SCIENTIFIC TEST & ANALYSIS TECHNIQUES
CENTER OF EXCELLENCE

# Case Studies on Model Validation Levels

March 2024

Corinne Stafford, Ctr
Nick Jones, Ctr
Kyle Provost, Ctr


To develop and apply independent, tailored Scientific Test & Analysis Techniques solutions to Test and Evaluation that deliver insight to inform better decisions.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, AFIT.ENS.STATCOE@us.af.mil
Call, 937-255-3636 x4736

Technical Reviewers:
Steve Oimoen
Gina Sigler


Scientific Test & Analysis Techniques Center of Excellence
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio



Version: 1, FY23

## Abstract

A Model Validation Level (MVL) is an objective, automatable metric scored from 0-9 that quantifies how much trust can be placed in the results of a model to represent the real world. This paper shows the calulation of an MVL for four different models: a physics model of a toy catapult, a regression model of hypersonic flow behaviors, a regression model of network timeliness, and a lab-scale simulator of an icebreaker ship. These case studies demonstrate how MVLs can be calculated in a variety of cases including when continuous or categorical factors and responses are present, when the model is deterministic, when multiple referents are available for validation, when multiple scopes are of interest, and when model and referent inputs are not matched. In each case study, the MVL and lower level metrics are calculated and possible actions to increase model trust are discussed.

*Keywords: model validation levels, validation, modeling and simulation, test and evaluation*

## Table of Contents

## Introduction

A Model Validation Level (MVL) is an objective, automatable metric scored from 0-9 that quantifies how much trust can be placed in the results of a model to represent the real world. The MVL framework, described in detail by Stafford et al. (2024), provides utility to both decision makers and model developers for using modeling & simulation (M&S) more effectively by providing an objective metric for model trust and measures to guide model improvement actions. Additionally, the MVL R package (Provost et al., 2024, Jones et al., 2024) provides a tool for practitioners to calculate MVLs for their own models. The case studies discussed here show how MVLs can be applied to a variety of different real world validation problems that could be encountered in Department of Defense (DOD) or Department of Homeland Security (DHS) testing.

This paper shows the calulation of an MVL for four different models: a physics-based model of a toy catapult, a regression model of hypersonic flow behaviors, a regression model of network timeliness, and a lab-scale simulator of an icebreaker ship. For each of these case studies, this paper gives background on the model being validated, the scope of intended use for the model, and the model and refererent data used for validation. Then the paper shows the results calculated for the model, including the MVL and lower level metrics, and discusses possible actions to increase model trust.

## Background

The MVL framework defines validation in terms of three key pillars: fidelity, referent authority, and scope. Fidelity is the level of consistency between a model and a referent, defined in the three dimensions of accuracy, repeatability, and resolution. A referent is defined to be a codified body of knowledge representing real system behavior. Referent authority refers to the strength of credibility of a referent's claim to be a high-fidelity representation of reality. Scope is the set of inputs, outputs, assumptions, and limitations representing the mission-relevant system parameters, environmental conditions, constraints, requirements, and their allowable values. The validity of a model is assessed over the scope of intended use. The MVL is mathematically derived from model and referent data within a defined scope of intended use. This process uses a combination of metrics based on fidelity, referent authority, and scope. Stafford et al. (2024) provide detailed descriptions of this calculation.

The MVL is interpreted on an absolute scale, where the MVL indicates the level of trust that can be placed in model results by mapping to quantified levels of trust that are placed in different data sources. This interpretation is shown in Table 1. For example, an MVL of 5 indicates that the model results are as trustworthy as lab-scale system test data. The MVL can be no higher than the highest authority level of data used to validate the model.

**Table 1**
*Interpretation of MVL in Terms of Trust Placed in Different Data Sources*

| MVL of: | Is as Trustworthy as: |
|---|---|
| 1 | Subject Matter Expert (SME) Judgement |
| 2 | First Principles/Physics Predictions |
| 3 | Component Lab Test Data |
| 4 | Integrated Component Lab Test Data |
| 5 | Lab-Scale System Test Data |
| 6 | Hardware-in-the-Loop (HWIL) & Software-in-the-Loop (SWIL) Data |
| 7 | Prototype Field Test Data |
| 8 | Live System Test Data |
| 9 | Operational Real-World Data |

### Case Study 1: Physics-based Model of a Toy Catapult

Physics-based models are commonly used in the DOD to simulate the performance of weapon systems and other DOD systems. This case study walks through the calculation of an MVL for a toy weapon system, a simple wooden catapult which uses rubber band tension to launch a foam ball. The toy catapult is pictured in Figure 1, which also shows various settings to control the launch behavior. This case study walks through the MVL calculation for a deterministic physics-based model, then shows how the MVL can be updated after the model was adapted to be stochastic.

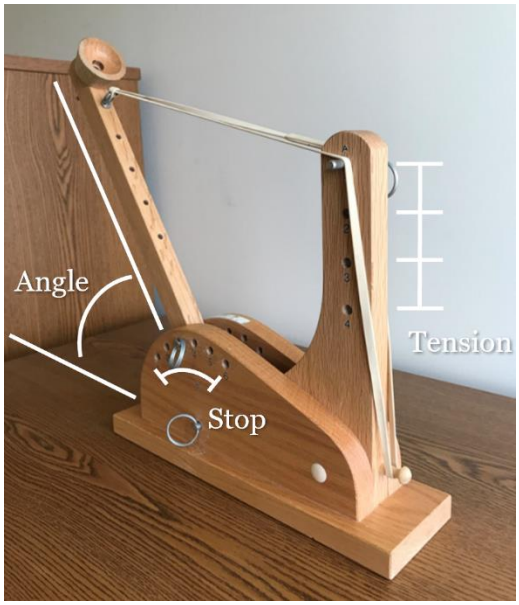

**Figure 1**
*Toy Catapult with Three Launch Factors*

### *Model Description*

The catapult model is a deterministic physics-based model implemented in Python. The Python model uses projectile motion physics equations from first principles combined with material properties of the catapult (e.g., rubber band elasticity) to predict the trajectory of the launched ball from the release point to first impact with the ground, assuming the catapult is on level terrain. The objective of the model is to estimate the catapult launch distance under different launch conditions.

### *Scope of Intended Use*

The scope of intended use for this model comprises one response, catapult launch distance, which varies as a function of three factors, launch angle, stop position, and rubber band tension, as pictured in Figure 1. Launch angle is a continuous factor while stop position and tension are discrete numeric factors since they can only be set to integer levels. However, for the purposes of MVL calculation, they were considered continuous (rather than categorical), since there is an associated physical measurement with each factor setting. Launch distance was measured from the front of the catapult to the point of first impact on level ground.

The factor ranges are summarized in Table 2.

**Table 2**
*Toy Catapult Factors and Ranges for Scope of Intended Use*

| Factor | Factor Range |
|---|---|
| Launch Angle | 160° to 180° |
| Stop Position | 1 to 4 |
| Tension | 1 to 4 |

### *Model and Referent Data*

The model was used to predict launch distance at all combinations of stop and tension settings for 5° increments of launch angle within the scope of intended use (80 unique combinations).

Two different types of referents were available for validation: catapult test data collected by students during design of experiment short courses and launch distance estimates provided by short course instructors. The test data was collected at 53 unique factor combinations, each with 1-15 replicates. To augment the student test data points, short course instructors familiar with the catapult were asked to predict the launch distance for the 9 points only containing one replicate.

Using Table 2 in the MVL Methods and Implementation paper, the student test data was assessed to have an authority level of 8 (operational test data), while the instructor estimates

were assessed be level 1 (SME judgement) (Stafford et al., 2024). Note this is Table is the same as Table 1 in this document except the column titles read Authority Level and Relevant Referent respectively.

### *MVL Calculation and Results*

Walking through the questions for determining MVL applicability (Stafford et al, 2024), it is determined that the model does predict system behavior, it meets the prerequisites of intended use, model data, and referent data, the model and the referent inputs are matched, and referent points are replicated. However, since the model is deterministic and does not include any quantification of uncertainty, only an $MVL_a$ may be calculated. For full details on the flowchart of determining the applicable MVL, please see Figure 3 in the MVL Methods and Implementation paper (Stafford et al, 2024). The $MVL_a$ assesses only the match in mean behavior between the model and referent and does not assess the match in variability. Using the MVL R tool, the $MVL_a$ was calculated, with results shown in Table 3 and Table 4 (Provost et al, 2024; Jones et al., 2024).

**Table 3**

*MVL Summary for Catapult Launch Distance*

| **$MVL_a$** | **6.75** |
|---|---|
| No. of validation points | 48 |
| Average $f_a$ – accuracy fidelity | 0.55 |
| Average authority level | 8.00 |
| Coverage | 0.98 |
| $C_V$ – volume coverage | 0.98 |
| $C_D$ – density coverage | 1.00 |

The $MVL_a$ score in Table 3 indicates that the deterministic catapult model predictions are close to as trustworthy as prototype field test data (level 7 in Table 1) for predicting average launch distance.

To break down this $MVL_a$ score, the high authority level 8 referent data means that the highest attainable $MVL_a$ with the available data is 8. The coverage score is also high and sees only a slight deduction in the $C_V$ score due to missing small corners of the scope. Thus, the loss of trust between the referent and the model is predominantly due to the poor fidelity score. Since the model is deterministic, only the accuracy component of fidelity, $f_a$, was calculated. The $f_a$ score of 0.54 indicates that on average, model predictions are about 1.1 referent standard deviations away from the average launch distances calculated from referent data.

**Table 4**
*Improvement Metrics for Catapult Launch Distance MVL*

**$MVL_a$ = 6.75**

| Metric | Improved MVL | Change |
|---|---|---|
| $f_a$ | 7.96 | 1.21 |
| Authority | 7.75 | 1.00 |
| Coverage | 6.79 | 0.04 |
| $C_V$ | 6.79 | 0.04 |
| $C_D$ | 6.73 | 0.00 |

The improvement table similarly shows that fidelity has the biggest impact on the $MVL_a$. Would fidelity have been perfect, an $MVL_a$ of 7.96 would be obtained. Validating with authority level 9 data would have the next highest impact on the $MVL_a$, while increasing coverage would have minimal impact as the current coverage is nearly perfect.

Since the $MVL_a$ only assesses model accuracy, the model has not been validated to predict variability of behavior. Since deterministic models do not account for variability, the model would need to be adapted to be stochastic for an MVL, not just $MVL_a$, to be calculated. The next section walks through the MVL calculation for the adapted stochastic catapult model.

### *Stochastic Model & MVL Results*

The deterministic catapult model was made to be stochastic by adding a randomly generated, normally distributed error term to the launch distance predicted by the Python model. While this is a simple implementation of a stochastic model, the MVL is able to assess the extent to which it can be considered valid.

Stochastic model predictions were generated for all the same input combinations as the deterministic model; however, this time 6 replicates were collected for each unique input.

Using this new model data, the MVL was calculated using the MVL R tool, with results shown in Table 5 and Table 6 (Provost et al, 2024; Jones et al., 2024).

**Table 5**

*MVL Summary for Catpapult Launch Distance with Stochastic Model*

| **MVL** | **5·74** |
|---|---|
| $MVL_a$ | 6.83 |
| $MVL_v$ | 6.72 |
| No. of validation points | 48 |
| Average fidelity | 0.33 |
| Average $f_a$ | 0.57 |
| Average $f_v$ – variability fidelity | 0.54 |
| Average authority level | 8 |
| Coverage | 0.98 |
| $C_V$ | 0.98 |
| $C_D$ | 1.00 |

First, the MVL for the stochastic model is actually lower than the $MVL_a$ for the deterministic model. This is because the MVL has stricter criteria for model validity: the model is rated on both accuracy and variability instead of just on variabiliy. Since the mean predicted by the model did not change when the model was made stochastic, the MVL is expected to be lower. This MVL of 5.76 indicates that the model was nearly as trustworthy as HWIL & SWIL data (see Table 1).

Comparing Table 5 and Table 3, the authority and scope coverage scores are unchanged. Comparing fidelity scores, the average $f_a$ goes up slightly, likely due to the small amount of error in calculating the model mean from 6 stochastic model samples. Table 5 also includes metrics which could not be calculated from the deterministic model. The average $f_v$ rates the average fidelity between model and referent variability and in this case indicates that on average, the model and referent standard deviation disagree by more than a factor of 2. Since both accuracy and variability fidelities can be calculated, $MVL_a$ and $MVL_v$ can both be calculated. These metrics report the MVL factoring in only the accuracy or variability component of fidelity, respectively. Again, the $MVL_a$ reported here is slightly higher than that reported in Table 3 due to random sampling error introduced by the stochastic model behavior.

**Table 6**

*Improvement Metrics for Catpapult Launch Distance MVL with Stochastic Model*

**MVL = 5.74**

| Metric | Improved MVL | Change |
|---|---|---|
| Fidelity | 7.96 | 2.21 |
| $f_a$ | 6.72 | 0.98 |
| $f_v$ | 6.83 | 1.09 |
| Authority | 6.74 | 1.00 |
| Coverage | 5.79 | 0.04 |
| $C_V$ | 5.78 | 0.04 |
| $C_D$ | 5.75 | 0.00 |

To improve this model further, Table 6 indicates that increasing fidelity would have the greatest impact on the MVL. The $f_a$ and $f_v$ scores indicate that the model needs to be improved both in terms of accuracy and in how it represents variability. The MVL R tool provides additional details on the fidelity scores across validation points to help users identify how fidelity may vary across the scope and where model developers can focus improvement efforts. Alternatively, no model improvement steps may be necessary if the MVL of 5.74 is deemed sufficient for the model's intended use case and acceptable risk level.

## Case Study 2: Regression Model of Hypersonic Flow Behaviors

Hypersonics are a current area of research and development critical for national defense. This case study demonstrates the MVL calculation for a statistical model which was constructed using data from a designed experiment studying hypersonic flow behaviors in wind tunnel testing. It also demonstrates how the MVL can be calculated and interpreted for a binary response, where both continuous and categorical factors are present.

Natoli et al. (2020) describe the design of experiment methodology used to collect wind tunnel data.

### *Model Description*

A statistical model was constructed to predict hypersonic flow behaviors in wind tunnel testing. The experimenters were interested in understanding flow properties of the inlet condition (start/unstart) for a hypersonic vehicle. Understanding the regimes where a hypersonic inlet "unstarts" was key for defining the operability envelope of an air breathing vehicle. A "started" inlet exhibits smooth mass flow, where most of the compressed gas behind the shock is directed into the inlet. An "unstarted" flow spills over the inlet boundaries, decreasing the portion of the flow that makes it into and through the inlet. Subject matter experts assigned a start/unstart condition to each test point based on their expertise and various pressure readings.

The STAT COE constructed a statistical model from this experimental data to be able to predict flow behavior throughout the envelope, as a quicker alternative to simulating flow behavior with computational fluid dynamics (CFD) simulations.

### *Scope of Intended Use*

The scope of intended use for this model comprises one binary response, start/unstart, which varies as a function of four factors, Reynolds number (Re), Angle of Attack (AoA), Angle of Sideslip (AoS), and sweep direction. Three of these factors are continuous (Re, AoA, and AoS) while sweep direction is a binary categorical factor (Up/Down).

The factor ranges are summarized in Table 7.

**Table 7**
*Hypersonic Flow Factors and Ranges for Scope of Intended Use*

| Factor | Factor Range |
|---|---|
| Reynolds number | $2.87 \times 10^6$ to $8.62 \times 10^6$ |
| Angle of Attack | $-5°$ to $12°$ |
| Angle of Sideslip | $-3°$ to $3°$ |
| Sweep Direction | Up, Down |

### *Model and Referent Data*

The wind tunnel data included 807 observations and was split into training and validation sets, where the training data was used to construct the regression model and the validation data served as the validation referent. The validation data set consisted of 72 points at 36 unique input combinations (2 replicates per combination). Validation data was selected randomly from unique combinations which contained at least two replicates. A logistic regression model was constructed to predict the probability of being in the start state, where the model was fit for main effects, second order interactions, and quadratic effects, then reduced to significant effects at the 0.05 significance level.

The logistic regression model was run at 36 unique factor combinations, aligned with the factor combinations present in the validation data set. The model was used to predict the probability of the unstart state occurring at a given input combination.

Using Table 2 in the MVL Methods and Implementation paper (or Table 1 from this paper with different column titles), the wind tunnel data was assessed to have an authority level of 5 (lab-scale system test data) (Stafford et al., 2024).

### *MVL Calculation and Results*

Since the model and referent contained matched inputs and the model predicted the probability of the unstart state occurring (model uncertainty was quantified), the MVL could be calculated. The results are shown in Table 8 and Table 9.

**Table 8**
*MVL Summary for Start/Unstart*

| **MVL** | **3.87** |
|---|---|
| $MVL_a$ | 4.18 |
| $MVL_v$ | 3.99 |
| No. of validation points | 36 |
| Average fidelity | 0.81 |
| Average $f_a$ | 0.94 |
| Average $f_v$ | 0.86 |
| Average authority level | 5 |
| Coverage | 0.70 |
| Average $C_V$ | 0.79 |
| Average $C_D$ | 0.89 |
| $C_1$ – categorical coverage | 1.00 |

The MVL in Table 8 indicates that the model is nearly as trustworthy as integrated component lab test data. In other words, the model is about one level less authoritative than the data which was used to validate it.

To break down where this loss of authority comes from, Table 8 shows that both fidelity and scope coverage are contributors to the drop in authority. The average fidelity score is 0.81, which is a good fidelity score, but results in a small drop in the MVL. The coverage score is 0.7, with a lower score in the volume coverage component. Table 8 reports the average $C_V$ because the presence of categorical factors (sweep direction) splits the scope into multiple continuous domains where the volume coverage is calculated individually. The score indicates that one or both continuous scope domains could not be completely interpolated with the validation points. The $C_1$ score assesses categorical coverage and indicates that both levels of the sweep direction (up and down) have been covered with validation points.

**Table 9**
*Improvement Metrics for Start/Unstart MVL*

**MVL = 5.58**

| Metric | Improved MVL | Change |
|---|---|---|
| Fidelity | 4.30 | 0.42 |
| $f_a$ | 3.99 | 0.11 |
| $f_v$ | 4.18 | 030 |
| Authority | 7.87 | 4.00 |
| Coverage | 4.58 | 0.70 |

Table 9 presents metrics to prioritize model improvement efforts, should they be deemed necessary. The table indicates that the greatest impacts to improving the MVL would be improvements in authority, coverage, and fidelity, in order of decreasing impact. If it is not feasible or desired to collect higher authority data, the model trust could be increased by increasing validation point coverage, either by collecting more wind tunnel data or reworking the model to be trained and validated on different data sets, with higher coverage in the validation set. Additionally, improving the fidelity would increase the MVL, and it could be improved by altering the model form. For instance, relevant terms might be missing, or another method such as neural networks or decision trees could provide better predictive capability. It's also possible that more replicates in the validation set could improve the fidelity, since the current fidelity score is based only on 2 replicates at each point.

## Case Study 3: Regression Model of Network Timeliness

DOD networks perform the critical task of transmitting messages or data to or between DOD systems. Messages must be timely for systems to work strategically together. This case study walks through the MVL calculation for a model of message send/receive time for the notional "NextGen Network", and it demonstrates differences in how the MVL is calculated and interpreted when only categorical factors are present.

### *Model Description*

For this case study, a regression model of message send/receive time was constructed from NextGen Network testbed data, which deployed the network's messaging system in a controlled test environment. The objective of this model is to be able to quickly predict message send/receive time under any conditions, alleviating the need to frequently run the testbed.

### *Scope of Intended Use*

The scope of intended use for this model comprises one response, the time for a message to be collected by a receiver after it has been sent. There are three factors which are believed to affect the time for the message to be received: the network configuration, the receiver, and the message priority. Notably, all these factors are categorical, meaning no interpolation can occur between factor levels.

The factor ranges are summarized in Table 10.

**Table 10**
*NextGen Network Factors and Ranges for Scope of Intended Use*

| Factor | Factor Range |
|---|---|
| Network Configuration | 16, 17, 18, 19, 20 |
| Receiver | 1, 2, 3 |
| Message Priority | High, Low |

### *Model and Referent Data*

The regression model was built from a portion of testbed data. The portion of testbed data not used to build the model serves as a validation referent. Notably, the testbed only contained a representation of receiver 2 since the test team believed the receiver would not have an effect and that receiver 2 would be representative of all receivers. Both the model and the testbed were run for all possible network configurations for high message priority cases, but not all configurations were tested for low priority cases. The testbed validation set contained 7-42 replicates for each tested combination. The regression model predicted the mean message time, and the root mean square error (RMSE) was used to estimate the standard deviation, assuming constant variance across the scope. Message times were measured to the nearest second for testbed data, so the resolution was set equal to 1 for the validation data. The regression model predicted to the nearest hundredth of a second (resolution = 0.01).

In addition to the testbed data, operational NextGen Network data was collected during an initial trial deployment. This trial used only configuration 16, however it indicated that the receiver did in fact influence the message time. Message times were measured to the nearest hundredth of a second for the trial deployment, so the resolution was set equal to 0.01.

The testbed referent was assessed to have an authority level of 4 (integrated component lab test data), while the operational data was assessed be level 9 (operational real-world data) (Stafford et al., 2024).

### *MVL Calculation and Results*

Since the model and referent contained matched inputs with model and referent uncertainty quantified, the MVL could be calculated. Using the MVL R tool, the MVL was calculated, with results shown in Table 11 and Table 12 (Provost et al, 2024; Jones et al., 2024).

**Table 11**
*MVL Summary for Message Time*

| **MVL** | **3.17** |
|---|---|
| $MVL_a$ | 3.54 |
| $MVL_v$ | 3.60 |
| No. of validation points | 7 |
| Average fidelity | 0.75 |
| Average $f_a$ | 0.87 |
| Average $f_v$ | 0.87 |
| Average authority level | 6.87 |
| Coverage | 0.23 |
| $C_1$ | 0.80 |
| $C_2$ | 0.45 |
| $C_3$ | 0.23 |

**Table 12**
*Improvement Metrics for Message Time MVL*

**MVL = 3.17**

| Metric | Improved MVL | Change |
|---|---|---|
| Fidelity | 3.96 | 0.79 |
| $f_a$ | 3.60 | 0.43 |
| $f_v$ | 3.54 | 0.38 |
| Authority | 5.51 | 2.34 |
| Coverage | 6.08 | 2.91 |

The MVL in Table 11 indicates that the model is about as trustworthy as component lab test data (level 3).

To break down this score, first looking at authority, the average authority level of the model is 6.87. Since the referents were a mix of level 9 and level 4, this average authority score indicates that not all validation points could be validated with level 9 data (only available for configuration 16) and level 4 data was the highest available data. This average authority also represents the highest obtainable MVL based on the authority of available data.

Looking at fidelity, the scores indicate good fidelity that results in a small drop in the MVL. The fidelity breakdown shows that accuracy and variability fidelity are about the same. Table 12 indicates that if fidelity had been perfect, the MVL would increase somewhat from 3.17 to 3.96.

Finally, Table 11 shows a very poor coverage score (0.23/1.00), and Table 12 shows that increasing this score would have the biggest effect on the MVL. To understand where this poor coverage score originates, Figure 2 shows the all the possible factor combinations in the scope of intended use and which ones are covered with a validation point (indicated by "X").

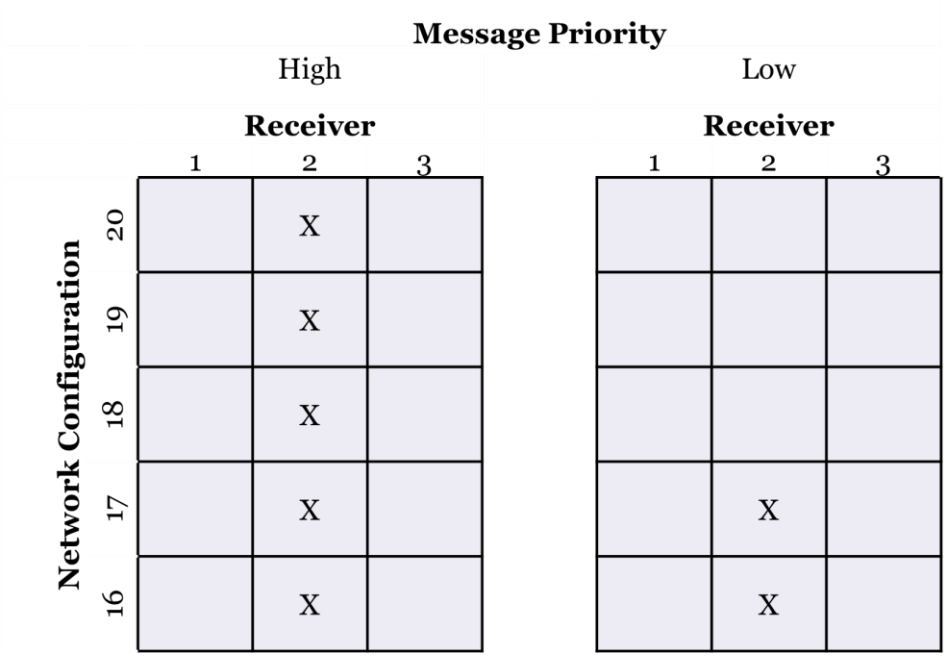


**Figure 2**
*NextGen Network Factor Combination Coverage*

Figure 2 shows that 7 out of 30 combinations have been covered with validation points (contains model and referent data). Note that 7/30 = 0.23 gives the coverage score in Table 11. Going back to the available data, the fact that model predictions were only made for receiver 2 severely limits the scope coverage. Additionally, since testing focused on higher priority messages, only 2 of the low message priority cases have been covered.

The lower-level coverage scores provided in Table 11 can help identify these sources of poor coverage. $C_1$ assesses the fraction of factor levels that are covered, not accounting for any combinations. Table 11 shows $C_1$=0.80, indicating that at least one factor has not been tested at all possible levels. Figure 2 shows that this was due to the lack of validation data for receivers 1 and 3. $C_2$ assess the fraction of two-way combinations that are covered; for example, all combinations of configuration and message priority. Table 11 shows $C_2$=0.45, which accounts for all the two-way interactions missed due to missing factor levels indicated with $C_1$ and new lack of coverage originating at the 2-way level, here being that not all network configurations are covered for all message priorities. Lastly $C_3$ accounts for all missed three-way combinations.

Considering the calculated MVL and additional metrics, if the current MVL was not suitable for

the intended use case, the best course of action to improve the MVL would be by improving coverage. In this case, the lack of matched model data for more than one receiver and lack of referent data for more than one receiver (excluding configuration 16 in the trial deployment) contributed most to lack of coverage, and more model and referent data would need to be collected for receivers 1 and 2 to improve coverage.

## Case Study 4: Lab-scale Icebreaker Simulator using a Referent Interpolator

The United States Coast Guard uses icebreakers to push through sea ice and provide safe passage to other ships, ultimately supporting the country's economic, commercial, maritime, and national security needs.

This case study demonstrates how referent interpolators can enable MVL calculation when data scarcity would not otherwise permit it. Additionally, it demonstrates how the MVL can be calculated for more than one scope of intended use, and how MVLs for different scopes can be interpreted.

The data for this case study was drawn form Su et al. (2010).

### *Model Description*

For this case study, the “model” is a physical simulator which uses a scale-model of the hull of an icebreaker ship to predict the speed at which the ship can travel through ice of various thicknesses. The goal of the physical simulator is to be able to predict icebreaker speed and verify the requirement that the icebreaker can move forward at six knots through one-meter-thick ice. Since ice thickness cannot be controlled in field testing, the simulator was key for assessing the requirement.

### *Scope of Intended Use*

The scope of intended use for this model comprises one response, icebreaker speed, and one factor, ice thickness. Due to the limited availability of field test data for thick ice, two different scopes of intended use are considered where the MVL may be of interest. First, the “test scope,” which is limited to where test data was available for validation, and second, the “operational scope” which includes the entire operational range with ice thicknesses up to 1.3 meters, where the icebreaker simulator speed drops to zero. These scopes are pictured in Figure 3 with the data from both the simulator and the field testing.

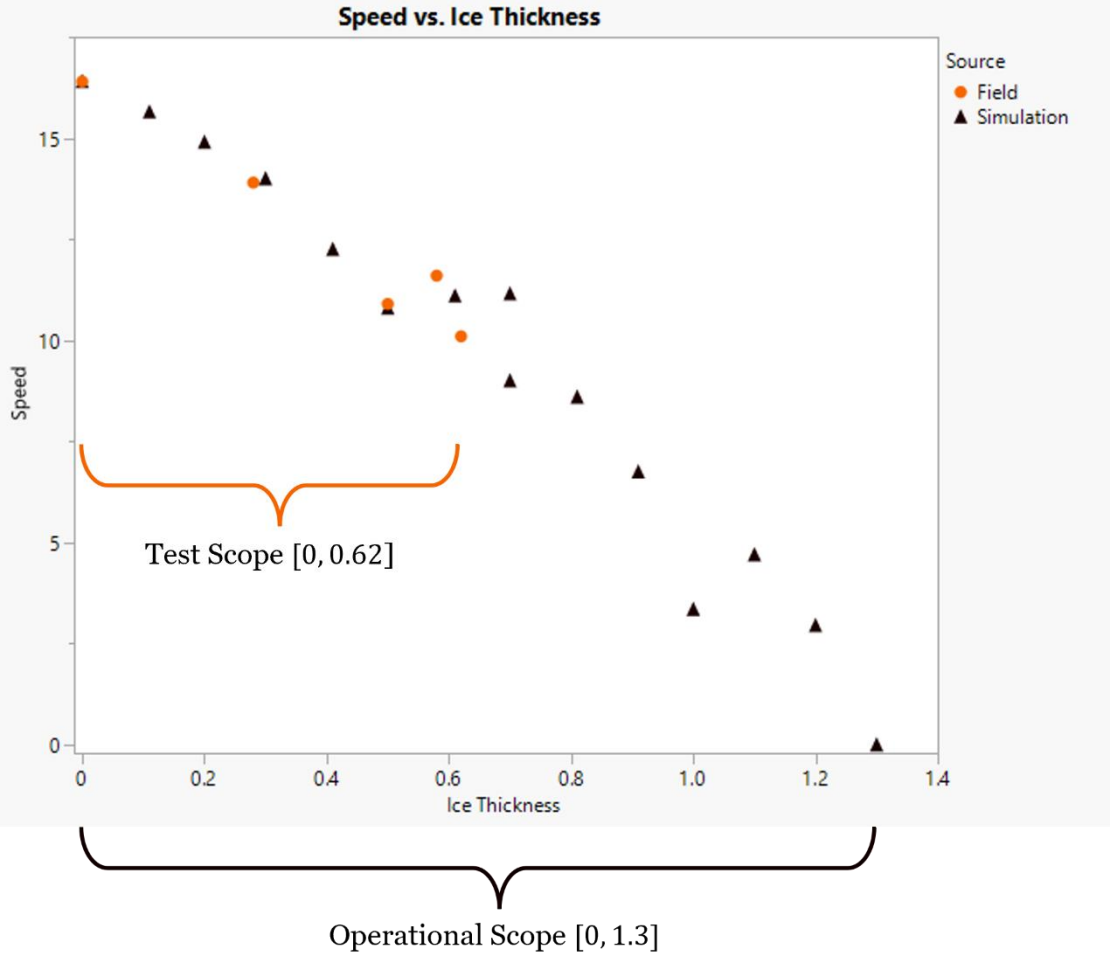


**Figure 3**
*Limited Scope and Whole Scope for Icebreaker Simualtor*

The test scope is based on conditions encountered during testing, and thus may be of interest for assessing icebreaker performance in similar settings. However, since the test did not encounter thick ice, the operational scope can be used to assess the validity of the simulator throughout the entire operational range. This operational range is needed to assess the speed requirement for one-meter-thick ice.

### *Model and Referent Data*

As seen in Figure 3, the simulator data consisted of 15 different observations with ice thickness varying between 0 and 1.3 meters thick. Only one of these observations (Ice Thickness = 0.7) was replicated.

The field test data consisted of five different observations from open water speed to an ice thickness of 0.62 meters. This referent was assessed to have an authority level of 8 (live system test data). The field data also did not contain any replicated observations, and furthermore, the field data points did not align with the simulator points. Since the model and referent inputs were not matched and the referent points were not replicated, a referent interpolator must be used to calculate the MVL, given no more data can be collected. For further information on the details of

how this was determined see Figure 3 in the MVL Methods and Implementation paper (Stafford et al., 2024).

The referent interpolator serves to predict referent behavior where referent data was not collected so that the referent and model can be directly compared. For this dataset, a linear regression model predicting speed from ice thickness was fit to the field test data. This regression had an $R^2$ equal to 0.95, indicating that the interpolator model explains 95% of the variation in icebreaker speed. In addition, the interpolator must include a measure of uncertainty. In this case, the root mean square error (RMSE) was used to approximate the standard deviation $\sigma$ of icebreaker speed, where the standard deviation was assumed to be constant across the ice thicknesses observed. The referent interpolator is pictured in Figure 4.

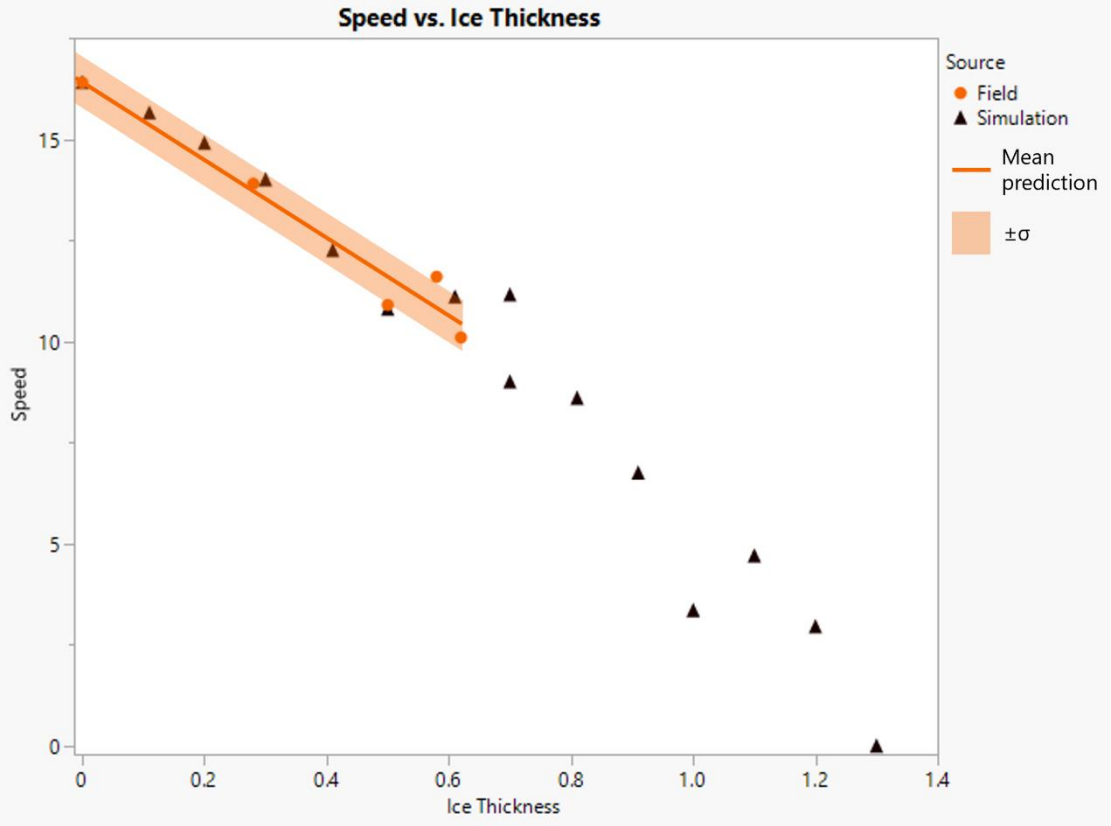


**Figure 4**
*Icebreaker Referent Interpolator*

As seen in Figure 4, the interpolator can only be used to interpolate between the referent data points used to train the interpolator and cannot extrapolate beyond the range of data. The interpolator can be used to compare referent behavior against the 7 simulator points observed within the interpolation region.

### *MVL Calculation and Results*

Since the simulator data did not contain replicated observations (model uncertainty is not quantified), only an $MVL_a$ may be calculated (for more details please see Figure F2 in the MVL Methods and Implementation paper) (Stafford et al, 2024). Recall, the $MVL_a$ assesses only the

match in mean behavior between the model and referent and does not assess the match in variability. Using the MVL R tool, the $MVL_a$ was calculated for both the test scope and operational scope, with results shown in Table 13 and Table 14 (Provost et al, 2024; Jones et al., 2024).

**Table 13**
*MVL Summary for Icebreaker Speed*

| | **Test Scope** | **Operational Scope** |
|---|---|---|
| **$MVL_a$** | **7.49** | **5.58** |
| No. of validation points | 7 | 7 |
| Average $f_a$ | 0.79 | 0.79 |
| Average authority level | 7.99 | 8 |
| Coverage | 0.98 | 0.38 |
| $C_V$ | 0.98 | 0.47 |
| $C_D$ | 1.00 | 0.80 |

The $MVL_a$ scores in Table 13 indicate that the mean behavior of the icebreaker simulator is at least as trustworthy as prototype field test data (level 7) within the test scope but only a bit more trustworthy than lab-scale system test data (level 5) for the operational scope. As the simulator produces Lab-Scale System Test Data (authority level 5), we would hope that any calculated MVL would be at least 5. In the case of both test scope and operational scope, we show that validating using higher authority referent data increases the trustworthiness of the simulator.

The difference in $MVL_a$ scores was expected due to the large difference in coverage between the two scopes: the test scope can be mostly interpolated with validation points, while the operational scope requires a large degree of extrapolation to validate. The low $C_V$ for the operational scope compared to the $C_D$ indicates that the loss of coverage was due more so to degree of extrapolation (scored by $C_V$) rather than lack of validation point density (scored by $C_D$).

The breakdown of lower-level metrics indicates that the fidelity and authority for the test and operational scopes are nearly the same, since the same 7 validation points are used to validate each scope region. There is, however, a slight reduction in the referent authority for the test scope case. This reduction is due to the interpolation required from field test data points to where interpolator predictions were compared to simulator data: the relative interpolation distance is greater when compared to the size of the scope for the smaller test scope, resulting in a deduction. The MVL Methods and Implementation paper discusses the mechanics of this authority reduction in Appendix F (Stafford et al., 2024).

To evaluate the requirement that the icebreaker was able to move forward at 6 knots through meter thick ice, decision makers can now trust the simulator data to be a bit more trustworthy than lab-scale system test data (level 5) for the operational scope. Whether the simulator indicates the requirement was met or not, the decision can now be made knowing the amount of trust that can be placed in the simulator. If the $MVL_a$ was not sufficiently high to assess the

requirement, actions can be taken to either (1) increase the $MVL_a$ and increase trust in the simulator or (2) collect higher authority data (e.g. live test with thick ice) to evaluate the requirement.

**Table 14**
*Improvement Metrics for Icebreaker Speed MVL*

| | **Test Scope** | | **Operational Scope** | |
|---|---|---|---|---|
| | **$MVL_a$ = 7.49** | | **$MVL_a$ = 5.58** | |
| Metric | Improved $MVL_a$ | Change | Improved $MVL_a$ | Change |
| $f_a$ | 7.96 | 0.47 | 6.05 | 0.47 |
| Authority | 8.50 | 1.01 | 6.58 | 1.00 |
| Coverage | 7.52 | 0.03 | 7.53 | 1.95 |
| $C_V$ | 7.52 | 0.03 | 7.09 | 1.51 |
| $C_D$ | 7.49 | 0.00 | 6.02 | 0.44 |

Table 14 gives improvement metrics, which can help prioritize actions for increasing the $MVL_a$ of the simulator. The table shows that improving coverage would have the largest effect on the $MVL_a$ for the operational scope. Coverage could be improved either by reducing scope of intended use or collecting data in uncovered scope regions. Since the scope cannot be reduced much if the simulator is to be used to evaluate the requirement, the better avenue would be to collect more data in uncovered regions. This data could be more level 8 field test data, or it could also be data of a lower authority. While using lower authority data decreases the average authority, the increase in coverage due to more data collection can result in a higher $MVL_a$ overall. Finally, if validation of the simulator variability was desired, more replicates of simulator data points would need to be collected so that the variability in icebreaker speed could be estimated and the MVL could be calculated.

## Conclusion

With the increased reliance on M&S in the DOD and DHS, MVLs can serve to provide an objective, cross-comparable metric for quickly quantifying model trust. This paper showed how MVLs can be calculated for a variety of models including physics-based models, statistical models, and physical lab-scale simulators. In addition, MVLs apply to different response and factor types (continuous and categorical). Depending on the data available, different types of MVLs can be calculated, including the MVL, $MVL_a$, or an MVL/$MVL_a$ from an interpolator, and multiple referent data sources can be pooled together to validate a model. MVLs can also be calculated for different scopes of intended use of interest. Lower-level MVL metrics provide insight into steps for model improvement, and the MVL can be used to quantitatively track model improvement over time. While this paper presents only a small sample of possible model validation cases, the MVL framework is designed to be broadly applicable, and can be widely applied across disciplines.

## References


Jones, N., Provost, K., & Stafford, C. (2024). Model Validation Level (MVL) R Tool User Guide. User Guide, Scientific Test & Analysis Techniques Center of Excellence.

Natoli, C. W., Sigler, G. S., Guldin, S. A., Harman, M. J., & Ahner, D. K. (2020). Design of Experiments in Characterizing Hypersonic Flow on a Wind Tunnel Model. *The ITEA Journal of Test and Evaluation*, *41*(3), 166–175.

Provost, K., Stafford, C., & Jones, N. (2024). MVL R Tool. Tool. Scientific Test & Analysis Techniques Center of Excellence.

Stafford, C., Jones, N., & Provost, K. (2024). Model Validation Levels: Methods and Implementation. Best Practice, Scientific Test & Analysis Techniques Center of Excellence.

Su, B., Riska, K., Moan, T., (2010). Numerical Simulation of Ship Turning in Level Ice. *Proceedings of the ASME 2010 29th International Conference on Ocean, Offshore and Arctic Engineering.* 751-758. https://doi.org/10.1115/OMAE2010-20110