WAVELET-BASED SIMULATION MODEL VALIDATION OF FUNCTIONAL DATA

DISSERTATION

Andrew D. Atkinson, Captain, USAF

AFIT-ENS-DS-17-S-034

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-DS-17-S-034

WAVELET-BASED SIMULATION MODEL VALIDATION OF FUNCTIONAL DATA

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

Andrew D. Atkinson, BA, MS

Captain, USAF

September 2017

AFIT-ENS-DS-17-S-034

WAVELET-BASED SIMULATION MODEL VALIDATION OF FUNCTIONAL DATA

Andrew D. Atkinson, BA, MS

Captain, USAF

Committee Membership:

Dr. Raymond R. Hill Jr.
Chair

Dr. Joseph J. Pignatiello Jr.
Member

Dr. G. Geoff Vining
Member

Dr. Edward D. White
Member

Dr. Eric Chicken
Member

ADEDEJI B. BADIRU, PhD
Dean, Graduate School of Engineering
and Management

AFIT-ENS-DS-17-S-034

Abstract

As computer hardware technology continues to advance, so does the scientific community's capability to develop high resolution computer models able to simulate complex systems and processes. This advancement has led to many challenges associated with verification and validation (V&V). These challenges include adapting methods to high-dimensional functional data, maintaining the necessary objectivity, and accounting for noisy data. Department of Defense (DoD) simulation models require validation techniques that are able to overcome these challenges before the models can be relied upon. Model validation substantiates that the model chosen sufficiently represents the system and that it produces results consistent with real-world data within the range of model applicability.

In this research, new statistical techniques will be proposed that improve upon existing simulation validation techniques. These techniques incorporate the use of wavelets to decompose the time-series data into the time-frequency spectrum allowing for objective and comprehensive assessment of the model. In addition, these techniques offer an improved method of analysis for noisy, high-dimensional data. These techniques are applied to assess the validity of simulation models, which will help ensure the accurate representation of the system they are meant to simulate.

With love to my family

Acknowledgments

I would first like to thank Dr. Ray Hill for all his support and guidance over the last three years. I am grateful to you for being not only my professor and research advisor, but also a professional and personal mentor. I would also like to express tremendous appreciation towards my entire research committee of Dr. Joseph Pignatiello, Dr. Geoff Vining, Dr. Edward White, and Dr. Eric Chicken. Finally, thank you to all my family and friends for their support over the years.

Andrew D. Atkinson

Table of Contents

Page

Page

List of Figures

List of Tables

WAVELET-BASED SIMULATION MODEL VALIDATION OF FUNCTIONAL DATA

I. Introduction

This dissertation considers problems associated with the validation of simulation models that produce functional data. Models representing systems that generate functional or time-series data present unique validation challenges. Given these challenges, this research provides new validation techniques that use wavelets for the data analysis of simulation models and represent objective assessments of model validity.

High-resolution computer models can effectively simulate complex systems and processes. Typically, a simulation can evaluate a solution more quickly, and at a fraction of the cost, compared to obtaining data from the real system. However, before relying upon the results of a computer model, verification and validation (V&V) is required to ensure model accuracy. Verification assesses whether the model's computer code is correct. The focus of this research, validation, ensures "that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of a model" [72].

The field of simulation V&V has been the focus of study and development since the advent of computers. Researchers have developed a variety of validation techniques that range from the informal comparison of system and model data to more rigorous, statistically-based methods. Today, many experts agree that emphasis should be placed on the statistical and analytical techniques that go beyond the subjective, visual comparisons of model and system results. However, many established statistical validation techniques are not well-suited for the validation of models that generate functional or time-series data.

Further, the transient phase of time-series data, which is typically characterized as non-stationary, provides another set of challenges, since standard validation techniques are not well-suited to address such non-stationary elements. However, there are situations where models must accurately represent transient behavior, thus requiring the development of new validation techniques.

This dissertation outlines the research contributions to address the unique challenges associated with the validation of simulation models that generate time-series data and offers promising solutions to this field. Chapter 2 reviews the literature on model validation, wavelets, and recently developed validation techniques. Chapter 3 describes a new model validation technique, designed to examine transient data. The technique uses wavelets to de-noise the data and then employs a model validation metric. Chapter 4 introduces a wavelet analysis of variance (WANOVA) approach to model validation that performs statistical inference in the time-frequency domain. This method reduces subjectivity in the model assessment by statistically accepting or rejecting a model as valid. Chapter 5 builds upon the WANOVA model validation approach by incorporating the technique as part of a bisection algorithm that locates the region(s) of discrepancy given an invalid model. Chapter 6 outlines the method of identifying individual wavelet coefficients that correspond to the areas of discrepancy between the system and model data. Chapters 3, 4, 5, and 6 are each stand-alone journal papers. Overall, this dissertation identifies an important area of study for the field of V&V and introduces a detailed and thorough plan to contribute to this discipline.

Specifically, Chapter 2 reviews the applicable literature for this research. This includes an introduction to the history of simulation and early validation efforts by seminal authors such as Balci [8] and Sargent [70]. These authors describe model validation approaches, principles, and techniques, which together form the basis of modern validation efforts. Special focus is given to the set of statistical validation techniques. Chapter 2 also surveys

recent work and developments in model validation, focusing on techniques for analyzing and validating functional data. One set of techniques includes model validation metrics, which typically quantify the discrepancy between system and model output to express the level of model validity. This chapter also describes wavelet analysis, which is a tool for analyzing time-series data by transforming data from the time domain to a time-frequency domain, similar to a Fourier transform. Wavelets offer opportunities to de-noise data sets and to conduct analysis outside the time domain. Lastly, the chapter introduces several wavelet-based analysis and validation techniques.

Chapter 3 presents a new wavelet-based model validation technique, designed to examine transient data. Transient pulses may be characterized by a large spike in magnitude followed by a sharp decrease in a short span of time. A specific concern associated with validating a model's transient phase is that the experimental system data are often contaminated with noise, due to the short duration and sharp variations in the data. Chapter 3 describes a validation approach that uses wavelet thresholding as an effective method for de-noising the system and model data signals to properly validate the transient phase of a model. The technique utilizes wavelet thresholded signals to calculate a validation metric that incorporates shape, phase, and magnitude error. The chapter then compares this technique to an approach that uses wavelet decompositions to de-noise the data signals. The chapter concludes by illustrating the advantages of the wavelet thresholding approach using a simulation study and empirical data from an automobile crash study.

Chapter 4 outlines a WANOVA approach to model validation. Many validation techniques described in the literature require some amount of subjective analysis in order to assess validity. This is particularly true with dynamic simulation output. To reduce or eliminate this subjectivity, a validation process that uses WANOVA is an effective method to statistically accept or reject a model as valid. This WANOVA validation approach performs statistical inference in the time-frequency domain to take advantage of wavelet

sparsity and decorrelation. This process uses a test statistic based on thresholded wavelet coefficients to test the null hypothesis that the set of system data and model data are statistically equivalent. The validation technique is illustrated using a simulation study and empirical data from an automobile crash study.

Chapter 5 introduces the WANOVA Bisection method for identifying simulation model bias. Current validation methods are able to assess simulation results, but when evaluating models that generate functional output it is useful to learn more than simply whether the model is valid or invalid. Specifically, if the model is deemed invalid, then what aspects of the model are incorrect? Is it possible to identify over what range the model data are a poor representation of the system data? The WANOVA Bisection method first assesses model validity and can then identify the interval(s) over which the model is biased. Specifically, this method establishes the signal region over which the model data are most biased compared to the system data. This information allows developers to correct the necessary model components. Several simulation studies demonstrate the technique.

Finally, Chapter 6 proposes a new concept for exposing the disparity and agreement between the system and model data. This method identifies individual wavelet coefficients associated with the discrepancy and uses an inverse wavelet transform to highlight the nature and scope of the disparity. This also reveals the areas of agreement between the system and model. This method provides information on the magnitude of the model bias and is illustrated with two examples.

In summary, this dissertation identifies the challenges associated with the validation of models that generate time-series data and offers novel solutions to these challenges. The work surveys the associated literature and presents new wavelet-based model validation approaches that are able not only to assess model validity but also isolate regions of model discrepancy and identify the magnitude of any bias. These new wavelet-based model

validation techniques serve to overcome many of today's V&V challenges and offer a valuable contribution to the field.

II. Literature Review

This chapter surveys the existing literature on topics related to the validation of simulation models.

2.1 Introduction

As computer hardware technology continues to advance, so does the scientific community's capability to develop high resolution computer models able to simulate complex systems and processes. In the realm of technology development, a computer simulation can offer a descriptive glimpse into the future of system capabilities. The simulation can evaluate a solution more quickly, at a fraction of the cost of developing a prototype. In other arenas, such as industrial engineering and combat modeling, simulations offer the possibility of conducting realistic what-if analysis on systems for a variety of settings and environments. In short, computer simulations compose a powerful tool in operations research, and yet it is a tool that does not come without its own set of conditions.

Simulation verification and validation, often called V&V, is a vital step in the simulation development process and one that must be executed before relying on the results of a computer model. V&V helps to ensure that a model is not only formulated correctly, but also is sufficiently representative of the system that it is meant to model. This literature review examines the historical development of the field of simulation validation, discusses key principles and techniques, examines challenges involved with the practice, and finally addresses recent work conducted in the field. This mostly chronological presentation of the material provides insight into how simulation validation began and how it has evolved to address new issues. This review provides a solid understanding of the current state of simulation validation and the future of the field.

2.2 Early Beginnings of Simulation Validation

2.2.1 Pre-Computer Simulation.

Before discussing the advent of simulation validation, it is first necessary to address the history of simulation. When most people think about simulation, they imagine a computer program with an animation of entities (e.g. people, machines) performing some actions (e.g., waiting in line, assisting a customer). In reality, the idea of simulation predates computers and many experts point to the Buffon Needle Experiment in 1777 as the origination of the Monte Carlo simulation method [53]. The Monte Carlo method involves sampling from a data set or through experimentation in order to gather information about some probabilistic distribution. Buffon's experiment involved throwing needles onto a surface with equally spaced parallel lines to estimate the value of π. Interestingly enough, Pierre-Simon Laplace discovered an error in Buffon's solution in 1812 and published a solution, perhaps demonstrating the first instance of simulation validation [36, 53].

In 1899, English statistician William Sealy Gosset applied simulation and statistical analysis as an employee of the Arthur Guinness & Son brewery in Dublin, Ireland. Gosset used his statistical expertise to help select the best varieties of barley and began developing a new probability distribution. Since Gosset's analytical results were incomplete, he used the brewing data and manual simulation to validate his ideas about the exact form of the probability density function and published his results in 1908. Since Guinness restricted the use of proprietary company data, Gosset published his results under a pseudonym, and the distribution he developed is now well-known as the Student's t-distribution. His work demonstrates the complementary nature of statistical analysis and simulation and also foreshadows their interdependence in simulation validation [36, 53].

2.2.2 Computer Era.

In the mid-1940s, the development of the first electronic computers opened up the field of simulation. In 1943, Polish-American mathematician Stanislaw Ulam began working on

the Manhattan Project to develop the first nuclear weapons for the United States. Ulam recognized that many of the hydrodynamical calculations were difficult or impossible to solve. Around the same time the first electronic general-purpose computer, known as the ENIAC, became available. Ulam realized that instead of solving the calculations explicitly, the team could instead use the ENIAC to conduct computer based simulations that would numerically estimate solutions to the intractable problems [36, 53].

Over the following decades, many more scientists and researchers contributed to the field of computer simulation. Keith Tocher developed the first general-purpose simulator to model an industrial plant and also wrote the first textbook on simulation, *The Art of Simulation* [80]. Additionally, the first version of SIMSCRIPT was released in 1963, which was a simulation program intended for more novice computer users [36, 53].

With the explosion of simulation code and programs came the need to engage the more theoretical aspects of computer simulation. In 1963, Richard Conway of Cornell University addresses some of these issues in his paper, "Some Tactical Problems in Digital Simulation." His paper addresses three phases in a simulation investigation: "1. Model Interpretation – description in a language acceptable to the appropriate computer. 2. Strategic Planning – design of an experiment that will yield the desired information. 3. Tactical Planning – determination of how each of the test runs specified in the experimental design is to be executed" [21]. This early work demonstrates both the need to consider the various aspects of simulation critically and a necessity to define a framework of steps to complete when conducting simulation [36, 53].

2.2.3 Types of Simulation.

The research focus is simulation model validation. However, basic components of a simulation require some discussion. Simulation output may include discrete forms or functional forms depending on the system being modeled. Discrete simulation output includes forms such as means and variances, while functional output includes time-series

data, such as number in service or in queue. In addition to the form of the output, the state of a simulation system must also be considered. A simulation system may be transient or steady state. Transient states are by definition, temporary, and typically last a short time. Often, transient states are found when systems start or reset, such as in a model of a queuing system. In contrast, the steady-state phase represents more long-term behavior and involves the system functioning during normal operations. Typically, transient output is removed from consideration when estimating and analyzing steady state performance measures. In practice, steady-state behavior is the focus with less emphasis given to transient-state behaviors.

2.3 Foundational Principles and Techniques

2.3.1 Simulation Framework.

A framework for conducting simulation studies consists of several steps that should be executed as part of the larger simulation process. There are a variety of such frameworks. Key among the steps in any framework is simulation verification and validation. Law [44] recommends the following steps as part of a sound simulation study, which illustrate how V&V fits into the simulation process.

1. Formulate Problem and Plan the Study
2. Collect Data and Define a Model
3. Is the Assumptions Document Valid?
4. Construct a Computer Program and Verify
5. Make Pilot Runs
6. Is the Programmed Model Valid?
7. Design Experiments

8. Make Production Runs

9. Analyze Output Data

10. Document, Present, and Use Results

According to Law's approach, there are separate steps for model verification and model validation. While both steps are meant to ensure that the model is generating accurate and useful data, it is necessary to distinguish the steps from each other. Most authors now define model verification as 'ensuring that the computer program of the computerized model and its implementation are correct' [72]. Model validation is defined as the 'substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model' [72]. Verification determines if the model is built correctly and validation evaluates whether the correct model is being used.

In 1979, Sargent [70] presented one of the first in a sequence of papers on simulation validation at the Winter Simulation Conference (WSC). Sargent [72], Balci [5], and Kleinjen [43] developed some of the foundational work on simulation validation.

Sargent outlined some of the principles and techniques for good simulation V&V in 1979, and shortly thereafter he began collaborating with Balci on the topic in the 1980 article "Bibliography on validation of simulation models" [8]. Both authors have continued to update and evolve their ideas over the years and consequently their work comprises the foundational aspects of simulation validation.

2.3.2 Basic Approaches.

Sargent [72] discusses three basic decision-making approaches for determining simulation model validity. One approach consists of the model development team evaluating the validity of the model based on a variety of tests conducted over the course of model development. However, this approach may be biased (since it involves the

developers) and thus it is preferable to follow one of the other two approaches. The second approach integrates the user of the simulation model into the process to help determine validity. The user is involved during the development process and reviews the model during each phase of V&V. Finally, the third approach is referred to as independent verification and validation (IV&V). As the name implies, the process involves using a third party to evaluate the model and can be conducted either concurrently with model development or after model development. IV&V is not only an effective and unbiased method but also aids in model credibility, which builds user confidence in the model.

2.3.3 Paradigms.

Sargent [72] outlines two paradigms that help illustrate the V&V process: a simple view and a detailed view. For brevity, this section focuses on the simplified paradigm. See Figure 2.1 for a diagram depicting the simplified version of the model development process.

There are three elements in this paradigm. The Problem Entity is the real-world system; the Conceptual Model is the mathematical or logical representation of the real-world system; the Computerized Model is based on the conceptual model and is developed in the computer programming or implementation phase. Linking each element in the paradigm are both a simulation development action and a V&V action. The simulation development actions involve analyzing and modeling the problem entity to create the conceptual model and then implementing the conceptual model in the computerized model. The team can then make inferences about the problem entity by conducting experiments using the computerized model [72].

The V&V actions include Conceptual Model Validation, which ensures that the conceptual model is an adequate representation of the system. Robinson [66] has written extensively on conceptual modeling for simulation and documents the need for conceptual model validation. Computerized Model Verification refers to ensuring that the computer

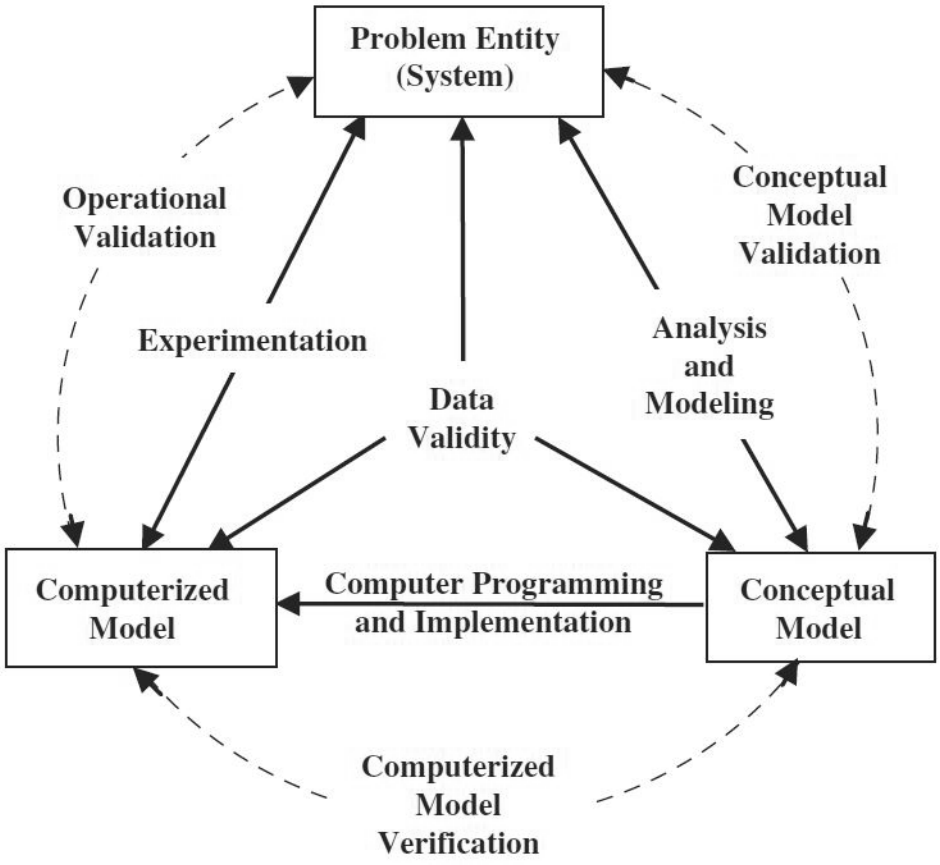


Figure 2.1: Simple Paradigm [72]

code properly reflects the intended conceptual model. Operational Validation compares the computer model to the system and checks whether the output of the computer model provides reasonably accurate results consistent with the problem entity. An additional V&V action described in the diagram is Data Validity, which is a step to ensure the data being used to build and test the model is both complete and accurate [72].

2.3.4 Principles.

Balci [5] identifies fifteen principles, which are listed in Table 2.1. These principles provide guidelines for conducting Verification, Validation, and Accreditation (VV&A) and help convey a better understanding of the process. For example, the principles reinforce

the necessity to conduct V&V throughout the modeling and simulation (M&S) life cycle, since the time and cost associated with conducting V&V at the end of the process could prove overwhelming. Several principles focus on the idea that the model is only applicable within a range of operating conditions and that the model's validity and credibility should be evaluated with this applicable range in mind. In Principle 9, Balci indicates that Type I, II, and III errors must be prevented. Type I error, α, is also known as model builder's risk and occurs when a valid model is rejected. Type II error, β, is model user's risk and occurs when an invalid model is accepted. Type III error refers to a model that provides the right answer to the wrong question.

2.3.5 Verification and Validation Techniques.

Both Sargent and Balci recommend a variety of V&V techniques, many of which overlap. Balci [5] separates his techniques into four categories: Informal, Static, Dynamic, and Formal, which are listed in Figure 2.2. Sargent's techniques are discussed in Section 2.3.5.5.

Table 2.1: Verification and Validation Principles [5]

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.

Figure 2.2: Verification and Validation Techniques [5]

2.3.5.1 Informal Techniques.

Informal techniques are a commonly used class of V&V techniques which are generally subjective and rely on human reasoning. Many of these techniques rely on the knowledge and experience of subject matter experts (SMEs), who can make assessments regarding the validity of the model. Although they are designated as "informal," these techniques include structured guidelines for implementation and can be very effective for model V&V. These informal techniques are summarized in Table 2.2 [5, 7].

Table 2.2: Informal Techniques

Audit	An Audit evaluates whether a simulation study is conducted in accordance with established standards and guidelines. This technique also seeks to establish traceability, or the ability to trace an error to its source.
Desk Checking	Desk Checking is the process of inspecting the simulation model for completeness and correctness and is ideally conducted by someone other than the developer.
Documentation Checking	Documentation Checking examines the correctness and completeness of the model documentation.
Face Validation	Face Validation is a process in which SMEs review the model and judge whether the behavior and output is reasonable.
Inspections	Inspections are conducted by a team of four to six members for any development phase of the model. The team inspects the model to locate and document any faults. An inspection is usually more rigorous and formal than a walkthrough.
Reviews	Reviews are similar to inspections and walkthroughs except that the team typically involves managers.
Turing Test	A Turing Test relies on SME opinions and challenges SMEs to distinguish between system data and simulation output in a blind test scenario.
Walkthroughs	Walkthroughs are conducted by teams where the majority of team members are not directly involved in the model's development. The goal is to detect and document errors.

2.3.5.2 Static Techniques.

A static model may be described as a time independent view of the system or alternatively a model where the state does not change. Static techniques focus on the accuracy of the static model design and do not require machine execution of the model. These techniques focus on the accuracy of the basic characteristics of the model and source code. These techniques may evaluate model structure, data flow, and syntactical accuracy. These static techniques are summarized in Table 2.3 [5, 7].

Table 2.3: Static Techniques

Cause-Effect Graphing	Cause-Effect Graphing identifies what actions or settings within the model causes which effects in order to assess model correctness.
Control Analysis	Control Analysis is used to analyze the control characteristics of the model. It includes Concurrent Process Analysis which analyzes the overlap of model components executed in parallel, and State Transition Analysis which identifies the transition of states during model execution.
Data Analysis	Data Analysis ensures that proper operations are applied to data objects and that the data used are properly defined. It includes Data Dependence Analysis and Data Flow Analysis, which are used to assess the dependence and flow of data within the model.
Fault/Failure Analysis	Fault/Failure Analysis is used to identify how the model might fail and under what conditions.
Interface Analysis	Interface Analysis evaluates both the sub-model to sub-model interface as well as the user-model interface.
Semantic Analysis	Semantic Analysis is often conducted by the simulation program's language compiler to check the structural accuracy of the code.
Structural Analysis	Structural Analysis examines the model's structure to check for anomalies, excessive levels of nesting, or other questionable practices.

Symbolic Evaluation	Symbolic Evaluation evaluates model accuracy by exercising the model with symbolic values. It can assist in showing path correctness.
Syntax Analysis	Syntax Analysis is similar to Semantic Analysis since it is usually conducted by the simulation programs language compiler. It assesses syntactical accuracy of the computerized model.
Traceability Assessment	Traceability Assessment matches model elements from the requirements specification of the model to the design specification of the model.

2.3.5.3 Dynamic Techniques.

In contrast to the static techniques, dynamic techniques require model execution. These techniques are meant to assess the model's execution behavior. This category contains the largest number of techniques, many of which are useful in model validation. These techniques are summarized in Table 2.4 [5, 7].

Table 2.4: Dynamic Techniques

Acceptance Testing	Acceptance Testing is conducted by the simulation model user to verify that contractual requirements have been met.
Alpha Testing	Alpha Testing involves the operational testing of the alpha version, or initial version of the complete model.
Assertion Checking	Assertion Checking verifies that what is happening during the simulation model execution matches what the modeler assumes should be happening.
Beta Testing	Beta Testing is the operational testing of the beta version, or the second, revised version of the model. It is usually conducted using realistic conditions.
Bottom-Up Testing	Bottom-Up Testing is used with bottom-up development, where base level sub-models are developed and tested first and then built upon and integrated.
Comparison Testing	Comparison Testing is used when there are multiple versions of a simulation model that are meant to represent the same system. These versions are executed and then compared against each other.
Compliance Testing	Compliance Testing evaluates the access authorization and security standards of the simulation model to ensure proper compliance with established regulations and standards.
Debugging	Debugging is an iterative process meant to identify model errors that lead to the simulation's failure.
Execution Testing	Execution Testing is used to collect and study data related to the execution behavior of the simulation model to uncover model representation errors.

Fault/Failure Insertion Testing	Fault/Failure Insertion Testing is the purposeful insertion of a fault or failure into the model to determine whether the anticipated invalid behavior is produced.
Field Testing	Field Testing is useful for military combat systems where the model is placed in an operationally representative environment.
Functional (Black-Box) Testing	Functional Testing inspects the accuracy of the transformation between the input and output of the model.
Graphical Comparisons	Graphical Comparisons use graphs of model variable values over time, which are compared with graphs of system variables to highlight similarities and differences.
Interface Testing	Interface Testing checks for errors with the data interface and model interface. Data interface testing assesses the accuracy of the data input and output, while model interface testing inspects for errors resulting from sub-model to sub-model interface problems.
Object-Flow Testing	Object-Flow Testing evaluates the life cycle of a model object or entity during execution to ensure accuracy.
Partition Testing	Partition Testing operates by decomposing the model into functional representatives or partitions and then comparing these partitions of the model specification and implementation.
Predictive Validation	Predictive Validation uses input and output data from the real system that is being modeled. This input data is used with the simulation model to generate output which is then compared to the system output.

Product Testing	Product Testing is conducted by the simulation model developer to evaluate the model prior to delivery to the user.
Regression Testing	Regression Testing ensures that any changes or corrections made to the model do not result in new, additional errors.
Sensitivity Analysis	Sensitivity Analysis involves changing the input parameters to the model and reviewing how that affects the model output; the nature in which the output changes in relation to the input should mirror that of the system.
Special Input Testing	Special Input Testing is meant to test the model using a variety of unique parameters, such as boundary values, extreme inputs, and invalid inputs.
Statistical Techniques	Statistical Techniques include hypothesis testing, simultaneous confidence intervals, Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA), goodness-of-fit tests, and regression analysis. Many of these statistical techniques are discussed further in Section 2.4.
Structural (White-Box) Testing	Structural Testing operates by evaluating the model's internal structure and assesses the accuracy of branches, conditions, loops, and internal logic of the simulation model.
Submodel/Module Testing	Submodel/Module Testing collects data on all input and output variables of each of the sub-models. This sub-model behavior is then compared to the corresponding sub-system behavior to determine validity.

Symbolic Debugging	Symbolic Debugging allows the modeler to conduct a step-by-step execution of the model while viewing the model at the source code level to determine correctness.
Top-Down Testing	Top-down testing is used with top-down model development, where high level sub-models are developed and testing, followed by lower level sub-model development and testing.
Visualization/Animation	Visualization/Animation uses a visual display to depict the simulation model's execution to aid in the identification of errors.

2.3.5.4 Formal Techniques.

Formal techniques are a category of techniques using mathematical proofs to demonstrate correctness and accuracy of the model. These proofs serve as an effective means of model V&V but are generally the most difficult and time-consuming to apply. These techniques are summarized in Table 2.5 [5, 7].

Table 2.5: Formal Techniques

Induction, Inference, Logical Deduction	Induction, Inference, and Logical Deduction refer to the act of justifying a conclusion based on some set of given premises. The argument should follow established rules of inference to reach the conclusion.
Inductive Assertions	Inductive Assertions is a three step process where input-output relations are identified; the relations are converted into assertion statements and placed along execution paths; and these paths are verified to be true.
Lambda-Calculus	Lambda-Calculus transforms the model into formal expressions so that mathematical proof of correctness techniques may be applied.
Predicate Calculus	Predicate Calculus provides rules for manipulating predicates, which is a combination of simple relations that are either true or false.
Predicate Transformation	Predicate Transformation formally defines the model with a mapping that transforms model output states to all possible model input states.
Proof of Correctness	Proof of Correctness techniques operate by expressing the model in a precise mathematical notation and then proving that the model satisfies requirements with sufficient accuracy.

2.3.5.5 Additional Techniques.

Sargent [72] discusses many of the same validation techniques as Balci, sometimes using different terminology. For example, Sargent's Degenerate Testing is a technique similar to Balci's Fault/Failure Insertion Testing. Sargent also discusses several unique techniques such as Internal Validation to determine the stochastic variability within the model by running multiple replications. Further, he also covers Philosophy of Science methods which include Rationalism, Empiricism, and Positive Economics. Rationalism requires a logically developed model from a set of assumptions; Empiricism requires every model assumption to be empirically validated; and Positive Economics simply requires that the model outcome is correct but not necessarily the assumptions or structure. Multistage Validation combines these three historical methods into a single validation technique.

2.3.6 Recommended Procedure.

Expanding upon the typical steps to follow in a sound simulation study, referenced in Section 2.3.1, Sargent [72] provides a recommended eight step procedure for conducting simulation validation. These steps ensure implementation of a thorough and agreed-upon validation process for the model.

1. The model development team and model users should agree on a validation approach.
2. Specify the necessary range of accuracy for the simulation models output variables prior to model development.
3. Test the model's assumptions and theories throughout the development process.
4. Perform face validity for each iteration of the conceptual model.
5. Explore the behavior of the computerized model for each model iteration.
6. Compare the model output data with the system behavior for several sets of experimental conditions.

7. Prepare V&V documentation to include with the simulation model documentation.

8. Develop a schedule for the periodic review of model validity.

2.4 Validation through Statistical Analysis

There are a variety of statistical validation techniques available to help ensure that simulation output accurately reflects system output. For example, Law [44] discusses results validation, a technique which compares the two sets of data to ensure they resemble each other. This technique relies upon having real-world data from the system to compare against simulation output. Balci [7], Sargent [10], and Kleijnen [43] all provide additional insight into some of the appropriate statistical analysis techniques for simulation validation.

2.4.1 Hypothesis Testing.

One of the primary statistical techniques for comparing two data sets is hypothesis testing. There are a variety of hypothesis tests available to use depending on the appropriate assumptions regarding the data. For example, parametric tests assume that the sample data come from a population that follows some underlying probability distribution. Alternatively, nonparametric tests do not make assumptions regarding the underlying distribution.

2.4.1.1 Parametric Tests.

Parametric tests commonly assume an underlying Normal distribution and that samples are independent (NID). One of the most basic and well-known tests is a two-sample t-test for a comparison of means. A two-sample t-test uses statistics calculated from sample data (y_1, y_2) from populations 1 and 2, respectively, to infer whether or not there is statistical evidence that the population means (μ_1, μ_2) differ. For the purpose of simulation validation, an appropriate method is to conduct a hypothesis test comparing the mean of system observations with the corresponding mean from the model output to determine if a significant difference exists. Additionally, this calculation incorporates the sample standard

deviations (s_1, s_2) and the number of observations in each sample (n_1, n_2). See Table 2.6 for information on the hypotheses, test statistic, and criteria for rejection [7, 44, 50, 79].

Table 2.6: Two Sample t-test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$	$t_0 = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$\lvert t_0 \rvert > t_{\alpha/2,v}$

The two-sample t-test leads to another important consideration in the validation process - the inspection approach. A basic inspection approach uses model input data that is sampled independent of the input data observed by the system. While it is possible to make comparisons regarding the model and system output, a risk to this approach involves the natural randomness of the input data. If the system and simulation experience different input data, one might expect to obtain different output data. To address this, a correlated inspection approach proves useful. A correlated inspection approach forces the simulation to experience the same observations as the system and results in a more statistically precise output. With this approach, a paired t-test determines if the average difference (μ_d) between the two sets of output data is statistically different from zero. This technique can result in a considerably smaller variance for the sample data, in contrast to the basic inspection approach. Calculate the difference for each replicate,

$$d_j = y_{1j} - y_{2j} \tag{2.1}$$

the mean difference,

$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j \tag{2.2}$$

and the variance of the difference,

$$S_d^2 = \frac{\sum_{j=1}^{n}(d_j - \bar{d})^2}{n-1}\,. \tag{2.3}$$

Then, use these values to perform the test in Table 2.7 [44, 50].

Table 2.7: Paired t-test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $\mu_d = 0$ H_1: $\mu_d \neq 0$	$t_0 = \frac{\bar{d}}{S_d/\sqrt{n}}$	$\lvert t_0 \rvert > t_{\frac{\alpha}{2},n-1}$

Analysis of variance (ANOVA) is a related statistical procedure that tests for differences among a group of population means by using sample data parameters. ANOVA generalizes the t-test to more than two groups, often referred to as treatments. Suppose there are a total of a treatments, with n replicates per treatment and N total observations. Then the Total Sums of Squares (SS_T) is calculated as

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 \tag{2.4}$$

where $\bar{y}_{i.}$ is the mean response at treatment i, and $\bar{y}_{..}$ is the mean response across all treatments. The Sums of Squares between treatments is calculated

$$SS_{Treatments} = n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{2.5}$$

and the Sums of Squares within treatments (error) is calculated

$$SS_{Error} = SS_T - SS_{Treatments} . \tag{2.6}$$

The Sums of Squares are used to calculate the value of F_0, which is the ratio of variability due to treatments versus error. The use of ANOVA requires the same assumptions as the t-test, including normality, independence, and constant variance. The standard ANOVA table and test are shown in Table 2.8 and Table 2.9 [50, 56].

Table 2.8: ANOVA Table

Source	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{Treatment}$	a-1	$MS_{Treatment} = \frac{SS_{Treatment}}{a-1}$	$F_0 = \frac{MS_{Treatment}}{MS_E}$
Error (within treatments)	SS_{Error}	N-a	$MS_E = \frac{SS_{Error}}{N-a}$	
Total	SS_T	N-1		

2.4.1.2 Nonparametric Tests.

As an alternative to parametric tests, nonparametric tests do not require any assumptions regarding the underlying distribution to be satisfied. One example is the

Table 2.9: ANOVA Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $\mu_1 = \mu_2 = \cdots = \mu_a$ H_1: $\mu_i \neq \mu_j$ $for\ some\ i,j$	F_0	$F_0 > F_{\alpha,a-1,N-a}$

Wilcoxon Rank Sum test, one of the most powerful nonparametric tests, since it is nearly as efficient as the t-test on normal distributions. This test determines whether two sets of sample data come from the same population, i.e. whether the difference between the two datasets (δ) is equal to zero. The Wilcoxon Rank Sum test (Table 2.10), or equivalently the Mann-Whitney test, uses ranks (R_j) assigned to the sample data (Y_j) to infer whether one population has larger values than the other. This test requires only that the observations are independent of one another [76].

Table 2.10: Wilcoxon Rank Sum Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $\Delta = 0$ H_1: $\Delta \neq 0$	$W = \sum_{j=1}^{n} R_j$	$W \geq w(\alpha_2, m, n)$ $\quad or$ $W \leq [n(m+n+1) - w(\alpha_1, m, n)]$

When there are more than two samples, the nonparametric Kruskal-Wallis test can detect whether the samples originate from some common distribution. The Kruskal-Wallis test (Table 2.11) equates to a nonparametric version of ANOVA. Similar to the Wilcoxon Rank Sum test, the Kruskal-Wallis test ranks the observations (y_{ij}) and then sums the ranks for each treatment using [50]

$$R_{ij}: \ rank\ of\ observation\ Y_{ij}$$

$$N: \ total\ number\ of\ observations$$

$$n_i: \ number\ of\ observations\ in\ the\ ith\ treatment.$$

Table 2.11: Kruskal-Wallis Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: *From same distribution* H_1: *From different distributions*	$H = \frac{12}{N(N+1)}\sum_{i=1}^{a}\frac{R_i^2.}{n_i} - 3(N+1)$	$H > \mathrm{X}^2_{\alpha,a-1}$

2.4.1.3 Multivariate Tests.

The hypothesis tests discussed in the last two sections prove very useful for univariate statistical tests on a single response variable of interest. However, in the case where there are multiple response variables of interest, these techniques are insufficient. Instead comparable statistical tests designed for handling multivariate response models are most useful. Two of the most prominent techniques are Hotelling's T^2 test and Multivariate Analysis of Variance (MANOVA). Of note, it is important to use one of these techniques when analyzing multivariate data rather than employing univariate techniques multiple times, since using a univariate technique multiple times inflates the type 1 error rate and does not account for correlation among the response variables. As such, the multivariate techniques discussed below offer more statistical power. Both techniques require the assumptions of normality, independence, and constant variance to be satisfied [9, 22].

Hotelling's T^2 test applies when there are k normally distributed response variables. This test operates in a manner similar to the two-sample t-test: the sample means calculated

for each response in both the model and system output are each defined by a *k x 1* vector ($\underline{\bar{X}}, \underline{\bar{Y}}$). The T^2 test statistic is calculated based on these mean vectors and on the pooled sample covariance matrix of the two sets of sample data,

$$S = \frac{(n_x - 1)S_X + (n_y - 1)S_Y}{(n_x - 1) + (n_y - 1)} . \tag{2.7}$$

In this case, the null hypothesis asserts that the population means are equal for each response of the two samples, as shown in Table 2.12 [9].

Table 2.12: Hotelling T^2 Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $\underline{\mu_X} = \underline{\mu_Y}$ H_1: $\underline{\mu_X} \neq \underline{\mu_Y}$	$T^2 = \left(\underline{\bar{X}} - \underline{\bar{Y}}\right)^T \left[S\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right]^{-1} \left(\underline{\bar{X}} - \underline{\bar{Y}}\right)$ $F_0 = \frac{n-k}{k(n-1)} T^2$	$F_0 > F_{\alpha,k,n-k}$

Where Hotelling's T^2 test is analogous to a two sample t-test, MANOVA is analogous to ANOVA. MANOVA is a statistical procedure for comparing multivariate means of several groups and may employ Hotelling's T^2 or an alternative test statistic, such as Wilk's Lambda (most common) or Pillai-Bartlett Trace. Instead of the standard sums of squares, MANOVA employs a sum of squares and cross products (SSCP) matrix to assess the variability within and among treatments (Table 2.13) to perform the test (Table 2.14) using [22]

$$Y: \ matrix\, of\, complete\, data\, (all\, collected\, results)$$

Y_i : *matrix of data associated with treatment i*

$\bar{Y}_i$: *matrix of means for each treatment*

$\bar{Y}$: *mean vector for all data.*

Table 2.13: MANOVA Table

Source	Sum of Squares	$Wilk's\ \Lambda$
Hypothesis SSCP Matrix (H)	$(\overline{Y_i} - 1 * \bar{Y})' * \left(\frac{n}{m}(\overline{Y_i} - 1 * \bar{Y})\right)$	$\frac{\lvert E \rvert}{\lvert T \rvert}$
Error SSCP Matrix (E)	$T - H$	
Total SSCP (T)	$(n-1) * cov(Y)$	

Table 2.14: MANOVA Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: *No difference in mean response among treatments* H_1: *Difference exists*	$F_0 = \frac{1 - \Lambda^{\frac{1}{b}}}{\Lambda^{\frac{1}{b}}} * \frac{df_1}{df_2}$	$F_0 > F_{\alpha, df_1, df_2}$

2.4.2 Confidence Intervals.

2.4.2.1 Univariate Confidence Intervals.

An alternative to the hypothesis test involves using confidence intervals (CIs). CIs prove beneficial since they provide the same information as a hypothesis test concerning whether a statistical difference exists. Additionally they provide information on the magnitude with which the two outputs differ. This additional information enables an assessment concerning whether the statistically significant difference is practically

significant. In other words, "is the magnitude of the difference large enough to invalidate any inferences about the system that would be derived from the model?" [44].

One approach uses the t-distribution to calculate a $100(1-\alpha)$ percent confidence interval around the sample difference between the model and system output,

$$\bar{d} \pm t_{\frac{\alpha}{2},n-1} \frac{S_d}{\sqrt{n}} \,. \tag{2.8}$$

If the two datasets are statistically equivalent, the CI contains zero. Otherwise, the CI indicates the magnitude of the difference. A CI approach must still consider whether the input data are independent or correlated in order to satisfy appropriate application of the inspection approach. Additionally, this CI technique requires the assumptions of normality, independence, and constant variance [44].

2.4.2.2 Simultaneous Confidence Intervals.

As with univariate hypothesis testing, modified techniques are necessary when applying a CI approach to datasets with multiple responses. These modified techniques prevent the inflation of the type 1 error probability. Balci and Sargent [10] discuss the use of simultaneous confidence intervals (SCI). They present four univariate techniques and three multivariate techniques for obtaining SCIs. The different techniques described are applicable based on whether the observations are independent (basic inspection) or correlated. The univariate techniques develop SCIs using the Bonferroni inequality, while the multivariate techniques use the Roy-Bose statistical method.

Both the univariate and multivariate SCI techniques offer different advantages. The univariate techniques allow the user to specify different confidence levels for each response variable CI, whereas the multivariate techniques require a single confidence level for the SCI. Univariate techniques also apply when there are unequal sample sizes among the response variables. However, univariate techniques do not provide exact confidence levels

because of the Bonferroni inequality, which instead provides bounds on the confidence level. Balci and Sargent state, "in summary, the selection of the specific statistical technique to be used should be based upon consideration of: (1) the satisfaction of underlying assumptions, (2) the size of the SCI, (3) the exactness of the confidence levels, (4) the equality or inequality of confidence levels for the ranges of accuracy, (5) the equality or inequality of sample sizes, and (6) the tradeoffs revealed by the schedules and graphs" [10].

A univariate SCI technique assumes there are k response variables from the model and the system, where $\underline{\mu_x}, \underline{\mu_y}$ represent the vectors of population means and $\underline{\bar{x}}, \underline{\bar{y}}$ represent the vectors of sample means. Then using the Bonferroni inequality,

$$(\underline{\bar{x}} - \underline{\bar{y}}) \pm t_{\frac{\alpha_j}{2}, n_x+n_y-2} * s_j \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \qquad \text{for } j = 1, 2, ..., k \tag{2.9}$$

gives a SCI with a confidence level of at least $(1 - \sum_{j=1}^{k} \alpha_j)$ [10].

A multivariate approach calculates the SCI directly using the Hotelling's T^2 distribution. For example, let $\underline{a}$ be a k dimensional vector defined to obtain the range of accuracy of the j^{th} model response with a value of 1 in the j^{th} element and zeroes in the others for $j = 1, 2, ..., k$, and let S be the pooled covariance matrix. Then

$$\underline{a}'(\underline{\bar{x}} - \underline{\bar{y}}) \pm \sqrt{\underline{a}' S \, \underline{a} * \frac{n_x + n_y}{n_x n_y}} * T^2_{\alpha, k, n_x+n_y-k-1} \tag{2.10}$$

gives a SCI with confidence level $1 - \alpha$ [10].

2.4.3 Regression Analysis.

Kleijnen [43] offered two new statistical tests that concern the idea of relative validity versus absolute validity. Absolute validity indicates that the numerical output from both the system and the model are statistically the same. For relative validity, while the output may

not be identical, an identical change in the input data or conditions for both the system and the model will result in similar changes in the output. Specifically, the change in output should be in the same direction and have similar magnitude for both the system and the model.

Kleijnen's first test operates to evaluate the relative validity or positive correlation of the two responses. Given that each observation results in an output for the system (v_i) and an output for the model (w_i), Kleijnen plots each of these n pairs and uses ordinary least squares to develop a simple linear regression model through the points,

$$\hat{w} = \beta_0 + \beta_1 v \,. \tag{2.11}$$

He then conducts a hypothesis test (Table 2.15) on the slope parameter for the model to determine if the parameter is greater than zero, indicating positive correlation [43, 51].

Table 2.15: Relative Validity Regression Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $\beta_1 \leq 0$ H_1: $\beta_1 > 0$	$t_0 = \frac{\beta_1}{se(\beta_1)}$	$t_0 > t_{\alpha,n-2}$

Kleijnen's second test focuses on absolute validity between the system and the model. For this technique, he follows a similar approach of developing a simple linear regression model through the pairs of points,

$$\hat{w}_i = \hat{\beta}_0 + \hat{\beta}_1 v_i \,. \tag{2.12}$$

In contrast to the test for relative validity, in this test he uses the sums of squares due to error,

$$SSE_{Full} = \sum_{i=1}^{n} (w_i - \hat{w}_i)^2 \tag{2.13}$$

and

$$SSE_{Reduced} = \sum_{i=1}^{n} (w_i - v_i)^2 \tag{2.14}$$

to perform a composite hypothesis test (Table 2.16) to check if the intercept parameter equals zero and if the slope parameter equals one. If this is the case, there is statistical evidence that the system output and the model output are identical [42, 43, 51].

Table 2.16: Absolute Validity Regression Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $\beta_0 = 0$ and $\beta_1 = 1$ H_1: $\beta_0 \neq 0$ or $\beta_1 \neq 1$	$F_0 = \frac{(SSE_{Reduced} - SSE_{Full})/2}{SSE_{Full}/(n-2)}$	$F_0 > F_{\alpha,2,n-2}$

2.4.3.1 Misuse of Regression for Model Validation.

Mitchell [49] critiques the use of regression for the empirical validation of models by citing five reasons why the method is inappropriate. The first reason is that it is a misapplication of regression, the purpose of which is to estimate a dependent variable, *y*, from a regressor variable, *x*. Second, the null hypothesis tests can be ambiguous due

to the variability of the data. Third, regression lacks the sensitivity to quantify how good the regression line truly is. Fourth, the fitted line is irrelevant to validation, since it shifts attention away from the actual model or model output. Last, the assumptions necessary for regression (normality, independence, constant variance) may be violated by the data. Instead of regression, Mitchell proposes an alternative method where the deviations between the system and model data are recorded and evaluated against an acceptable deviation interval. He recommends that at least 95% of points must fall within this interval if the model is to be considered valid. While this proposed method is effective at shifting focus to the deviations between the system and model data, it is also important to recognize the subjectivity needed to designate the acceptable deviation interval and the required proportion of points necessary for validity.

2.4.4 Goodness-of-Fit Tests.

Goodness-of-fit tests comprise a category of hypothesis tests used to determine whether observations fit a particular reference or theoretical probability distribution. The most common goodness-of-fit tests are the Chi-Square test and the Kolmogorov-Smirnov (K-S) test. The Chi-Square test facilitates simulation validation by testing whether the set of simulation output data has the same distribution as the set of observed historical data. Law states, "a chi-square test may be thought of as a more formal comparison of a histogram with the fitted density or mass function" [44]. One disadvantage of this test is that it is necessary to select appropriate, unbiased categories to separate the data in order to perform this test. Similarly, the K-S test operates to determine the level of agreement between the model output and the theoretical distribution of the system output [44].

Computation of the Chi-Square test statistic requires dividing the range of the data into k adjacent intervals, $[a_{i-1}, a_i]$, and then calculating the number of observations in each interval (N_j). The next step comprises determining the expected proportions (p_j)

of observations that would fall in each interval based on the surmised distribution ($\hat{F}$) and executing the test shown in Table 2.17 [44].

Table 2.17: Chi-Square Test

Hypothesis	Test Statistic	Criteria for Rejection
H_0: $X_i's$ *are IID with distribution* $\hat{F}$ H_1: $X_i's$ *do not have distribution* $\hat{F}$	$X^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$	$X^2 > X^2_{k-1,1-\alpha}$

2.4.5 Theil's Inequality Coefficient.

The use of Theil's inequality coefficient (TIC) is another technique to assess the validity of a model and applies when comparing time-series data. TIC is a metric that describes the amount of error between the system (x) and model (y) output. The calculation is based on the standard deviation of the error and the second moments of the output data,

$$TIC = \frac{\sqrt{E[(x-y)^2]}}{\sqrt{E[x^2]} + \sqrt{E[y^2]}}. \tag{2.15}$$

TIC varies between zero and one, where zero indicates perfect agreement and a value of one represents no agreement. It is also possible to use TIC to compare time-series data, where collected system and model data vary over time [55, 67].

2.4.6 Time Series Analysis.

2.4.6.1 Correlation Analysis.

Correlation analysis is a useful tool for evaluating time series data. In general, correlation refers to some linear form of dependence between two data sets. This level of dependence is typically measured using a correlation coefficient, often denoted

ρ (population) or r (sample). There are a variety of dependency measures, and the Pearson product moment coefficient, ρ, is a popular choice. The Pearson product moment coefficient measures the degree of linear dependence between two variables, x and y. This coefficient,

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{2.16}$$

uses the covariance of the two variables (σ_{xy}), as well as the standard deviations of each variable, σ_x and σ_y [13, 15].

Time series data are often highly autocorrelated; the response is correlated with itself at different points in time, referred to as autocorrelation. In this case, output values are a function of the time lag between them. The autocorrelation function,

$$R_{xx}(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T x(t)x(t+\tau)dt \tag{2.17}$$

calculates the level of autocorrelation based on the value of the time lag, τ [13, 15].

The autocorrelation function is a special case of the more general cross-correlation function. The cross-correlation function,

$$R_{xy}(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T x(t)y(t+\tau)dt \tag{2.18}$$

evaluates the correlation between two sets of output, x and y, while accounting for a time lag [13].

When autocorrelation is present, many of the statistical techniques discussed in prior sections will underestimate the sample variances leading to inaccurate statistics and conclusions. Correlation analysis is a more effective technique for comparing two sets of

time series data. Additionally, the Pearson correlation coefficient measures correlation by using pairs of observations over time. Therefore, the degree of correlation is based on how system and model output vary at identical time steps. To account for the possibility of a time lag between the two datasets, one should use the correlation coefficient function,

$$\rho_{xy}(\tau) = \frac{R_{xy}(\tau) - \mu_x \mu_y}{\sqrt{[R_{xx}(0) - \mu_x^2][R_{yy}(0) - \mu_y^2]}} \tag{2.19}$$

which provides a more effective measure of similarity between two time series. For simulation validation, this technique provides the level of correlation between system (x) and model (y) time series output and the degree to which one lags behind the other [13, 15, 55].

2.4.6.2 Spectral Analysis.

Spectral analysis is another technique for analyzing autocorrelated time series data and operates by transforming the data from the time domain to the frequency domain. Spectral analysis allows for the objective comparison of time series data in order to validate a simulation model. This method uses a Fourier transform on the correlation function of the time series data to develop a continuous spectral density function,

$$G_{xy}(f) = 2 * \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-(i2\pi f \tau)} d\tau \, . \tag{2.20}$$

The equation uses the cross-correlation function, R_{xy}, the time lag, τ, and frequency, f [13, 28, 29].

More specifically, Equation 2.20 is the one-sided spectral density function of the data. For simulation validation, the spectral density functions enable comparison of system and model time series data. Given that the cross-correlation function and autocorrelation

functions transform into spectral density functions, it is possible to generate the coherence function,

$$\gamma_{xy}^2(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)} \qquad 0 \leq \gamma_{xy}^2(f) \leq 1\ . \tag{2.21}$$

This coherence function is similar to the correlation coefficient function given in Equation 2.19 and gives a measure of similarity between the system and model output. To apply spectral analysis to a time series, the data must be stationary [13, 28, 29].

2.5 Challenges Encountered

Over the years, new challenges have arisen associated with modeling and simulation V&V. Pace [60] summarized some of the challenges discussed by experts at the 2002 Foundations Workshop. Pace separates the challenges identified into two categories: Management Challenges and Research Challenges.

2.5.1 Management Challenges.

The management challenges Pace discusses concern "how to do what we know how to do" [60]. Specifically, the three major management challenges are Qualitative Assessment, Formal Assessment, and Costs and Resources. The Qualitative Assessment challenge addresses the role of human judgment in validity assessments. For example, Face Validity or model walkthroughs rely on expert opinion to determine model validity. The key for good Qualitative Assessment is that evaluators are true experts with not only the appropriate credentials to make the assessments but also the necessary objectivity to ensure an accurate assessment. The Formal Assessment challenge considers the difficulty associated with implementing statistical or other mathematical tests for validation. Lastly, the Costs and Resources challenge recognizes the need to correctly estimate the required resources necessary for simulation validation, to include data requirements.

2.5.2 Research Challenges.

Research Challenges involve "areas that we need to understand better in order to find viable technical solutions." Pace identifies four major research challenges: Inference, Adaptation, Aggregation, and Human involvement and representation [60].

The challenge associated with Inference considers the availability of data to support the assessment of simulation predictions. Pace states, "there are currently no scientifically rigorous methods for making inferences about simulation results (predictions) elsewhere in the application domain (i.e. in those regions where data do not exist)" [60].

Adaptive programming includes artificial intelligence, expert systems, genetic algorithms, and machine learning. However, "no scientifically rigorous methods currently exist to ensure that future modeling and simulation performance involving adaptive programming will be as good as or better than past performance" [60].

Aggregation is a technique of combining different simulation sub-processes into a single, larger process. Since different elements of a simulation may be represented in varying levels of detail or resolution, this can result in problematic issues present in the aggregated model. The challenge concerns identifying better methods for determining how the differences in resolution affect the overall simulation results in order to prevent inaccurate simulation results [60].

Lastly, it is becoming increasingly important to represent human behavior in simulations. The complexity of human behavior, however, makes it extremely challenging to adequately represent in a computer simulation [60]. These four challenges comprise new research areas within the field of simulation validation which justify further attention.

2.5.3 Distributed Simulation Challenges.

In addition to the Management and Research Challenges introduced by Pace, unique simulation environments, such as Distributed Simulation, introduce new challenges. Page *et al*. [61] offers a case study of VV&A for Advanced Distributed Simulation (ADS), which

was an initiative within the defense M&S community. Distributed simulation, frequently used by military organizations for war-gaming exercises, allows for a real-time simulation to occur across multiple computers. In the case study scenario, there are differences between traditional and ADS V&V requirements. The case study uses the Joint Training Confederation (JTC) designed to aggregate many individual tank simulators into a single, larger simulation called a confederation. Some of the challenges that Page *et al*. discuss include that "ADS environments, particularly those that support training often include human-in-the-loop aspects. Additionally, training simulations, ADS or otherwise, rarely contain a formulation of the kind of output process that is typical in, say, the discrete event simulation world" [61]. To manage these challenges, the Defense Modeling and Simulation Office (DMSO) developed a VV&A process for a Distributed Interactive Simulation (DIS) exercise.

1. Develop VV&A Plans
2. Verify Standards
3. Perform Conceptual Model Validation
4. Perform Architectural Design Verification
5. Perform Detailed Design Validation
6. Perform Exercise Validation
7. Perform Accreditation
8. Prepare VV&A Reports

2.5.4 Challenges in Evaluating a Validation Process.

As simulation V&V has grown within the M&S field, a growing cadre of scientists and engineers have adopted V&V processes in simulation development. However inadequate

or incorrect application of V&V techniques poses a growing challenge. Many scientists publish articles detailing their validation efforts in order to demonstrate the effectiveness of their model, but they often fail to apply statistical rigor to their analysis. For example, Park and Schneeberger [62] attempt to validate a traffic simulation model using newly collected real-world data. Although the authors spend significant effort planning and executing the study, their published validation results consist solely of a graph that compares a single response of field and simulation data, shown in Figure 2.3. Many articles [63, 82] discuss validation but only provide visual evidence of their process, which can be useful when validated by a subject matter expert, but otherwise fails to meet the standard of robust statistical validation.

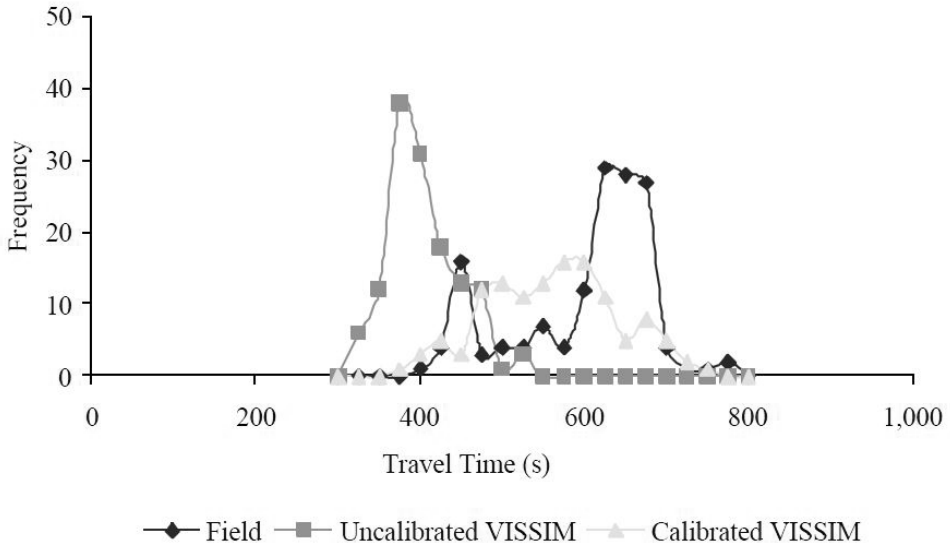


Figure 2.3: Traffic Model Validation [62]

To address this growing challenge, Harmon and Youngblood [38] offer a solution. They structure simulation validation processes into varying levels of maturity to develop a more organized and systematic approach that inspires confidence in model validation

efforts. The authors note that while the process maturity model shares similarities with general software engineering maturity models, there are important distinctions. Thus, they develop a six-level process maturity model based on validation criteria, referents, conceptual models, development products, and simulation results. The lowest level, Level 0, indicates no validation performed, while Level 1 indicates an informal technique such as face validation. Level 2 begins reducing the dependency on the SME, while Levels 3 and 4 "progressively improve the objectivity of the conceptual modeling and results validation component processes" [38]. The highest level, Level 5, applies to a fully automated validation process where informal user need statements are transformed into formal validation criteria. Table 2.18 outlines the main characteristics of each process maturity level. In summary, the authors assert that a more mature validation process should demonstrate more objective and higher quality processes.

Table 2.18: Process Maturity Level Characteristics [38]

Level	Judge of Validity	Validation Criteria	Confidence Estimates
0	None	None	No
1	SME	SME knowledge of user needs	No
2	SME	Explicit specification of needed entities, entity properties, and behaviors together with SME knowledge of user needs for accuracy	No
3	Independent observer	Explicit specification of needed entities, entity properties, behaviors, minimum needed accuracies, and property ranges over which minimum accuracies can be guaranteed	No
4	Independent observer	Explicit specification of needed entities, entity properties, behaviors, minimum needed accuracies, property ranges over which minimum accuracies can be guaranteed, and desired confidences in validation evidence	Yes
5	Formal proof	Formal specification of needed entities, entity properties, behaviors, minimum needed accuracies, property ranges over which minimum accuracies can be guaranteed, and desired confidences in validation evidence	Yes

2.5.5 Challenges in Validation of Transient Data.

When validating model output of time-series data, there are a variety of new challenges associated with the transient phase of the data. Many accepted validation techniques are designed for assessing the validity during the steady state phase of a process. These techniques are not necessarily well-suited for data collected during the transient phase, which typically includes the initialization period of the simulation. Transient pulses may be characterized by a large spike in magnitude followed by a sharp decrease in a short span of time. Oftentimes, simulation validation processes handle this transient or initialization phase by excluding it from the data analysis. However, in some circumstances analysis and validation of this portion of the data are necessary. For example, consider the scenario where a large pulse of energy causes an electronic equipment malfunction. It is imperative that the simulation is able to accurately model this energy pulse that could occur in the system [39]. Recent work to address this challenge is presented in Section 2.6.2.7.

2.6 Recent Work and Developments

The challenges presented in Section 2.5 create the opportunity for new work and innovation in the field of simulation validation. Section 2.6 summarizes efforts to address these challenges.

2.6.1 Relative and Absolute Validity.

Godley *et al*. [35] perform a validation study on a driving simulator. The authors evaluate how a human operator behaves or reacts to rumble strips in a real vehicle versus a simulator. Driving speed is the primary response of interest, and the experiment involves two different evaluations to assess relative validity and absolute validity.

In the experiment, the normal road defines the control area whereas the rumble strips represents the treatment area. To assess relative validity, the authors use two procedures - Averaged Relative Validity and Interactive Relative Validity. For Averaged Relative Validity, they calculate the difference between the average speed in the control area and the

average speed in the treatment area for both sites: the vehicle and simulator observations. A two-factor analysis of variance (ANOVA) determines if there was a significant interaction between the site main effect (vehicle or simulator) and the treatment main effect (control or rumble strips) [35].

The second procedure used, Interactive Relative Validity, takes the speed profile across the data collection area and examines the correlation between sites. The authors use a slightly modified canonical correlation analysis approach to evaluate the correlation in speed as participants progressed through the test track [35].

For absolute validation, the authors again take the average speeds for the control area and the treatment area at both sites. For this validation they perform two one-way ANOVAs to assess if there was a statistically significant difference in average speeds between the real vehicle and the simulator [35].

2.6.2 Model Validation Metrics.

Model validation metrics are another technique for assessing model validity, particularly those models with functional output such as time-series data. A model validation metric is typically a quantified value or set of values that express the level of model validity and is developed by comparing computational model results with experimental measurements. This commonly takes the form of an error metric that measures the discrepancy between the system and model output. When evaluating functional data, this discrepancy may involve magnitude, phase, shape, or a combination of three. The next few sections discuss some of the desired qualities in a model validation metric as well as several examples of previously developed validation metrics.

2.6.2.1 Developing Model Validation Metrics.

Oberkampf and Trucano [58] discuss validation methodology and the relatively new concept of validation metrics. They deem inadequate methods such as graphical validation, where the system and model output are compared on a graph. They also discuss the need

to estimate the experimental and computational uncertainty in a validation experiment. It is important to identify and quantify the experimental error because model accuracy is "measured in relation to experimental data, our best measure of reality" [58]. Similarly, the estimation of computational uncertainty stems from the need to distinguish between computational uncertainty and the error attributed to the model. These steps aid in the development of a more effective validation metric.

Oberkampf and Barone [57] extend prior work by recommending several features of validation metrics. The first feature of a validation metric is that it should include an estimate of the numerical error resulting from the computational simulation. Second, the metric should be a quantitative evaluation of the overall accuracy of the model. Third, the metric should include an estimate of the error resulting from post-processing of the experimental data. Fourth, a metric should include an estimate of the measurement errors from the experimental data. Fifth, the metric should depend on the number of experimental measurements that are made. Last, the metric should exclude any indications of the level of adequacy in agreement between the system and model data. The authors also note that if the data have a periodic character or a complex mixture of many frequencies, then the data require sophisticated time-series analysis and/or mapping to the frequency domain. They refer readers to validation metrics constructed by Geers [31], Russell [68], and Sprague and Geers [78]. These validation metrics will be discussed further in the following sections.

2.6.2.2 The Geers Metric.

Geers [31] was one of the first to introduce an error measure for the comparison of calculated and measured transient response histories. The Geers metric assigns a single numerical value to the discrepancy between two sets of time-series data. The metric is based on two correspondence histories, each of which approaches a constant value. One correspondence history is sensitive to magnitude, while the other is sensitive to phase. These two correspondence histories comprise a comprehensive error factor.

Calculate the Geers metric as follows: let *c(t)* be the calculated response history or model data and let *m(t)* be the measured response history or system data. Then calculate the factors,

$$\psi_{cc} = T^{-1} \int_0^T c^2(t)dt \,, \tag{2.22}$$

$$\psi_{mm} = T^{-1} \int_0^T m^2(t)dt \,, \tag{2.23}$$

and

$$\psi_{cm} = T^{-1} \int_0^T m(t)c(t)dt \,. \tag{2.24}$$

Calculate the magnitude (M), phase (P), and comprehensive (C) error factors as [77]

$$M = \sqrt{\frac{\psi_{cc}}{\psi_{mm}}} - 1 \,, \tag{2.25}$$

$$P = 1 - \frac{\psi_{cm}}{\sqrt{\psi_{cc}\psi_{mm}}} \,, \tag{2.26}$$

and

$$C = \sqrt{M^2 + P^2} \,. \tag{2.27}$$

Sprague and Geers [78] further refine the Geers metric because the phase error factor was found to be insufficiently sensitive to phase errors. The change in the Sprague and

Geers metric relative to the original Geers metric is that the phase error factor is now calculated as

$$P = \frac{1}{\pi} arccos\left(\frac{\psi_{cm}}{\sqrt{\psi_{cc}\psi_{mm}}}\right). \tag{2.28}$$

The following sections incorporate a discussion of the effectiveness of the Sprague and Geers metric and how it compares to other time-series error metrics.

2.6.2.3 Russell's Error Measure.

Russell [68] develops a new set of magnitude, phase, and comprehensive error measures which can be used to assess the difference between two sets of data. Russell's contribution seeks to remedy the problem with other error measures which are biased towards one set of data. This biasing is the result of assuming that one data set, typically the system data, is absolutely correct. Since it is not necessarily true that the experimental system data are 100% correct, the error measure could be biased towards incorrect data. Russell's error measure is based on a relative magnitude error measure and a phase correlation, which in turn are used to develop a magnitude, phase, and comprehensive error factor. Calculate the relative magnitude error,

$$m = \frac{\sum_{i=1}^{N}(f_1(i)^2 - f_2(i)^2)}{\sqrt{\sum_{i=1}^{N} f_1(i)^2 * \sum_{i=1}^{N} f_2(i)^2}} \tag{2.29}$$

where f_1 and f_2 are the two datasets being compared. Determine the phase correlation,

$$p = \frac{\sum_{i=1}^{N} f_1(i) f_2(i)}{\sqrt{\sum_{i=1}^{N} f_1(i)^2 * \sum_{i=1}^{N} f_2(i)^2}}. \tag{2.30}$$

Then calculate the magnitude error factor using the signum function, sgn,

$$\epsilon_m = sgn(m) Log_{10}(1 + |m|)\,, \tag{2.31}$$

the phase error factor,

$$\epsilon_p = \frac{cos^{-1}(p)}{\pi}\,, \tag{2.32}$$

and the comprehensive error factor,

$$\epsilon_c = \sqrt{\frac{\pi}{4}(\epsilon_m^2 + \epsilon_p^2)}\,. \tag{2.33}$$

Using these calculations, the magnitude error factor is unbiased and independent of any phase shift. The phase error factor is independent of the magnitudes of the data and is bound between 0 and 1. The two error factors are on roughly the same scale and the comprehensive error measure is scaled to compromise between the two error factors. Russell suggests that an acceptable limit for all three error factors is 0.2, but he emphasizes this is only a suggested value and that acceptable error levels must be based upon the goals of a specific evaluation [68].

2.6.2.4 Whang's Inequality Index and Zilliacus' Error Index.

Whang *et al*. [85] propose two new visually meaningful correlation measures for comparing calculated and measured response histories. Similar to Russell, the authors critique the practice of assuming one of the response histories to be true when comparing datasets. Then, any deviation from the presumed true dataset is referred to as error. In a model validation experiment, it is inappropriate to assume that the experimental system measurements are without error. Therefore, the authors propose an inequality index, as well as an error index to accommodate the possible discrepancies. First, Whang's Inequality

Index (W) compares the difference between the two histories to the sum of the two, without assuming one of the histories to be true. It is a simplification of Theil's inequality coefficient. Calculate Whang's Inequality Index as

$$W = \frac{\sum |c_i - m_i|}{\sum |c_i| + \sum |m_i|} \tag{2.34}$$

where c_i represents the calculated values and m_i represents the measured values.

Zilliacus' Error Index,

$$Z = \frac{\sum |c_i - m_i|}{\sum |m_i|} \tag{2.35}$$

compares the difference between two histories to the one assumed to be true and is a simplification of the root-sum-square (RSS) error factor [85].

2.6.2.5 Comparison of Time-Series Error Metrics.

Many authors compare the viability and effectiveness of the various error metrics discussed in Sections 2.6.2.2 - 2.6.2.4. Russell [69] examines the Geers metric, Whang's Inequality, and Zilliacus' Error against his own error factor. By examining the error measures for multiple case studies, Russell concludes that the Geers, Whang, and Russell metrics are the recommended candidate error measures. He notes that the Geers metric is biased towards under-predicting the response and that Whang's Inequality Index is very sensitive to phase errors. Unsurprisingly, he recommends his own measure as robust and unbiased. Schwer [75] evaluates the performance of validation metrics for response histories and highlights the performance of the Sprague and Geers metric on several case studies. Schwer uses waveforms from a case study and compares SME order of preference with error metric values. The SMEs also graded the level of agreement between waveforms using a zero-to-one assessment. The Sprague and Geers metric agreed with the SMEs

in order of preference and were generally within one standard deviation of the SME assessments.

2.6.2.6 EARTH Method.

Sarin *et al*. [73] describe a new validation method with a focus on vehicle safety applications. The authors develop an error metric to provide an overall quantitative value of the discrepancy between time series data. They discuss several existing error measures and metrics including vector norms, average residual, coefficient of correlation, the Sprague and Geers metric, and others. The authors use Dynamic Time Warping (DTW), a technique used in speech recognition which allows identification of whether two time histories with time shifts are a match. DTW aligns the peaks and valleys of the data as much as possible by expanding and compressing the time axis. Figure 2.4 depicts the effect of DTW on time histories.

The authors propose phase error, magnitude error, and slope error as the three primary error measures and note that since these three error measures are often strongly coupled, it is necessary to minimize their influence on each other. In other words, they want to ensure that the error is not double counted in multiple error measures. They use a cross-correlation technique to quantify phase error with a penalty function to weight small and large time step differences appropriately. To measure magnitude error, they first apply DTW to compensate for phase error and then use an L1 norm to measure the relative magnitude error. For slope error, they compute the derivative of the time-shifted histories again to minimize the effect of global phase error [73].

These three error measures comprise the Error Assessment of Response Time Histories (EARTH) metric for validation. The authors apply the technique to a case study involving crash testing, in which they compare three crash test dummy experiments to three different simulation models. They collect results on three responses of interest including head displacement, acceleration, and impact force. If the error measures between the three

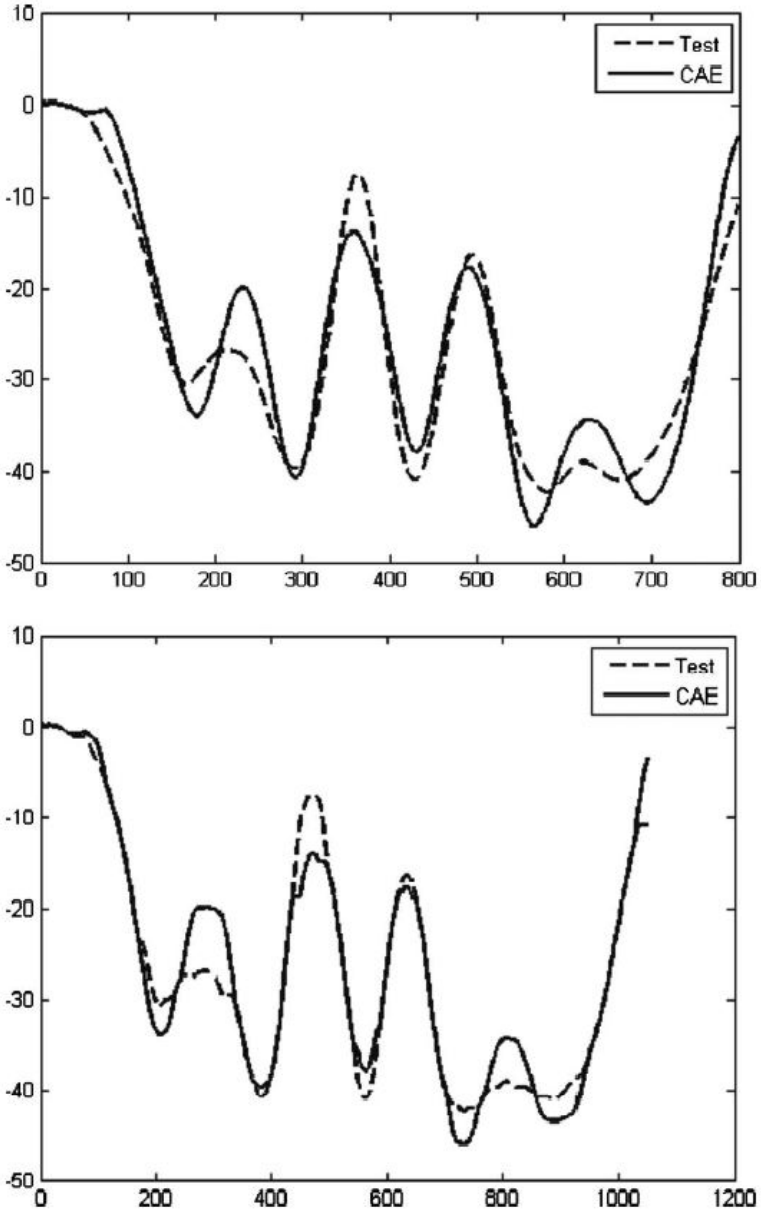


Figure 2.4: Effect of DTW: Before (top) and After (bottom) [73]

experimental data sets are greater than the errors between the model and test data, then the authors infer that the simulation model is adequate [73].

An alternative technique using SME input involves assigning weights to the three EARTH error measures which helps to provide an overall EARTH metric. After conducting this technique, the authors compare the performance of this overall EARTH metric to other validation techniques including wavelet decomposition, step function, and corridor violation with these results provided in Figure 2.5. The EARTH metric performs well at predicting SME evaluations and thus appears capable of recognizing key features and providing an overall error measure [73].

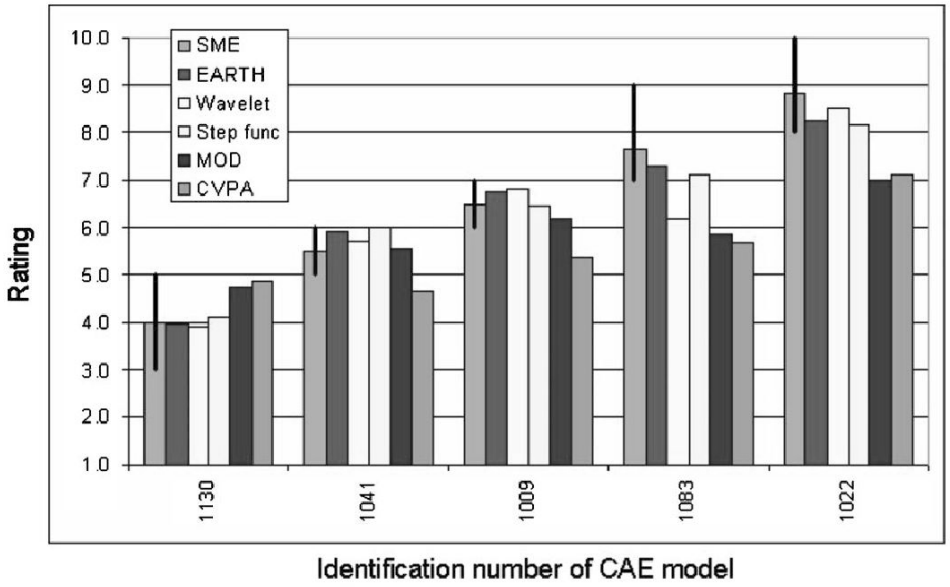


Figure 2.5: Comparison of EARTH to Other Metrics [73]

2.6.2.7 Transient Time Domain Validation.

It can be difficult and yet critical to validate models during the transient phase of a process. Jauregui *et al*. [39] analyze the performance of the Feature Select Validation (FSV) method and propose a new method better equipped to handle transient data. Their article focuses on the validation of Finite Difference Time Domain (FDTD) simulations.

The authors apply the FSV method by separating the difference in output between the model and the system into two separate measures: Amplitude Difference Measure (ADM) and Feature Difference Measure (FDM). Similar to the EARTH method, ADM is a measure of the magnitude difference and FDM is a measure of the shape difference. Together, the two measures make up the Global Difference Measure (GDM). In the authors' review of how FSV performs with transient data, however, they conclude that the ADM leads to an error in the GDM and causes a misinterpretation of the results.

The authors propose a new method called Transient Time Domain Validation (TTDV) which uses five indicators to assess different characteristics of transient data. These five indicators are: Feature Difference Measure (FDM), Maximum Amplitude Levels (APL), Maximum Rate Time (MRT), Energy Contained in the Signals (ECS), and Total Error Average (TEA). FDM is calculated the same way as in the FSV method, since this measure did not generate problems with transient data validation. APL is a magnitude measure and measures the difference between the maximum amplitude of the signals. MRT is a form of slope error, and ECS is calculated by taking the integral of the data function. Ultimately, TEA combines the previous metrics to provide an overall error assessment [39].

The authors apply both the FSV and TTDV methods to electromagnetic (EM) voltage data for both the system and the model in three different configurations. The level of agreement between the system and model data varies with each configuration, from a wide discrepancy (Figure 2.6) to nearly imperceptible variation as shown in Figure 2.8 [39].

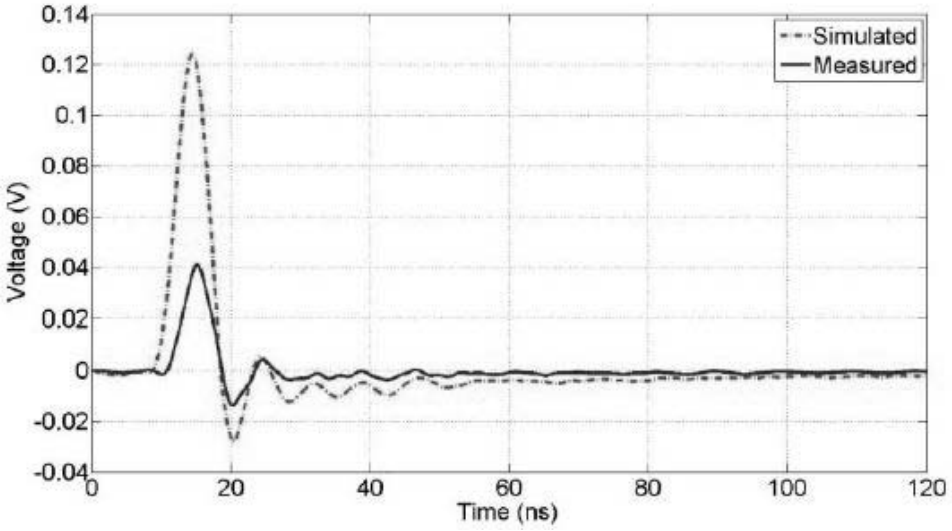


Figure 2.6: Configuration 1 [39]

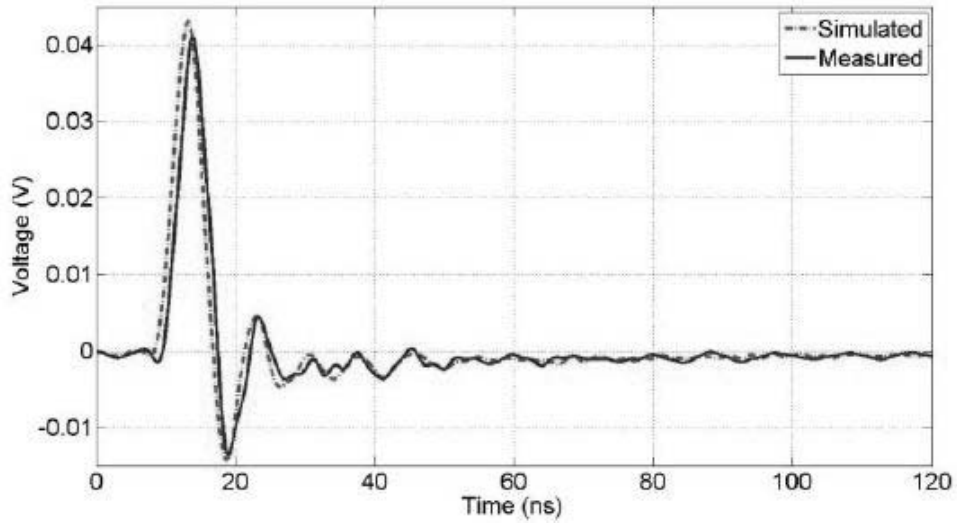


Figure 2.7: Configuration 2 [39]

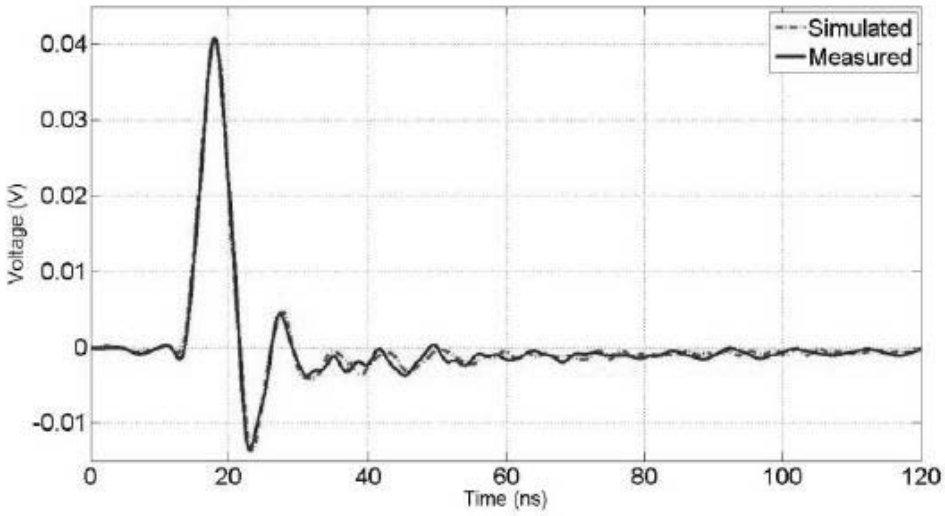


Figure 2.8: Configuration 3 [39]

The FSV method is unable to distinguish a difference in model quality between the second and third configurations, rating each as "Very Good." The GDM indicator for these two configurations differs by just 1.1%. However, the TTDV can successfully distinguish between these last two configurations, since the TEA indicator differs by 26% between the two. This demonstrates that the TTDV method is more effective as model validation assessment for the transient time phase than is the FSV method [39].

2.6.3 Wavelets.

Wavelets are a relatively recent development in the field of mathematics but have had a significant impact thus far. Wavelets apply in a variety of situations, and there is potential use of wavelets as a method for simulation validation. Ogden states, "broadly defined, a wavelet is simply a wavy function carefully constructed as to have certain mathematical properties. An entire set of wavelets is constructed from a single 'mother wavelet' function, and this set provides useful 'building block' functions that can be used to describe any in a large class of functions" [59]. In short, wavelets are a family of functions that serve as

basis functions and may express either discrete or continuous signals, similar to a Fourier transform. The following sections introduce wavelet theory and applications. For a more detailed explanation of wavelets, the reader is referred to Burrus *et al*. [16], Chui [20], and Ogden [59].

2.6.3.1 Fourier Transforms.

Wavelets involve data transforms. Transforms engage data signals, often time series, to aid in their description and analysis. The most well-known transform is the Fourier transform, which decomposes the function into a frequency representation. This process transforms the function from the time domain to the frequency domain [59].

Although Fourier transforms have been used for hundreds of years, they possess limitations: namely, the inability to detect changes in frequency in the original signal. The Fourier transform can detect and specify which frequencies are contained within a signal, but cannot express when these frequencies occur in time. For this reason, the traditional Fourier transform requires stationary data [14].

To handle situations where the data are not stationary, a Windowed Fourier Transform (WFT), also known as the Short Time Fourier Transform, offers a potential solution. The WFT divides the signal into segments for which the signal is assumed stationary, and the Fourier transform is then performed on each of these stationary segments. However, this technique leads to other problems such as the tradeoff between the frequency and time resolution [20, 59].

This resolution tradeoff is captured by the Heisenberg Uncertainty Principle, which applies to both Fourier transforms and wavelet transforms in the following sense: the exact frequency at an exact instant in time cannot be known. For this reason, there is a tradeoff between the frequency and time resolutions of the signal. In the time domain, there is perfect time resolution since the exact value of the signal is known at every instant in time, but perfect frequency resolution is absent. After a Fourier transform, the function is in the

frequency domain, resulting in perfect frequency resolution, but the time resolution is zero. For the WFT the window is of finite length, which causes a loss in frequency resolution. In other words, frequency and time resolutions are restricted based on the necessary size of the window [59].

2.6.3.2 Wavelet Theory.

The wavelet transform offers a solution to the problems associated with the Fourier transform and the WFT. Wavelets are still bound by the uncertainty principle that the exact frequency and time cannot be known simultaneously, but wavelets make it possible to know what frequency bands exist at what time intervals. This time-frequency localization enables the wavelet transform to provide information on the frequency and time content simultaneously and can handle changes in frequency within the data. This renders wavelets suitable to transform non-stationary data [59].

Wavelets are a special type of function, in particular a basis function, which are localized in both time and frequency. Wavelets are typically defined using a mother wavelet (ψ) or wavelet function, which characterizes the basic wavelet shape. Figure 2.9 illustrates a few examples of some commonly used wavelet functions [30].

The wavelet function generates an entire family of wavelets through dilations and translations according to

$$\psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^j t - k) \quad j,k \in \mathbb{Z} \tag{2.36}$$

with dilation index (j) and translation index (k). The dilation index is also known as a scale factor, and the translation index is also known as a shift factor [59].

A linear combination of these shifted and scaled versions of the mother wavelet may represent a signal or function similar to a Fourier series. This representation is one form

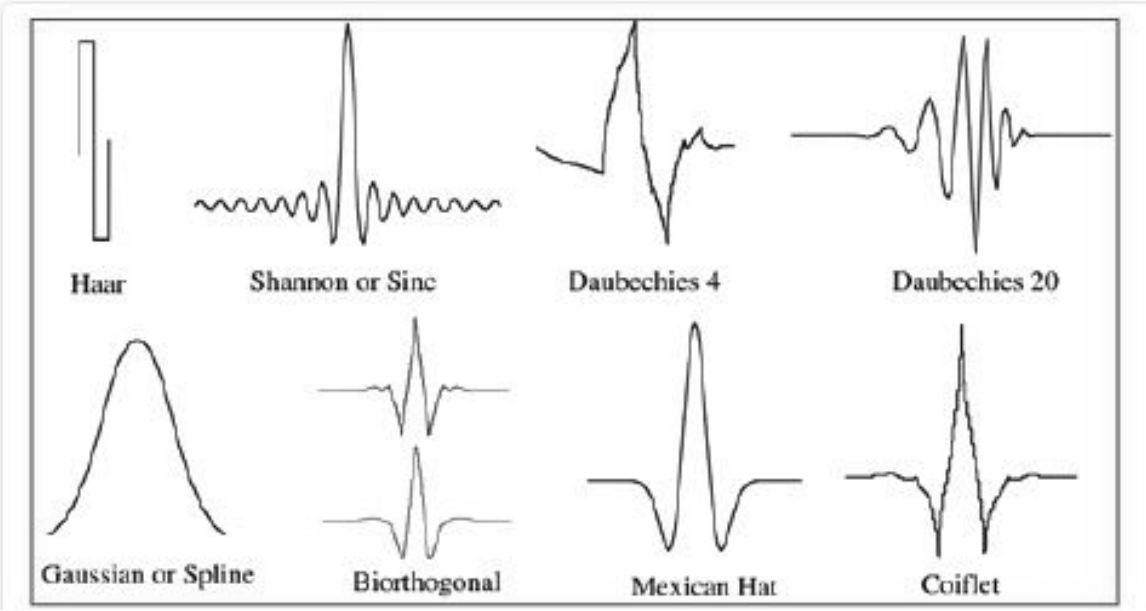


Figure 2.9: Wavelet Function Examples [30]

of the Discrete Wavelet Transform (DWT). Prior to delving further into the DWT, it is first useful to discuss the multiresolution analysis (MRA) property of wavelets.

2.6.3.3 Multiresolution Analysis (MRA).

Wavelets have a multiresolution property that allows analyzing a function at different levels of resolution, which results from the time-frequency localization ability of wavelets. To better understand MRA, first define a scaling function,

$$\phi_{j,k}(t) = 2^{\frac{j}{2}}\phi(2^j t - k) \quad j,k \in \mathbb{Z} \tag{2.37}$$

with scale factor (j) and shift factor (k) according to Equation 2.37. The scaling function is also known as the father wavelet [16].

Next, define a sequence of closed subspaces of $L^2(\mathcal{R})$,

$$\{0\} \subset ... \subset V_{-1} \subset V_0 \subset V_1 \subset ... \subset L^2(\mathcal{R}) \,. \tag{2.38}$$

A subspace of $L^2(\mathcal{R})$, V_j, is spanned by the set of scaling functions,

$$V_j = \overline{Span_k(\phi_{j,k}(t))} \,. \tag{2.39}$$

Therefore, a function $f(t) \in V_j$ can be expressed using a linear combination of the scaling functions [16],

$$f(t) = \sum_k a_k \phi_{j,k} \,. \tag{2.40}$$

For $j > 0$, the span can be larger since $\phi_{j,k}(t)$ is narrower, so it can represent finer detail or higher resolution. For $j < 0$, $\phi_{j,k}(t)$ is wider, so the space the scaling functions span is smaller. The wider scaling functions can represent only coarse or low resolution information. The nature of the nested subspaces in Equation 2.38 show that the space which contains high resolution signals will contain those of lower resolution also. Figure 2.10 depicts these spanned spaces [16].

Based on the definition of V_j, the spaces satisfy the scaling condition,

$$f(t) \in V_j \;\leftrightarrow\; f(2t) \in V_{j+1} \tag{2.41}$$

where the elements of a space are just scaled versions of elements in the next space. Therefore the scaling function, $\phi(t)$ may be expressed as a linear combination of shifted versions of $\phi(2t)$,

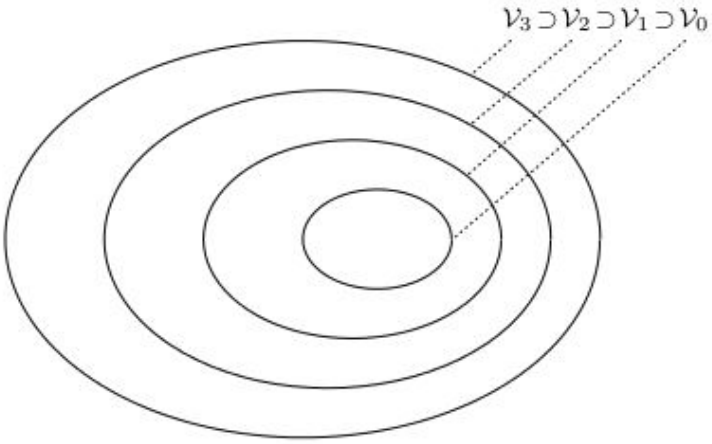


Figure 2.10: Nested Vector Spaces Spanned by Scaling Functions [16]

$$\phi(t) = \sum_n h(n)\,\sqrt{2}\phi(2t - n), \quad n \in \mathbb{Z} \tag{2.42}$$

where *h(n)* are the scaling function coefficients [16].

The wavelet function, $\psi(t)$, may aid the scaling function in the description of a signal. The wavelet function spans the differences between the spaces spanned by the scaled versions of the scaling function. Set the wavelet functions to be orthogonal to the scaling function and define the orthogonal complement of V_j in V_{j+1} as W_j. W_j is a wavelet spanned subspace such that

$$V_{j+1} = V_j \oplus W_j\,. \tag{2.43}$$

Figure 2.11 depicts these spanned subspaces [16].

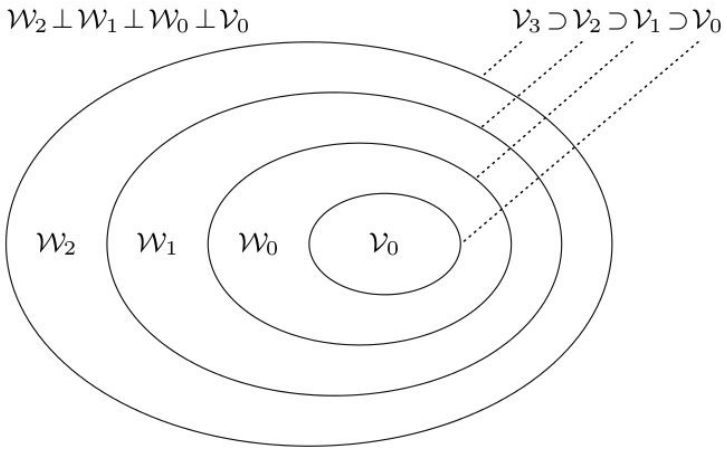


Figure 2.11: Scaling Function and Wavelet Vector Spaces [16]

Linear combinations of shifted versions of $\phi(2t)$ may represent the wavelets that span W_j, since they also reside in the space V_{j+1}, which is the next narrower scaling function space, using coefficients $h_1(n)$,

$$\psi(t) = \sum_n h_1(n)\,\sqrt{2}\phi(2t-n), \quad n \in \mathbb{Z}\,. \tag{2.44}$$

Since wavelet functions span the orthogonal complement spaces, the wavelet function coefficients are related to the scaling function coefficients according to [16],

$$h_1(n) = (-1)^n h(1-n)\,. \tag{2.45}$$

Together, the wavelet and scaling functions may span all of $L^2(\mathcal{R})$, so that

$$f(t) = \sum_k c_{j_0,k}\phi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{\infty} d_{j,k}\psi_{j,k}(t) \tag{2.46}$$

represents any function $f(t) \in L^2(\mathcal{R})$. The choice of j_0 sets the coarsest scale whose space is spanned by the scaling functions [16]. This leads into the Discrete Wavelet Transform.

2.6.3.4 Discrete Wavelet Transform (DWT).

The DWT of a signal generates the coefficients in Equation 2.46, $c_{j,k}$ and $d_{j,k}$. Typically, this signal is a discrete sample from a function, f. These wavelet coefficients, which are estimated via inner products

$$c_{j,k} = \langle f(t), \phi_{j,k} \rangle = \int f(t)\phi_{j,k}(t)dt \tag{2.47}$$

and

$$d_{j,k} = \langle f(t), \psi_{j,k} \rangle = \int f(t)\psi_{j,k}(t)dt \tag{2.48}$$

completely describe the original signal [16]. The DWT plays an important role in the wavelet decomposition of a signal, which is described in Section 2.6.3.5.

2.6.3.5 Wavelet Decomposition.

A useful tool of wavelet analysis is the separation of the high-frequency content and low-frequency content of a signal. This is often referred to as the wavelet decomposition of a signal. This wavelet decomposition acts by breaking up a signal into the low-frequency approximation, which is described by the scaling function, and the high-frequency details, which are described by the wavelet function. A signal, $f(t)$, that exists in the space V_{j+1} can be expressed using solely scaling functions at a scale of $j + 1$. The act of lowering the resolution by one level decomposes the signal into an approximation component and a detail component. Define the approximation component by the scaling functions at a scale of j, while the scale j wavelet functions describe the rest of the signal. Therefore, describe the signal using [16],

$$f(t) = \sum_k c_{j,k}\phi_{j,k}(t) + \sum_k d_{j,k}\psi_{j,k}(t)\,. \tag{2.49}$$

This example of a single level decomposition leads to another issue, however. This process doubles the number of data points in the original signal. To prevent the uncontrolled growth in data size, downsampling ensures that the number of data points is maintained. It is now possible to decompose the signal further, resulting in coarser approximations of the original signal, as illustrated by Figures 2.12 and 2.13 [16].

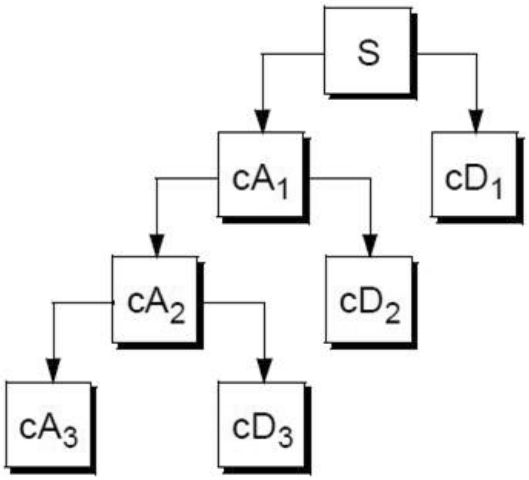


Figure 2.12: Wavelet Decomposition Process [48]

Therefore, further decompositions of a signal are possible until the desired resolution level is attained. A signal, S, may be decomposed into the level j approximation (A_j) which lives in the space V_j and the signal details (D_j), according to

$$S = A_j + \sum D_j\,. \tag{2.50}$$

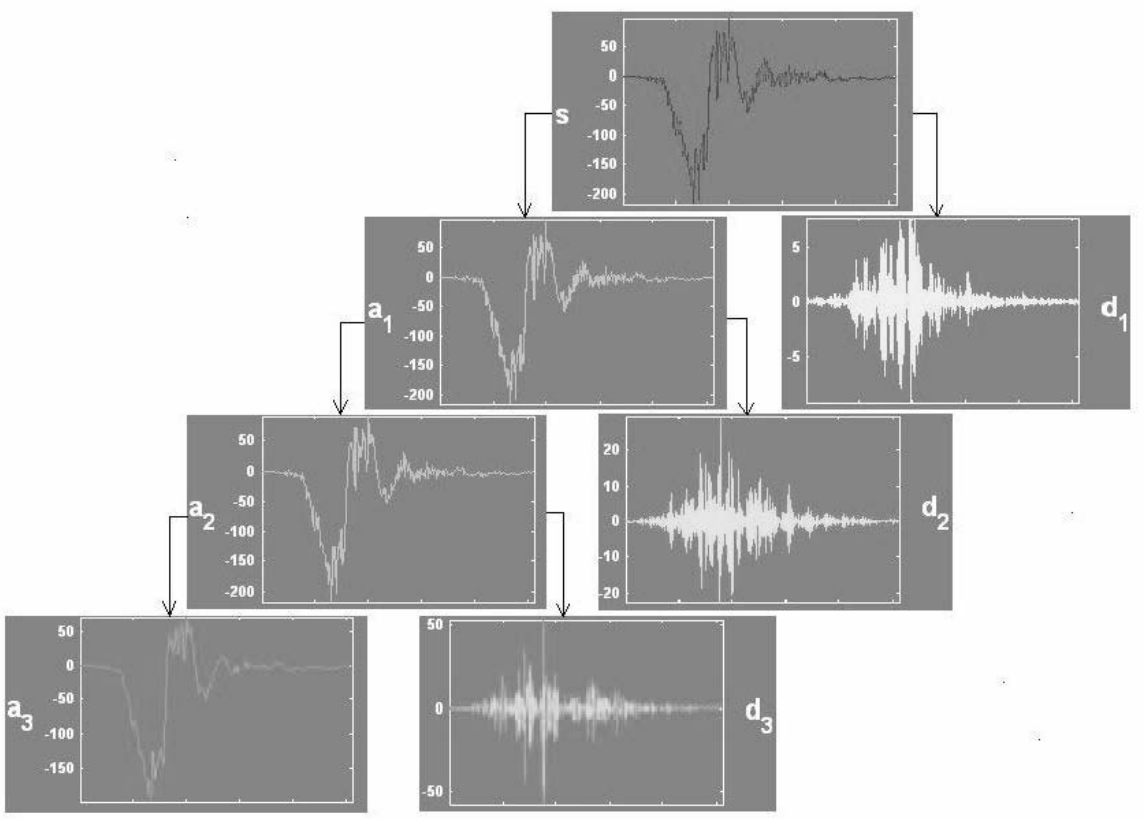


Figure 2.13: Wavelet Decomposition of Signal Example [19]

It is appropriate to consider j a resolution level, or a decomposition level. Define the signal approximation,

$$A_j(t) = \sum_k c_{j,k}\phi_{j,k}(t) \tag{2.51}$$

and detail,

$$D_j(t) = \sum_k d_{j,k}\psi_{j,k}(t) \tag{2.52}$$

using the scaling and wavelet functions [16, 59].

2.6.3.6 Properties of Wavelet Analysis.

Several properties promote the popularity of wavelets for data analysis. First, their multiresolution property allows users to focus on the local, high-frequency content of a signal, making them better suited to analyze non-stationary or transient data. Second, they are a computationally efficient technique, with a transform requiring *O(N)* operations. Third, they offer the flexibility of selecting a family of wavelet functions suited to the particular analysis objectives. Fourth, users may reverse the wavelet decomposition of a signal into wavelet coefficients via the inverse wavelet transform and reconstruct the original signal with no loss of information. Fifth, since wavelets operate in the time-frequency domain, they decorrelate autocorrelated observations in a signal. Last, the sparsity property of wavelets implies that the wavelet representation of a signal or function requires only a small number of wavelet coefficients. This makes wavelets suitable as a data compression technique. Further, this sparsity property combined with the orthogonality of the transform render wavelets effective as a de-noising technique by a process called wavelet thresholding, as described further in the next section [16, 59].

2.6.3.7 Wavelet Thresholding.

Wavelet thresholding is a technique to compress or de-noise a signal. Donoho and Johnstone [25] introduce the concept, proposing a technique of uncovering the true underlying function from a noisy sample. The noisy sample has normal, independently distributed random error whose variance may be known or unknown. Their technique relies on the idea of wavelet sparsity, where most of a clean signal's energy is concentrated in a small subset of the wavelet coefficients and the remaining coefficients are zero. If the signal is contaminated with noise, the linearity of the DWT results in noisy estimated coefficients whose errors are transformations of the original observed errors. These wavelet

coefficients that were previously equal to zero are now primarily nonzero. By identifying a value representing the wavelet coefficient noise, the wavelet coefficients may be modified or thresholded resulting in a de-noised signal.

Researchers have suggested several thresholding methods to optimally de-noise a signal. A crude de-noising approach takes approximations of the signal as the de-noised representation of the signal [19, 48]. However, this technique discards all the high-frequency information in the signal, causing the loss of many of the original signal's sharpest features. Donoho and Johnstone propose multiple thresholding techniques including RiskShrink, which is a procedure that uses a minimax threshold to reduce the "risk" or mean squared error between the true underlying function and the de-noised reconstruction [25]. They also develop a technique called SureShrink that uses a level-dependent threshold to minimize the Stein Unbiased Estimate of Risk (SURE) [24]. Nason [54] develops a cross-validation technique that forms two subsequences from the original noisy data set and uses a cross-validation score to minimize the integrated square error (ISE) of a reconstruction. Cai and Silverman [17] devise Neighblock, which increases estimation precision by utilizing information about neighboring wavelet coefficients. McGinnity *et al*. [45] improve upon Nason's method by developing a distribution-free method which employs block thresholding and level dependence. Further information regarding these thresholding techniques may be found in the references provided. This review will focus on the most common method, VisuShrink, developed by Donoho and Johnstone [25].

VisuShrink uses what Donoho and Johnstone [25] refer to as a universal threshold,

$$\lambda = \hat{\sigma}\sqrt{2log(n)} \tag{2.53}$$

where $\hat{\sigma}$ is an estimate of the standard deviation of the noise and n is the sample size of the

data. The universal threshold is based on principle that if z_i is a white noise sequence that is independent and identically distributed *N(0,1)*, then as $n \to \infty$,

$$pr\left(\max_i |z_i| > \sqrt{2log(n)}\right) \to 0\,. \tag{2.54}$$

Therefore, using this technique there is a high probability that each wavelet coefficient that should be zero in the absence of noise is correctly estimated as zero. Estimate the noise, $\hat{\sigma}$, using the Median Absolute Deviation (MAD) of the finest scale detail coefficients divided by 0.6745, since these finest scale detail coefficients are essentially pure noise. Once the universal threshold is calculated, apply a soft thresholding approach so that the estimated coefficients, $\tilde{\theta}$, are replaced with the thresholded coefficients [25],

$$\hat{\theta} = \begin{cases} 0, & if\ |\tilde{\theta}| \le \lambda \\ \tilde{\theta} - \lambda, & if\ \tilde{\theta} > \lambda \\ \tilde{\theta} + \lambda, & if\ \tilde{\theta} < -\lambda\,. \end{cases} \tag{2.55}$$

Donoho [23] states that this de-noised reconstruction is at least as smooth as the true, underlying function in any of a wide variety of smoothness measures and that it comes nearly as close in mean square to the function as any measurable estimator can come. In summary, the VisuShrink thresholding method which uses universal, soft thresholding acts as an effective de-noising process on a signal contaminated with pure error.

2.6.3.8 Wavelet Packets.

In addition to the traditional wavelet analysis, wavelet packets offer further analysis opportunities. Wavelet packets are a generalization of the traditional wavelet basis. Linear combinations of wavelet functions form wavelet packets. Wavelet packets inherit the properties of orthonormality and smoothness from wavelet functions. In addition, they

offer additional flexibility since a large number of possible wavelet packet bases may be constructed allowing the selection of a best basis for an application. This "best basis" or best representation is typically chosen with a user-defined criterion function, such as an entropy criterion [59].

A wavelet packet function,

$$w^m_{j,k}(t) = 2^{\frac{j}{2}} w^m(2^j t - k) \tag{2.56}$$

has three indices: j, k, and m (Equation 2.56). The indices j and k represent the scale and shift factor, respectively, while the index m is called the modulation parameter or oscillation parameter [59].

The set of all wavelet packet functions contains too many elements to form an orthonormal basis, so a subset of this collection is chosen. Let this set I represent the appropriate indices, so that the decomposition of a function into its wavelet packet components is given by

$$f(t) = \sum \sum_{(m,j)\in I} \sum_{k\in\mathbb{Z}} a^m_{j,k} w^m_{j,k}(t) \, . \tag{2.57}$$

The coefficients, $a^m_{j,k}$ are calculated as [59]

$$a^m_{j,k} = \int_{-\infty}^{\infty} f(t) w^m_{j,k}(t) dt \, . \tag{2.58}$$

A best basis algorithm selects the most appropriate set of basis functions to represent a function. In traditional wavelet analysis, when a function is decomposed using wavelets it generates a wavelet decomposition tree. In this wavelet decomposition tree, only the approximation of the signal is decomposed with each successive resolution level. In

wavelet packet analysis, the details of the signal are also successively decomposed. The appropriate branches of the decomposition tree are selected to identify the best basis. There are a variety of methods for choosing the best basis or "best tree." Perhaps most common is to use an entropy criterion, such as the Shannon entropy measure [59]. Today, many software packages automatically calculate this entropy measure and identify the best wavelet packet basis for a given function or signal.

Wavelet packets are also useful in that they may be used to calculate the wavelet packet component energy, which is a measure of the signal energy content contained in some specific frequency band. In a wavelet transform, the total energy of a signal, E^f, is partitioned into various time-frequency components, E_m^f, and this energy may be realized in the wavelet packet coefficients, according to Parseval's theorem [16, 40],

$$E^f = \sum_{m=1}^{2^j} E_m^f = \sum_{m=1}^{2^j} \int_{-\infty}^{\infty} [a_{j,k}^m(t)]^2 dt \, . \tag{2.59}$$

This wavelet packet component energy can help in de-noising or compression using wavelet packets. The wavelet packet component energy can serve as a feature to help determine whether a branch of the wavelet decomposition tree is worth retaining in a compressed or de-noised reconstruction. A component with large energy may be thought of as containing much of a signal's content, while a component with low energy may feature very little signal content.

2.6.4 Wavelet-Based Simulation Validation.

Wavelets are a valuable tool for simulation model validation in a variety of ways. One use is the de-noising capability of wavelets to remove the pure error from the system and model data prior to performing a validation assessment. A second way is to use a wavelet coherence measure to assess the similarity of two signals. Third, the wavelet packet component energy may be used as a feature to compare two data signals. Finally, wavelets

may serve to decorrelate autocorrelated data so that hypothesis testing or ANOVA may be performed outside the time-domain.

2.6.4.1 Validation Metrics Based on Wavelet Approximations.

Using many of the concepts discussed in Section 2.6.3, Cheng *et al*. [19] use wavelets to validate their simulation model. Their work models biodynamic responses, such as impact responses, notable for their short duration and transient nature. The authors judge conventional correlation analysis as an insufficient technique for model validation due to the responses' transient nature and contamination with noise. The authors propose wavelets as a method to represent the data and perform correlation analysis.

The authors use wavelets to decompose the test data and simulation data to a specified decomposition level j, as shown in Figure 2.14. Next, they calculate the maximum cross-correlation coefficient between these two wavelet-decomposed signals. They combine this value with measures of magnitude error and phase error to develop an overall validity metric,

$$R^j = \left[\alpha_1(1-\rho^j) + \alpha_2\left|\frac{\tau^j}{T}\right| + \alpha_3\left|\frac{A_s^j - A_m^j}{A_s^j}\right|\right] \times 100\% \; . \tag{2.60}$$

The α_i's represent weighting coefficients, ρ is the correlation coefficient, τ is the time lag, T is the pulse duration, and A represents the amplitude for system (s) and model (m) [19].

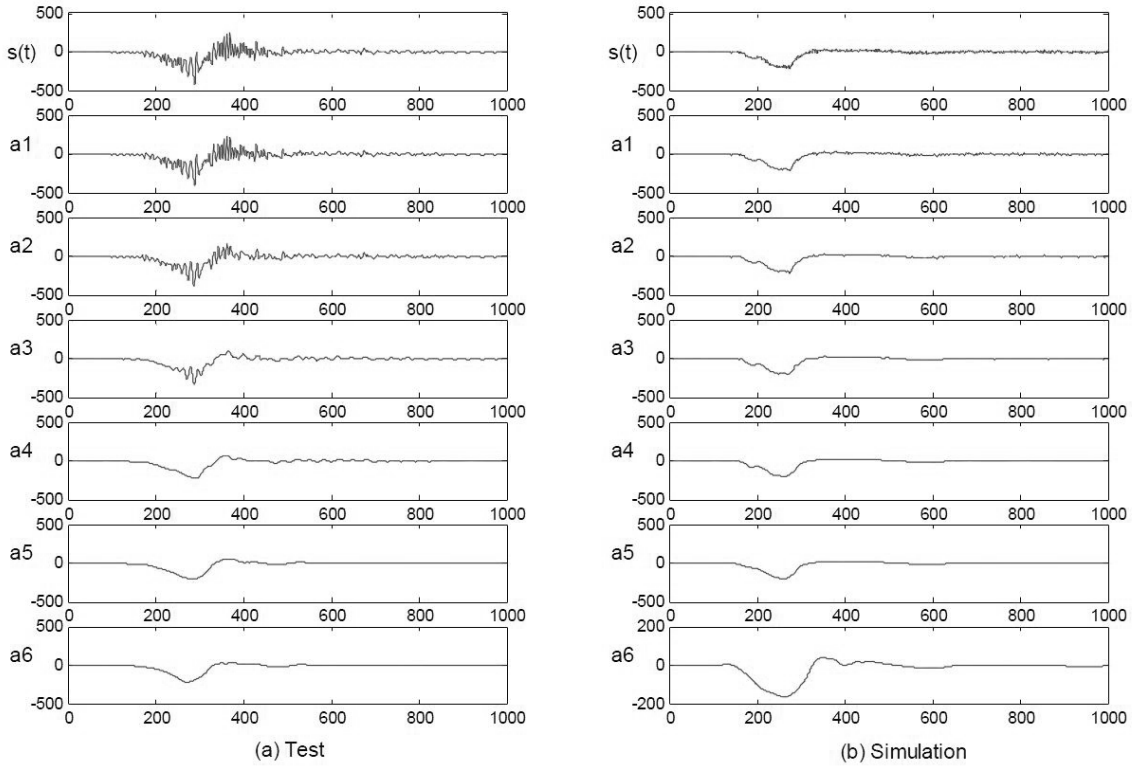


Figure 2.14: Wavelet Validation [19]

Note that this R^j value may be calculated for each level of decomposition, j. A smaller value of R^j represents better agreement between the system and model results. Intuitively, a highly decomposed signal (larger value of j) will result in a greater level of agreement between results, i.e. a smaller value of R^j. Consequently, the authors develop an algorithm (Figure 2.15) which sets a threshold value for R^j and a maximum decomposition level, L. If the R^j value is lower than the threshold before decomposition to level L, then the authors conclude that the model is valid [19].

2.6.4.2 Wavelet Coherence.

Articles by Torrence and Compo [81] and Grinsted *et al*. [37] introduce new wavelet analysis techniques applied to time-series data associated with weather patterns. In

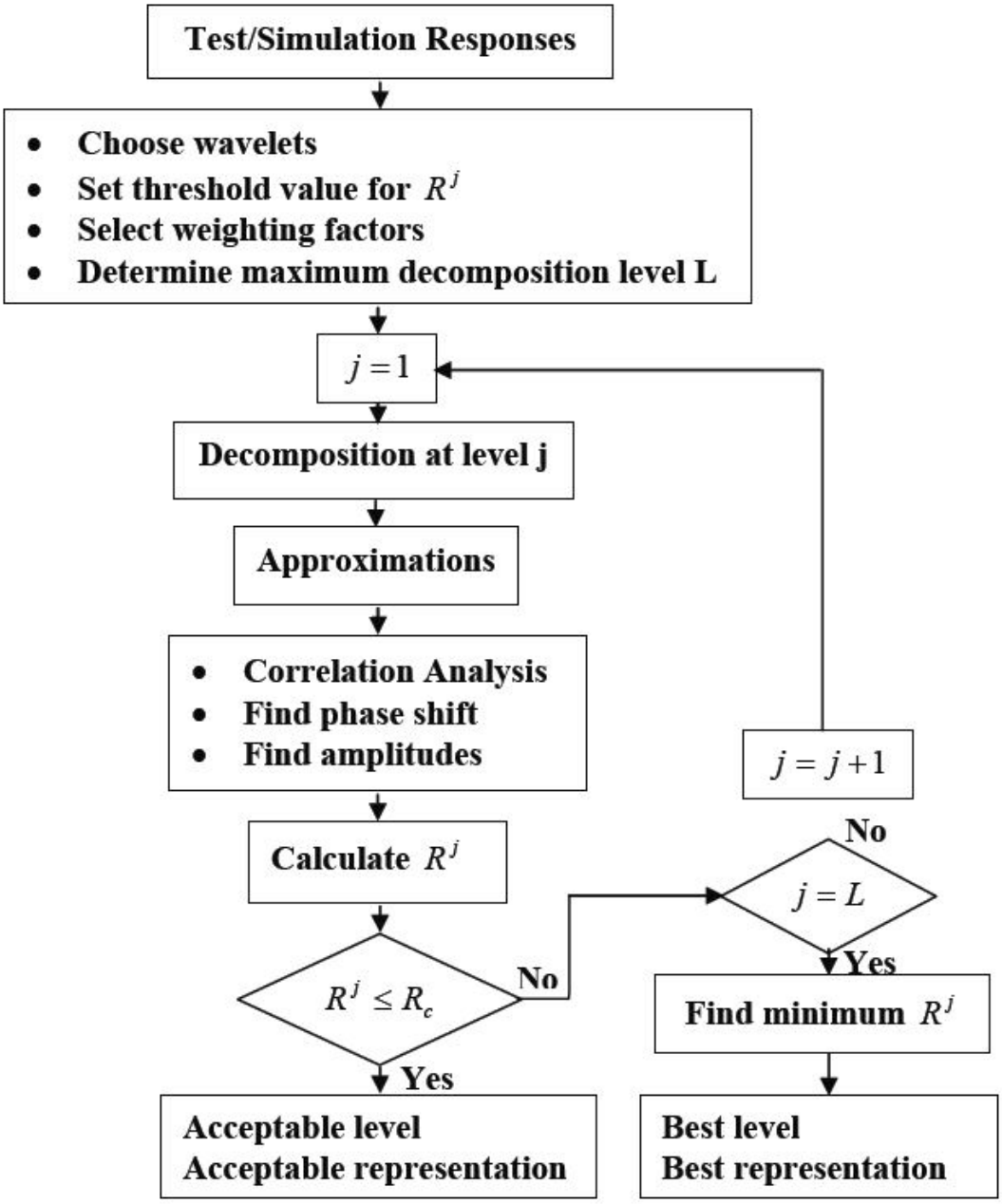


Figure 2.15: Wavelet Validation Algorithm [19]

particular, they use wavelet coherence, a quantitative measure of the similarity between two sets of time-series data. Jiang and Mahadevan further develop this analysis technique

for the purpose of model validation. Jiang and Mahadevan [41] identify coherence as a direct measure of the correlation between the spectra of two processes, which can be used to quantify the synchronization of their amplitude and phase. Traditionally, this coherence measure is calculated using Fourier analysis. However, Fourier-based coherence analysis has the following limitations: it assumes stationarity; the effects of amplitude and phase are not separated; and it has no time resolution. For these reasons, the authors propose a wavelet-based coherence measure, which is a quantitative measure of the correlation in the time-frequency domain.

Jiang and Mahadevan [41] use a complex, Morlet wavelet to transform the system data ($y(t_j)$) and the model data ($z(t_j)$) to obtain the associated wavelet coefficients, $W_y(a,b)$ and $W_z(a,b)$, respectively. Calculate the wavelet power spectrum for the system data,

$$S_s(a,b) = |W_s(a,b)|^2 \,. \tag{2.61}$$

Obtain the time-frequency cross-spectrum,

$$S_{yz}(a,b) = W_y(a,b)W_z(a,b) \,. \tag{2.62}$$

The time-frequency coherence,

$$C_{yz}^2(a,b) = \frac{|\bar{S}_{yz}(a,b)|^2}{[\bar{S}_y(a,b)]^2} \,, \tag{2.63}$$

is the squared absolute value of the smoothed cross-wavelet spectrum between the system and model data, normalized by the squared, smoothed wavelet power spectrum of the system data. The authors then perform a significance test on the wavelet time-frequency coherence to determine whether the coherence is equal to one or not. Monte Carlo

simulations are performed to estimate the distribution. If the data support the hypothesis that the coherence is equal to one, then the model is assessed as valid [41].

2.6.4.3 Wavelet Packet Based Validation.

Several authors use wavelet packets in their validation efforts, taking advantage of their added flexibility. Cheng *et al*. [18] modify their validation approach described in Section 2.6.4.1 by incorporating wavelet packets into their analysis. Once the system and model data are transformed using wavelet packets, calculate the total signal energy,

$$\|S\|^2 = \sum \sum_{(j,n)\in I} \sum_{k} q_{jnk}^2 \tag{2.64}$$

where q_{jnk} represent the wavelet packet coefficients. The authors then use the wavelet packet coefficient indices to understand the distribution of energy across the signal; this energy distribution describes the variation of amplitude with respect to frequency and time. By comparing the energy distributions of the system and model data, they accomplish one step in evaluating simulation results to validate a model. The authors also suggest combining this energy distribution comparison with the calculation of the cross-correlation coefficient and lag.

Jiang and Mahadevan [40] also develop a wavelet packet-based validation technique for dynamic systems. The authors use a wavelet packet transform to decompose both the system and model data and then compute the wavelet packet component energies from each of the two datasets to be used as a signal feature. As a signal feature, the wavelet packet component energy identifies where the majority of a signal's content is located (in both time and frequency), so that the signals may be reconstructed using only the significant features. These reconstructions are first assessed using cross-correlation and cross-coherence to ensure that the features effectively represent the original signals. Then, the authors use a Bayesian hypothesis test to determine whether the extracted features from

the system and model are the same. In summary, the use of the wavelet packet component energy as a feature extraction method is very similar to a signal compression or de-noising method using traditional wavelets.

2.6.5 Functional Data Analysis.

Ramsay and Silverman [64] provide extensive coverage on the field of functional data analysis in their text. Functional data typically consist of functional observations of a dependent variable paired with an independent variable, such as time-series data. The authors' goals for functional data analysis include studying patterns and variations in the data; explaining the variation of a dependent variable by using independent variable information; comparing two or more sets of data with respect to certain types of variation; and confirmatory analyses such as hypothesis testing. For simulation validation, functional data analysis provides opportunities to analyze and compare two sets of time-series data. The authors describe this type of analysis as a Functional Analysis of Variance (FANOVA). This section introduces the FANOVA technique, followed by variations on the technique in which the analysis is conducted outside of the time domain.

2.6.5.1 Functional Analysis of Variance (FANOVA).

FANOVA falls under the category of functional linear models, which also includes functional multiple regression. Linear models may be functional in two ways: the response variable, y, with argument, t, is functional, and/or the independent variable(s), x, is functional. FANOVA will focus on a functional response with a categorical or scalar independent variable [64].

Ramsay and Silverman [64] use an example to illustrate the FANOVA approach. There are multiple weather stations in different climate zones in Canada, all recording the temperature over time, t. The climate zone, i, is the treatment, while each recorded temperature function in a climate zone is a replicate, l. The traditional effects model for the temperature function is

$$y_{il}(t) = \mu(t) + \alpha_i(t) + \epsilon_{il}(t) \tag{2.65}$$

where $i = 1, ..., p$ and $l = 1, ..., n_i$. The function μ is the grand mean function, representing the average temperature profile for Canada. The treatment effects, α_i, represent the deviation from the mean due to the climate zone. ϵ_{il} represents the residual of replicate l of treatment i, and these errors are assumed to be normal, independently distributed.

The standard least square estimators are

$$\hat{\mu}(t) = \overline{y}_{\cdot\cdot}(t) \tag{2.66}$$

and

$$\hat{\alpha}_i(t) = \overline{y}_{i\cdot}(t) - \overline{y}_{\cdot\cdot}(t) \, . \tag{2.67}$$

The estimators are obtained by minimizing the functional version of the residual sum of squares,

$$LMSSE = \sum_t \sum_{i,l} [y_{il}(t) - (\mu(t) + \alpha_i(t))]^2 \tag{2.68}$$

subject to the constraint [64],

$$\sum_i n_i \alpha_i(t) = 0 \, . \tag{2.69}$$

The fundamental ANOVA identity involving the decomposition of the total sum of squares remains the same. The total sums of squares,

$$SST(t) = \sum_{i,l} [y_{il}(t) - \hat{\mu}(t)]^2 \tag{2.70}$$

and error sums of squares,

$$SSE(t) = \sum_{il} [y_{il}(t) - \bar{y}_{i.}(t)]^2 \tag{2.71}$$

are used to calculate the mean squared for treatment,

$$MSTr(t) = \frac{SST(t) - SSE(t)}{p - 1} \tag{2.72}$$

with the F-ratio equal to [64]

$$F_0 = \frac{MSTr}{SSE/(n-p)} \,. \tag{2.73}$$

This F-ratio is compared to a critical F-value for a specific level of significance and the appropriate degrees of freedom. This comparison evaluates whether the climate zone has a significant effect on the temperature at time, t. Of particular note is that this FANOVA process is essentially a univariate ANOVA problem for each specific value of t. This technique does not provide an overall assessment on the significance of treatment effect on the overall temperature profile [64].

This somewhat naïve approach of executing the standard univariate ANOVA for each specific t followed by a series of pointwise F-tests can lead to several issues. If the functional data have a large dimension, this leads to a large number of hypothesis tests causing a serious multiplicity problem. This can result in an uncontrolled Type I error or false positive rate. Meanwhile, a Bonferroni procedure would yield an extremely low power

of the test [1]. Other authors have proposed methods to solve this problem, as described in the following sections.

2.6.5.2 FANOVA using a Multivariate Statistic.

Girimurugan *et al*. [32] develop a FANOVA model based on a multivariate statistic instead of the univariate statistic reviewed in the preceding section. This multivariate statistic has a Hotelling T^2 distribution, similar to a MANOVA approach. Given a multivariate response of dimension n, let the noise $\boldsymbol{\epsilon}_{ij} \in \mathbb{R}_n$ have a multivariate normal random distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ is defined as $\sigma^2 \boldsymbol{I} \in \mathbb{R}^{nxn}$. Then the response Y_{ijk}, for treatment, $i = 1, 2, ..., t$, and replicate, $j = 1, 2, ..., r_i$, is

$$Y_{ijk} = \mu + \alpha_{ijk} + \boldsymbol{\epsilon}_{ijk} \, . \tag{2.74}$$

The ANOVA decomposition of the total sum of squares is

$$\sum_{i=1}^{t} \sum_{j=1}^{r_i} (Y_{ij} - \overline{Y}_{..})'(Y_{ij} - \overline{Y}_{..}) = \sum_{i=1}^{t} (\overline{Y}_{i.} - \overline{Y}_{..})'(\overline{Y}_{i.} - \overline{Y}_{..}) + \sum_{i=1}^{t} \sum_{j=1}^{r_i} (Y_{ij} - \overline{Y}_{i.})'(Y_{ij} - \overline{Y}_{i.}) \tag{2.75}$$

with the corresponding degrees of freedom,

$$n(\varphi - 1) = n(t - 1) + n(\varphi - t) \tag{2.76}$$

where $\varphi = \sum_{i=1}^{t} r_i$ [32].

Then let the Hotelling-FANOVA statistic,

$$\vartheta = \frac{(\varphi - t) \sum_{k=1}^{n} (\overline{Y}_{i.k} - \overline{Y}_{...})'(\overline{Y}_{i.k} - \overline{Y}_{...})}{(t - 1) \sum_{i=1}^{t} \sum_{j=1}^{r_i} (Y_{ij} - \overline{Y}_{i.})'(Y_{ij} - \overline{Y}_{i.})} \tag{2.77}$$

be used to test for differences in a functional response. ϑ follows an F distribution with

degrees of freedom $n(t-1)$ and $n(\varphi - t)$ under the null hypothesis that the functional responses are statistically equivalent. Therefore a significant test would lead to the conclusion that the treatment yields a significant difference in the functional data [32].

2.6.5.3 High-Dimensional Analysis of Variance (HANOVA).

Fan [26] and Fan and Lin [27] propose a new test of significance for comparing functional data based on the adaptive Neyman test and wavelet thresholding techniques. The authors first discuss some of the weaknesses of previous FANOVA approaches, citing low test power and an inability to handle high-dimensional data. They propose two approaches for testing sets of curves based on data transforms, which they call HANOVA. The first approach uses a Fourier transform to decorrelate the data and isolate the existing frequencies. They next apply an adaptive Neyman test to a subset of the calculated Fourier coefficients. The adaptive Neyman test statistic identifies whether there is any statistically significant difference among the set of curves. However, since the Neyman test statistic does not follow a known distribution, the critical value must be determined via simulation. Their second approach uses a DWT on the data and thresholds the resulting wavelet coefficients to reduce the dimensionality of the data. The authors calculate a thresholding test statistic which follows an approximate standard normal distribution. The authors conclude by recommending the adaptive Neyman test if the underlying mean curves are reasonably smooth and the wavelet thresholding procedure otherwise.

2.6.6 Wavelet-Based ANOVA Models.

Researchers have begun investigating functional analysis techniques in the wavelet domain. These wavelet-based ANOVA models offer several benefits, as well as solutions to the potential problems associated with functional analysis in the time domain. Ramsay and Silverman [64] discuss the importance of smoothing when implementing functional data analysis. Smoothing, or regularization, helps to remove some of the observational error associated with a discrete set of measured data. The second issue, previously alluded to, is

dimensionality. Functional data are typically high-dimensional and can lead to hypothesis testing with extremely low power [83]. The last issue is that time-series data are typically autocorrelated, which may lead to problems with the statistical analysis if not properly accounted for. Performing the functional analysis in the wavelet domain offers solutions to all the aforementioned problems. The following sections review a few of these analysis techniques.

2.6.6.1 Wavelet ANOVA (WANOVA) proposed by Vidakovic.

Vidakovic [83] develops a wavelet ANOVA (WANOVA) technique based on the FANOVA framework described by Ramsay and Silverman [64]. He outlines the FANOVA model, described previously in Equations 2.65 to 2.73. Next, let the vector **d** represent the DWT of the set of observations, **y**; **d** is the vector of wavelet coefficients, with elements *d(j,k)* representing the wavelet coefficient with scale factor *j* and shift factor *k*. Due to the linearity and orthogonality of the wavelet transform, the wavelet coefficients for treatment *i* and replicate *l* are

$$d_{il}(j, \boldsymbol{k}) = \theta_i(j, \boldsymbol{k}) + \epsilon'_{il}(j, \boldsymbol{k}) = \theta(j, \boldsymbol{k}) + \tau_i(j, \boldsymbol{k}) + \epsilon'_{il}(j, \boldsymbol{k}) \, . \tag{2.78}$$

If $\hat{\theta}$ and $\hat{\tau}_i$ are least-square estimators for θ and τ_i, then

$$\hat{\mu} = \mathbb{W}^{-1}\hat{\theta} \tag{2.79}$$

and

$$\hat{\alpha}_i = \mathbb{W}^{-1}\hat{\tau}_i, \quad i = 1, ..., p \, . \tag{2.80}$$

The energy preservation of orthogonal wavelet transforms also imply

$$\sum_{\boldsymbol{t}\in\boldsymbol{T}} MSE(\boldsymbol{t}) = \sum_{j,\boldsymbol{k}\in\tilde{\mathcal{I}}} WMSE(j,\boldsymbol{k}) \tag{2.81}$$

and

$$\sum_{\boldsymbol{t}\in\boldsymbol{T}} MSTr(\boldsymbol{t}) = \sum_{j,\boldsymbol{k}\in\tilde{\mathcal{I}}} WMSTr(j,\boldsymbol{k})\,. \tag{2.82}$$

It is now possible to test the hypothesis

$$H_0: \;\; \theta_i(j,\boldsymbol{k}) = \theta(j,\boldsymbol{k}), \quad i = 1,...,p \tag{2.83}$$

to determine if there is a significant overall difference in sets of functional data [83].

Prior to developing the statistic to test the hypothesis, Vidakovic recognizes the need for regularization and dimension reduction of the data. Therefore, he recommends identifying a subset of the wavelet coefficients with the m largest magnitudes, since they contain the most energy. Obtain this subset, $\tilde{\tilde{\mathcal{I}}}$, using wavelet thresholding techniques. Then using this subset of wavelet coefficients, calculate the test statistic,

$$T_m^* = \frac{1}{\sqrt{2m(p-1)}}\left[\sum_{(j,\boldsymbol{k})\in\tilde{\tilde{\mathcal{I}}}}\sum_{i=1}^{p}\left(\frac{\overline{d}_i(j,\boldsymbol{k}) - \overline{d}(j,\boldsymbol{k})}{\hat{\sigma}^2(j,\boldsymbol{k})/\sqrt{n_i}}\right)^2 - m(p-1)\right] \tag{2.84}$$

where

$$\overline{d}_i(j,\boldsymbol{k}) = \frac{1}{n_i}\sum_{l=1}^{n_i} d_{il}(j,\boldsymbol{k})\,, \tag{2.85}$$

$$\overline{d}(j,\boldsymbol{k}) = \frac{1}{n}\sum_{i=1}^{p} n_i\overline{d}_i(j,\boldsymbol{k})\,, \tag{2.86}$$

and

$$\hat{\sigma}^2(j, \boldsymbol{k}) = WMSE(j, \boldsymbol{k}) \,. \tag{2.87}$$

This test statistic follows a χ^2-distribution under the null hypothesis. Reject the null hypothesis if [83]

$$T_m^* \geq \frac{1}{\sqrt{2m(p-1)}} \left(\chi^2_{m(p-1)}(1-\alpha) - m(p-1) \right) \,. \tag{2.88}$$

This WANOVA framework is an effective technique for identifying whether there is a statistically significant overall difference among sets of functional data while executing the test in the wavelet domain to take advantage of several wavelet properties.

2.6.6.2 WANOVA proposed by Girimurugan et al..

Girimurugan *et al*. [32] also propose a WANOVA procedure for detecting differences among functional data, but they develop a different test statistic with a different analysis objective in mind. The authors' primary interest is developing a technique that detects local and global profile changes and focus on the performance of the statistic in profile monitoring and change point detection problems. The basis for their approach is derived from the Hotelling-FANOVA statistic. The intent is to first adjust the Hotelling-FANOVA statistic by estimating the variation, σ, using the Median Absolute Deviation (MAD) instead of Mean Square Error (MSE). Second, the statistic is adjusted by estimating the sum of squares in the wavelet domain.

First, estimate σ using MAD,

$$\hat{\sigma} = \frac{1}{\Phi^{-1}(3/4)} \mathbb{M} \left[\left| W_{(J-1)k} - \mathbb{M}[W_{(J-1)k}] \right| \right] \tag{2.89}$$

where $\mathbb{M}$ is the median operator and only the finest scale detail coefficients, $W_{(J-1)k}$, are

used to represent the noisy components of the signal. Next, modify the Hotelling-FANOVA statistic by representing the sum of squares in the wavelet domain. This modified statistic is

$$\vartheta = \left(\hat{\sigma}^2(t-1)\right)^{-1} \sum_{i=1}^{t} \frac{1}{s_i} \mathbb{W}[\overline{Y}_{i.} - \overline{Y}_{..}]' \mathbb{W}[\overline{Y}_{i.} - \overline{Y}_{..}] \tag{2.90}$$

where $\mathbb{W}$ represents the DWT [32].

Let the coefficients for the treatment i effect be defined by

$$\theta_i = \hat{\sigma}^{-1} \mathbb{W}\left[\overline{Y}_{i.} - \overline{Y}_{..}\right] \tag{2.91}$$

and let $\tilde{\theta}_i = (\tilde{\theta}_{i1}, \tilde{\theta}_{i2}, ..., \tilde{\theta}_{iT_i})$ represent the thresholded version of these coefficients. The authors' proposed statistic,

$$\kappa_\eta = \sum_{i=1}^{t} \sum_{k=1}^{T_i} \tilde{\theta}_{ik}^2 \tag{2.92}$$

may be thought of as the "net energy" content corresponding to signal content among profiles [32].

The test statistic is used to test the null hypothesis that the set of t profiles corresponding to different treatments are statistically equivalent. This test statistic follows a complicated closed form distribution and is simulated. The statistic and associated WANOVA procedure is demonstrated to offer higher power with good control of Type I error [32]. More recently, Girimurugan *et al.* [34] modified this technique making it nonparametric and implements covariance shrinkage estimation and a block cross-validation method of thresholding.

2.6.6.3 Statistically Significant Contrasts Based on WANOVA.

McKay *et al*. [47] develop a procedure termed wavelet-based functional ANOVA (wfANOVA) to reveal statistically significant contrast curves between electromyographic (EMG) waveforms. Their goal is to compare neurophysiological signals. For example, the method of binning the time-series data to decrease the number of comparisons results in a loss of temporal resolution. However, performing ANOVA on every data point sacrifices the statistical power of the test. The authors reason that because the differences between curves may be represented by a small number of wavelet functions, then performing ANOVA in the wavelet domain may be most effective. The differences found in the wavelet domain are then transformed back to the time domain for visualization.

For this procedure, the collected time-series data should come from two or more treatments or conditions. Next, transform these signals from the time domain to the wavelet domain via the DWT. Due to the sparsity property of wavelets, these signals are well-represented by a small number of wavelet coefficients. These wavelet coefficients are then averaged for each treatment and compared among treatments via a traditional fixed-effects ANOVA model. The wavelet coefficients with a statistically significant difference are consolidated and transformed back into the time domain via the inverse wavelet transform. This provides a smooth, statistically significant contrast curve between treatments. Figure 2.16 depicts this procedure alongside a more traditional ANOVA approach (tANOVA) [47].

The authors apply their method to previously published EMG data to demonstrate its ability for determining differences in time-series data. They compare their wfANOVA technique to a traditional ANOVA method used to generate contrast curves. Figure 2.17 summarizes these results [47].

Unlike mean difference curves, the wfANOVA contrast curves identify only statistically significant differences in EMG signals. The tANOVA contrast curves also identify statistically significant differences in time periods generally similar to wfANOVA but

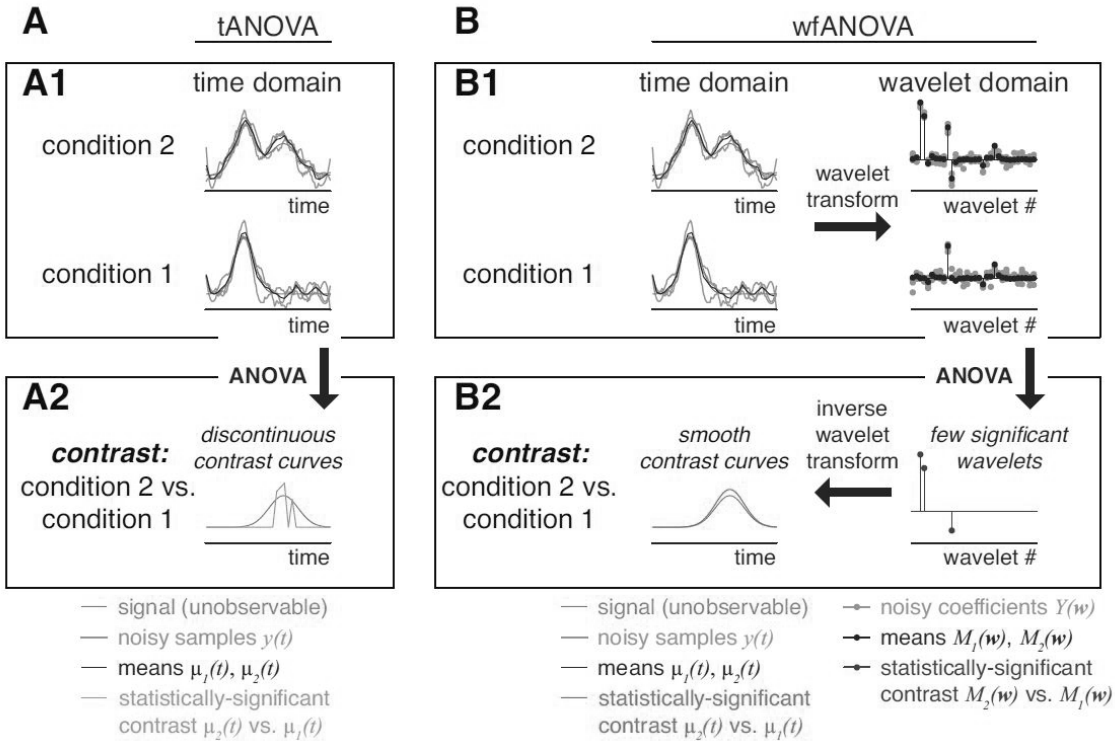


Figure 2.16: wfANOVA Procedure [47]

with features that are discontinuous. In summary, this wfANOVA approach is an effective technique for revealing the statistically significant contrast between signals from different treatments [47].

2.7 Summary and Future Work

This review of the literature introduces the need to conduct simulation validation efforts prior to relying on the results of a computer model. Early pioneers include Osman Balci and Robert Sargent, who define model validation as the "substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" [72]. While numerous authors have compiled a long list of model validation techniques, many agree that emphasis should be placed on

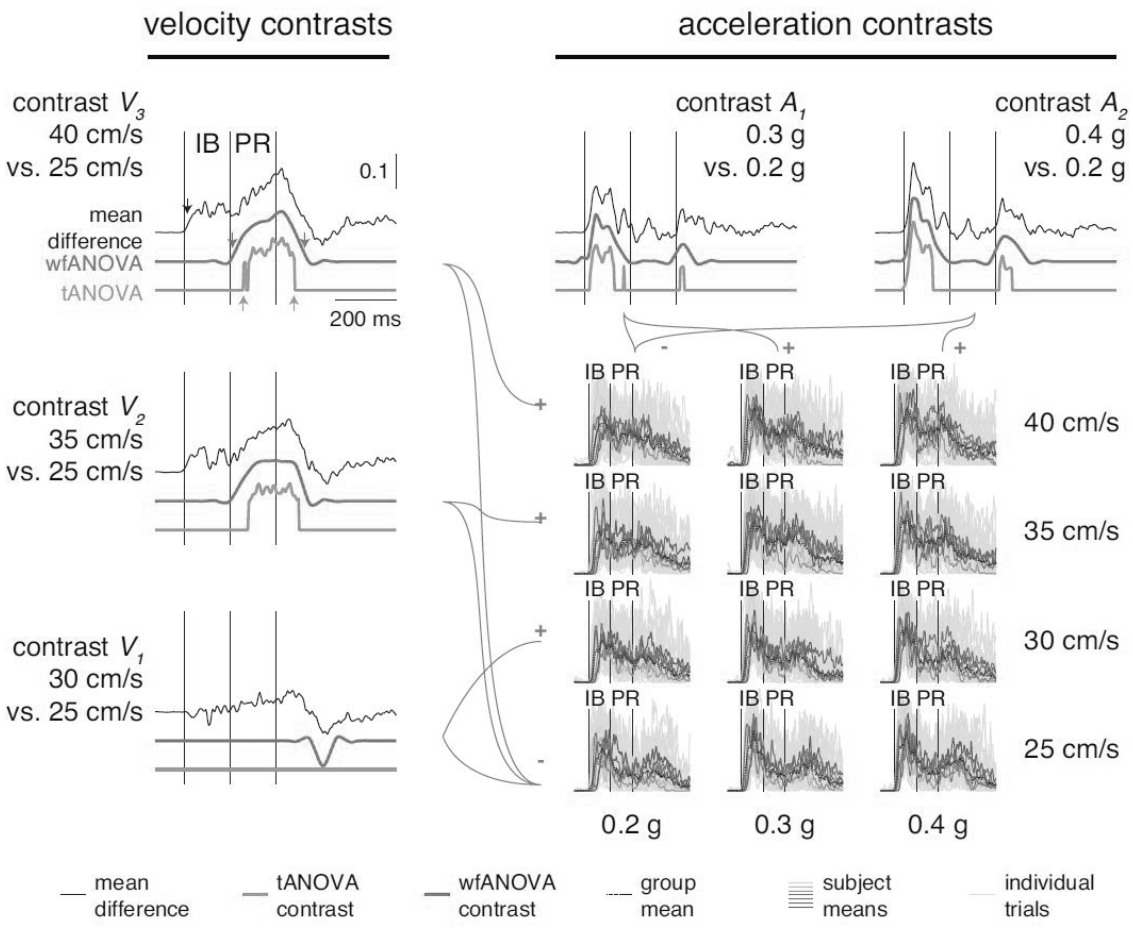


Figure 2.17: Analysis Results from EMG Study [47]

statistical and analytical techniques that go beyond the subjective, visual comparisons of results. Some of these basic statistical techniques include: hypothesis testing, confidence intervals around the difference, and goodness-of-fit tests.

In addition to the statistical techniques described above, this review introduced new techniques for time-series and functional data analysis. Many systems that require validation generate a stream of data which may be categorized as time-series data or functional data. New techniques are necessary to properly analyze and validate models that generate this type of data. A sub-category of this functional data are transient data,

which are a form of time-series data that are non-stationary and are typically characterized by large spikes in magnitude followed by a sharp decrease in a short span of time. The validation of transient phase data provides another set of challenges.

Techniques for validating time-series data include correlation analysis, which identifies the level of correlation between two data sets via a correlation coefficient, and spectral analysis, which compares the data in another domain (specifically the frequency domain) via a Fourier transform. Model validation metrics are also used to compare this type of data. Model validation metrics often encapsulate the discrepancy between signals and should account for experimental and computational uncertainty and error. Examples of these metrics include: Sprague and Geers metric [78], Russell's error measure [68], Whang's inequality index [69], the EARTH method [73], and the Transient Time Domain Validation method [39]. These metrics typically include magnitude, phase, and shape error, which combine to form a comprehensive error measure. Finally, functional data analysis techniques include the functional ANOVA (FANOVA), which tests for a statistically significant difference among curves [64]. Variations on FANOVA include statistics based on Hotelling's T^2 [32], or based on Fourier coefficients [27].

This review also introduced wavelets as an orthogonal transform of data into a time-frequency representation that offers certain advantages over Fourier transforms, such as the ability to transform non-stationary data. Wavelets are a computationally efficient tool with benefits such as a multiresolution property that allows focus on local, high-frequency content of a signal along with the ability to decorrelate autocorrelated data. In addition, the sparsity property of wavelet coefficients combined with the orthogonality of the transform make wavelets an effective tool for data compression and de-noising via a process called thresholding [16, 59].

Wavelets introduce opportunities to serve as a basis for new model validation techniques. Cheng *et al*. [19] develop a validation metric based on wavelet approximations

of noisy, transient signals. Other authors develop wavelet ANOVA (WANOVA) models which perform statistical inference in the wavelet domain. Girimurugan *et al*. [32] construct a test statistic using the set of wavelet coefficients to test for an overall difference among functional data. McKay *et al*. [47] establish a model that tests individual wavelet coefficients for statistically significant differences and converts them back into the time domain to reveal the statistically significant contrast curves.

Based on the challenges discussed in Section 2.5 and the recent work surveyed in Section 2.6, there is still room for improvement and innovation in the field of simulation validation. A recurring theme in the literature is the validation of models that generate time series data, which require special consideration due to autocorrelation and non-stationary data. Additional adjustments are needed when assessing the transient phase of time series data, typified by apparent erratic short term behavior. Many of the standard time series analysis techniques are unable to properly evaluate the validity of this type of data.

Wavelets offer potential solutions to the problem of validating the transient phase of a model. Wavelets allow the analyst to decompose the system and model output data in order to filter out noise and retain the important components of the response. It is then possible to apply error analysis to these decomposed wavelet representations to obtain several quantified error measures, including amplitude error, phase error, shape error, and signal energy error. These errors may combine to generate an overall error metric, capable of assessing the validity of a model. It is also possible to perform statistical inference in the wavelet domain. This allows the preservation of statistical power due to the sparsity of wavelet coefficients. The analyst may then conclude whether a statistically significant difference exists between the system and model data and could also take steps to identify additional information regarding the discrepancy to aid in the validity assessment.

A simulation model requires verification and validation (V&V) before proving reliable. Model validation substantiates that the model chosen truly represents the system

and that it produces results consistent with real-world data within the range of model applicability. It is imperative to develop analytical techniques capable of assessing all forms of data, including functional data, to build credibility and confidence in the developed model.

III. Dynamic Model Validation Metric Based on Wavelet Thresholded Signals

3.1 Introduction

Model validation is a vital step in the simulation development process and one that must be executed before relying on the results of the model for decision making purposes. Validation helps to ensure that a model is sufficiently representative of the system that it is meant to model. Sargent [72] defines validation as the 'substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model.'

There is a vast literature detailing a variety of model validation techniques. Many of these validation techniques are designed for assessing model validity during the steady state phase of a process. In these situations, statistical techniques such as hypothesis testing or regression analysis may be used to compare the system and model response in order to assess validity. However, one aspect of model validation that deserves special attention is when validation is required for the transient phase of a process. In this case, different techniques are necessary to analyze the time series data generated by the system and model. The techniques used to validate steady-state processes are not necessarily well-suited for data collected during the transient phase, which typically includes the initialization period of the system or simulation and ends when the process reaches stationary, steady-state behavior. Transient pulses may be characterized by a large spike in magnitude followed by a sharp decrease in a short span of time.

An additional concern associated with validating dynamic system and model data is that the experimental system data is often contaminated with noise, due to the short duration and sharp variations in the data. We assume this noise is normally distributed and could be attributed to measurement error or it could be inherent in the transient phase of the system. Since system observations are often limited, it is critical that this noise in the

data is properly accounted for and the system response signal clearly depicted, lest the experimental noise impact the results of a validity assessment. Oberkampf and Trucano [58] and more recently the American Society of Mechanical Engineers (ASME) Standard for V&V in Computational Fluid Dynamics and Heat Transfer [46] emphasize the need to identify and estimate the uncertainty and error in both the computational model and the experiment during a model validation process. This includes the experimental random error present during an observation of the system.

Often, simulation validation processes handle this transient or initialization phase by excluding it from the data analysis. However, in some circumstances analysis and validation of this portion of the data is necessary. For example, consider the scenario where a large pulse of energy causes an electronic equipment malfunction. It is imperative that the simulation accurately model this energy spike that could occur in the system [39].

The key contribution of this paper is a new methodology that is capable of assessing noisy, dynamic signals as part of a model validation assessment. We address many of the aforementioned concerns associated with properly validating the transient phase of a process by using wavelet thresholding as an effective method for eliminating the normally distributed noise in the signal. This de-noising process aids in controlling the experimental random error present in the system observations and the pure error included as a stochastic component in the simulation model. Consequently, this exposes the underlying system response signal and ensures that the signal noise does not interfere with the next step in our validity assessment, which is the calculation of a validation metric. This validation metric assesses the discrepancy between the system and model data. Therefore, our overall validation methodology provides accurate results for evaluating simulation models.

The paper proceeds as follows: Section 3.2 includes a review of the relevant literature on model validation. Section 3.3 introduces wavelet analysis and thresholding. Section 3.4 outlines the proposed validation approach, including how it deviates from a method

that uses wavelet decomposition. Finally, Section 3.5 compares the performance of the thresholding method to the decomposition method using a simulation study and empirical data.

3.2 Literature Review

Model verification and validation (V&V) has been pioneered by several authors who discuss the need to assess whether a simulation model is appropriate for use [6, 7, 43, 44, 72]. Verification ensures that the conceptual model is correctly implemented into a computerized model, while validation assesses whether the model is truly representative of the system. Authors such as Balci and Sargent provide a framework and set of techniques to guide the analyst through the validation process. Validation techniques include those for time series analysis, such as correlation analysis, which other authors expand upon in their texts [13, 15, 55].

The validation of computer models with functional outputs, such as time series data, is also a subject many authors have explored. Bayarri *et al*. [11] provide a framework for the validation of computer models with functional output using Bayesian statistics and likelihood methodology to assess validity. Jiang and Mahadevan [41] use wavelet analysis to validate a model by examining wavelet coherence, which is a measure that quantifies the amplitude and phase synchrony of two signals. They later use an energy-based Bayesian wavelet method to validate a multivariate model of a dynamic system [40].

The calculation of a model validation metric is another technique for assessing the validity of models with functional output. The ASME Guide for V&V in Computational Solid Mechanics [74] describe the use of a validation metric to compare experiment and simulation results. The metric may take the form of simple binary metric or a more complex comparison of the magnitude and phase difference in wave forms. Oberkampf and Barone [57] include several recommended features of validation metrics. One example of a validation metric measures the discrepancy between the system and model output and

is sometimes called an error metric. Many time-series error metrics have been developed over the years, including the Sprague and Geers' metric [78], Russell's error factors [68], Whang's inequality [85], Zilliacus' error [85], the Knowles and Gear metric [75] and the Error Assessment of Response Time Histories [73]. Many of these time-series error metrics include a magnitude error component and a phase error component, but vary in the manner in which each are calculated. The different error components may then be combined into a comprehensive error component. The use of these validation metrics is helpful for situations in which there is interest in quantifying which model among a set of models is most accurate, given the experimental data. However, the use of these metrics requires subjective input to designate a value for the validation metric through which the model may be judged valid or invalid.

Cheng *et al*. [19] provide the inspiration behind the work presented in this paper, as they combine wavelet analysis and their own time-series error metric to validate a model. They conduct a validity assessment of a biodynamical model by performing a wavelet decomposition of the test and simulation signals and then compare the signal approximations. The correlation coefficient, lag, and amplitude difference between the wavelet decomposed signals comprise the three components for an overall validation metric. The rationale behind this approach is that the wavelet decomposition process separates the low frequency content or "approximation" from the high frequency content or "details." Therefore, by comparing the low frequency approximations, a validity assessment is made which discards the noisy, high-frequency signal content. The authors then apply this validation methodology to a case study analyzing the performance of a 1997 Honda Accord finite element (FE) crash model versus the corresponding actual crash test data from the National Highway Traffic Safety Administration (NHTSA).

The weaknesses of the Cheng *et al*. [19] approach include the subjectivity involved in selecting a decomposition level, as well as the somewhat indiscriminate nature in which

high frequency content is removed from the signal. With their approach, there is the risk of removing not just noise, but also important signal content inherent to the real system. This paper proposes an alternative method called thresholding, which selectively removes the signal content that is judged to be noise.

3.3 Wavelet Analysis

A full overview of wavelets is beyond the scope of this paper, but works by Ogden [59], Burrus *et al*. [16], and Chui [20] offer further instruction. Generally, wavelets are a family of functions that serve as basis functions and may express either discrete or continuous signals. Wavelet analysis is closely related to Fourier analysis, which is used to transform data from the time domain into the frequency domain to aid in analysis. Wavelet analysis overcomes many of the limitations associated with a Fourier transform, including the inability to detect changes in frequency over time. Wavelets are localized in both the frequency and time domains and are thus suitable to transform non-stationary data.

The foundation for discrete wavelet analysis begins with a mother wavelet (ψ) and father wavelet (ϕ), which are functions with certain mathematical properties. The pair of wavelet functions are used to develop an entire family of wavelets by a scale factor expressed with subscript j and shift factor with a subscript k. This family of wavelets acts as basis functions so that a function, $f(t)$, may be expressed as a linear combination of these wavelets,

$$f(t) = \sum_{k} c_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k} d_{j,k}\psi_{j,k} \, . \tag{3.1}$$

The discrete wavelet transform (DWT) is used to estimate the wavelet coefficients, c_{jk} and d_{jk}, from a discrete sample of data by calculating the inner products of the signal and wavelet functions.

The wavelet representation of a function depends on the value selected for j_0, which is sometimes called the resolution level or the decomposition level. The resolution level is limited by the number of observations in the data set, which is ideally a dyadic number. Often, the first part of Equation 3.1 is referred to as the approximation of the function at level j_0 (A_{j_0}), while the second part is referred to as the details at level j_0 (D_{j_0}). The approximation and details are defined as

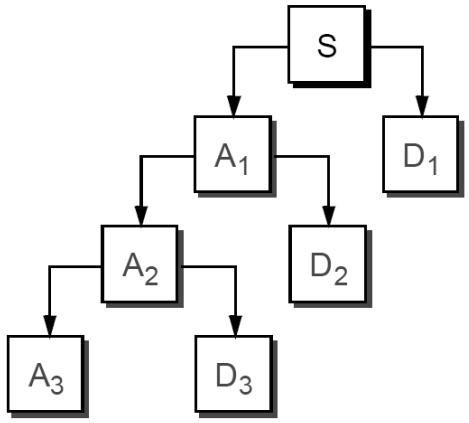
$$A_{j_0} = \sum_k c_{j_0,k}\phi_{j_0,k} \tag{3.2}$$

and

$$D_j = \sum_k d_{j,k}\psi_{j,k}\,. \tag{3.3}$$

A signal may be decomposed into the low-frequency approximation and high-frequency details. Additionally, the approximation may be subsequently decomposed into further approximations and details, as shown in Figure 3.1, via a recursive filtering and downsampling process. As the approximation is decomposed further, it represents a progressively coarser version of the original signal. This process may also be reversed, such that the approximations and details are synthesized back into the original signal with no loss of information.

Wavelet functions are developed with certain mathematical properties in mind. One useful property of wavelet analysis is that the wavelets comprise an orthogonal basis. Therefore, the orthogonal wavelet transform implies any noise in the original signal is transformed into noise in the transformed data. This noise may be observed in the wavelet coefficients of the transformed signal. Since the wavelet transform of a noise-free signal is sparse, these two properties mean wavelets are an effective tool for de-noising and compression.



$$S = A_1 + D_1$$
$$= A_2 + D_2 + D_1$$
$$= A_3 + D_3 + D_2 + D_1$$

Figure 3.1: Decomposition of signal S into Approximation and Details [48]

Wavelets are used for de-noising and compression by transforming the signal using the DWT and then reconstructing a de-noised or compressed version of the signal by using only a subset of the calculated wavelet coefficients. Several methods exist to accomplish this process. A crude de-noising approach involves simply taking the approximations of the signal as the de-noised representation of the signal. However this technique discards all the high-frequency information in the signal, causing the loss of many of the original signal's sharpest features. A more effective de-noising technique requires a more selective approach called thresholding. Wavelet thresholding was introduced by Donoho and Johnstone [25], who describe the wavelet transform of a noise-free signal as sparse, where many wavelet coefficients are equal to zero. If the signal is contaminated with noise, the orthogonal wavelet transform converts the signal noise into noise in the coefficients. These wavelet coefficients that were previously equal to zero are now primarily nonzero. By identifying a value which represents the wavelet coefficient noise, the wavelet coefficients may be modified or thresholded resulting in a de-noised signal.

When thresholding is applied, wavelet detail coefficients below the threshold value are set to zero. Donoho and Johnstone propose a universal threshold,

$$\lambda = \hat{\sigma}\sqrt{2log(n)} \tag{3.4}$$

where $\hat{\sigma}$ is an estimate of the standard deviation of the noise and n is the sample size. The universal threshold assumes that the noise is normally and independently distributed. The noise estimate, $\hat{\sigma}$, is traditionally calculated using the Median Absolute Deviation (MAD) of the finest scale detail coefficients scaled under normality assumptions according to

$$\hat{\sigma} = \frac{1}{\Phi^{-1}(\frac{3}{4})}\mathbb{M}[|W_{(J-1)k} - \mathbb{M}[W_{(J-1)k}]|] \tag{3.5}$$

where Φ references the normal distribution, $\mathbb{M}$ is the median operator, and W are wavelet coefficients. Once the universal threshold is calculated, a soft thresholding approach may be used so that the estimated coefficients, $\tilde{\theta}$, are replaced with the thresholded coefficients, $\hat{\theta}$, as

$$\hat{\theta} = \begin{cases} 0, & \text{if } |\tilde{\theta}| \leq \lambda \\ \tilde{\theta} - \lambda, & \text{if } \tilde{\theta} > \lambda \\ \tilde{\theta} + \lambda, & \text{if } \tilde{\theta} < -\lambda\,. \end{cases} \tag{3.6}$$

3.4 Validation Approach

As described in Section 3.2, a validation metric can serve as an effective tool in the model validation process. Ideally, a validation metric offers a comprehensive comparison between system and model data, but is expressed by a single value. However, to provide a comprehensive comparison between any two sets of data, it is necessary to identify what aspects of the signal should be compared. In the ensuing discussion, system data is modeled

as the x variable, model data as the y variable. Cheng *et al*. [19] develop a validation metric based on the magnitude, phase, and shape errors, where they measure the difference in shape via the correlation coefficient function,

$$\rho_{xy}(\tau) = \frac{R_{xy}(\tau) - \mu_x \mu_y}{\sqrt{[R_{xx}(0) - \mu_x^2][R_{yy}(0) - \mu_y^2]}} \tag{3.7}$$

with time lag, τ. $R_{xy}(\tau)$ represents the cross-correlation function, while R_{xx} and R_{yy} represent the autocorrelation functions of x and y, respectively. For reference, the cross-correlation function is

$$R_{xy}(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T x(t) y(t + \tau) dt \, . \tag{3.8}$$

This correlation coefficient function provides a measure of the linear relationship between two sets of time-series data, while accounting for a possible time lag between the datasets. We assume that a valid model will yield values close to unity, of course this need not be true when comparing highly non-linear models. Therefore, the maximum value of the correlation coefficient,

$$\rho_{xy} = \max_{\tau}(\rho_{xy}(\tau)) \tag{3.9}$$

provides a measure of the shape error, while the corresponding time lag,

$$\tau = arg \max_{\tau}(\rho_{xy}(\tau)) \tag{3.10}$$

provides a measure of the phase error. Finally, the magnitude error is calculated by taking

the relative difference in amplitude (A_x, A_y) between the two signals. We expect that a valid model will return a phase error and magnitude error close to zero.

Cheng *et al*. [19] combine these three error components into a single validation metric,

$$R = \left[\alpha_1(1-\rho_{xy}) + \alpha_2\left|\frac{\tau}{T}\right| + \alpha_3\left|\frac{A_x - A_y}{A_x}\right|\right] \times 100\% \tag{3.11}$$

where α_1, α_2, and α_3 represent weighting coefficients, such that $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$. These weighting coefficients should be balanced to ensure even consideration is given to each error component, but may be varied to give more or less emphasis to an error factor, based on importance. A smaller value of R represents higher model validity.

Our validation approach calculates the validation metric using wavelet thresholded system and model data signals. We first use wavelet thresholding to control the noise present in the system and model data. This may include noise inherent in the system and captured during the experiment, as well as any noise present in the simulation model, potentially from a stochastic component. We then compare the de-noised signals using the validation metric described above. This two-step approach renders our validation methodology suitable and appropriate for validating dynamic signals such as those exhibited during the transient phase of a process. Previously established validation methodologies that use a metric [68, 78, 85] calculate the validation metric based on the original data signals obtained from experimentation and simulation. While these may be effective where noise or transient behavior are not a concern, they are not as effective when that behavior must be addressed. Thus, our validation methodology operates as a comprehensive comparison between the system and model data and as an indicator of model validity.

This approach differs from that proposed by Cheng *et al*. [19], who recommend the iterative wavelet decomposition of the original signals into coarser approximations, followed by the calculation of the validation metric. If the validation metric meets an acceptable value, the model is declared valid. Otherwise, the signals are decomposed to an additional level and then compared, continuing until some maximum decomposition level has been reached, a level specified in advance by the analyst. If the validation metric does not meet the acceptable value by this point, the model is declared invalid. Their approach presents several potential problems including the subjectivity involved in determining the maximum decomposition level. The analyst must choose a maximum decomposition level *a priori* without any reasonable justification; however this selection impacts whether the model is assessed to be valid or invalid. Second, the use of the approximations to represent the original signal involves the indiscriminate removal of high frequency content from the signal. Lastly, a signal that has been decomposed multiple times in such a way may bear very little resemblance to the original signal that it is supposed to approximate, resulting in the comparison of two signals that do not truly represent the original system and model data.

The validation technique proposed in this paper solves these problems by using wavelet thresholding. Wavelet thresholding is a single step instead of a multi-level decomposition process and does not require subjective input from the analyst. The threshold value is determined via the universal threshold, which is based on an estimate of the signal noise. Therefore, the threshold value is signal-specific and applicable only to the signal being analyzed. Since the threshold is determined via a process that is specific to the particular signal, it is both more objective and more precise than de-noising based on the subjective determination of a maximum decomposition level. In addition, the universal threshold is ideal for de-noising applications since it is both an effective and computationally efficient technique. The next section compares the validation

metric results calculated using wavelet thresholding as a de-noising approach to a wavelet decomposition approach.

3.5 Illustration of Approach

3.5.1 Simulation Study.

A simulation study demonstrates the effectiveness of wavelet thresholding as a de-noising approach yielding effective validation metric results. For the first part of this study, a series of random signals were generated, developed from a series of cosine waves with randomly generated frequency and phase parameters. Each base signal is the sum of 500 random cosine waves so that a large variety of signals are evaluated. For a given iteration, two normally distributed random error vectors were created and added to a constructed base signal to create two different noisy signals. A random lag component was also incorporated. These two noisy signals represent the system data and the model data. Since the two noisy signals are constructed from the same base signal and differ only by a random noise and lag component, the validation methodology should result in a small validation metric value and therefore indicate a high level of agreement between the two signals. The study simulated 1000 iterations and calculated the correlation coefficient, lag, amplitude difference, and validation metric for the original signals, thresholded signals, and approximations at different decomposition levels. The fourth order Daubechies wavelet, db4, was used for consistency with Cheng's work [19]. These results are summarized in Table 3.1. The column with the header *Original* provides the error components calculated using the original, unmodified system and model data. The *Thresholded* column uses the wavelet thresholding approach proposed in this paper. The *Level 1 - Level 6 Approximations* columns use the wavelet decomposition approach of Cheng [19].

Table 3.1: Simulation Study Measures (Correlation Coefficient, Lag, Amplitude Difference)

			Approximations					
	Original	**Thresholded**	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
Correlation	0.4175	0.8696	0.5169	0.6188	0.7126	0.7922	0.8464	0.8694
Lag	0.0214	0.0231	0.0224	0.0213	0.0219	0.0215	0.0228	0.0236
Amplitude	0.0829	0.1169	0.0833	0.0950	0.0970	0.0961	0.1054	0.1220

Table 3.2 displays the average validation metric values, which were calculated using the following weighting coefficient values, proposed by Cheng *et al*. [19]: $\alpha_1 = 0.5, \alpha_2 = 0.2, \alpha_3 = 0.3$.

Table 3.2: Simulation Study Validation Metric, R

			Approximations					
	Original	**Thresholded**	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
Metric (R)	32.04	10.49	27.10	22.33	17.71	13.71	11.30	10.66

The results from the simulation study indicate that the thresholding method is very effective at removing the artificial noise inserted into the original signals. Once this noise is removed, the correlation coefficient, lag, and amplitude difference are calculated using the de-noised signals and show a high level of agreement. As a result, the validation metric associated with the thresholded signals is very low, indicating higher level of validity. In comparison, the validation metric calculated using the original, unmodified signals is three times as high, while the different levels of wavelet approximation show varying validation metric values. As expected, the validation metric value decreases as

the wavelet decomposition levels increase, but even the Level 6 approximation does not yield an average validation metric value as low as the thresholded signals. Further, in a real validation study, the maximum decomposition level would be subjectively determined prior to the analysis. A poor choice for this decomposition level may result in a valid simulation model being incorrectly rejected as invalid. Our thresholding technique circumvents this problem by eliminating the need to choose a decomposition level *a priori* and assesses the validity using the wavelet thresholded signals.

For the second part of this simulation study, confusion matrices are used to illustrate the accuracy of the various validation methodologies. A confusion matrix indicates how well a classification model performs on a dataset for which the true class is known. In this case, the classes are a valid or invalid model, and the classification model is the validation methodology. The dataset used is constructed in the same manner as the first part of this study, where two noisy signals that originate from the same base signal form the valid class of the dataset. In contrast, two noisy signals that originate from different base signals form the invalid class of the dataset. This study uses a dataset of 1000 replications, split evenly between the valid and invalid classes. The main assumption with this part of the study is that if the system and model data share the same base signal structure, then it indicates the model is valid. Otherwise, the model is declared invalid.

Before the confusion matrix is constructed, a validation rule is established for the different methodologies. In particular, a validation metric value is designated to determine whether a model is declared valid or invalid. However, as is often the case with validation metrics, the designation of such a value is both difficult and highly subjective. For this reason, results for several different validation rules are examined. To maintain consistency among the methods, the same validation rule is used for all validation methodologies.

The following validation rules are examined: accept model as valid if the calculated validation metric value, R, is less than 10, 20, 30, or 40. Note that a validation metric value

of $R = 0$ represents perfect agreement of system and model data. Therefore, a validation rule of $R < 10$ represents a more stringent validation requirement, while a validation rule of $R < 40$ corresponds to a more relaxed requirement. The confusion matrices for the validation rule of $R < 20$ are provided in Tables 3.3 through 3.7, where Table 3.4 is our method and Tables 3.5 - 3.7 are three levels of the decomposition method.

Table 3.3: Confusion Matrix for Original Signals, $R < 20$

		Predicted	
		Valid	**Invalid**
Actual	**Valid**	119	381
	Invalid	3	497

Table 3.4: Confusion Matrix for Thresholded Signals, $R < 20$

		Predicted	
		Valid	**Invalid**
Actual	**Valid**	429	71
	Invalid	19	481

Table 3.5: Confusion Matrix for Level 1 Approximations, $R < 20$

		Predicted	
		Valid	**Invalid**
Actual	**Valid**	166	334
	Invalid	6	494

Table 3.6: Confusion Matrix for Level 3 Approximations, $R < 20$

		Predicted	
		Valid	**Invalid**
Actual	**Valid**	295	205
	Invalid	12	488

Table 3.7: Confusion Matrix for Level 5 Approximations, $R < 20$

		Predicted	
		Valid	**Invalid**
Actual	**Valid**	414	86
	Invalid	16	484

Table 3.3 shows that the use of a validation metric is ineffective at categorizing noisy signals, as it declares 76% of valid models to be invalid. Table 3.4 indicates that our thresholding method is the most effective at correctly assessing model validity with a 91% overall accuracy rate. Tables 3.5 - 3.7 demonstrate varying levels of accuracy using a wavelet decomposition approach. In general, higher decomposition levels yield increased classification accuracy. The classification accuracy is provided in Table 3.8 for all validation rules. These results show that even for varying cases of a validation rule, the highest classification accuracy stems from calculating the validation metric using the thresholded signals, as opposed to different wavelet approximation levels.

3.5.2 Automobile Crash Study.

The next comparison replicates Cheng *et al.'s* [19] validation study on a 1997 Honda Accord FE crash model using actual crash test data from the NHTSA. This study analyzes the crash signals for a full frontal impact, specifically the acceleration responses recorded

Table 3.8: Classification Accuracy

			Approximations					
Validation Rule	**Original**	**Thresholded**	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
10	56.6%	79.5%	59.4%	62.7%	66.2%	71.7%	76.1%	79.8%
20	61.6%	91.0%	66.0%	71.7%	78.3%	84.7%	89.8%	90.8%
30	63.7%	90.1%	69.8%	77.2%	83.6%	88.4%	89.9%	90.0%
40	69.3%	85.6%	78.2%	82.3%	85.0%	85.8%	85.8%	85.8%

by an accelerometer positioned at the top of the vehicle engine (Engine Top) and on the right-rear cross member (RRCM) of the automobile. The response data contains 1,000 data points with a sampling rate of 0.1 ms for a total time duration of 100 ms. These signals are displayed in Figure 3.2.

The test and simulation signals are compared by calculating the correlation coefficient, lag, and amplitude difference for the original signals, thresholded signals, and approximations at different decomposition levels. The wavelet analysis used the fourth order of Daubechies wavelet, db4. These results are summarized in Tables 3.9 and 3.10. The validation metric values were calculated using the weighting coefficient values: $\alpha_1 = 0.5, \alpha_2 = 0.2, \alpha_3 = 0.3$.

In contrast to the simulation study, the validation metric values calculated using the wavelet decomposed approximations are generally smaller than those calculated using the thresholded signals. For the engine top (Table 3.9), the validation metric value for the thresholded signals falls between a Level 1 and Level 2 approximation. For the right-rear cross member (Table 3.10), even the Level 1 approximation results in a smaller validation metric value than the thresholded signal. Based on these observations, this might indicate

Table 3.9: Engine Top Analysis

			Approximations					
	Original	**Thresholded**	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
Correlation	0.65	0.71	0.66	0.72	0.77	0.82	0.83	0.92
Lag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amplitude	0.49	0.50	0.48	0.45	0.32	0.09	0.01	0.25
Metric (R)	32.0	29.4	31.3	27.5	22.7	11.7	8.6	11.5

Table 3.10: Right-Rear Cross Member Analysis

			Approximations					
	Original	**Thresholded**	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
Correlation	0.20	0.30	0.31	0.40	0.50	0.72	0.81	0.88
Lag	0.003	0.004	0.003	0.003	0.002	0.005	0.00	0.006
Amplitude	4.12	3.89	0.90	0.39	0.01	0.23	0.19	0.23
Metric (R)	163.9	151.9	61.5	41.8	25.4	21.3	15.2	13.1

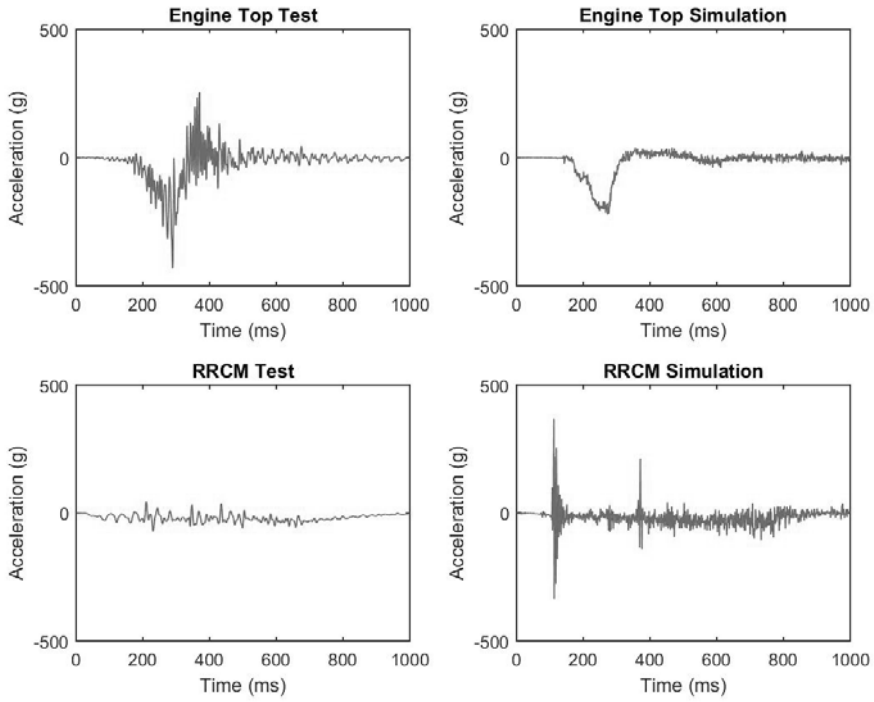


Figure 3.2: Crash Signals

that the wavelet decomposition method is more effective at identifying a valid model and therefore more effective at de-noising. However upon closer inspection, several key observations arise. First, we note that the correlation and lag values for the thresholded signals are mostly in line with those calculated using the various decomposition levels. In fact, most of the discrepancy between the thresholded and decomposition validation metrics are attributed to the amplitude difference component of the metric. The decomposition method results in a much smaller difference in amplitude (i.e. magnitude error), resulting in a smaller validation metric value. However, if the intent is to validate the transient phase of a model which often contains sharp spikes as part of the process, then wavelet decomposition may not be the appropriate choice. To further illustrate, consider Figure

3.3, which presents how the wavelet decomposition technique affects the original RRCM signal.

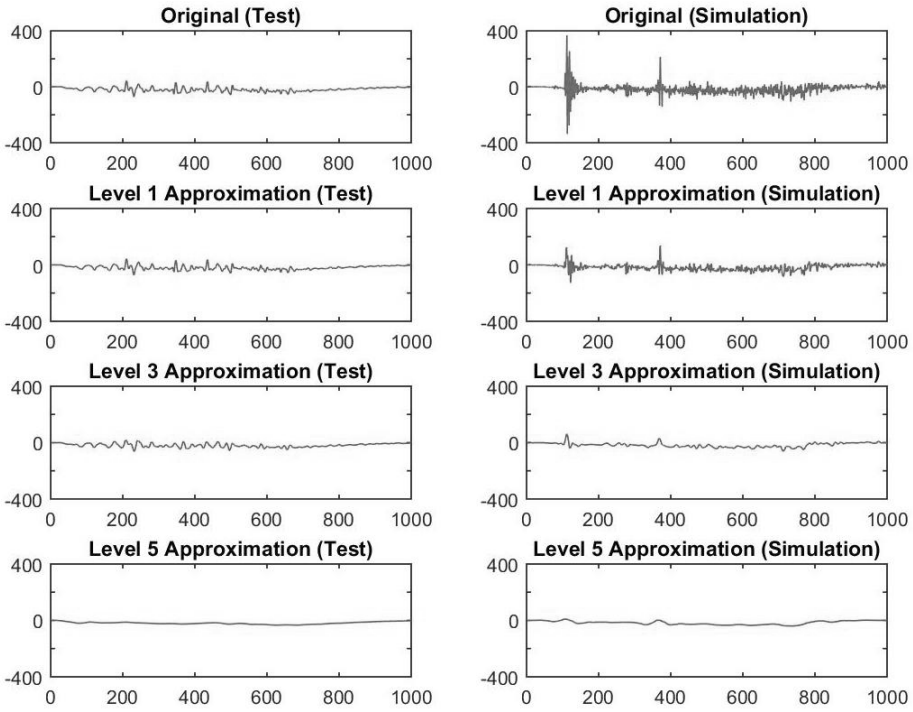


Figure 3.3: Decomposed Signals (Right-Rear Cross Member)

Figure 3.3 highlights one of the potential dangers associated with the wavelet decomposition approach of comparing the wavelet approximations of the test and simulation data. Too much of the high frequency content is indiscriminately removed resulting in the comparison of two signals that exhibit very little similarity to the original signal. While this lack of resemblance is evident in the graphs, it can be further shown by comparing the original signal to its various approximations. Table 3.11 displays the correlation coefficient between the original RRCM simulation signal and the approximated versions of that signal. The table shows that the correlation between the original signal and a wavelet decomposed approximation is as low as 0.35. This illustrates some of the risk

in the subjective selection of a maximum decomposition level, since a highly decomposed signal may be significantly altered from the original signal.

Table 3.11: RRCM Simulation Signal vs. Approximations

			Approximations					
	Original	**Thresholded**	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
Correlation	1.00	0.79	0.63	0.50	0.44	0.39	0.37	0.35
Amplitude	367.35	323.43	137.30	100.97	61.77	47.07	39.48	40.41

Table 3.11 also includes the signal amplitudes, which is the maximum absolute value of the signal. These values, considered along with Figure 3.3, show the decompositions' near removal of the sharp spikes from the original simulation data signal. This removal in particular is what allows the decomposition method to generate a lower validation metric value. However, often it is imperative that these sharp spikes or pulses in data are properly characterized and compared in order to validate a simulation model. Their exclusion may result in the incorrect validation of an inaccurate simulation model. In comparison to the wavelet decomposition approach, consider the effect of wavelet thresholding on the original data signals. Although wavelet thresholding does remove the signal content that it evaluates as noise, the overall integrity of the signal is unaffected as Table 3.11 and Figure 3.4 illustrate the retention of the peaks in the original signal.

3.5.3 Follow-On Simulation Study.

A follow-on simulation study further illustrates the risks of using a wavelet decomposition to approximate the original signal. Random base signals are generated in a manner similar to those constructed in Section 3.5.1. For this study, the system and model data originate from the same base signal with normally distributed random error added. In

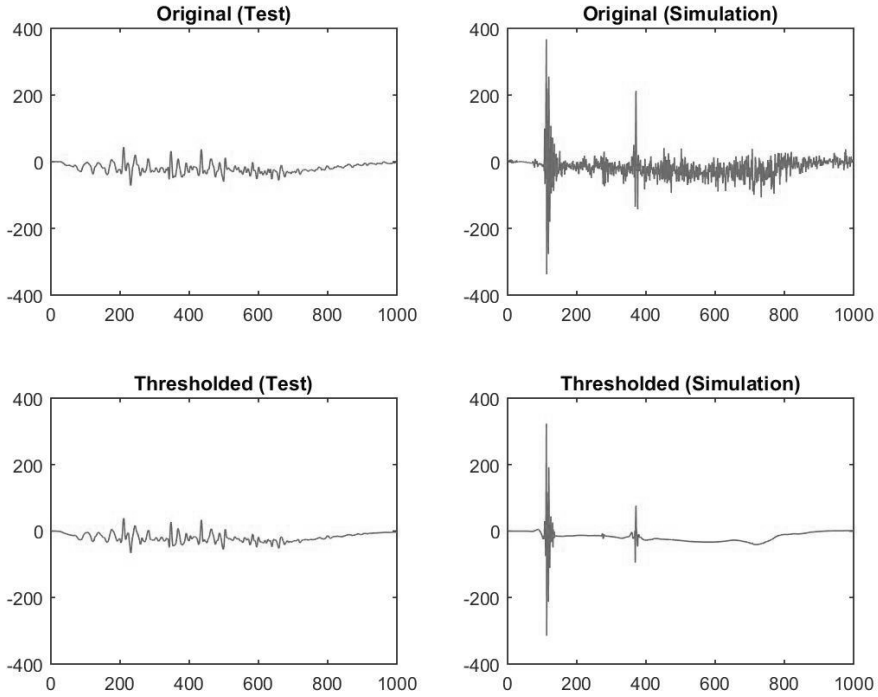


Figure 3.4: Thresholded Signals (Right-Rear Cross Member)

the previous section, this led to the assumption that the model was valid. However, this follow-on study incorporates additional sharp variations into the system data which are meant to represent a system process characteristic, such as an energy surge. The magnitude of this data spike and its location within the signal are randomly selected. Figure 3.5 shows an example of two noisy signals that originate from the same base signal, but the system data includes a sharp spike in the data during the process. The simulation of data of this form expands upon the data exhibited in the automobile crash study, therefore offering an additional comparison between the thresholding and decomposition approaches.

Although both the system and model data originate from the same base signal, there is clearly a discrepancy between the two signals. Therefore, our validation technique should recognize this discrepancy and declare the model invalid. The study simulated

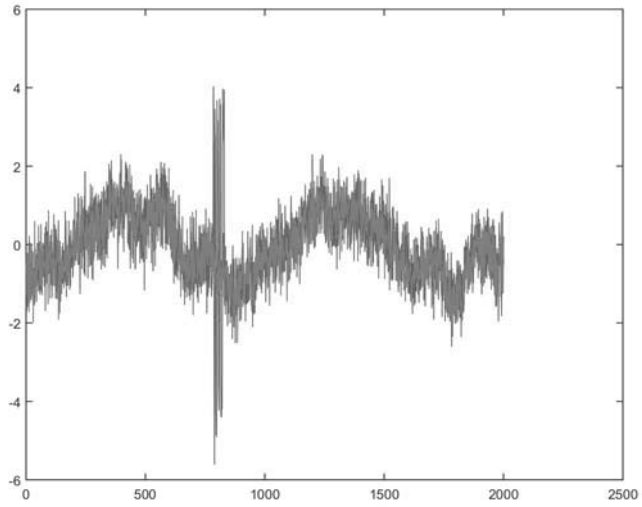


Figure 3.5: Example Data for Follow-On Study; System (Blue) and Model (Red)

1000 iterations, where each iteration generated a unique base signal, a unique random noise vector, a unique spike location, and a unique spike magnitude. For each iteration, we calculated the correlation coefficient, lag, amplitude difference, and validation metric for the original signals, thresholded signals, and approximations at different decomposition levels. These results are summarized in Table 3.12.

Table 3.12: Follow-On Simulation Study Results

			Approximations					
	Original	**Thresholded**	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**	**Level 6**
Correlation	0.56	0.87	0.64	0.73	0.85	0.92	0.95	0.96
Lag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amplitude	0.42	0.61	0.52	0.60	0.39	0.13	0.08	0.08
Metric (R)	34.65	24.68	33.50	31.54	19.25	8.22	4.96	4.47

These results are similar to those determined using the real crash test data, where the validation metric values calculated using the wavelet decomposed approximations are generally smaller than those calculated using the thresholded signals. The average thresholding validation metric is approximately 25, while the validation metrics obtained using the wavelet decompositions yield average values as low as 4.5. This indicates that at higher-level decompositions, Cheng's technique eliminates the discrepancy caused by the spikes in the data, which ultimately results in the validation technique providing inaccurate assessments of model validity. In contrast, the thresholding technique preserves the integrity of the original signals and is thus still able to identify the discrepancy. Therefore, it generates an accurate assessment of the model and demonstrates that thresholding is more precise at de-noising a signal.

3.5.4 Improved Validation Metric.

This paper used the validation metric proposed by Cheng *et al*. [19] in order to directly compare the performance of our validation methodology that utilizes wavelet thresholding versus a technique that uses wavelet approximations. However, the actual validation metric may be improved upon so that it more effectively characterizes the discrepancy between the system and model data. The three error components - shape, phase, and magnitude - should be standardized and balanced so that the comprehensive validation metric is not biased towards any one error component. Additionally, the use of the amplitude difference, i.e. the difference in the maximum absolute values of the two signals, is not the ideal measure to describe the magnitude error. We propose two changes to our metric and then compare the performance of our validation methodology to two other validation metrics well-known in the literature.

First, our validation metric may be improved by standardizing the three error components. The legacy version of the metric includes a shape error component that could range in value from zero to two. The phase error component ranges from zero to

one. The magnitude error factor is unbounded. To balance and standardize the three error components, we modify the shape and magnitude error components to achieve a range of zero to one, which matches the phase error component. This ensures the overall metric is not overwhelmed by any one error source.

Second, the magnitude error component in the legacy metric calculates the amplitude error at a single point in the signal, and not over the entire signal. We modify the component so that it accurately reflects a measure of magnitude difference across the full signal. Russell's magnitude error factor [68] offers one solution for a relative magnitude error between two signals that is insensitive to phase. However, this magnitude error factor is also unbounded. Thus, we modify this factor to be bounded between zero and one and obtain a new magnitude error component,

$$m = \frac{|\sum_{i=1}^{N} x(i)^2 - \sum_{i=1}^{N} y(i)^2|}{\max\left(\sum_{i=1}^{N} x(i)^2, \sum_{i=1}^{N} y(i)^2\right)}\,, \tag{3.12}$$

where x represents the system data and y represents the model data, each with dimension N.

We also remove the multiplication by 100% from the equation because it is extraneous to the metric formulation. These modifications lead to our new proposed validation metric,

$$R^* = \alpha_1 \left(\frac{1-\rho_{xy}}{2}\right) + \alpha_2 \left|\frac{\tau}{T}\right| + \alpha_3(m)\,. \tag{3.13}$$

The weighting coefficients $(\alpha_1, \alpha_2, \alpha_3)$ are retained to give the flexibility to add more emphasis to a specific error component to reflect importance. However, as a default, we recommend that the weights are set equal to one another to ensure balance among the three error components.

We demonstrate the performance of this new metric as part of our overall validation methodology alongside two other validation metrics that are well-established in the literature: Russell's error factor [68] and the Sprague and Geers' metric [78]. While it is slightly subjective to compare validation metrics side-by-side, we believe it is worthwhile to examine our performance against other metrics. We perform a simulation study using the same parameters established in Section 3.5.1. We use a dataset of 1000 replications, split evenly between the valid and invalid classes, to assess classification accuracy. We examine results for validation rules of 0.15 and 0.30. Table 3.13 summarizes the results of our analysis. The results indicate that our thresholding method with new validation metric is most effective at correctly assessing model validity with up to a 90.6% overall accuracy rate. Russell's error factor achieves up to 64.2% accuracy, while the Sprague and Geers' metric is up to 62.4% accurate. This discrepancy highlights that our methodology is more accurate at validation assessments, particularly when examining noisy data. In addition, this improved validation metric serves as a more robust measure of discrepancy for comparing system and model data to test validity.

Table 3.13: Classification Accuracy Comparison

Validation Rule	**New Metric (R^*)**	**Russell**	**Sprague & Geers**
0.15	90.6%	55.1%	54.2%
0.30	80.2%	64.2%	62.4%

3.6 Conclusion and Recommendations

This paper illustrates that wavelet thresholding is very effective at removing the noise from a signal in order to make a more accurate model validity assessment. The validation approach that de-noises via wavelet thresholding results in an overall higher classification accuracy than the approach that relies on wavelet decomposed approximations. In addition,

the wavelet thresholding process preserves the integrity of the original signal, while the wavelet decomposition process may significantly alter the original system and model data. For these reasons, wavelet thresholding is a preferred method when validating transient phase data.

While the use of wavelet thresholding to de-noise a signal prior to calculating a validation metric offers great utility, there are a few notable limitations to the methodology outlined in this paper. One limitation is our distributional assumption for the system and model noise. If normality cannot be assumed, we recommend a nonparametric approach to wavelet thresholding [45]. A second limitation is the subjective designation of an acceptable metric value that indicates model validity. Future work will include eliminating the use of a validation metric altogether and instead using some form of hypothesis test that accepts or rejects a model as valid. It will also be worthwhile to examine alternative wavelet de-noising techniques to include the use of wavelet packets.

Despite the limitations highlighted above, wavelets offer the ability to transform data and use the sparse property of wavelet coefficients to calculate a threshold and de-noise the signal. A wavelet transform is better suited than a Fourier transform since it is not limited by stationary signal requirements, which can be critical when analyzing data from the transient phase of a process. Then, by calculating the magnitude, phase, and shape errors between de-noised versions of the system and model data, the level of discrepancy between the signals may be evaluated in the form of a comprehensive validation metric. While a validation rule of $20 - 30$ is appropriate for the legacy metric, R, we find that a value of 0.15 represents an effective rule for the new metric, R^*. Since the magnitude error component is calculated differently, it is not practical to compare the two metrics, but it is worth noting that R contains a multiplicative factor of 100 compared to R^*.

This method of wavelet thresholding is shown to be more effective than the technique of using wavelet decomposed approximations to represent de-noised signals for several

reasons. First, it eliminates the subjectivity of selecting a maximum decomposition level for which an approximation may represent the signal. Next, wavelet thresholding assesses the noise present in the signal and selectively removes the appropriate signal content. Last, this study shows that the wavelet decomposition approach involves the removal of high frequency content from the signal and that multiple decompositions can result in an approximation that is actually very different from the original signal, while wavelet thresholding retains the overall integrity of the original signal. Thus, wavelet thresholding combined with the calculation of a comprehensive validation metric is a recommended method for the validation of dynamic models.

IV. Wavelet ANOVA Approach to Model Validation

4.1 Introduction

In the field of modeling and simulation, verification and validation (V&V) is required to ensure accuracy before relying on the results of a computer model to make system or process inferences. Verification assesses whether the conceptual model is correctly implemented into a computerized model. Validation ensures "that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of a model" [72]. Verification ensures the model was built correctly, while validation ensures the correct model was built.

Although there are numerous validation techniques described in the literature [44, 72], the quality of these techniques can vary tremendously. The most basic techniques [5] include subjective, informal comparisons of system and model data, while more rigorous methods [7] include statistically-based analytical techniques. Ostensibly, more precise and statistically-justified validation methods are desired. Today, many systems that require validation generate a stream of data which may be categorized as time-series data or functional data. However, many established statistical validation techniques are not well-suited for the validation of models that generate this functional or time-series data. Specifically, some of the techniques that are designed for time-series analysis may lack statistical rigor. For example, the class of model validation metrics used to compare two sets of time-series data [57] can measure the discrepancy between the system and model, but it is often the responsibility of a subject matter expert (SME) to determine if the metric value indicates model validity. This SME input interferes with the objectivity necessary in a model validation exercise.

An additional challenge associated with validating a model's functional output is the dimension of the data. The functional output often has a large dimension which may greatly

complicate the analysis. A traditional analysis of variance (ANOVA) approach is ineffective at handling functional data in higher dimensions, since the large number of hypothesis tests results in an uncontrolled Type I error or false positive rate. A multiple comparison procedure, such as the Bonferroni correction, would yield an extremely low power of the test. Thus, new techniques are necessary to properly analyze and validate models that generate functional or time-series data.

This paper presents a validation methodology that resolves the aforementioned challenges associated with the objective validation of a model that generates time-series data. This validation process uses wavelets to aid in the analysis and assesses whether the system and model data are statistically equivalent using wavelet analysis of variance (WANOVA). If statistically equivalent, the model is declared a valid representation of the system. If a statistically significant difference exists, then the model is declared invalid. The WANOVA validation process provides an objective assessment of validity that is applicable to models which generate time-series data.

The paper proceeds as follows: Section 4.2 includes a review of the relevant literature on model validation and functional data analysis. Section 4.3 outlines wavelet analysis and thresholding. Section 4.4 introduces WANOVA and the proposed test statistic. Section 4.5 describes the distribution of the proposed test statistic under the null hypothesis and illustrates this via a simulation study. Finally, Section 4.6 demonstrates the performance of the validation method using both a simulation study and empirical study.

4.2 Literature Review

Model V&V has been pioneered by several authors who discuss the need to assess whether a simulation model is appropriate for use [5, 7, 21, 43, 44, 72]. Authors such as Balci [5] and Sargent [72] provide frameworks and discuss techniques to guide the analyst through the validation process. They include informal techniques, which are generally subjective and rely on human reasoning; static techniques that focus on the accuracy of

the static model design; dynamic techniques, which require model execution; and formal techniques that use mathematical proofs to demonstrate model correctness. Within the class of dynamic techniques is the set of statistical techniques such as hypothesis testing, simultaneous confidence intervals, and analysis of variance (ANOVA).

However, many of these proposed statistical validation techniques are applicable only to simulation models that generate discrete output. Models that generate functional output, such as time-series data, require additional consideration. Some analysis methods distill the functional data down to a single parameter, such as the mean or variance, for analysis. However, this process discards most of the important information contained in the signal. Other functional or time-series analysis techniques use methods such as correlation analysis, which evaluates the level of linear dependence between two data sets, or spectral analysis, which first transforms the data from the time domain to the frequency domain via the Fourier transform.

Model validation metrics offer another technique for assessing the validity of models with functional output. Oberkampf and Barone [57] discuss the need to quantitatively compare computational results with experimental measurements in the form of a validation metric. A common form of validation metric measures the discrepancy between the system and model output and is sometimes called an error metric. Many time-series error metrics have been developed over the years, including Geers' measure [31] and Russell's error factors [68]. More recently, a validation method that uses wavelet thresholding to de-noise a signal prior to calculating a validation metric based on correlation, lag, and amplitude error was proposed [4].

Validation metrics are useful for situations where it is necessary to identify a most accurate model among some set of simulation models. However, one drawback of the validation metric approach is the need to subjectively designate an acceptable metric limit that indicates model validity. Model validation techniques that require subjective input

carry risks. It may not be possible to find an unbiased SME to assess the model, while someone independent of the process may not have the necessary expertise. Additionally, some systems and models are too complex for an individual to provide a comprehensive assessment. Sargent [71] expresses concerns with validation metrics stating, "this author does not believe in the use of scoring models for determining validity because (1) a model may receive a passing score and yet have a defect that needs to be corrected, (2) the subjectiveness of this approach tends to be hidden and thus this approach appears to be objective, (3) the passing scores must be decided in some (usually) subjective way." Ideally, subjectivity is eliminated when performing a model validation assessment.

Functional data analysis offers potential for eliminating subjectivity in an assessment of whether two sets of functional or time-series data are statistically equivalent. Ramsay and Silverman [64] provide extensive coverage on the subject of functional data analysis, which includes functional analysis of variance (FANOVA). FANOVA tests the null hypothesis that the treatments have no effect on the functional response versus the alternative hypothesis that one or more treatments has an effect on the response. However, this FANOVA method is essentially a univariate ANOVA problem for each specific value of time. This series of pointwise F-tests leads to a large number of hypothesis tests, causing a multiplicity problem and resulting in an uncontrolled Type I error or false positive rate.

Girimurugan *et al*. [32] control this Type I error by developing a FANOVA model based on a multivariate statistic instead of a univariate statistic. Similarly, Fan [26] and Fan and Lin [27] propose a test of significance for comparing functional data based on the adaptive Neyman test and wavelet thresholding techniques. Perhaps most promising are functional analysis techniques in the wavelet domain [32, 47, 83]. These wavelet-based ANOVA models help smooth noisy data, reduce the dimensionality, and decorrelate time-series data. These techniques have been developed for general testing of significance when the data are curves or for more specific testing in profile monitoring. To the authors'

knowledge, WANOVA has not yet been used for the validation of simulation models. Yet, WANOVA offers significant potential in testing for statistical equivalence of system and model time-series output while simultaneously addressing the aforementioned concerns associated with subjectivity in model V&V.

4.3 Wavelet Analysis

Wavelet analysis is a relatively recent development in applied mathematics. Wavelets serve as basis functions, which are able to represent discrete or continuous signals similar to Fourier analysis. However, while Fourier analysis transforms data from the time domain to the frequency domain, wavelets are localized in both the frequency and time domains. This representation of the data in the time-frequency domain renders wavelets well-suited for transforming non-stationary data, which is one of the major limitations of Fourier analysis. In addition, wavelet analysis is a computationally efficient technique, requiring only $O(n)$ operations, compared to $O(n\ log\ n)$ operations for the Fast Fourier Transform. The ability of wavelets to transform non-stationary data and their computational efficiency have led to the growing use and popularity of wavelet analysis in modern applications. Works by Burrus *et al*. [16], Chui [20], and Ogden [59] serve as excellent resources for interested readers.

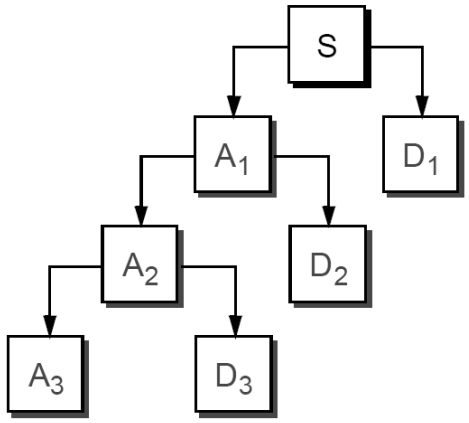
Wavelets are developed with certain mathematical properties in mind, such as orthonormality, smoothness, and compact support. Over the years, a variety of wavelet families have emerged, including the Morlet, Daubechies, and Coiflet wavelet families. In any wavelet family, there is a wavelet function (ψ), also known as the mother wavelet, and a scaling function (ϕ), also called the father wavelet. The rest of the wavelet family derives from these two functions by scaling and shifting the mother and father wavelets. Most wavelet literature expresses the scale factor with subscript j and the shift factor with subscript k. As a result, a function, $f(t)$, may be described using this family of wavelet basis functions as,

$$f(t) = \sum_{k} c_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k} d_{j,k}\psi_{j,k} \, . \tag{4.1}$$

Equation 4.1 includes the coefficients, $c_{j,k}$ and $d_{j,k}$, also called the wavelet coefficients. The discrete wavelet transform (DWT) is used to obtain these coefficients by estimating the inner products of the signal and wavelet functions. Often the first part of Equation 4.1 is referred to as the approximation of the function and represents the low-frequency content of the signal, while the second part is referred to as the details and represents the high-frequency content.

The DWT may be performed on a signal as part of an iterative process known as wavelet decomposition. This process decomposes the low-frequency approximation into further approximations and details using a recursive filtering and downsampling process. Figure 4.1 illustrates this process with the level i Approximation (A_i) and level i Details (D_i), each of which correspond to the appropriate subset of wavelet coefficients. A heavily decomposed approximation of the signal will appear as a very coarse version of the original signal, since only the low frequency components are present. Wavelet reconstruction is the process of reversing wavelet decomposition, where the approximations and details are united back into the original signal with no information loss.

As previously mentioned, wavelet functions possess certain desirable mathematical properties. First, wavelets comprise an orthogonal basis and the DWT is an orthogonal linear transform matrix. The linearity of the DWT results in noisy estimated coefficients whose errors are transformations of the original observed errors. A second wavelet property is sparsity, where a noise-free signal may be well-represented with a small number of nonzero wavelet coefficients. These two properties make wavelets an effective tool for de-noising and compression applications. A de-noised and compressed representation of the signal reduces the dimension of the original signal.



$$S = A_1 + D_1$$
$$= A_2 + D_2 + D_1$$
$$= A_3 + D_3 + D_2 + D_1$$

Figure 4.1: Decomposition of signal S into Approximation and Details [48]

Wavelet thresholding is the process for de-noising and compressing a signal. Thresholding operates by transforming the signal using the DWT and then reconstructing a de-noised or compressed version of the signal by using a subset or modified version of the calculated wavelet coefficients. Donoho and Johnstone [25] introduced this effective technique for de-noising, taking advantage of wavelet sparsity and assuming independent, identically distributed normal errors. For example, the wavelet transform of a noisy signal will have many nonzero coefficients as opposed to a noise-free signal, which has only a few nonzero coefficients. By identifying a value which represents the wavelet coefficient noise, the noisy signal's wavelet coefficients may be modified or thresholded, resulting in a de-noised signal. Wavelet coefficients that fall beneath the designated threshold are set to zero, and the remaining coefficients are retained in a hard thresholding approach. One of the most common thresholds is the universal threshold, λ, proposed by Donoho and Johnstone,

$$\lambda = \hat{\sigma}\sqrt{2log(n)} \tag{4.2}$$

where $\hat{\sigma}$ is a consistent estimate of the standard deviation of the noise and n is the sample size. The noise estimate, $\hat{\sigma}$, is traditionally calculated using the Median Absolute Deviation (MAD) of the finest scale detail coefficients.

4.4 WANOVA

Several authors have introduced work on WANOVA or related topics [1, 26, 27, 32, 47, 65, 83]. This paper specifically extends WANOVA procedures developed by Girimurugan *et al*. [32] as they are well-suited for adaptation as a model validation tool. The WANOVA procedure developed by Girimurugan *et al*. may be used to test a null hypothesis that a system and model time-series data are statistically equivalent.

The WANOVA methodology by Girimurugan *et al*. [32] was developed for detecting differences among functional data with a focus on profile monitoring and change point detection problems. Their research begins with the FANOVA model outlined by Ramsay and Silverman [64], which is then extended into an ANOVA model that uses a multivariate test statistic based on the Hotelling T^2 statistic. Given a multivariate response of dimension n, let the noise $\boldsymbol{\epsilon}_{ij} \in \mathbb{R}_n$ have a multivariate normal random distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ is defined as $\sigma^2 \boldsymbol{I} \in \mathbb{R}^{n \times n}$. Then the functional response Y_{ijk}, for treatment, $i = 1, 2, ..., t$, replicate, $j = 1, 2, ..., r_i$, and response, $k = 1, 2, ..., n$ is

$$Y_{ijk} = \mu + \alpha_{ijk} + \epsilon_{ijk} \, . \tag{4.3}$$

Then the Hotelling-FANOVA statistic is

$$\vartheta = \frac{(\varphi - t) \sum_{k=1}^{n} (\overline{Y}_{i.k} - \overline{Y}_{...})' (\overline{Y}_{i.k} - \overline{Y}_{...})}{(t-1) \sum_{i=1}^{t} \sum_{j=1}^{r_i} (Y_{ij} - \overline{Y}_{i.})' (Y_{ij} - \overline{Y}_{i.})} \tag{4.4}$$

where the subscript "." represents the sum across the applicable parameter, an overbar

represents an average, and $\varphi = \sum_{i=1}^{t} r_i$. The statistic, ϑ, is used to test for differences in a functional response. ϑ follows an F distribution with degrees of freedom $n(t-1)$ and $n(\varphi - t)$ under the null hypothesis that the functional responses are statistically equivalent. Therefore a significant test would lead to the conclusion that the treatment yields a significant difference in the functional data [32].

Girimurugan *et al*. adjust this Hotelling-FANOVA statistic by estimating the variation using the Median Absolute Deviation (MAD) instead of Mean Square Error (MSE) in the denominator of Equation 4.4 and by estimating the sum of squares in the wavelet domain. Therefore, σ is estimated using MAD,

$$\hat{\sigma} = \frac{1}{\Phi^{-1}(3/4)} \mathbb{M}\left[\left|W_{(J-1)k} - \mathbb{M}[W_{(J-1)k}]\right|\right] \tag{4.5}$$

where $\mathbb{M}$ is the median operator and only the finest scale detail coefficients, $W_{(J-1)k}$, are used to represent the noisy components of the signal. Next, the Hotelling-FANOVA statistic is modified by representing the sum of squares in the wavelet domain. This modified statistic is

$$\vartheta = \left(\hat{\sigma}^2(t-1)\right)^{-1} \sum_{i=1}^{t} \frac{1}{\varsigma_i} \mathbb{W}[\overline{Y}_{i.} - \overline{Y}_{..}]' \mathbb{W}[\overline{Y}_{i.} - \overline{Y}_{..}] \tag{4.6}$$

where $\mathbb{W}$ represents the DWT and $\varsigma_i = \frac{1}{r_i} + \frac{1}{t}\sum_{j=1}^{t} \frac{1}{r_j}$ [32].

Let the coefficients for the treatment i effect be defined by

$$\Theta_i = \hat{\sigma}^{-1} \mathbb{W}\left[\overline{Y}_{i.} - \overline{Y}_{..}\right] \tag{4.7}$$

where $\Theta_i = \{\theta_{i1}, \theta_{i2}, ..., \theta_{in}\}$. Then let $\tilde{\Theta}_i = \{\tilde{\theta}_{i1}, \tilde{\theta}_{i2}, ..., \tilde{\theta}_{iT_i}\}$ represent the thresholded version of these coefficients. The proposed statistic,

$$\kappa_\eta = \sum_{i=1}^{t} \sum_{k=1}^{T_i} \tilde{\theta}_{ik}^2 \tag{4.8}$$

may be thought of as the "net energy" content corresponding to signal content among profiles. The test statistic is used to test the null hypothesis that the set of t profiles corresponding to different treatments is statistically equivalent. The distribution of the test statistic follows a complicated closed form distribution and is simulated in [32]. The statistic and associated WANOVA procedure results in fewer hypothesis tests than FANOVA and is demonstrated to offer higher power with good control of Type I error.

4.5 Model Validation Test using WANOVA

4.5.1 Methodology and Distribution.

For model validation problems, there are two sets of signals: one representing the system data and one representing the model data. We wish to compare our system data signal, $\mathbf{s}$, to our model data signal, $\mathbf{m}$. We want the model data signal to match the system data signal. These signals have a dimension of size n. Then we use WANOVA to test the hypotheses that,

$$H_0 : \mathbf{s} = \mathbf{m}$$

$$H_1 : \mathbf{s} \neq \mathbf{m}.$$

Since $\kappa_\eta \in [0, \infty)$, our test is right-tailed and we reject the null hypothesis if the WANOVA test statistic exceeds a critical value,

$$\kappa_\eta \geq \kappa_\eta(\alpha), \tag{4.9}$$

where α represents the level of significance. Otherwise, we fail to reject the null hypothesis that the model is valid.

In [32], the authors state that the test statistic follows a complicated closed form distribution and choose to estimate it via simulation. However, in subsequent work [33], Girimurugan explicitly determines the test statistic's distribution under the null hypothesis. In particular, the distribution of the κ_η test statistic is a χ^2 distribution convolved with a truncated χ^2 distribution. The justification for this distribution is due to the orthogonal nature of the wavelet transform, where normally distributed noise in the signal is transformed into normally distributed noise in the wavelet coefficients.

We assume that our signals include normally distributed errors. As part of the WANOVA technique, we apply a wavelet transform to the signals which generates normally distributed wavelet coefficients. Under the null hypothesis, the system data and model data are statistically equivalent, so the signal difference has mean zero. Therefore, under the null hypothesis, the elements of the set $\Theta_i = \hat{\sigma}^{-1}\mathbb{W}\left[\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right]$ follow a standard normal distribution, $N(0,1)$ when scaled by the appropriate estimator of functional variance [33]. These θ_i values are the wavelet coefficients associated with the scaled treatment effect before thresholding. If our WANOVA test statistic is calculated without thresholding, then

$$\kappa_\eta = \sum_{i=1}^{t}\sum_{k=1}^{n}\theta_{ik}^2 \tag{4.10}$$

follows a χ^2 distribution with degrees of freedom equal to the number of wavelet coefficients. However, since the technique calls for thresholding of the wavelet coefficients, we must adjust the distribution accordingly.

Let n_t be the number of wavelet coefficients not considered for thresholding; these are typically the approximation coefficients. Since these coefficients are not modified, they

follow a standard normal distribution and their sum of squares follows a χ^2 distribution with n_t degrees of freedom. The remaining coefficients, $n - n_t$, are thresholded and therefore follow a truncated standard normal distribution, bounded below by the amount of the threshold. Their sum of squares generates a truncated χ^2 distribution with $n - n_t$ degrees of freedom. Therefore, under the null hypothesis the WANOVA test statistic, κ_η, follows the χ^2 distribution convolved with a truncated χ^2 distribution. In particular, the distribution of κ_η is

$$\kappa_\eta \sim \chi^2_{n_t} \ * \ \left]\chi^2_{n-n_t}\right[_\lambda, \tag{4.11}$$

where $*$ represents the convolution operator,] [represents a truncated distribution, and λ is the amount of threshold [33].

4.5.2 Simulation Study.

A simulation study further illustrates the distribution of the test statistic. A set of random signals were simulated using a series of cosine waves with randomly generated frequency and phase parameters. For a given iteration, two normally distributed random error vectors were created and added to a constructed base signal to create two different noisy signals, which represent the system data and the model data. Since the two noisy signals are constructed from the same base signal and differ only by a random noise component, we assume that the null hypothesis of statistically equivalent system and model data holds true.

The study simulated three hundred iterations to generate random noisy signals under the null hypothesis and compare them using the WANOVA model validation procedure. The system and model data signals have a dimension of 1024 and are analyzed using the fourth order Daubechies wavelet with four vanishing moments. Each iteration of noisy signals is thresholded using a hard thresholding approach with the universal threshold,

where the approximation coefficients are retained. The sum of the squared thresholded coefficients is calculated to evaluate an empirical κ_η test statistic value for each iteration. This empirical distribution of κ_η statistic values is compared to the theoretical distribution described in Section 4.5.1. We used the theoretical distribution described by Equation 4.11 with parameters estimated using the dimension of the signal, the number of unthresholded approximation coefficients, and the amount of the threshold. A graphical comparison of the normal kernel function of the empirical test statistic results and the theoretical distribution is shown in Figure 4.2. The empirical and theoretical distributions are further compared using the Kolmogorov-Smirnov (K-S) Goodness-of-Fit test. This resulted in a p-value of 0.4408, so we fail to reject the null hypothesis that the two distributions are statistically equivalent.

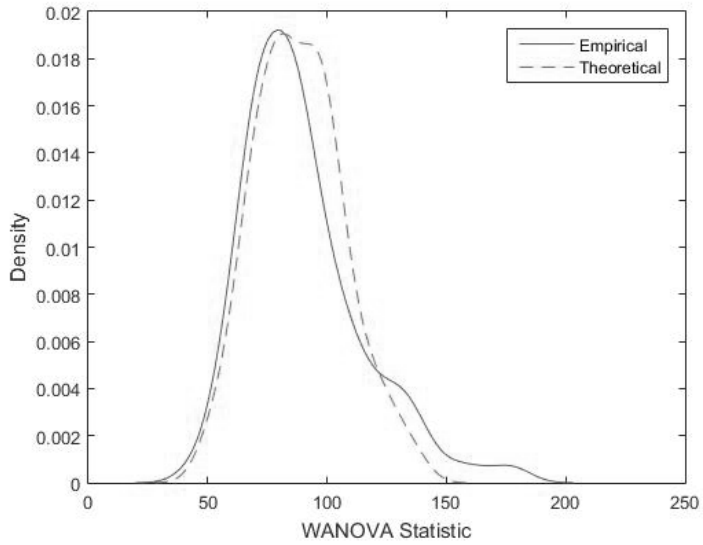


Figure 4.2: Comparison of Empirical and Theoretical Distributions

It is also meaningful to understand the distribution of the test statistic under the alternative hypothesis, that the system and model data are not statistically equivalent. Under

the alternative hypothesis, κ_η follows a non-central χ^2 distribution. Since under the null hypothesis the functional difference between the system and model data has mean zero, the scaled treatment effect wavelet coefficients follow a standard normal distribution. Then under the alternative hypothesis the coefficients have a nonzero mean and are distributed $N(\mu_i, 1)$. Therefore, the sum of squares of these coefficients are distributed according to a non-central χ^2 distribution. The next examples exploit this information.

4.6 Illustration of Approach

4.6.1 Simulated Example.

This simulated example illustrates the application of the WANOVA model validation approach. Another random base signal of dimension 1024 is simulated with two normally distributed random error vectors to create two different noisy signals. They represent the system and model data, and since they differ only by a noise component, the WANOVA model validation test should conclude that the model is valid. Figure 4.3 displays a plot of the simulated system and model data.

We wish to test the following hypotheses,

$$H_0 : \mathbf{s} = \mathbf{m}$$

$$H_1 : \mathbf{s} \neq \mathbf{m},$$

to assess the model. Applying the WANOVA model validation test yields a test statistic value of $\kappa_\eta = 75.82$. The critical value is obtained by evaluating the theoretical distribution of a χ^2 convolved with a truncated χ^2 distribution. The degrees of freedom associated with the χ^2 distribution are the number of unthresholded approximation coefficients (64), and the degrees of freedom associated with the truncated χ^2 distribution are the number of thresholded wavelet coefficients (960). Further, the truncated χ^2 distribution is developed using the amount of the threshold (2.91). Using a level of significance of $\alpha = 0.05$, the

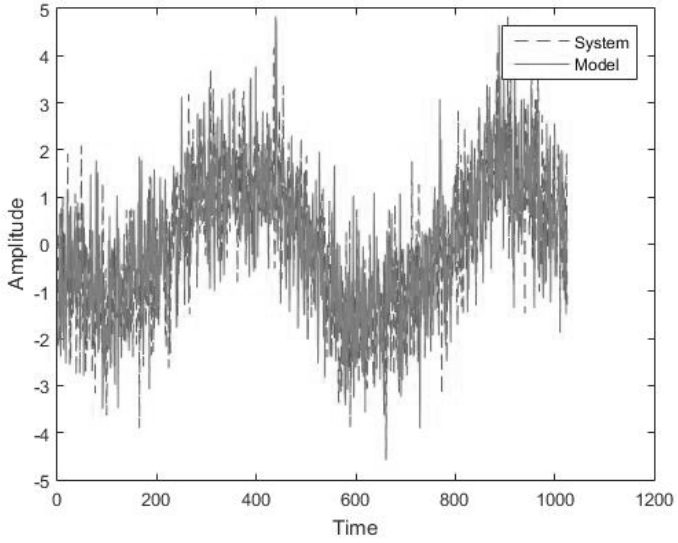


Figure 4.3: Simulation Study System and Valid Model Data

critical value is $\kappa_\eta^* = 111.13$. Since our observed statistic value is less than our critical value, we fail to reject the null hypothesis and may conclude that the model is valid, as expected.

For an alternative case, consider an example where the true system and model data are distinct. Let the underlying signal from the previous example remain the same for the system data, but the signal's frequencies are slightly modified for the model data. Since the true underlying signal for the system and model are distinct, our validation technique should conclude that the model is invalid. It is notable that the presence of noise makes it difficult to visually determine the validity of the model, as shown in Figure 4.4.

We again apply the WANOVA model validation technique to test the hypothesis that the model is valid. This yields a test statistic value of $\kappa_\eta = 292.22$. Since the signal dimension remains unchanged, we may use the same critical value calculated previously of $\kappa_\eta^* = 111.13$. Since our observed statistic value is greater than our critical value, we reject

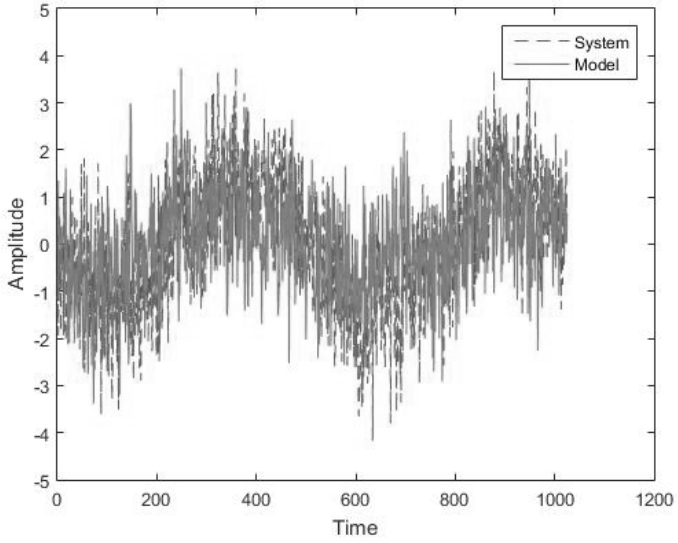


Figure 4.4: Simulation Study System and Invalid Model Data

the null hypothesis and may conclude that the model is invalid, as expected. Additional simulation studies may be found in [33].

4.6.2 Automobile Crash Test Study.

The next comparison uses data from a validation study of a 1997 Honda Accord finite element crash model using actual crash test data from the National Highway Traffic Safety Administration (NHTSA). This study was originally presented in [19] and analyzes the crash signals for a full frontal impact, specifically the acceleration responses of the engine top and right-rear cross member (RRCM) of the automobile. The response data contains 1,000 data points with a sampling rate of 0.1 ms for a total time duration of 100 ms. These signals are displayed in Figure 4.5.

Cheng *et al*. determined that the computational model is a reasonable representation of the actual vehicle by using a wavelet decomposition to compare the experimental data with the model data via a validation metric. However, recent work by [4] has shown that the

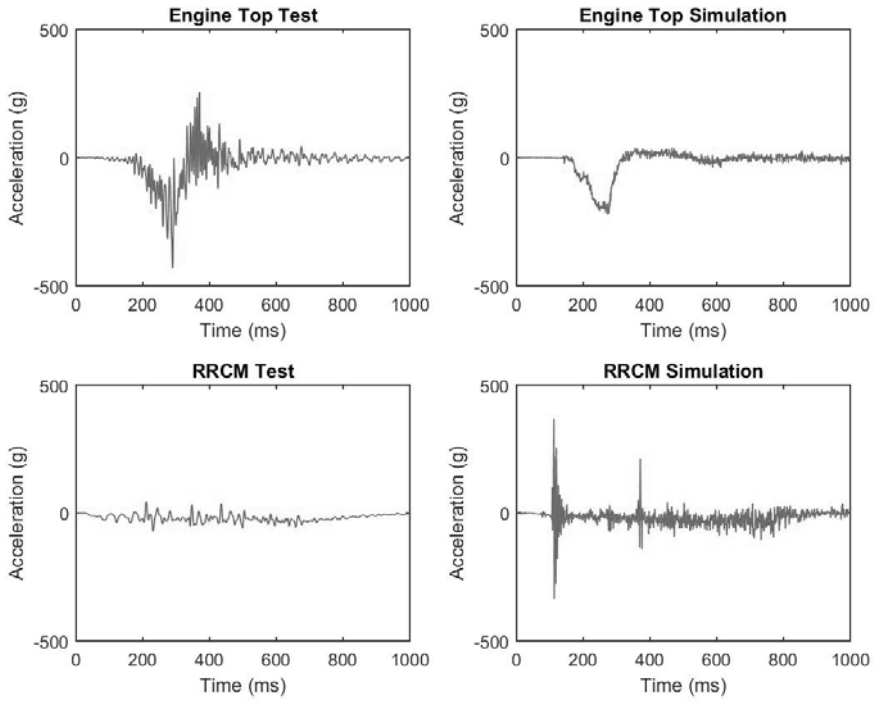


Figure 4.5: Crash Signals

use of wavelet thresholding as a more effective de-noising technique yields different results which suggest the model is not a valid representation of the system. In either case, the use of a model validation metric requires subjective analysis to determine validity, as alluded to in Section 4.1. The WANOVA model validation methodology can instead provide a definitive assessment of model validity without SME input.

In this analysis, the engine top test signal is compared to the engine top simulation signal, while the RRCM test signal is compared to the RRCM simulation signal. The fourth order Daubechies family of wavelets are again used for analysis. The engine top comparison yields a test statistic value of $\kappa_\eta = 36054$, while the RRCM comparison yields $\kappa_\eta = 6142$. In both cases, this greatly exceeds the critical value of $\kappa_\eta^* = 90.42$. Therefore, we reject the null hypothesis that the signals are statistically equivalent and conclude that

the model is invalid for simulating the acceleration response of the engine top and RRCM during an automobile crash.

4.7 Conclusion and Recommendations

This paper presents a new model validation technique that uses wavelets as part of a statistical test. Our WANOVA validation methodology assesses whether the model is an adequate representation of the system by transforming the system and model data signals using wavelets. The resulting thresholded wavelet coefficients are used to calculate a statistic to test the hypothesis that the data are statistically equivalent. The test statistic is compared to a critical value based on a χ^2 distribution convolved with a truncated χ^2 distribution. The approach was illustrated via a simulation study and an empirical study using automobile crash test data. The studies demonstrate that our WANOVA validation technique is an effective method for substantiating model time-series data.

While the WANOVA validation method offers great utility, there are a few limitations as well. First, the wavelet thresholding technique utilized relies upon the assumption of normally distributed errors, which may not always be appropriate. Second, some people question the use of hypothesis testing in general for model validation, contending that a statistically significant difference does not imply an invalid model and a statistically insignificant difference does not imply model validity. These supporters believe that the SME should not be removed from the validation process since they may assess whether the difference between the system and model data is *practically significant* [44]. Finally, if our validation method concludes that the model is invalid, we still do not have any information regarding the scope or location of the discrepancy, nor any insight to implement improvements to the model.

Despite these concerns, this method represents an important step forward in the field of model validation, particularly for proponents of eliminating subjectivity from validation assessments. WANOVA offers an objective assessment that removes an individual's bias

or lack of expertise from the validation exercise. If the model is deemed invalid, it is still possible to incorporate SME input to evaluate what model components must be corrected. Future work may evaluate the robustness of the procedure to violations of the normality assumption or provide a nonparametric solution that does not assume a distributional form for the errors associated with the observed data. In addition, future research will focus on the appropriate steps to take if the test concludes that the model is invalid. One option is to divide the system and model data signals into segments and apply the WANOVA validation technique to each segment. This would highlight the location(s) of the discrepancy in the functional data. Major differences with respect to particular wavelet coefficients may also provide information on both location and scale of model inconsistency. Work from McKay *et al*. [47] that reveals the statistically significant contrast between signals using wavelet-based functional ANOVA may also be useful for visualizing the disparity between system and model data. These follow-on steps can help model developers correct and improve the appropriate model components.

Based on current and future work, WANOVA offers great potential as a model validation technique. First, it is well-suited for assessing time-series or functional data. Many systems and models now generate functional data, which creates unique analysis challenges associated with dimensionality. WANOVA circumvents the issues of low statistical power or high Type I error associated with a traditional ANOVA approach by taking advantage of wavelet sparsity and thresholding. Second, the technique does not rely upon subjective assessments that may bias the results of the analysis. Techniques such as model validation metrics or error metrics still depend upon a SME to determine what metric value constitutes a valid model. WANOVA provides a more objective, statistical approach to assess the model. Lastly, the wavelet thresholding step assesses and removes the noise inherent in the signals, making it capable of evaluating noisy system and model data. For

these reasons, WANOVA is a powerful tool for validating a simulation model and shows promise as an integral step in the future of model V&V.

V. Wavelet ANOVA Bisection Method for Identifying Simulation Model Bias

5.1 Introduction

Advances in computer hardware technology have allowed the scientific community to build high-resolution computer models capable of simulating complex systems and processes. These computer models can not only evaluate a solution quickly and inexpensively, but also produce dynamic functional output, such as a set of time-series data generated during a process. Since computer simulation technology has quickly advanced, it is critical that the set of verification and validation (V&V) techniques similarly progresses. V&V is an integral part of the simulation development process, one that assesses the accuracy and suitability of the model before relying upon the results.

V&V techniques vary both in quality and applicability to certain models. Often, the quality of the technique may be judged by the amount of subjectivity involved. Basic V&V approaches [6] include subjective, visual comparisons of system data to model data. More advanced methods [7] utilize statistical comparisons of the data that are very complete and more objective. The applicability of a particular V&V technique may depend on the nature of the simulation output. For example, simulation output may include discrete forms and functional forms depending on the system being modeled. Discrete simulation output includes measures such as means and variances, while functional output includes time-series data.

It is clear that while there are a wide variety of V&V techniques available, it is important to select an approach that meets both quality and applicability requirements. This paper focuses on objective, statistical validation techniques used to evaluate models that generate functional output. There are several types of validation methods that meet this criteria [2, 4, 26, 27, 57]. However, once these validation techniques are applied, if the model is assessed as invalid, analysts are still limited in both knowledge and understanding

as to the exact nature of the problem leading to the conclusion of an invalid model. The logical, follow-up question to an assessment of invalidity is, "what is wrong with the model?" If the model generates functional output, such as time-series data, it would be very valuable to identify over what range the model data are a poor representation of the system data. Alternatively, over what range is the model data a good representation of the system? Current techniques stop before answering these resulting questions.

This paper presents a sequential validation methodology that answers the resulting questions associated with an invalid model based on functional output. This method first assesses the validity of a model using wavelet analysis of variance (WANOVA). If the model is declared invalid, the wavelet-based test statistic is used in conjunction with a traditional bisection univariate search approach to compare the system and model data and identify the interval with the largest discrepancy. This establishes the region in the signal over which the model data are most biased in relation to the system data. The identification of this biased region in the signal then allows developers to correct the appropriate components of the model.

The paper is organized as follows: Section 5.2 surveys the available literature on model validation and wavelet-based functional data analysis. Section 5.3 reviews wavelet analysis and WANOVA as a model validation technique. Section 5.4 presents the WANOVA Bisection method for identifying simulation model bias. Section 5.5 provides a detailed example of the method applied to a simulation study and the results from a large number of simulations. Finally, Section 5.6 identifies several distinct invalid model scenarios and assesses the performance of the algorithm under these conditions.

5.2 Literature Review

The concept of simulation can be traced back to sampling theory demonstrated with the Buffon Needle Experiment in 1777 in what would become the Monte Carlo simulation method [53]. Since then, the advent of computer technology opened new doors in the

field of computer simulation. In 1943, Ulam used one of the first electronic general-purpose computers to conduct computer based simulations that would numerically estimate solutions to intractable problems associated with the Manhattan Project and actually coined the phrase Monte Carlo for the statistical sampling approach [36, 53]. With the rise of computer based simulations, some recognized the need to assess the simulation process critically and define a framework of steps to follow to ensure the quality of the resulting simulation. These steps included evaluating the model for both correctness and suitability. In 1979, Sargent [70] presented one of the first in a sequence of papers on simulation validation. Over time, Sargent, Balci [6], and Kleijnen [43] developed some of the foundational work on simulation validation. Today, Balci [6] describes verification as "building the model right," whereas validation evaluates "building the right model."

Over the years, a wide range of validation techniques have emerged. For example, Balci [7] describes informal techniques that rely on human judgment and dynamic techniques that utilize statistical analysis such as hypothesis testing and confidence intervals. However, one needs to recognize that many established statistical techniques are designed for use with models that generate discrete output. Alternative techniques are required to assess models that generate functional output, such as time-series data. Performing analysis on a single parameter, such as the mean, of the functional data is an oversimplification of the system and model results.

Model validation metrics provide a comprehensive technique for evaluating models that generate time-series data. Validation metrics measure the discrepancy between system and model data by calculating the error associated with different signal components, such as correlation, lag, and magnitude. Together, these errors comprise an overall validation metric that describes the level of agreement between two data signals. Oberkampf and Barone [57] discuss the construction of validation metrics and some recommended features. Several authors including Atkinson *et al*. [4], Geers [31], Russell [68], and Sarin *et al*. [73]

introduce different versions of validation metrics. However, an important shortcoming with the use of validation metrics is that they still require a subjectively chosen metric value to declare model validity. Accordingly, Sargent [71] expresses concerns with the use of validation metrics and the subjectivity required in their use.

More objective model validation techniques exist within the field of functional data analysis. Functional data analysis is the statistical study of functional data and includes Functional Analysis of Variance (FANOVA). Ramsay and Silverman [64] describe FANOVA as a statistical test on whether a treatment has an effect on the functional response. For time-series data, this basic FANOVA method evaluates a univariate ANOVA for each value of time. Unfortunately, a drawback to this approach is that the dimension of the response can lead to a large number of hypothesis tests and a compounding Type I error rate. Other authors [26, 27, 32] have introduced methods to control this Type I error via multivariate statistics and wavelet thresholding. Wavelets may offer benefits in this regard, as they are known for their data compression capabilities.

Wavelets transform data from the time domain to the time-frequency domain. They offer the benefits of smoothing, dimension reduction, and decorrelation of data [16, 20, 59]. Several authors [32, 47, 83] explore wavelet-based functional data analysis, or WANOVA, an approach whose models operate by transforming the data to the wavelet domain and calculating an appropriate test statistic. This test statistic is applied to general tests of significance with functional data. Section 5.3 of this paper discusses the dynamics of wavelet analysis in further detail.

Wavelet-based model validation is well-suited for assessing models that generate functional data for the reasons given above. There are wavelet validation methods based on model validation metrics [4, 19] and wavelet coherence [41]. Recently, a WANOVA based validation effort has been proposed [2] which uses a WANOVA test statistic [32] to test for a statistically significant difference between system and model data. This technique offers

an objective evaluation of the model that is capable of examining data of large dimension. However, if any of the aforementioned techniques conclude that the model is invalid, there is still little to no information regarding the extent or location of the disparity, nor any insight for correcting the model. A technique that not only assesses model validity but also provides information on the region of model bias would be quite valuable and is presented in this work.

5.3 Wavelet Analysis and WANOVA

5.3.1 Wavelets.

As introduced above, wavelet analysis transforms data signals from the time domain to the time-frequency domain via a set of wavelet basis functions. Ogden [59] states, "broadly defined, a wavelet is simply a wavy function carefully constructed as to have certain mathematical properties. An entire set of wavelets is constructed from a single 'mother wavelet' function, and this set provides useful 'building block' functions that can be used to describe any in a large class of functions." Wavelets operate in a manner similar to a Fourier transform, but add several advantages, including computational efficiency and the ability to transform non-stationary data. Several textbooks on wavelets [16, 20, 52, 59, 84] are available for further instruction.

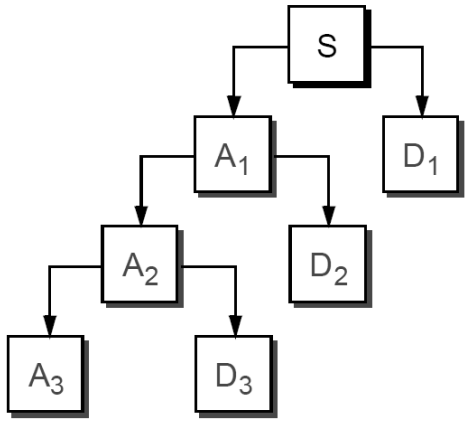
Wavelets are typically defined using a mother wavelet (ψ) and father wavelet (ϕ). They may be chosen from a group of established and commonly used wavelets, such as those developed by Daubechies, Meyer, or Shannon. The parent wavelet functions generate an entire family of wavelets through dilations and translations. The dilation index or scale factor is expressed using subscript j and the translation index or shift factor with subscript k. A linear combination of these shifted and scaled versions of the wavelet functions represent a signal as,

$$f(t) = \sum_{k} c_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k} d_{j,k}\psi_{j,k} \,, \tag{5.1}$$

where $c_{j,k}$ and $d_{j,k}$ represent the wavelet coefficients. These wavelet coefficients are estimated via the Discrete Wavelet Transform (DWT), which calculates the inner products of the signal and wavelet functions. In Equation 5.1, the summation containing the father wavelet is known as the low-frequency "Approximation," while the summation containing the mother wavelet is the high-frequency "Detail."

The wavelet decomposition process separates the high and low frequency content of the signal through an iterative application of the DWT. Each level of signal approximation is successively decomposed into another level of approximation and details until the desired resolution level is attained. Therefore, the level i approximation (A_i) would be decomposed into the level $i + 1$ approximation and details. Figure 5.1 illustrates this process, where A_3 is $\sum_k c_{j_0,k}\phi_{j_0,k}$, D_3 is $\sum_k d_{j_0,k}\psi_{j_0,k}$, D_2 is $\sum_k d_{j_0+1,k}\psi_{j_0+1,k}$, and D_1 is $\sum_k d_{j_0+2,k}\psi_{j_0+2,k}$. Further, the inverse wavelet transform may be applied to perfectly reconstruct the original signal from these approximations and details. The wavelet decomposition process plays a role in the data compression and de-noising capabilities of wavelets, most notably in a process called wavelet thresholding.

Wavelet thresholding, or wavelet shrinkage, is a technique used to compress or de-noise a data signal. It utilizes two key properties of wavelet transforms: orthogonality and sparsity. First, the DWT is an orthogonal linear transform matrix such that the original observed errors are transformed into noisy estimated wavelet coefficients. Second, wavelet sparsity asserts that most of a clean signal's energy is concentrated in a small subset of wavelet coefficients and the remaining coefficients are zero. Donoho and Johnstone [25] introduce thresholding while assuming normal, independent errors. In this case,



$$S = A_1 + D_1$$
$$= A_2 + D_2 + D_1$$
$$= A_3 + D_3 + D_2 + D_1$$

Figure 5.1: Decomposition of signal S into Approximation and Details [48]

orthogonality ensures the wavelet-transformed errors retain their normality. They define a universal threshold,

$$\lambda = \hat{\sigma}\sqrt{2log(n)}, \tag{5.2}$$

where $\hat{\sigma}$ is a consistent estimate of the standard deviation of the noise and n is the sample size. This universal threshold represents a reasonable upper bound on wavelet coefficient noise, so all wavelet coefficients that fall below this threshold are set to zero. The de-noised signal may then be reconstructed with the remaining nonzero wavelet coefficients. Wavelet thresholding is thus an effective technique for de-noising and dimension reduction, and is also an important step in the WANOVA process discussed in the next section.

5.3.2 WANOVA.

WANOVA consists of a statistical hypothesis test performed in the wavelet domain. These wavelet-based ANOVA models offer several benefits, such as smoothing the functional data and reducing dimensionality. Several authors [1, 26, 27, 32, 47, 65, 83]

introduce work on WANOVA or related topics. Specifically, this paper advances the methods presented by Girimurugan *et al*. [32] and Atkinson *et al*. [2].

Girimurugan *et al*. develop a WANOVA procedure for detecting differences among functional data aimed at profile monitoring applications. The authors first develop a FANOVA model based on the multivariate Hotelling T^2 statistic. This model considers a functional response Y_{ijk}, for treatment, $i = 1, 2, ..., t$, replicate, $j = 1, 2, ..., r_i$, and response, $k = 1, 2, ..., n$. The noise is assumed multivariate normal, $N(\mathbf{0}, \mathbf{\Sigma})$, with covariance matrix $\mathbf{\Sigma}$ equal to $\sigma^2 \boldsymbol{I} \in \mathbb{R}^{n \times n}$ [32].

Girimurugan *et al*. modify the Hotelling-FANOVA statistic by estimating the sum of squares in the wavelet domain and the variation using the Median Absolute Deviation (MAD) of the finest scale detail coefficients. This results in the modified Hotelling-FANOVA statistic,

$$\vartheta = \left(\hat{\sigma}^2(t-1)\right)^{-1} \sum_{i=1}^{t} \frac{1}{\varsigma_i} \mathbb{W}[\overline{Y}_{i.} - \overline{Y}_{..}]' \mathbb{W}[\overline{Y}_{i.} - \overline{Y}_{..}] \,, \tag{5.3}$$

where $\varsigma_i = \frac{1}{r_i} + \frac{1}{t} \sum_{j=1}^{t} \frac{1}{r_j}$, $\mathbb{W}$ represents the DWT, subscript "." represents the sum across the applicable parameter, and an overbar represents an average. Then let the set of wavelet coefficients for the treatment i effect be defined by

$$\Theta_i = \hat{\sigma}^{-1} \mathbb{W} \left[\overline{Y}_{i.} - \overline{Y}_{..} \right], \tag{5.4}$$

and let $\tilde{\Theta}_i = \{\tilde{\theta}_{i1}, \tilde{\theta}_{i2}, ..., \tilde{\theta}_{iT_i}\}$ represent the thresholded version of these coefficients. The proposed test statistic,

$$\kappa_\eta = \sum_{i=1}^{t} \sum_{k=1}^{T_i} \tilde{\theta}_{ik}^2, \tag{5.5}$$

is used to test the null hypothesis that the set of t profiles corresponding to different treatments is statistically equivalent [32].

Atkinson *et al*. [2] adapt this WANOVA methodology to solve model validation problems. The system data signal, $\mathbf{s}$, is compared to the model data signal, $\mathbf{m}$, each with dimension n. If the model is valid, the model data signal should match the system data signal. WANOVA tests the hypotheses that,

$$H_0 : \mathbf{s} = \mathbf{m}$$

$$H_1 : \mathbf{s} \neq \mathbf{m}.$$

The test statistic, κ_η, is nonnegative, and at the α level of significance we reject the null hypothesis if the statistic exceeds a critical value,

$$\kappa_\eta \geq \kappa_\eta(\alpha). \tag{5.6}$$

Otherwise, we fail to reject the null hypothesis that the model is valid. Under the null hypothesis, the κ_η test statistic is distributed as a χ^2 distribution convolved with a reverse truncated χ^2 distribution with degrees of freedom based on the signal dimension and the number of thresholded wavelet coefficients. In particular, the distribution of κ_η is

$$\kappa_\eta \sim \chi^2_{n_t} \ * \ \left]\chi^2_{n-n_t}\right[_\lambda, \tag{5.7}$$

where n is the signal dimension, n_t is the number of wavelet coefficients not considered for thresholding, $*$ represents the convolution operator,] [represents a reverse truncated distribution, and λ is the amount of threshold. Girimurugan *et al*. [33] and Atkinson *et al*. [2] describe the distribution of κ_η under the null and alternative hypotheses in greater detail.

5.4 WANOVA Bisection Method

The WANOVA validation methodology performs a statistical test on functional system and model data to determine whether the data are statistically equivalent. If a statistically significant difference exists, then the model is deemed invalid. This assessment of invalidity invites questions as to the nature of the difference, such as the scope and location of a discrepancy. The following WANOVA Bisection method answers these questions by identifying the interval over which the model data differ the most from the system data.

The WANOVA Bisection method operates using an iterative application of the WANOVA process. Once the model is assessed as invalid, the system and model data signals are bisected. Next, the WANOVA test statistic is calculated for each half of the signal and compared against each other. The signal half with the larger test statistic value is selected and the procedure is repeated on the selected half. This process continues until a desired interval length is achieved. The resulting interval represents the most biased region of the model data in relation to the corresponding system data.

The steps below summarize the formal WANOVA Bisection method, which mimics a traditional root-finding bisection search method [12].

- Initialization Step
  - Let $[a_1, b_1]$ be the signal interval and let ℓ be the allowable final interval of uncertainty. Let q be the smallest positive integer such that $(\frac{1}{2})^q \leq \frac{\ell}{b_1 - a_1}$. Let $p = 1$ and proceed to the Main Step.
- Main Step
  1. Let $\lambda_p = \frac{a_p + b_p}{2}$. Perform WANOVA over $[a_p, \lambda_p]$ and $[\lambda_p, b_p]$ to calculate $\kappa_{\eta a}$ and $\kappa_{\eta b}$. If $\kappa_{\eta a} > \kappa_{\eta b}$, proceed to Step 2, else proceed to Step 3.
  2. Let $a_{p+1} = a_p$ and $b_{p+1} = \lambda_p$. Proceed to Step 4.
  3. Let $a_{p+1} = \lambda_p$ and $b_{p+1} = b_p$. Proceed to Step 4.

4. If $p = q$, stop; the model discrepancy lies in the interval $[a_{q+1}, b_{q+1}]$. Otherwise, replace p by $p + 1$ and repeat Step 1.

The analyst may also seek to identify any region(s) in the signal where there is little to no bias. This region may correspond to valid sections of the model data that do not require correction. To identify this portion of the signal, simply modify Step 1 of the WANOVA Bisection method so that, "If $\kappa_{\eta a} < \kappa_{\eta b}$, proceed to Step 2, else proceed to Step 3."

5.5 Simulation Study

5.5.1 Example of Method.

A detailed example illustrates the WANOVA Bisection method for identifying the region of greatest model discrepancy. Through simulation, we generate a random signal of dimension 1024 for analysis. A series of cosine waves with randomly selected frequency and phase parameters comprise this random signal, which represents the system data without noise. An additive bias is incorporated into the signal over the interval $[128, 256]$ to represent invalid model data with a specific region of discrepancy. Last, we add normally distributed noise to both the system and model data signals to obtain representative, noisy data. Of note, the noisy signals being analyzed have magnitude ranging from approximately $(-5, 5)$, whereas the model bias is set to $+0.75$. Figure 5.2 presents the resulting system (blue) and model data (red) signals. Note that the presence of noise makes it difficult to identify whether the model is invalid, let alone allowing identification of a specific region of model discrepancy.

Before applying the WANOVA Bisection method, use WANOVA to assess whether the model is valid. Compare our calculated κ_η test statistic value to a critical value, κ_η^* at the $\alpha = 0.05$ level of significance. We obtain $\kappa_\eta = 207.50 > 144.64 = \kappa_\eta^*$ and therefore reject the null hypothesis and deem the model invalid.

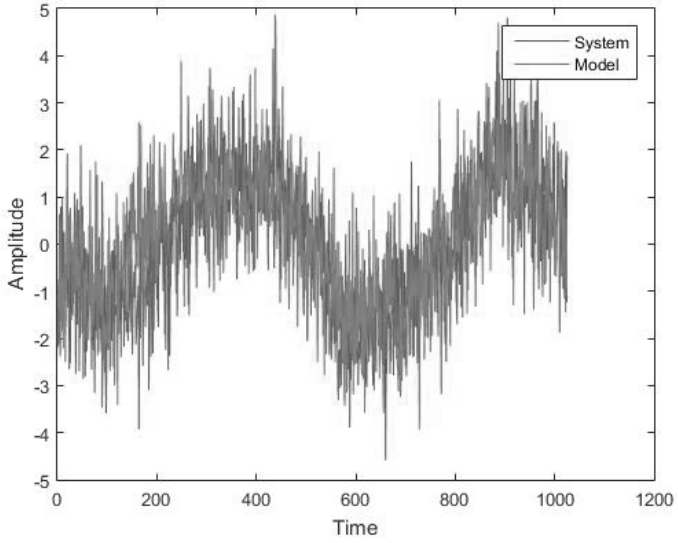


Figure 5.2: Simulation Study System and Invalid Model Data

To initialize the WANOVA Bisection method, where $[a_1, b_1] = [0, 1024]$, set $\ell = 128$ as the allowable final interval of uncertainty. In the first iteration of the main step, perform WANOVA over $[0, 512]$ and $[512, 1024]$ to calculate $\kappa_{\eta a}$ and $\kappa_{\eta b}$. In this case, $\kappa_{\eta a} = 129.33 > 88.24 = \kappa_{\eta b}$, so proceed to Step 2 where $[a_2, b_2] = [a_1, \lambda_1] = [0, 512]$. The second iteration performs WANOVA over $[0, 256]$ and $[256, 512]$. Now $\kappa_{\eta a} = 100.41 > 97.46 = \kappa_{\eta b}$, so $[a_3, b_3] = [0, 256]$. In the final iteration, analyze $[0, 128]$ and $[128, 256]$ and calculate $\kappa_{\eta a} = 15.18 < 122.28 = \kappa_{\eta b}$. Therefore, the interval is $[a_4, b_4] = [128, 256]$, and since the algorithm has achieved the allowable final interval of uncertainty, the method stops. The WANOVA Bisection method identifies the interval $[128, 256]$ as the region of largest model discrepancy, which correctly matches the region of inserted bias.

5.5.2 Large Simulation Study.

To assess the accuracy of the method over a large number of instances, a large number of system and model data signals of length 1024 are simulated. An additive bias equal to 30% of the noisy signal magnitude is incorporated into the model data with a bias duration

of 128. Additionally, the system and model data signals include normally distributed noise according to $N(0, 1)$.

The WANOVA Bisection method is applied to 500 instances where the WANOVA null hypothesis is rejected. The allowable final interval of uncertainty is set to 128. Over the 500 instances, the bisection method correctly assesses the interval of discrepancy 491 times, for an overall accuracy rate of 98.2%. The nine incorrect assessments occur when the noise overpowered the bias in the model signal. Nonetheless, the overall accuracy supports the effectiveness of the WANOVA Bisection method at identifying the region of model discrepancy.

The results obtained are positive, but not fully conclusive. Overall, the scenarios considered are relatively benign. The bias inserted into the model signal falls completely within the allowable final interval of uncertainty, and both the bias duration and allowable interval of uncertainty have length 128. Furthermore, the bias is only inserted in one location in the model signal, which allows the bisection method to focus on that particular region of discrepancy. Section 5.6 considers more complex scenarios that may be more representative of real-world system and model data. These more challenging biased model conditions serve to demonstrate the effectiveness of the WANOVA Bisection method in a variety of invalid model scenarios.

5.6 Invalid Model Scenarios

5.6.1 Incorrect Specification of Interval.

The analyses from the previous section presumed knowledge of the model bias duration. In real-world applications the analyst does not have prior knowledge of the duration of the model bias. Thus, it is important to assess the effectiveness of the validation technique when the allowable final interval of uncertainty is incorrectly specified. Towards this purpose, two scenarios are evaluated: the acceptable interval is larger than the actual region of bias, and the acceptable interval is smaller than the actual region of bias.

In the first scenario, the duration of the inserted model bias is 64, but the allowable final interval of uncertainty remains at 128. The study assesses whether the WANOVA Bisection method identifies an interval that contains the smaller region of model bias. All other study parameters remain the same from the previous simulations. Over 500 instances, the bisection method correctly identifies the interval that contains the bias region 94.9% of the time. The slightly lower accuracy rate is attributable to the smaller bias region having a smaller effect compared to the signal noise.

In the second scenario, the duration of the inserted model bias is 192 data points long with an allowable interval of uncertainty of 128. Since the bias now spans two search regions, the search region that is completely biased is the *majority region of bias*, while the search region that contains the remaining 64 biased data points is the *minority region of bias*.

The study evaluates whether the WANOVA Bisection method identifies the interval with the highest proportion of bias, i.e. the majority region of bias. All other study parameters remain the same. Over 500 instances, the bisection method correctly identifies the majority region of bias 77.8% of the time. Additionally, the bisection method identifies the minority region of bias 22.0% of the time. The algorithm was incorrect for 0.2% of iterations. Ideally, the method would identify the majority region of bias more consistently. This is due to the algorithm detecting the bias in the minority region. Nevertheless, overall the method was accurate in identifying one of the bias regions 99.8% of the time. Ultimately, these two scenarios show that despite the incorrect specification of the interval with respect to the true bias duration, the WANOVA Bisection method is still extremely effective at identifying the intervals in the model data that are biased.

5.6.2 Multiple Bias Regions.

The previous analyses assessed whether the bisection method could identify the single region of model discrepancy. In practice, the region of bias may not be limited to just

one interval. If there are two or more regions of bias in the model data, it is useful to understand which region the algorithm will identify first and what steps should be taken to find the other areas of invalidity.

The next study analyzes model data with two separate regions of varied model bias, a strong bias region where the bias is equal to 30% of the noisy signal magnitude and a weak bias region where the bias is equal to 15% of the noisy signal magnitude. All other parameters are unchanged. Over 500 instances, the algorithm identifies the strong bias region 86.5% of the time, the weak bias region 6.0% of the time, and was incorrect 7.4% of the time. The results indicate that the algorithm is more likely to identify the more biased region first. Further, a greater bias discrepancy between the two regions results in a higher likelihood that the bisection method will identify the strong bias region. Under similar levels of bias in the two regions, the technique finds each region about half of the time, as appropriate and expected.

The last study retains the two separate bias regions, but varies the bias length in each region. In particular, the long bias region has a duration of 100 data points, while the short bias region has a duration of 50. While the length of the bias is different, the magnitude of bias is equivalent. Over 500 instances, the WANOVA Bisection method identifies the long bias region on 79.8% of iterations, the short bias region on 12.2% of iterations, and an incorrect region on 8.0% of iterations. Overall, the technique generally identifies the region of highest discrepancy first, whether it is due to a larger magnitude of bias or a larger duration of bias. These results align with the preferred order of identification. Table 5.1 summarizes the results from all five of the computational analyses.

In the majority of analyses, the WANOVA Bisection method correctly identifies the more biased interval first. However, one must also account for any other regions of bias. These other problem areas need to be identified. There are two ways to initially account for this situation.

Table 5.1: WANOVA Bisection Summarized Results

Scenario Considered	Correct Identification of Biased Region			Incorrect Identification
	Primary	*Secondary*	*Total*	
Large Run Study	98.2%	-	98.2%	1.8%
Incorrect Interval (oversized)	94.9%	-	94.9%	5.1%
Incorrect Interval (undersized)	77.8%	22.0%	99.8%	0.2%
Multiple Bias Regions (magnitude)	86.5%	6.0%	92.5%	7.5%
Multiple Bias Regions (duration)	79.8%	12.2%	92.0%	8.0%

The first option is to extract the biased interval from the system and model data. Then, re-assess the remaining data signals using the suite of WANOVA model validation techniques we have described. A second option is to inform the model developers of the originally identified interval of bias allowing them to correct the necessary components of the simulation model. Then re-assess the improved version of the model using the WANOVA model validation technique. Ideally, the improved version will not show the original interval as biased. The analysts and developers can iterate this test-and-fix process using our WANOVA validation methods.

5.7 Conclusion and Recommendations

This paper presents a new model validation methodology that identifies the interval(s) in model data that are most biased in relation to associated system data. This procedure first evaluates the simulation model by executing a WANOVA validation assessment to determine if the system and model functional data are statistically equivalent. If a statistical difference exists, then the WANOVA Bisection method identifies the region of greatest model discrepancy. It may also be modified to identify the region of least discrepancy. The paper illustrates the approach via several simulation studies which demonstrate that

the WANOVA Bisection method is an effective technique for identifying the most biased interval in the model data.

Although the WANOVA Bisection method is very accurate, there are several considerations and limitations. First, the bias in the model data must be significant enough to be detected among the signal noise in the data. Second, the wavelet thresholding and WANOVA technique rely upon the assumption of normally distributed noise. Third, the method identifies the region of model discrepancy but does not provide information on the scope or nature of the bias—positive or negative. Last, as discussed in Section 5.6, the method performs best when the bias is isolated to a single location in the model data and the final interval of uncertainty is correctly specified.

In spite of these considerations, the WANOVA Bisection method is quite robust and effective at identifying regions of model discrepancy. The studies in this paper show that a bias as low as 30% of the noisy signal magnitude is sufficient for the algorithm to be correct on over 98% of problem instances. In addition, the method is quite robust to more challenging out-of-control conditions, evidenced by accuracy rates above 92% in a variety of invalid model scenarios. Future work will consider nonparametric solutions to the problem and also develop methods that not only identify the location of model discrepancy, but also provide information on the scope and nature of the bias.

Based on the studies in this paper, the WANOVA Bisection method is a very effective technique for assessing regions of model discrepancy during the validation process. This validation procedure can objectively evaluate functional system and model data through the WANOVA validation process. If the model is deemed invalid, the methodology helps answer some of the resulting questions that typically arise. In particular, the bisection method calculates the WANOVA test statistic value over different intervals of the data and performs a series of comparisons to identify the region of largest model discrepancy in relation to the system data. This process provides solutions to the questions that current

validation techniques fall short of answering. The identification of the biased interval(s) in the model data assists model developers to determine what aspects of the simulation model must be corrected. Thus, the WANOVA Bisection method is a valuable technique for identifying model bias and represents a critical step in the model validation process for functional data output.

VI. Exposing System and Model Disparity and Agreement using Wavelets

6.1 Introduction

Test and evaluation of real-world systems grows increasingly expensive and time-consuming in today's technology-driven world. When several test parameters must also be considered and varied, this can result in hundreds of necessary test runs to achieve a comprehensive understanding of the system's performance characteristics. Modeling and simulation (M&S) offers a relatively fast and inexpensive means of conducting large numbers of test runs or process executions. However, the results from a model that has not been properly verified and validated cannot be relied upon. Therefore, verification and validation (V&V) are critical steps in the simulation model development process.

Balci [5] defines V&V as follows: "model verification is substantiating that the model is transformed from one form into another, as intended, with sufficient accuracy. Model validation is substantiating that the model, within its domain of applicability, behaves with satisfactory accuracy consistent with the M&S objectives." In summary, V&V ensures the accuracy and suitability of simulation models, which is a critical and necessary step to solving today's M&S challenges.

There are several challenges and considerations associated with the V&V techniques that, once implemented, ensure the desired accuracy and suitability of the simulation models. One consideration is the amount of subjectivity required in a validation assessment. Informal validation techniques include a simple, visual comparison of system and model data that relies heavily on human reasoning and subjectivity. Alternatively, dynamic validation techniques may apply more statistical rigor [7]. A second consideration is the nature of simulation output, whether discrete measures such as means and variances, or functional output such as time-series data. Last, it may be necessary to validate system and model data that are contaminated with noise, such as experimental random error.

This noise must be identified and accounted for during the validity assessment. It is challenging to apply a statistically rigorous validation assessment to models that generate noisy, functional output.

The challenges associated with applying objective and statistically-based validation procedures to systems and models that generate noisy, functional data has led to the recent development of new validation techniques that address these challenges. One technique applies a model validation metric to signals that have been de-noised via wavelet thresholding [4]. A second technique uses Wavelet Analysis of Variance (WANOVA) to test for statistical equivalence between functional system and model data [2]. A follow-on to this technique uses a WANOVA Bisection method to identify a specific interval or region of discrepancy within the functional model data [3].

Building upon these techniques, this paper introduces a new concept of identifying the individual wavelet coefficients that contribute the most to the WANOVA test statistic. These large magnitude wavelet coefficients are used to identify and isolate the areas of discrepancy and provide information on the nature and scope of the discrepancy. This also reveals the areas of agreement between the system and model. Thus, these wavelet-based techniques offer great utility for analyzing and validating functional system and model data.

This paper is organized according to these sections: Section 6.2 summarizes the literature on model validation, functional data analysis techniques, and relevant validation case studies. Section 6.3 provides an overview of wavelet analysis and wavelet-based model validation techniques. Section 6.4 introduces the concept of examining individual wavelet coefficients to reveal the nature of the discrepancy between the system and model data.

6.2 Literature Review

Law [44] details the basics of simulation modeling by first identifying several key concepts. The *system* is "the facility or process of interest" while the *model* includes the set

of assumptions about how the system works. A *simulation* uses a computer to evaluate a model and gather data to estimate the true characteristics of the model. One of the proposed steps in a sound simulation study asks, "is the programmed model valid?" This validation assessment seeks to answer whether the simulation model is an accurate representation of the actual system. Seminal works on V&V may be found in [5, 28, 43, 66, 72, 76].

Many validation techniques have been established over the years that range from informal, subjective comparisons to formal proofs of correctness [7]. An informal comparison relies heavily on the input from a subject matter expert (SME), such as Face Validation where the SMEs review the model and judge whether the behavior and output are reasonable. Dynamic techniques require model execution and include the class of statistical techniques, such as hypothesis testing and ANOVA. However, many of these traditional statistical techniques apply only to discrete data.

The assessment of models that generate functional output requires alternative validation methods. Time-series analysis techniques, such as correlation analysis [15], offer one option. Validation metrics may incorporate the correlation coefficient as part of a more comprehensive measure of model validity [4, 31, 68]. However, validation metrics still require the subjective designation of an acceptable metric value to declare model validity [71].

Functional data analysis provides an objective, statistical approach for assessing functional data. Ramsay and Silverman [64] introduce Functional Analysis of Variance (FANOVA) as a test to identify whether a treatment has a statistically significant effect on a functional response. However, this test operates via multiple applications of a univariate ANOVA, based on the dimension of the response. Therefore, high-dimensional data lead to a large number of ANOVA tests resulting in an overwhelming Type I error rate. Alternatively, McKay *et al*. [47] describe a wavelet-based functional ANOVA technique for revealing the statistically significant contrasts between electromyographic (EMG) signals.

This wavelet-based technique provides a useful tool for identifying differences in functional data.

There are many model validation case studies that can be found in the literature, including those that seek to assess models that generate functional or time-series data. The most basic validation efforts [62] consist solely of a graph that compares a single response of field and simulation data to evaluate a traffic simulation model. Godley *et al*. [35] perform a validation study on a driving simulator that assesses the speed profile by examining the correlation. The Error Assessment of Response Time Histories (EARTH) method [73] uses a validation metric to assess models used in vehicle safety applications. Lastly, Cheng *et al*. [19] use a validation metric on wavelet-decomposed signals to assess the noisy data signals associated with automobile crash testing. These methods all vary in quality and the level of subjectivity required. Harmon and Youngblood [38] evaluate validation methods critically by providing characteristics on validation process maturity levels. They classify more mature processes as those that do not require SME input. Therefore, the best validation techniques are objective and also capable of assessing functional data.

6.3 Wavelet Analysis and Model Validation

6.3.1 Wavelets.

Wavelets may offer solutions to some of the challenges associated with high-dimensional noisy data. Wavelets can transform signals or functions from the time domain to the time-frequency domain. They operate similarly to a Fourier transform, but instead of sine and cosine waves, a wavelet function is used. The nature of wavelets leads to several advantages over the traditional Fourier transform, such as the ability to transform non-stationary data and increased computational efficiency. See [16, 20, 52, 59] for additional information regarding wavelets.

A wavelet may be selected from one of several wavelet families, such as those developed by Haar or Daubechies. Each wavelet family has slightly different characteristics, offering different advantages and disadvantages. Once a family is selected, a mother (ψ) and father (ϕ) wavelet generate an entire set of wavelets through dilations (j) and translations (k). A linear combination of this set of wavelets can express a signal,

$$f(t) = \sum_{k} c_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0} \sum_{k} d_{j,k}\psi_{j,k} \, , \tag{6.1}$$

where $c_{j,k}$ and $d_{j,k}$ are the wavelet coefficients. These coefficients are obtained via the Discrete Wavelet Transform (DWT), which calculates the inner products of the signal and wavelet functions. The iterative application of the DWT is called the wavelet decomposition of a signal. This wavelet decomposition results in different levels of low-frequency "Approximation" and high-frequency "Details." The inverse discrete wavelet transform (IDWT) recovers the signal from the wavelet coefficients and reverses the DWT.

The DWT enables wavelet thresholding, which is the technique used to de-noise a noisy signal or compress a high-dimensional signal. Donoho and Johnstone [25] first introduce wavelet thresholding and analyze a signal assuming normal, independent errors. Since the DWT is an orthogonal linear transform matrix, the signal noise transforms into noisy estimated wavelet coefficients. This orthogonal transform is used in conjunction with the wavelet sparsity property, where most of a clean signal's energy is represented by a small number of wavelet coefficients, to enable thresholding. Donoho and Johnstone define a universal threshold,

$$\lambda = \hat{\sigma}\sqrt{2log(n)}, \tag{6.2}$$

where $\hat{\sigma}$ is a consistent estimate of the standard deviation of the noise and n is the sample

size. Any estimated coefficients that fall below this threshold are set to zero, while the others are retained. This modified set of wavelet coefficients represents the de-noised signal. The wavelet thresholding process has been shown to be very effective in de-noising and data compression applications and is utilized in the WANOVA process.

6.3.2 WANOVA.

WANOVA is a technique for performing statistical inference in the time-frequency domain of wavelets. Girimurugan *et al*. [32] present a WANOVA methodology that tests for statistical differences among functional data. They adapt a traditional FANOVA model with functional response Y_{ijk}, for treatment, $i = 1, 2, ..., t$, replicate, $j = 1, 2, ..., r_i$, and response, $k = 1, 2, ..., n$, while assuming multivariate normal noise. This is developed into a wavelet representation with test statistic,

$$\kappa_\eta = \sum_{i=1}^{t} \sum_{k=1}^{T_i} \tilde{\theta}_{ik}^2. \tag{6.3}$$

In Equation 6.3, $\tilde{\theta}_{ik}$ represents the thresholded wavelet coefficients associated with the treatment i effect. The κ_η test statistic is compared to a critical value to test the null hypothesis that the t sets of functional data are statistically equivalent.

Atkinson *et al*. [2] use WANOVA for model validation assessments by comparing the system data signal, s, to the model data signal, m. This process tests the hypotheses that,

$$H_0 : \mathbf{s} = \mathbf{m}$$

$$H_1 : \mathbf{s} \neq \mathbf{m},$$

where we assume a valid model data signal should be statistically equivalent to the system data. If the test statistic exceeds a critical value, the null hypothesis is rejected and the model is deemed invalid. Further details on the distribution of κ_η and the associated critical value can be found in [2, 33].

6.3.3 WANOVA Bisection Method.

Most validation methods can assess a model as valid or invalid, but these methods fail to provide further information on the nature of the discrepancy associated with an invalid model. Atkinson *et al*. [3] introduce a technique that identifies a range in the functional data over which the model is most biased in relation to the system. This approach uses WANOVA and a bisection univariate search approach to identify the interval of greatest model discrepancy, which aids developers in correcting the necessary elements of the model.

Given an invalid model, the WANOVA Bisection method first bisects the system and model data signals and calculates the WANOVA test statistic on each half of the signal. Since the test statistic represents a value of the net signal energy between the system and model, the half with the larger statistic value is identified as the half containing the greater model bias. This process may continue until the desired interval length is reached. The resulting interval represents the region in the model data that differs most from the corresponding system data. The steps to the formal WANOVA Bisection method are detailed in [3].

6.4 Assessing Wavelet Coefficients to Expose Disparity and Agreement

6.4.1 Methodology.

Although the WANOVA Bisection method can isolate the interval of greatest model discrepancy, it still does not provide information on the scope or the nature of the bias. This leaves the analyst unclear concerning whether the model produces results greater than or less than the system at a particular signal location. The magnitude of the discrepancy is also unclear. Another limitation of the WANOVA Bisection method is that it identifies only a single interval of model discrepancy at a time. It would be valuable to understand the nature of the bias and to identify any biases at multiple points in the signal. This information

would also convey where the system and model data agree. This paper introduces a technique to accomplish this.

It is possible to expose the disparity between the system and model data by examining the individual wavelet coefficients that contribute the most to the WANOVA test statistic described in Section 6.3.2. These individual wavelet coefficients represent the difference in signal energy between the system and model for a given time/location and frequency. Therefore, by examining the large magnitude wavelet coefficients, the analyst gains insight into the location, scope, and nature of any disparities.

It is important to consider which wavelet function to use when implementing this approach. Most wavelet functions overlap within a resolution level, which refers to the state of time-frequency resolution. This may affect how the net signal energy is distributed among wavelet coefficients. The Haar wavelet [59] offers a straightforward solution, since these wavelets are non-overlapping within a resolution level.

This technique is applicable in cases where the WANOVA model validation assessment deems a model is invalid. The technique functions effectively as a follow-on step, in conjunction with the WANOVA Bisection method. First estimate the thresholded Haar wavelet coefficients associated with the mean difference signal. Then, identify coefficients that exceed a certain value depending on the resolution level. These large coefficients expose the disparity between the system and model data via the IDWT. This approach also reveals where the signals agree. The next section provides two examples, demonstrating the effectiveness and contributions of this novel process.

6.4.2 Examples.

This section provides two examples to illustrate the method of assessing wavelet coefficients to expose the system and model disparity. First a random signal of dimension 1024 is simulated containing a large variety of frequencies. This represents the base system signal. The signal includes an additive bias of one over the interval $[128, 192]$ to act as the

model data. It also contains normally distributed noise with a mean of zero and a standard deviation of one to the signals and develop two replicates of system data and two replicates of model data. Figure 6.1 displays a system data replicate in blue and a model data replicate in red.

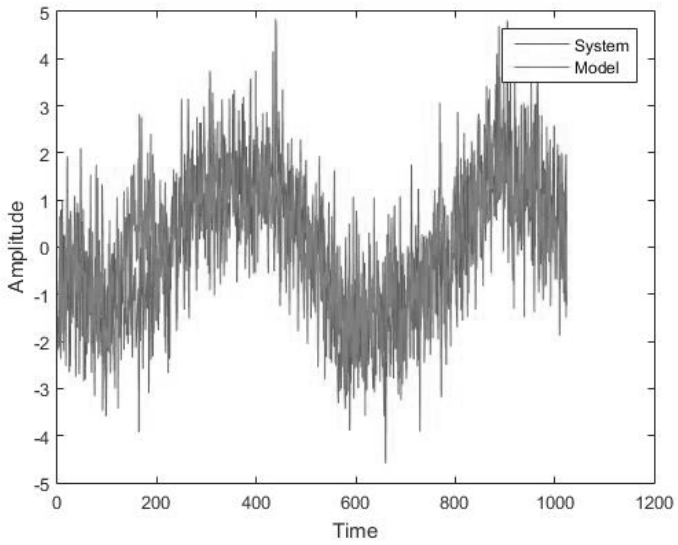


Figure 6.1: System and Model Data, Example 1

Calculate the DWT using the Haar wavelet with the mean difference signal and threshold using the universal, hard thresholding approach of Donoho and Johnstone. Analyze the resulting coefficients from four levels of wavelet decomposition. The 95th percentile of the nonzero wavelet coefficients is approximately four, so identify any wavelet coefficients that exceed a value of four. The largest 5% of nonzero wavelet coefficients offered a good representation of the model discrepancy, however this is not a hard rule. A larger percentage, such as 10% will be more sensitive to detecting discrepancies, while a smaller percentage will be less sensitive. Transform these identified coefficients back to the time domain via the IDWT to obtain Figure 6.2.

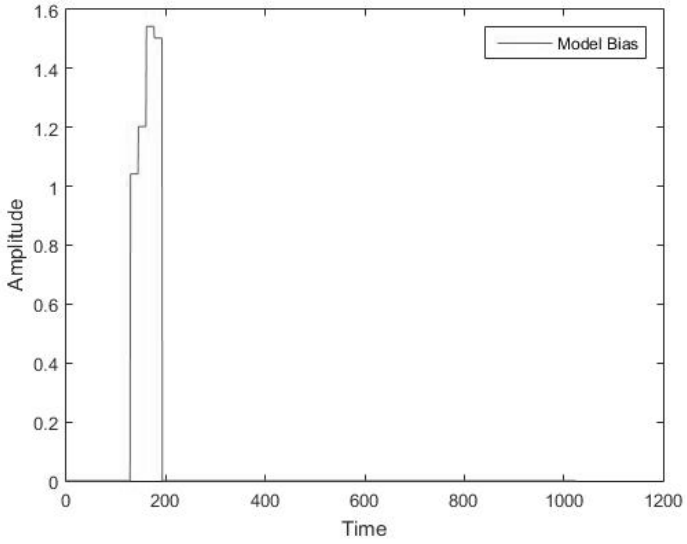


Figure 6.2: System and Model Disparity, Example 1

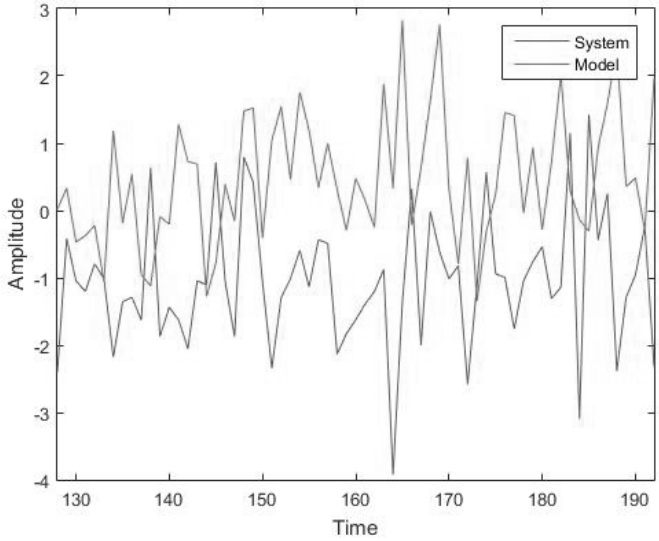


Figure 6.3: Magnification of System and Model Data Disparity

Figure 6.2 exposes the range of disparity between the system and model data that occurs over the interval $[128, 192]$, as appropriate. Note how the disparity in Figure 6.2 corresponds to the original signals in Figure 6.1. Figure 6.3 zooms in on this identified region of disparity. Next, this exposed disparity also indicates a positive model bias, as expected. We see that this positive model bias has an estimated magnitude of approximately 1 to 1.5. This is a slight overestimation of the actual bias, due to the influence of the signal noise. Finally, while Figure 6.2 depicts a positive model bias over the interval $[128, 192]$, it is equal to zero outside this interval. This indicates that the system and model data signals are in agreement over the intervals $[0, 128]$ and $[192, 1024]$.

A second example tests the method's ability to locate a negative bias in two separate locations. Generate a different random signal of dimension 1024 and add a negative bias of one over two separate intervals, $[128, 192]$ and $[768, 832]$. After including noise, there are two replicates of system and model data. One replicate of each is shown in Figure 6.4.

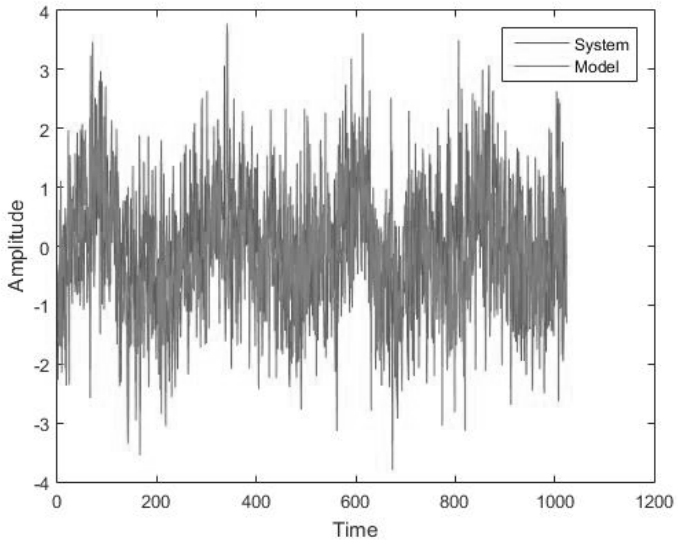


Figure 6.4: System and Model Data, Example 2

Assess the thresholded Haar wavelet coefficients, and in this example the 95th percentile of nonzero coefficients corresponds to a value of approximately two. Therefore, capture all wavelet coefficients greater than two and apply the IDWT to obtain Figure 6.5.

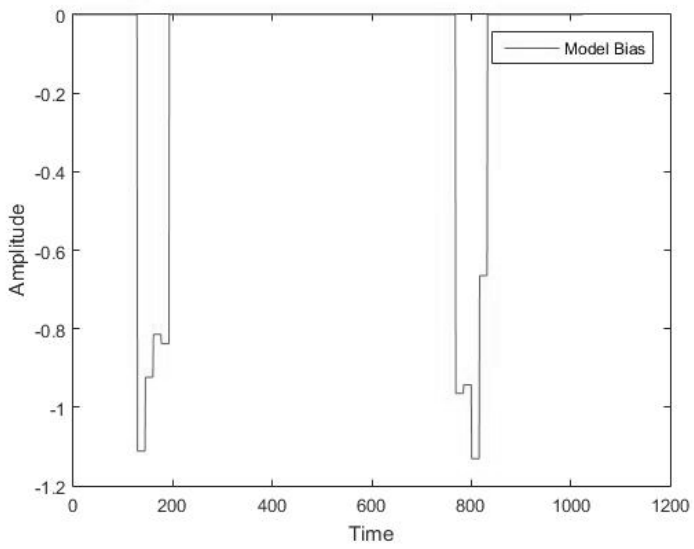


Figure 6.5: System and Model Disparity, Example 2

This second example shows that the method is capable of capturing a negative bias at multiple points in the signal. The method correctly identifies negative model bias over the intervals $[128, 192]$ and $[768, 832]$. Figure 6.5 shows the magnitude of the bias to range between -0.8 and -1.1, which is very close to the actual bias of -1. Overall, these two examples show that the large magnitude wavelet coefficients associated with the mean difference signal offer valuable insight into the disparity and agreement between the system and model data.

6.5 Conclusion

This paper contributes a novel methodology for uncovering the areas of discrepancy and accord between system and model data signals. The approach considers the WANOVA test statistic and identifies the specific wavelet coefficients that contribute the most to the statistic. These large magnitude, Haar wavelet coefficients represent the difference in signal energy between the system and model. Therefore, by implementing an IDWT, the analyst gains insight into the location and magnitude of any model biases. The results also illustrate the regions where the system and model are in agreement. Two examples demonstrate the ability of this technique to identify bias as part of a model validation assessment.

Although this technique is useful, there are some limitations to consider. First, the Haar wavelet operates as a type of step function and does not provide a smooth representation of the disparity. If the actual bias varies continuously, the Haar wavelet is limited in representing this bias by the resolution level. However, the Haar wavelet is utilized because it is non-overlapping within a resolution level, enabling the net signal energy to be distributed to individual wavelet coefficients. Another limitation is that this method assumes normally distributed noise. Finally, the wavelet-based model validation techniques described in this paper are effective when assessing time-series data. It would also be valuable to adapt the techniques to apply to models that require validation over a design space. Future work may address these limitations.

Overall, the process of assessing individual wavelet coefficients is an effective procedure for exposing the disparity and agreement between system and model data signals. The examples in this paper show the method's capability of accurately revealing the location and magnitude of model biases. This improves upon previous techniques which are limited to providing information solely on location and can only identify a single region of bias at a time. The identification of the magnitude and the positive or negative nature

of the bias aids model developers in correcting the components of the simulation model. Therefore, this technique is a valuable tool in the simulation model development process.

VII. Conclusion

This dissertation presents four novel methods for overcoming the challenges associated with the objective validation assessment of functional model data. The first approach uses wavelet thresholding in conjunction with a model validation metric to assess simulation model validity. This approach offers an innovative solution for assessing the noisy data typically associated with the transient phase of a process. The second technique applies wavelet analysis of variance (WANOVA) to model validation. This technique performs statistical inference in the wavelet domain and is capable of analyzing high-dimensional data. The third method uses WANOVA and a bisection univariate search to identify the interval of the time-series data containing the greatest model bias. This may be performed as a follow-on method to the WANOVA model validation technique. Last, the dissertation introduces a procedure that exposes areas of system and model disparity and agreement in the data. This procedure may also be applied as a follow-on to the WANOVA model validation technique.

Together, these new validation techniques use wavelet analysis to solve today's complex simulation model validation problems. When studying real-world systems, it is usually possible to collect a stream of data during a system process. Therefore, many simulation models seek to emulate this system process and similarly produce a stream of functional data, often time-series data. However, data of this form present problems that traditional verification and validation (V&V) techniques are unable to handle.

Specifically, computer models that generate functional data present many challenges that traditional V&V techniques are unable to overcome. It is first notable that most traditional V&V techniques are designed for models that generate discrete data. This includes the class of statistical techniques, such as hypothesis testing, confidence intervals, and goodness-of-fit tests. These statistical techniques typically compare the means or

variances of the data. However, when assessing functional or time-series data, the application of such a technique requires that the data be distilled down to a single parameter. This distillation discards most of the valuable information contained within a set of functional data and should be avoided.

There are other V&V techniques that do not require the data to be simplified into a discrete parameter. These may include informal techniques such as Face Validation or Turing Tests, which rely on subject matter experts (SMEs) to evaluate system and model data. These SMEs then judge whether the simulation model is an accurate representation of the real-world system. However, methods such as these rely too heavily on subjective input to make the assessment. For example, any SME biases would undermine the validation assessment. Furthermore, noisy data may make such a subjective assessment infeasible. Ultimately, the most effective validation techniques should be both objective and statistically rigorous.

There are still other V&V techniques that are both objective and able to assess functional data, yet they prove limited as well. FANOVA is an analysis method that performs a test on two or more sets of functional data to assess whether they are statistically equivalent. However, this method is essentially a univariate ANOVA for each data point, which causes a multiplicity problem and results in an uncontrolled Type I error rate. A statistically based model validation test must be able to handle high-dimensional time-series data.

Therefore, this dissertation presents four new wavelet-based validation techniques that act as powerful analysis tools when both the real-world system and the corresponding simulation model produce functional data. This dissertation is in the *k*-paper format, with four methodological papers providing contributions to the body of knowledge in simulation model verification and validation. Each of these papers is summarized in the chapter descriptions below.

Chapter 3 introduces the first wavelet-based validation approach and is designed for the examination of transient phase data. This approach utilizes wavelet thresholding and a model validation metric. Wavelet thresholding is a technique for de-noising a data signal. Since the transient phase of a process is characterized as the dynamic portion of a signal, the data are often contaminated with noise that must be addressed. Therefore, the wavelet thresholding removes this noise from the system and model data prior to further analysis. These de-noised signals are then examined via a model validation metric that measures the discrepancy between the system and model data. The validation metric incorporates shape, phase, and magnitude error. Additionally, a simulation study and empirical data from an automobile crash study illustrate the advantages of this validation approach.

Chapter 4 presents a more statistically objective validation technique using WANOVA. While a model validation metric still requires the subjective designation of an acceptable metric value, the WANOVA technique is a statistical test of model validity. The WANOVA test statistic is based on the wavelet coefficients associated with the data signals and is used to determine whether the system and model data are statistically equivalent. If the data are statistically equivalent, the model is accepted as valid. Otherwise, it is rejected as invalid. The WANOVA technique provides several advantages over a traditional FANOVA approach, including wavelet sparsity. Thresholding and wavelet sparsity allow WANOVA to overcome the dimensionality problem associated with FANOVA and prevent an uncontrolled Type I error rate.

Chapter 5 outlines the WANOVA Bisection method for identifying simulation model bias. Legacy validation techniques simply answer whether a simulation model is valid or invalid. The WANOVA Bisection method aims to advance the analysis a step further by determining the region(s) of model bias. This method first assesses model validity using the WANOVA technique described in the previous chapter. If the model is invalid, the method then calculates the WANOVA test statistic as part of a bisection algorithm that identifies the

signal interval containing the largest amount of model bias in relation to the system data. The WANOVA Bisection method yields valuable information that assists model developers in isolating problems with the model so that they make the appropriate corrections.

Chapter 6 describes a final procedure for revealing the contrast between the system and model data. The previous chapter's WANOVA Bisection method is able to identify an interval of greatest model discrepancy, but was unable to provide information on the magnitude of the bias and can only identify one interval at a time. To improve upon these limitations, this chapter presents a concept that estimates the wavelet coefficients associated with the system and model net signal and focuses on the individual, large magnitude coefficients. These wavelet coefficients are transformed back to the time domain via the inverse discrete wavelet transform (IDWT), which exposes the system and model disparity and agreement. Any exposed disparities indicate whether the model is positively or negatively biased and the magnitude of this bias. This procedure is also able to locate multiple areas of bias in a single iteration.

There are a few limitations associated with the techniques described in this dissertation. First, these methods all assume additive normally-distributed noise. This assumption is part of the universal threshold chosen for our wavelet thresholding procedure. Gaussian noise is also assumed as part of our calculation of the WANOVA test statistic critical value. Second, the dynamic model validation metric based on wavelet thresholded signals requires the subjective designation of an acceptable metric value. A passing score for a validation metric must be determined in a subjective manner. Third, some criticize the use of hypothesis testing for model validation purposes. They argue that it is more important to consider whether the system and model difference is of *practical significance*. Last, the WANOVA Bisection method requires that the model biases be significant enough to be detected among the signal noise. The bisection method is also most effective when the

final interval of uncertainty is correctly specified. These limitations should be considered when implementing one of the techniques.

There are many opportunities to improve upon or extend the work developed in this dissertation. This includes a nonparametric approach that does not make any distributional assumptions regarding the signal noise. Nonparametric wavelet thresholding techniques exist to accomplish this. Another opportunity is the extension of these methods to validate a model across a design space. This enables the validation of a model that generates functional data at different design points or system configurations. Finally, it is worthwhile to apply these validation techniques to real-world case studies. There are a plethora of Department of Defense (DoD) simulation models that require V&V. One potential application is to the Air Force Research Laboratory's (AFRL) Integrated Vehicle and Energy Technology (INVENT) program that seeks to develop a robust model-based design method for next generation aircraft. Another application is to the Naval Undersea Warfare Center's (NUWC) efforts towards developing new torpedo models as part of a broadband-capable weapons analysis facility. These DoD simulation models generate functional data that require validation, thus proving ideal cases for the application of these techniques.

Despite the limitations outlined above, these new validation techniques offer valuable analysis tools whose impact may be extended even further with future work. First, the validation metric of wavelet thresholded signals is a powerful technique for the analysis of transient phase data. The second technique presents an objective, statistical test for model validation using WANOVA. This may be followed by a step for identifying regions of model bias using the WANOVA Bisection method. Further improvements render a procedure that provides additional information regarding the bias, including the magnitude. These novel techniques are an important resource for the V&V of functional data.

The V&V of functional data is a problem that must be addressed to validate today's simulation models. The techniques in this dissertation serve to solve these problems. V&V

is an imperative step that ensures not only that a model concept is correctly implemented into a computerized model but also assesses whether the model is truly representative of the system. This step is necessary to build confidence in the model before relying upon the results. These novel validation techniques contribute to a more rigorous V&V methodology, which results in a more effective simulation model development process.

Bibliography

[1] Abramovich, F., A. Antoniadis, T. Sapatinas, and B. Vidakovic (2004). Optimal testing in a fixed-effects functional analysis of variance model. *International Journal of Wavelets, Multiresolution and Information Processing 2*(4), 323–349.

[2] Atkinson, A., R. Hill, J. Pignatiello, G. Vining, E. White, and E. Chicken. Wavelet anova approach to model validation. *under review in Simulation Modelling Practice and Theory*.

[3] Atkinson, A., R. Hill, J. Pignatiello, G. Vining, E. White, and E. Chicken. Wavelet anova bisection method for identifying simulation model bias. *under review in Simulation Modelling Practice and Theory*.

[4] Atkinson, A., R. Hill, J. Pignatiello, G. Vining, E. White, and E. Chicken (2017). Dynamic model validation metric based on wavelet thresholded signals. *Journal of Verification, Validation and Uncertainty Quantification 2*(2), 021002.

[5] Balci, O. (1997). Verification validation and accreditation of simulation models. In S. Andradottir, K. Healy, D. Withers, and B. Nelson (Eds.), *Proceedings of the 1997 Winter Simulation Conference*, pp. 135–141. Atlanta, GA: Institute of Electrical and Electronics Engineers Computer Society.

[6] Balci, O. (2003). Verification, validation, and certification of modeling and simulation applications. In S. Chick, P. Sanchez, D. Ferrin, and D. Morrice (Eds.), *Proceedings of the 2003 Winter Simulation Conference*, pp. 150–158. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

[7] Balci, O. (2013). Verification, validation, and testing. In *Encyclopedia of Operations Research and Management Science* (Third ed.)., pp. 1618–1627. New York: Springer.

[8] Balci, O. and R. G. Sargent (1980). Bibliography on validation of simulation models. *ACM Simuletter 15*(3), 15–27.

[9] 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.

[10] Balci, O. and R. G. Sargent (1984). Validation of simulation models via simultaneous confidence intervals. *American Journal of Mathematical and Management Sciences 4*(3-4), 375–406.

[11] Bayarri, M. J., J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu (2012). A framework for validation of computer models. In *Proceedings of the Workshop on Foundations for Verification and Validation in the 21st Century*. Johns Hopkins University/Applied Physics Lab.

[12] Bazaraa, M. S., H. D. Sherali, and C. M. Shetty (2006). *Nonlinear Programming Theory and Algorithms* (Third ed.). Hoboken, NJ: John Wiley & Sons, Inc.

[13] Bendat, J. S. and A. G. Piersol (1980). *Engineering Applications of Correlation and Spectral Analysis* (First ed.). New York: Wiley-Interscience.

[14] Benedetto, J. J. (1993). *Wavelets: Mathematics and Applications* (First ed.). Boca Raton, FL: CRC press.

[15] Box, G. E., G. M. Jenkins, and G. C. Reinsel (2008). *Time Series Analysis: Forecasting and Control* (Fourth ed.). New York: John Wiley & Sons, Inc.

[16] Burrus, C. S., R. A. Gopinath, and H. Guo (1998). *Introduction to Wavelets and Wavelet Transforms* (First ed.). Upper Saddle River, New Jersey: Prentice Hall, Inc.

[17] Cai, T. T. and B. W. Silverman (2001). Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyā: The Indian Journal of Statistics, Series B 63*, 127–148.

[18] Cheng, Z., J. Pellettiere, and A. Rizer (2004). Wavelet-based validation methods and criteria for finite element automobile crashworthiness modeling. In *Proceedings of the 22nd IMAC Conference and Exposition: A Conference & Exposition on Structural Dynamics, Dearborn, MI*, pp. 1753–1769.

[19] Cheng, Z., J. Pellettiere, and N. Wright (2006). Wavelet-based test-simulation correlation analysis for the validation of biodynamical modeling. In *Proceedings of the 24th Conference and Exposition on Structural Dynamics, St. Louis, MO*, pp. 2124–2132.

[20] Chui, C. K. (1992). *An Introduction to Wavelets* (First ed.). Boston, MA: Academic Press.

[21] Conway, R. W. (1963). Some tactical problems in digital simulation. *Management Science 10*(1), 47–61.

[22] Dillon, W. R. and M. Goldstein (1984). *Multivariate Analysis: Methods and Applications* (First ed.). New York: Wiley.

[23] Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory 41*(3), 613–627.

[24] Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*(432), 1200–1224.

[25] Donoho, D. L. and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425–455.

[26] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association 91*(434), 674–688.

[27] Fan, J. and S.-K. Lin (1998). Test of significance when data are curves. *Journal of the American Statistical Association 93*(443), 1007–1021.

[28] Fishman, G. S. and P. J. Kiviat (1967). The analysis of simulation-generated time series. *Management Science 13*(7), 525–557.

[29] Fitzsimmons, J. A. (1974). The use of spectral analysis to validate peanning models. *Socio-Economic Planning Sciences 8*(3), 123–128.

[30] Fugal, D. L. (2009). *Conceptual wavelets in digital signal processing: an in-depth, practical approach for the non-mathematician* (First ed.). Space & Signals Technical Pub.

[31] Geers, T. L. (1984). An objective error measure for the comparison of calculated and measured transient response histories. *The Shock and Vibration Bulletin 54*, 99–108.

[32] Girimurugan, S., E. Chicken, J. J. Pignatiello Jr, and M. S. Zeisset (2013). Wavelet anova for detection of local and global profile changes. In A. Krishnamurthy and W. Chan (Eds.), *Proceedings of the 2013 Industrial and Systems Engineering Research Conference*, pp. 3235–3244. San Juan, PR: Institute of Industrial Engineers.

[33] Girimurugan, S. B. (2014). *Nonlinear Multivariate Tests for High-Dimensional Data Using Wavelets with Applications in Genomics and Engineering*. Ph. D. thesis, Florida State University.

[34] Girimurugan, S. B., K. Hillebrandt, E. Chicken, and J. J. Pignatiello Jr (2015). Nonparametric detection of profile treatment differences. In S. Cetinkaya and J. Ryan (Eds.), *Proceedings of the 2015 Industrial and Systems Engineering Research Conference*, pp. 118–127. Nashville, TN: Institute of Industrial Engineers.

[35] Godley, S. T., T. J. Triggs, and B. N. Fildes (2002). Driving simulator validation for speed research. *Accident Analysis & Prevention 34*(5), 589–600.

[36] Goldsman, D., R. E. Nance, and J. R. Wilson (2009). A brief history of simulation. In M. Rossetti, R. Hill, B. Johansson, A. Dunkin, and R. Ingalls (Eds.), *Proceedings of the 2009 Winter Simulation Conference*, pp. 310–313. Austin, TX: Institute of Electrical and Electronics Engineers Computer Society.

[37] Grinsted, A., J. C. Moore, and S. Jevrejeva (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics 11*(5/6), 561–566.

[38] Harmon, S. and S. M. Youngblood (2005). A proposed model for simulation validation process maturity. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology 2*(4), 179–190.

[39] Jauregui, R., P. J. Riu, and F. Silva (2010). Transient FDTD simulation validation. In *2010 IEEE International Symposium on Electromagnetic Compatibility (EMC)*, pp. 257–262.

[40] Jiang, X. and S. Mahadevan (2008). Bayesian wavelet method for multivariate model assessment of dynamic systems. *Journal of Sound and Vibration 312*(4), 694–712.

[41] 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.

[42] Kleijnen, J. and W. S. van Groendall (1992). *Simulation: A Statistical Perspective*. New York: John Wiley & Sons, Inc.

[43] Kleijnen, J. P. (1995). Verification and validation of simulation models. *European Journal of Operational Research 82*(1), 145–162.

[44] Law, A. M. (2013). *Simulation Modeling and Analysis* (Fifth ed.). New York: McGraw Hill.

[45] McGinnity, K., R. Varbanov, and E. Chicken (2017). Cross-validated wavelet block thresholding for non-Gaussian errors. *Computational Statistics & Data Analysis 106*, 127–137.

[46] McHale, M., J. Friedman, and J. Karian (2009). Standard for verification and validation in computational fluid dynamics and heat transfer. Technical report, The American Society of Mechanical Engineers, Three Park Avenue, New York, NY 10016.

[47] McKay, J. L., T. D. Welch, B. Vidakovic, and L. H. Ting (2013). Statistically significant contrasts between EMG waveforms revealed using wavelet-based functional anova. *Journal of Neurophysiology 109*(2), 591–602.

[48] Misiti, M., Y. Misiti, G. Oppenheim, and J.-M. Poggi (1997). *Wavelet Toolbox Getting Started Guide* (First ed.). Natick, MA: The Mathworks, Inc.

[49] Mitchell, P. (1997). Misuse of regression for empirical validation of models. *Agricultural Systems 54*(3), 313–326.

[50] Montgomery, D. C. (2008). *Design and Analysis of Experiments* (Sixth ed.). New York: John Wiley & Sons, Inc.

[51] Montgomery, D. C., E. A. Peck, and G. G. Vining (2012). *Introduction to Linear Regression Analysis* (Fifth ed.). New York: John Wiley & Sons, Inc.

[52] Najmi, A.-H. (2012). *Wavelets A Concise Guide* (First ed.). Baltimore, MD: Johns Hopkins University Press.

[53] Nance, R. E. and R. G. Sargent (2002). Perspectives on the evolution of simulation. *Operations Research 50*(1), 161–172.

[54] Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 463–479.

[55] Naylor, T. H. and J. M. Finger (1967). Verification of computer simulation models. *Management Science 14*(2), B92–B101.

[56] Naylor, T. H., K. Wertz, and T. H. Wonnacott (1967). Methods for analyzing data from computer simulation experiments. *Communications of the ACM 10*(11), 703–710.

[57] 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.

[58] Oberkampf, W. L. and T. G. Trucano (2000). Validation methodology in computational fluid dynamics. Technical Report 2000-2549, American Institute of Aeronautics and Astronautics.

[59] Ogden, T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis* (First ed.). Boston: Birkhauser.

[60] Pace, D. K. (2004). Modeling and simulation verification and validation challenges. *Johns Hopkins APL Technical Digest 25*(2), 163–172.

[61] Page, E. H., B. S. Canova, and J. A. Tufarolo (1997). A case study of verification, validation, and accreditation for advanced distributed simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 7*(3), 393–424.

[62] Park, B. and J. Schneeberger (2003). Microscopic simulation model calibration and validation: case study of VISSIM simulation model for a coordinated actuated signal system. *Transportation Research Record: Journal of the Transportation Research Board* (1856), 185–192.

[63] Park, S., C. Shah, J. Kwak, C. Jang, J. Pitarresi, T. Park, and S. Jang (2007). Transient dynamic simulation and full-field test validation for a slim PCB of mobile phone under drop/impact. In *57th Proceedings 2007 Electronic Components and Technology Conference*, pp. 914–923. IEEE.

[64] Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (Second ed.). New York: Springer Science Business Media, Inc.

[65] Raz, J. and B. I. Turetsky (1999). Wavelet anova and fMRI. In M. Unser, A. Aldroubi, and A. Laine (Eds.), *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pp. 561–570. International Society for Optics and Photonics.

[66] Robinson, S. (2004). *Simulation: The Practice of Model Development and Use* (First ed.). Chichester, UK: Wiley.

[67] Rowland, J. R. and W. M. Holmes (1978). Simulation validation with sparse random data. *Computers & Electrical Engineering 5*(1), 37–49.

[68] Russell, D. M. (1997a). Error measures for comparing transient data: part I: development of a comprehensive error measure. In *Proceedings of the 68th Shock and Vibration Symposium*, pp. 175–184. Hunt Valley, MD: Shock and Vibration Exchange.

[69] Russell, D. M. (1997b). Error measures for comparing transient data: part II, error measures case study. In *Proceedings of the 68th Shock and Vibration Symposium*, pp. 3–6. Hunt Valley, MD: Shock and Vibration Exchange.

[70] Sargent, R. G. (1979). Validation of simulation models. In *Proceedings of the 1979 Winter Simulation Conference*, pp. 497–503. San Diego, CA: Institute of Electrical and Electronics Engineers Computer Society.

[71] Sargent, R. G. (2003). Verification and validation of simulation models. In S. Chick, P. Sanchez, D. Ferrin, and D. Morrice (Eds.), *Proceedings of the 2003 Winter Simulation Conference*, pp. 37–48. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

[72] Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation 7*(1), 12–24.

[73] 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.

[74] Schwer, L., H. Mair, and R. Crane (2006). Guide for verification and validation in computational solid mechanics. Technical report, The American Society of Mechanical Engineers, Three Park Avenue, New York, NY 10016.

[75] Schwer, L. E. (2007). Validation metrics for response histories: perspectives and case studies. *Engineering with Computers 23*(4), 295–309.

[76] Shannon, R. E. (1975). *Systems Simulation: The Art and Science* (First ed.). Englewood Cliffs, NJ: Prentice-Hall.

[77] Sprague, M. A. and T. L. Geers (1999). Response of empty and fluid-filled, submerged spherical shells to plane and spherical, step-exponential acoustic waves. *Shock and Vibration 6*(3), 147–157.

[78] Sprague, M. A. and T. L. Geers (2004). A spectral-element method for modelling cavitation in transient fluid-structure interaction. *International Journal for Numerical Methods in Engineering 60*(15), 2467–2499.

[79] Teorey, T. J. (1975). Validation criteria for computer system simulations. In *Proceedings of the 3rd Symposium on Simulation of Computer Systems*, pp. 161–173. IEEE Press.

[80] Tocher, K. D. (1975). *The Art of Simulation* (First ed.). London: English Universities Press.

[81] Torrence, C. and G. P. Compo (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society 79*(1), 61–78.

[82] Underwood, G., D. Crundall, and P. Chapman (2011). Driving simulator validation with hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour 14*(6), 435–446.

[83] Vidakovic, B. (2001). Wavelet-based functional data analysis: theory, applications and ramifications. In T. Kobayashi (Ed.), *Proceedings of the 3rd Pacific Symposium on Flow Visualization and Image Processing*. Maui, HI.

[84] Walker, J. S. (2008). *A primer on wavelets and their scientific applications* (Second ed.). Boca Raton, FL: CRC press.

[85] Whang, B., W. E. Gilbert, and S. Zilliacus (1994). Two visually meaningful correlation measures for comparing calculated and measured response histories. *Shock and Vibration 1*(4), 303–316.

Vita

Capt Andrew Atkinson graduated from General H.H. Arnold High School in Wiesbaden, Germany in 2004. He graduated from the University of Virginia in May 2008 with a Bachelor of Arts in Mathematics and was commissioned through Air Force Reserve Officers' Training Corps in May 2008.

Capt Atkinson served his first assignment as a Test Engineer for the 46th Test Wing, followed by an assignment as an Alternative Navigation Research Analyst at the Air Force Research Laboratory, Eglin Air Force Base, Florida. In 2011, he completed his Masters of Science in Industrial and Systems Engineering from the University of Florida. Shortly thereafter, he became the Chief of Directed Energy Analysis at the 711th Human Performance Wing, Fort Sam Houston, Texas.

In 2014, Capt Atkinson began his doctoral studies in Operations Research at the Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio. He conducted his dissertation research in the area of simulation model validation under the guidance of Dr. Raymond R. Hill. Upon graduation, he will be assigned to Pacific Air Forces Headquarters, Joint Base Pearl Harbor-Hickam, Hawaii as an operations analyst.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
14–09–2017	Doctoral Dissertation	Oct 2014–Sep 2017

4. TITLE AND SUBTITLE

Wavelet-Based Simulation Model Validation of Functional Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Atkinson, Andrew D., Captain, 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-034

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

10. SPONSOR/MONITOR'S ACRONYM(S)

OSD DOT&E

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

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A:
Approved For Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

As computer hardware technology continues to advance, so does the scientific community's capability to develop high resolution computer models able to simulate complex systems and processes. This advancement has led to many challenges associated with verification and validation (V&V). These challenges include adapting methods to high-dimensional functional data, maintaining the necessary objectivity, and accounting for noisy data. Department of Defense (DoD) simulation models require validation techniques that are able to overcome these challenges before the models can be relied upon. Model validation substantiates that the model chosen sufficiently represents the system and that it produces results consistent with real-world data within the range of model applicability.

In this research, new statistical techniques will be proposed that improve upon existing simulation validation techniques. These techniques incorporate the use of wavelets to decompose the time-series data into the time-frequency spectrum allowing for objective and comprehensive assessment of the model. In addition, these techniques offer an improved method of analysis for noisy, high-dimensional data. These techniques are applied to assess the validity of simulation models, which will help ensure the accurate representation of the system they are meant to simulate.

15. SUBJECT TERMS

Wavelets, simulation validation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Raymond R. Hill Jr., AFIT/ENS
U	U	U	UU	204	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636 x7469 raymond.hill@afit.edu