Approximations of tolerance intervals for normally distributed data. *Quality and Reliability Engineering International* 25(5), 571–580.
- [56] Jiang, X. and S. Mahadevan (2008a). Bayesian validation assessment of multivariate computational models. *Journal of Applied Statistics* 35(1), 49–65.
- [57] Jiang, X. and S. Mahadevan (2008b). Bayesian wavelet method for multivariate model assessment of dynamic systems. *Journal of Sound and Vibration* 312(4), 694–712.
- [58] Jiang, X. and S. Mahadevan (2011). Wavelet spectrum analysis approach to model validation of dynamic systems. *Mechanical Systems and Signal Processing* 25(2), 575–590.
- [59] Jiang, X., R.-J. Yang, S. Barbat, and P. Weerappuli (2009). Bayesian probabilistic PCA approach for model validation of dynamic systems. *SAE International Journal of Materials and Manufacturing* 2(2009-01-1404), 555–563.
- [60] John, S. (1968). A central tolerance region for the multivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 599–601.
- [61] Johnson, M. E., L. M. Moore, and D. Ylvisaker (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* 26(2), 131–148.
- [62] Kleijnen, J. and W. van Groenendall (1992). *A statistical perspective*. Chichester: Wiley and Sons.
- [63] Kleijnen, J. P. (1992). Regression metamodels for simulation with common random numbers: comparison of validation tests and confidence intervals. *Management Science* 38(8), 1164–1185.
- [64] Kleijnen, J. P. (1995a). Sensitivity analysis and optimization of system dynamics models: regression analysis and statistical design of experiments. *System Dynamics Review* 11(4), 275–288.
- [65] Kleijnen, J. P. (1995b). Verification and validation of simulation models. *European Journal of Operational Research* 82(1), 145–162.
- [66] Kleijnen, J. P. and D. Deflandre (2006). Validation of regression metamodels in simulation: Bootstrap approach. *European Journal of Operational Research* 170(1), 120–131.
- [67] Kleijnen, J. P., A. J. Feelders, and R. C. Cheng (1998). Bootstrapping and validation of metamodels in simulation. In *Proceeding of the 1998 Winter Simulation Conference*, Volume 1, pp. 701–706. Washington D.C.: Institute of Electrical and Electronics Engineers Computer Society.
- [68] Kleijnen, J. P. and R. G. Sargent (2000). A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research* 120(1), 14–29.

- [69] Kline, P. (2014). *An easy guide to factor analysis*. New York: Routledge.
- [70] Knepell, P. L. and D. C. Arango (1993). *Simulation validation: a confidence assessment methodology*, Volume 15. Chichester: John Wiley & Sons.
- [71] Krishnamoorthy, K. and T. Mathew (2009). *Statistical tolerance regions: theory, applications, and computation*. John Wiley & Sons, Hoboken.
- [72] Krishnamoorthy, K. and S. Mondal (2006). Improved tolerance factors for multivariate normal distributions. *Communications in Statistics-Simulation and Computation* 35(2), 461–478.
- [73] Lamb, D., M. Castanier, H. Pan, M. Kokkolaras, and G. Hulbert (2012). Model validation for simulations of vehicle systems. Technical report, DTIC Document.
- [74] Law, A. M. (2007). *Simulation modeling and analysis* (4 ed.). New York: McGraw-Hill.
- [75] Law, A. M. (2008). How to build valid and credible simulation models. In *Proceedings of the 2008 Winter Simulation Conference*, pp. 39–47. Miami, FL: Institute of Electrical and Electronics Engineers Computer Society.
- [76] Liu, Y., W. Chen, P. Arendt, and H.-Z. Huang (2011). Toward a better understanding of model validation metrics. *Journal of Mechanical Design* 133(7), 071005.
- [77] Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.
- [78] McFarland, J. and S. Mahadevan (2008). Multivariate significance testing and model calibration under uncertainty. *Computer Methods in Applied Mechanics and Engineering* 197(29), 2467–2479.
- [79] McKay, M. D., R. J. Beckman, and W. J. Conover (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21(2), 239–245.
- [80] Miham, G. A. (1972). The modeling process. *IEEE Transactions on Systems, Man and Cybernetics* (5), 621–629.
- [81] Mitchell, P. (1997). Misuse of regression for empirical validation of models. *Agricultural Systems* 54(3), 313–326.
- [82] Mitchell, P. and J. Sheehy (1997). Comparison of predictions and observations to assess model performance: a method of empirical validation. In *Applications of Systems Approaches at the Field Level*, pp. 437–451. Springer.

- [83] Montgomery, D. C. (2009). *Design and analysis of experiments* (7 ed.). Hoboken: John Wiley & Sons.
- [84] Montgomery, D. C., E. A. Peck, and G. G. Vining (2015). *Introduction to linear regression analysis*. Hoboken: John Wiley & Sons.
- [85] Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application* 2, 321–359.
- [86] Myers, R. H., D. C. Montgomery, and C. M. Anderson-Cook (2009). *Response surface methodology: process and product optimization using designed experiments* (3 ed.). Hoboken: John Wiley & Sons.
- [87] Naylor, T. H. and J. M. Finger (1967). Verification of computer simulation models. *Management Science* 14(2), B92–B101.
- [88] Oberkampf, W. L. and M. F. Barone (2006). Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics* 217(1), 5–36.
- [89] Oberkampf, W. L. and T. G. Trucano (2000). Validation methodology in computational fluid dynamics. *AIAA Journal* 2549, 19–22.
- [90] Oberkampf, W. L. and T. G. Trucano (2002). Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences* 38(3), 209–272.
- [91] Ogden, T. (1996). *Essential wavelets for statistical applications and data analysis*. Boston: Birkhauser.
- [92] O’Keefe, R. M., O. Balci, and E. P. Smith (1986). *Validation of expert system performance*. Department of Computer Science, Virginia Polytechnic Institute and State University.
- [93] Patnaik, P. (1949). The non-central χ^2 and F-distribution and their applications. *Biometrika* 36(1/2), 202–232.
- [94] Rebba, R., S. Huang, Y. Liu, and S. Mahadevan (2005). Statistical validation of simulation models. *International Journal of Materials and Product Technology* 25(1-3), 164–181.
- [95] Rebba, R. and S. Mahadevan (2006a). Model predictive capability assessment under uncertainty. *AIAA Journal* 44(10), 2376–2384.
- [96] Rebba, R. and S. Mahadevan (2006b). Validation of models with multivariate output. *Reliability Engineering & System Safety* 91(8), 861–871.
- [97] Rebba, R. and S. Mahadevan (2008). Computational methods for model reliability assessment. *Reliability Engineering & System Safety* 93(8), 1197–1207.

- [98] Rebba, R., S. Mahadevan, and S. Huang (2006). Validation and error estimation of computational models. *Reliability Engineering & System Safety* 91(10), 1390–1397.
- [99] Reckhow, K. H., J. T. Clements, and R. C. Dodd (1990). Statistical evaluation of mechanistic water-quality models. *Journal of Environmental Engineering* 116(2), 250–268.
- [100] Rowland, J. R. and W. M. Holmes (1978). Simulation validation with sparse random data. *Computers & Electrical Engineering* 5(1), 37–49.
- [101] Rucker, A. (2014). Improving statistical rigor in defense test and evaluation: use of tolerance intervals in designed experiments. Technical report, DTIC Document.
- [102] Russell, D. M. (1997). Error measures for comparing transient data: Part i, development of a comprehensive error measure. In *Proceedings of the 68th Shock and Vibration Symposium, Hunt Valley, MD*, pp. 175–184.
- [103] Rutherford, A. (2011). *Introducing ANOVA and ANCOVA: a GLM approach*. Hoboken: John Wiley & Sons.
- [104] Sargent, R. (2015). An interval statistical procedure for use in validation of simulation models. *Journal of Simulation* 9(3), 232–237.
- [105] Sargent, R. G. (1981). An assessment procedure and a set of criteria for use in the evaluation of computerized models and computer-based modeling tools. Technical report, DTIC Document.
- [106] Sargent, R. G. (1984). A tutorial on verification and validation of simulation models. In *Proceedings of the 1984 Winter Simulation Conference*, pp. 114–121. Dallas, TX.: Institute of Electrical and Electronics Engineers Computer Society.
- [107] Sargent, R. G. (1996). Verifying and validating simulation models. In *Proceedings of the 1996 Winter Simulation Conference*, pp. 55–64. Coronado, CA: Institute of Electrical and Electronics Engineers Computer Society.
- [108] Sargent, R. G. (2010). Verification and validation of simulation models. In *Proceedings of the 2010 Winter Simulation Conference*, pp. 166–183. Baltimore, MD.: Institute of Electrical and Electronics Engineers Computer Society.
- [109] Sargent, R. G., D. M. Goldsman, and T. Yaacoub (2015). Use of the interval statistical procedure for simulation model validation. In *Proceedings of the 2015 Winter Simulation Conference*. Huntington Beach, CA: Institute of Electrical and Electronics Engineers Computer Society.
- [110] Sarin, H., M. Kokkolaras, G. Hulbert, P. Papalambros, S. Barbat, and R.-J. Yang (2010). Comparing time histories for validation of simulation models: error measures and metrics. *Journal of Dynamic Systems, Measurement, and Control* 132(6), 061401–1–061401–10.

- [111] Shannon, R. E. (1975). *Systems simulation: the art and science*. Englewood Cliffs: Prentice-Hall.
- [112] Sprague, M. and T. Geers (2004). A spectral-element method for modeling cavitation in transient fluid-structure interaction. *International Journal for Numerical Methods in Engineering* 60(15), 2467–2499.
- [113] Thunert, C. (2012). Cora release 3.6 users manual. *Partnership for Dummy Technology and Biomechanics*.
- [114] Tsatsanis, M. K. and G. B. Giannakis (1993). Time-varying system identification and model validation using wavelets. *IEEE Transactions on Signal Processing* 41(12), 3512–3523.
- [115] Untaroiu, C., J. Shin, and Y.-C. Lu (2013). Assessment of a dummy model in crash simulations using rating methods. *International Journal of Automotive Technology* 14(3), 395–405.
- [116] Van Horn, R. L. (1971). Validation of simulation results. *Management Science* 17(5), 247–258.
- [117] Wackerly, D. D., W. Mendenhall III, and R. L. Scheaffer (2008). *Mathematical Statistics with Applications* (7 ed.).
- [118] Wald, A. (1943). An extension of wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics* 14(1), 45–55.
- [119] Wallis, W. A. (1951). Tolerance intervals for linear regression. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 43–51. University of California Press. Berkeley.
- [120] Wang, S., W. Chen, and K.-L. Tsui (2009). Bayesian validation of computer models. *Technometrics* 51(4), 439–451.
- [121] Weissberg, A. and G. H. Beatty (1960). Tables of tolerance-limit factors for normal distributions. *Technometrics* 2(4), 483–500.
- [122] Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics* 12(1), 91–96.
- [123] Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *The Annals of Mathematical Statistics* 13(4), 400–409.
- [124] Young, D. S. (2013). Regression tolerance intervals. *Communications in Statistics-Simulation and Computation* 42(9), 2040–2055.
- [125] Young, I. T. (1977). Proof without prejudice: use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. *Journal of Histochemistry & Cytochemistry* 25(7), 935–941.

- [126] Youngblood, S. M., D. K. Pace, P. L. Eirich, D. M. Gregg, and J. E. Coolahan (2000). Simulation verification, validation, and accreditation. *Johns Hopkins APL technical digest* 21(3), 359–367.
- [127] Zhan, Z., Y. Fu, and R.-J. Yang (2011). Enhanced error assessment of response time histories (EEARTH) metric and calibration process. Technical report, SAE Technical Paper.
- [128] Zhan, Z., Y. Fu, R.-J. Yang, and Y. Peng (2011). An enhanced bayesian based model validation method for dynamic systems. *Journal of Mechanical Design* 133(4), 041005.
- [129] Zorn, M. E., R. D. Gibbons, and W. C. Sonzogni (1997). Weighted least-squares approach to calculating limits of detection and quantification by modeling variability as a function of concentration. *Analytical Chemistry* 69(15), 3069–3075.

REPORT DOCUMENTATION PAGE					<i>Form Approved</i> <i>OMB No. 0704-0188</i>	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.						
1. REPORT DATE (DD-MM-YYYY) 14-09-2017		2. REPORT TYPE Doctoral Dissertation			3. DATES COVERED (From — To) Oct 2014-Sep 2017	
4. TITLE AND SUBTITLE Validation of Discrete and Functional Simulation Responses Over Experimental Regions Using Response Surfaces				5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER		
6. AUTHOR(S) Storm, Scott M., Major, USAF				5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way WPAFB, OH 45433-7765					8. PERFORMING ORGANIZATION REPORT NUMBER AFIT-ENS-DS-17-S-044	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of the Secretary of Defense ATTN: Catherine Warner 1700 Defense Pentagon Washington D.C., 20301 (703) 697-7247 Catherine.Warner@osd.mil OSD DOT&E					10. SPONSOR/MONITOR'S ACRONYM(S) OSD DOT&E	
12. DISTRIBUTION / AVAILABILITY STATEMENT Distribution Statement A: Approved for Public Release; Distribution Unlimited					11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
13. SUPPLEMENTARY NOTES This work is declared a work of the U.S. Government and is not subject to copyright protection in the United States.						
14. ABSTRACT The Director of Test and Evaluation has placed great emphasis on the use of experimental design to properly collect and analyze data from military assets. This emphasis on statistical based data collection plays a large role in the validation of simulation models. It is prudent to validate simulation models across these experimental regions to ensure accuracy at all desirable configurations, to include points within the region where no experimental data has been collected. This research proposes the use of response surfaces to capture the behavior of the response across the experimental region. The first paper develops a methodology for point-wise validation of model response output at unique points within an experimental region using regression based tolerance intervals. The second paper develops an F-statistic for validation of the simulation model over the entire region based on a ratio of mean square errors constructed from the model and system data with regards to a common response surface. The third paper proposes an extension to the Correlation Analysis (CORA) objective rating system that uses functional regression techniques to allow for assessments of functional response data when no system data is available. The fourth paper applies each methodology to the Sandia thermal challenge problem.						
15. SUBJECT TERMS simulation, validation, response surfaces, functional response, experimental region						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT		18. NUMBER OF PAGES	
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	UU		217	
					19a. NAME OF RESPONSIBLE PERSON Dr. Raymond R. Hill, AFIT/ENS	
					19b. TELEPHONE NUMBER (include area code) (937) 255-3636 raymond.hill@afit.edu	