12-2023

Evaluating Model Fidelity to Aid Model Selection

Evan Hybl

Clemson University, ehybl@g.clemson.edu

Evaluating Model Fidelity to Aid Model Selection

Evan Hybl, Clemson University, 105 Sikes Hall, Clemson, SC 29634

Abstract

In simulation, fidelity has become a topic of interest in determining how well a simulation is able to represent its referent situation. In many cases, the *true referent* is the real-world scenario in which the system will exist. However, the fidelity of a simulation may be computed in comparison to other referents including other simulation models or tests. Several metrics have been proposed to evaluate a model based on qualitative or subjective parameters. These proposed metrics offer possible solutions for the quantification of model fidelity, however their inability to compare features relative importance greatly limits their applicability to models and introduces ambiguity in model evaluation. Frist, previously presented metrics are introduced and evaluated. A new metric is then proposed to address concerns presented in the existing metric evaluation. The proposed metric uses model accuracy to a referent case to both determine feature weights and total model fidelity. The proposed metric is then applied to a simulation case and the results are used to make model selection decisions given hypothetical application scenarios. The proposed relative metric aims to compare similar models' level of fidelity with the end goal of aiding in model selection. By combining the proposed metric with model computational cost, decisions on feature fidelity and inclusion can be made to meet the needs of a given simulation's application.

1 Introduction

Simulation is a powerful tool to aid in the design process of ground vehicles. Simulation can support early and late stages of design, however the requirements of a simulation's fidelity, accuracy, and cost change depending on a model's application.

Fidelity describes how well a simulation can represent a referent situation. In the real world every conceivable situation for a given key performance indicator (KPI) cannot be tested; requiring a referent to be generated to represent a test case scenario. A referent is an abstraction from reality, which is meant to encompass the overarching phenomena in a test case. A model is then made to represent the referent situation. One important item of note in this process is that a loss of information occurs at every step. This is demonstrated in figure 1 Where oval areas correspond to the amount of information encompassed. When a referent is made every phenomenon and situation which contributes to the real-world case cannot be considered. When a model is made, often not every phenomenon considered in the referent can be included. This leads to the term of fidelity which is meant to give a measure of how closely a simulation can represent the phenomena present in the referent situation.

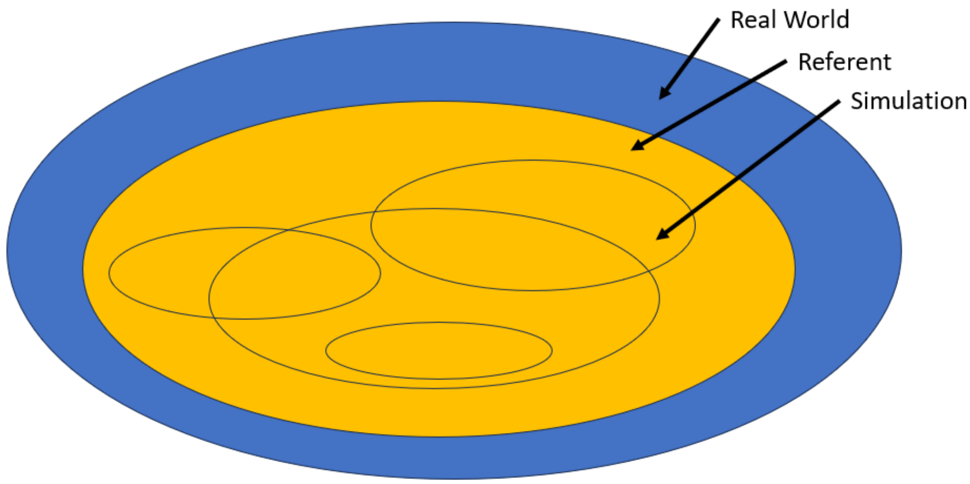


Figure 1. Information Encompassed in KPI Evaluation

Generally, a higher fidelity model is expected to give a higher accuracy result, as compared to a lower fidelity model. This statement is contingent on the fact that phenomena included in a model are implemented correctly and relevant to the referent. Another general expectation is that a higher fidelity model will also have a higher computational cost and require more input data (data hungry). These two drawbacks, with a higher fidelity model, mean that using the highest fidelity model for all stages of design does not make sense from a resource and time perspective.

Accuracy and cost generally do not have a linear relationship. At a certain point adding more, and higher fidelity, features to a model will lead to much higher computational costs with minimal increase to accuracy. The fidelity of a model was increased through feature inclusion and feature refinement (see Figure 2). Each point in the below graph is representative of a different model, with varying degrees of fidelity. The higher accuracy models in this figure are associated with higher fidelity models.

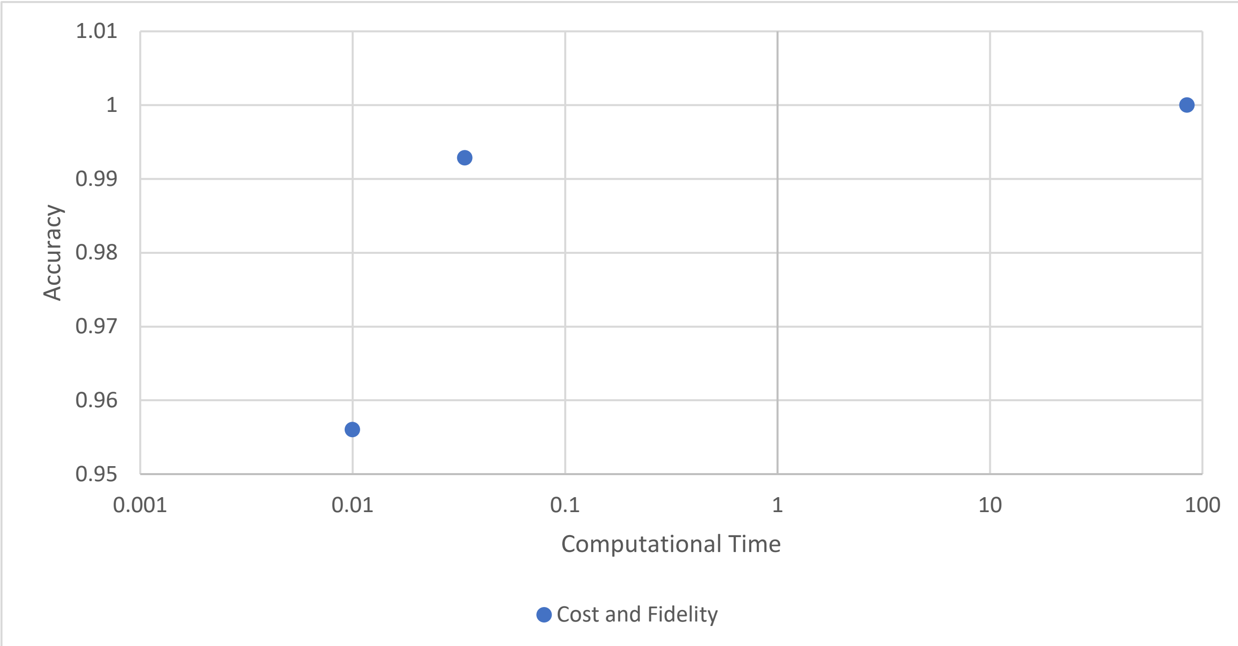


Figure 2. Computational Time Vs. Model Accuracy

Being able to quantify which features should be included in a model and at what level is essential to optimizing a model choice for a given application. Understanding how different phenomena's inclusion in a model changes the expected outcome can be vital to making decisions on the required features and refinement to reach the desired level of accuracy, at a desirable cost. This relationship between model fidelity and cost is generally referred to as utility.

2 Existing Metrics of Simulation Fidelity

Several researchers have made efforts to create a metric for model fidelity. These metrics differ greatly in their classification methods and are worth mentioning to give background on existing solutions.

2.1 SAE

One fidelity metric proposed by SAE works based on a level scale to classify level refinement. The fidelity scale ranged from level 0 to level 7, shown in Figure 3.

level 0 - "**Null**" - no model exists -
No representation yet exists in electronic form.

level 1 - "**Place**" - a placeholder with identifying attributes but no connectivity or behavior -
This could be a database representation of a component. It may include name, description, part numbe cost, etc. Such a representation could be used to generate bills of material, cost reports, size or reliabilit estimates of a system.

level 2 - "**Pins**" - a model with named *interfaces* to the external system, but no internal *features* -
A representation including the component's *interface* elements or pinouts. This could take the form of a schematic symbol of the component or a pinout table. Such a representation could be used to build a *system* design, a schematic diagram, to perform placement and routing, or for complexity analysis.

level 3 - "**Paths**" - a model with identification of internal *states* and connectivity, but no behavior -
This is the first level that allows the most basic forms of simulation, such as sneak circuit analysis. A model at this level contains qualitative information about the component's internal structure, its possible *states* and *flows*, but without quantitative definitions. For example, a relay has ON and OFF *states*, and conduction paths for the contacts, but no electrical resistance or *behavioral* coupling between coil and contacts.

level 4 - "**Static**" - a model with time-invariant, *steady-state* internal behavior -
A model of this level has primary quantitative *properties*. It is typically useful for dc or *steady-state* ac analysis, but is not sufficient for transient analysis. For example, a motor armature might be treated as a resistor, or a wire could be treated as a resistance.

level 5 - "**Dynamic**" - a model with time-varying behavior -
A model of this level is suitable for transient analysis, and may include *non-linear* characteristics. This level captures the principal time-dependent behavior of the component, but may ignore more subtle effects. For example, a motor may have inductance, as well as mechanical effects like inertia and friction, but not cogging or backlash. A wire's resistance could vary with self-heating.

level 6 - "***Precision***" - a model with significant secondary behavior patterns -
A model of this level goes beyond the primary requirements for time-varying behavior. A motor model might have core saturation, cogging, brush arcing, bearing wobble, etc. A wire could react to external thermal loading. A switch could have bounce, arcing, wetting current and aging effects.

level 7 - "**Vector**" - a model with directional or spatial *interfaces* -
A model of this level goes beyond one-dimensional lumped *connection points* to *interface* with neighboring components. It may use multiple *connection points* or a *distribution* function to achieve this. For example, a wire model may have an axial heat *flow* for each end plus a radial heat *flow* from the center. A lamp or antenna may have a *radiation* pattern.

Figure 3. SAE Levels of Model Refinement [2]

The lowest fidelity is level zero (the lowest form of fidelity), which corresponds to a "Null" case where no model exists. level 7 (the highest form of fidelity), referred to as the "Vector" case, is classified by a model with several connection points with different components in a model. This level system attributes a higher fidelity to time variant models as opposed to time invariant models. The SAE level system is a step in the right direction and undertakes a large task of creating a metric which can be applied to any dynamic model. It considers the type of model, the feedback included in model features, along with the complexity of the features within a model.

One major drawback to this level system is that it looks at models through a black and white lens. Oftentimes models contain features which have varying levels of fidelity. Using this system, it becomes challenging to identify a model's fidelity classification if it does not have all features included at the same level of fidelity. The SAE level system states that a model's fidelity is determined by its most central feature's fidelity [2]. This method can become problematic when several features contribute significantly to the performance of a model. In this instance determining which feature is most central would be challenging and may introduce subjectivity. Additionally, if the most central feature has a higher level of refinement than other contributing features, the resulting fidelity score could lose its usefulness. Consider two models with their most central feature at level 5. One model has all other supporting features at level 3, whereas the other would have all other supporting features at level 4. Both models would be given identical fidelity scores using this system. This would prevent modelers from distinguishing the higher fidelity variant from the two and make model selection problematic.

Another pitfall of this metric is that it does not consider how many features are included in a model. A time invariant model that contains 90% of the phenomena from the referent would be scored with a lower fidelity than a time variant model only containing 10% of the phenomena (assuming the central feature was included in both). The metric does not consider supporting features inclusion in its fidelity rating. The metric does, however, give some insight into how feature fidelity should be evaluated. For this method to overcome these hypothetical situations a method needs to be constructed to evaluate feature importance to a model in a non-subjective manner.

2.2 Experimental Frame Metric

Another metric is proposed by Kim and colleagues [1], referred to henceforth as the Kim fidelity metric. This metric uses a model's inputs and outputs, along with a model's level of abstraction to quantify its fidelity. In this fidelity metric a model which requires more inputs and generates more outputs is considered the higher fidelity model. This metric uses inputs as a quantifier of feature inclusion. In general, a higher fidelity model with more feature inclusion will demand more inputs than a lower fidelity model with less features. Additionally, a higher fidelity model will also have the potential to provide more information in the form of an output. If evaluating a model's inputs and outputs cannot discern between two models' level of fidelity the metric, then requires the modeler to evaluate the level of abstraction present. The model with a lower level of abstraction is then considered the higher fidelity model.

One issue with this proposed metric, in the context of vehicle design, is how an output is quantified. Often in vehicle design a model is created to provide a single KPI output to guide a design team in a decision. In a higher fidelity model, more outputs could conceivably be extracted from the model, however determining this number does introduce ambiguity.

Another drawback of this metric is that using number of inputs and outputs to determine fidelity could be misleading. Consider two simulations which aim to take an input force and predict a suspension member's displacement. Both models would have a single input and a single output, however one model would use a linear spring constant, whereas the other would use curve fitting to model a nonlinear spring. Evaluating the model with the Kim fidelity metric would lead to both models having the same fidelity. However, the nonlinear spring model would have a higher fidelity than the linear spring model only using a singular K value. One counterargument to this example could be that the nonlinear model would require more data than the linear model in its generation. Although this is true, removing hard coded data from a model and expressing it as an input would lead to a loss in utility to the modeler as it requires more effort to use. The next option to evaluate the model would be to define each model's level of abstraction. In this example the nonlinear model would have a lower level of abstraction than the linear spring constant. The issue with relying on level of abstraction to define model fidelity is that its qualitative in nature, and can introduce subjectivity.

Lastly, this proposed metric does not have a way to compare similar models with features represented at varying degrees of fidelity without introducing subjectivity. Consider similar models with several different phenomena represented at varying levels of refinement. To use this metric there would need to be some form of weight factor in determining phenomena's relative level of importance in the referent case. Like the SAE metric, it is missing a method to determine individual feature weights.

2.3 Gross and Friedman

Gross and Freidman [3] also have proposed a metric to quantify the fidelity of a given model. They propose an equation which measures a model fidelity based on the sum of its individual features level of importance, multiplied by each individual features level of refinement.

$$F = \sum W_i F_i$$

$$F_i = Fidelity\ of\ each\ refernt\ phenomena$$

$$W_i = Importance\ rate\ of\ phenomena$$

This metric considers the relative importance of a specific phenomenon in a referent case. It also considers the refinement of the represented phenomena. The main issue with the proposed metric is determining the values of F_i and W_i. Gross and Freidman proposed a subjective assignment of these values when evaluating a model. If these values are chosen subjectively by the creator of a model it could lead to inconsistent fidelity scores between similar models. Conceivably the above two metrics could be used to determine the fidelity of each referent phenomena, however a method needs to be determined to non-subjectively find the importance rate of phenomena.

3 Proposed Metric

To quantify a model's fidelity, for the purpose of making a design choice, it is important that variant models can be compared relative to one another. Trying to quantify models' fidelity which are not meant to measure the same KPI is a task which often leads to ontologies and metrics that are far too broad to be useful in model selection. Another challenge is directly measuring model fidelity. Directly measuring the fidelity of a model is near impossible without the inclusion of subjective or qualitative metrics. One idea is to use model accuracy as an indirect measure of model fidelity.

To evaluate a model's fidelity the referent results must be identified to compare model results too. The best referent data will generally take the form of real-life test data of the KPI a given simulation is meant to output. Oftentimes referent data for a specific KPI cannot be obtained or may require experimental testing to attain. Experimental testing can be time intensive and expensive, which is often not feasible for model evaluation. The next best option is to define a referent model. This referent model will take the form of the highest fidelity model with all available referent features included. It would be the highest cost simulation with the highest fidelity features included to model the specific situation. Comparing different models, with differing phenomena and fidelity representations of those phenomena, to this referent model can be useful in determining the relative importance of features in a simulation for specific KPI. This can give a way to remove bias and subjectivity from determining a weight value for a specific feature (w_i).

To evaluate similar models' fidelity relative to a model referent the simplest representation of a referent must also be found. This simplest form of a referent should model the base situation, with no added features. For example, consider a critical angle of tip simulation. This simulation would be designed to determine the angle at which a static vehicle would be expected to roll over. The simplest model representation would take the form of a point mass. Features could include phenomena such as suspension relaxation of a vehicle on a hill, tire deformation due to normal forces, or even fluids shifting in the vehicle body changing its center of gravity. Comparing this simplest model to the model referent can give a baseline for the information left to gain through feature implementation.

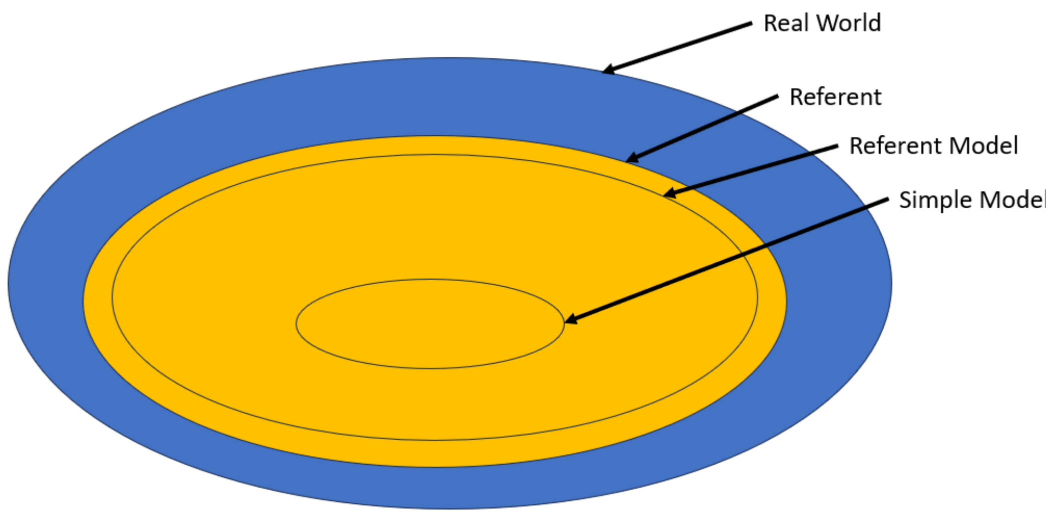


Figure 4. Referent Model Relationship to True Referent and Simple Model

This comparison can be done based on model response to a range of input variables. The range of input variables can be determined by the modeler, based on the range of input values the simulation is being designed to evaluate. Looking at the precent difference between the results of the simplest model and the referent model over a range of input values can then give some indication of the error that is attributed to a lack of features. Averaging this percent difference, over the range of input values for the output of the simplest model compared to the model referent, can then give a relative fidelity based on the accuracy of the models. In this comparison it is assumed that the model referent has a fidelity of 100%, as it is the highest fidelity representation of the situation available to the modeler. Although its actual fidelity is not perfect, this assumption does allow for the comparison of models relative to one another. Subtracting this average percent difference from the model referent fidelity of 100% can then give the simplest models fidelity score (F_{SM}).

Features w_i values must then be determined. The features w_i values are what is left to improve accuracy of the simplest model representation. The simplest model must be outfitted with individual features to then evaluate an individual features effect on improving the accuracy of a model, with respect to the model referent. By evaluating individual model features, their cost and contribution to the accuracy of a model can be evaluated.

Consider, from the example outlined above, a model referent which considers the deformation of vehicle tires in addition to the relaxation of suspension members. To evaluate these features the modeler should configure the simplest model with an individual feature. Following this, the modeler can then compare the results of the simple model with the individual feature included to the simplest model's response. Then, by calculating the percentage difference between the simple model and the simple model with an individual feature, a feature weight can be computed. This would give an estimate of the improvement in accuracy of the model in getting closer to the referent.

It should be noted that when evaluating the importance rate of a particular phenomenon the highest fidelity feature representation should be used. This can give a baseline for the importance of the phenomenon's effect on the simulation results. Evaluating lower fidelity representations of the phenomenon can be done in the same way; however, this calculated value will not simply be the importance rate of a phenomena (w_i). It will be the product of the importance rate of a phenomenon and the fidelity score for the representation. The highest fidelity feature is assumed to have a fidelity score (f_i) of 1.

The resulting fidelity score can then be computed using the following modified equation:

Equation 1.

$$F = \sum f_i w_i + F_{SM}$$

$$F = Total\ model\ fidelity$$

$$F_{SM} = Simple, zero\ feature, model\ fidelity$$

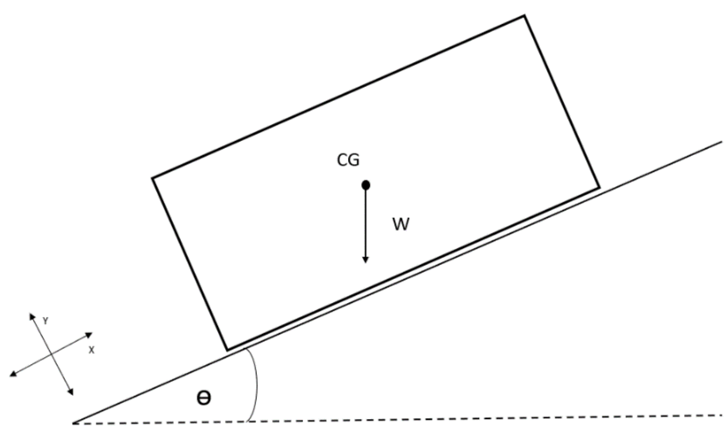
$$f_i = Fidelity\ of\ feature$$

$$w_i = Importance\ rate\ of\ feature$$

4 Example Case

The critical angle of tip example mentioned above was created using an algebraic MATLAB code to model an M1152 HMMVW. A simple model was created which modeled a point mass with no suspension or tire deformation phenomena included (see Figure 5).

Figure 5. Simplest Model Configuration



Additional variant models were created which included either tire deformation or suspension relaxation features (see Figure 6).

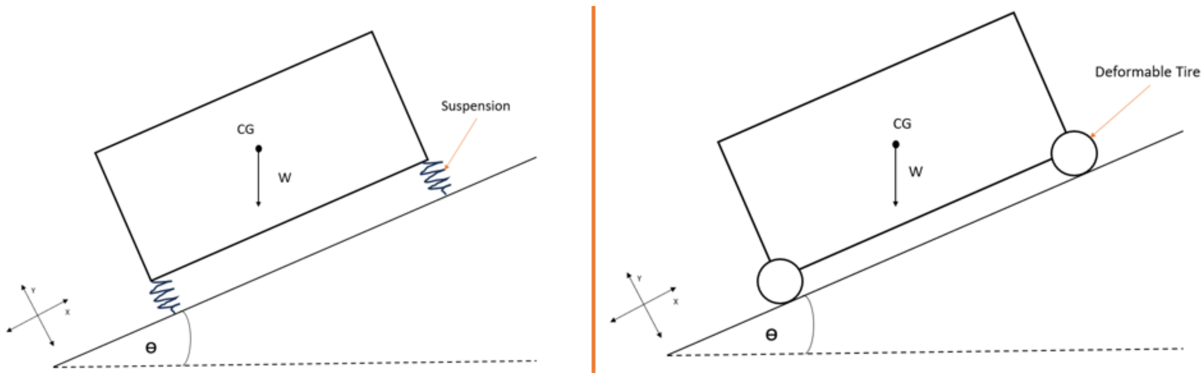


Figure 6. Variant Model Configuration

Two feature variants were created, a low fidelity feature (using a singular k value to represent tire deformation and suspension relaxation), and a high-fidelity feature (using a lookup table and a spline function to represent the nonlinear behavior of tire deformation and suspension relaxation). Finally, a referent model was created which included both tire deformation and suspension relaxation (using the highest fidelity feature representations, see Figure 7).

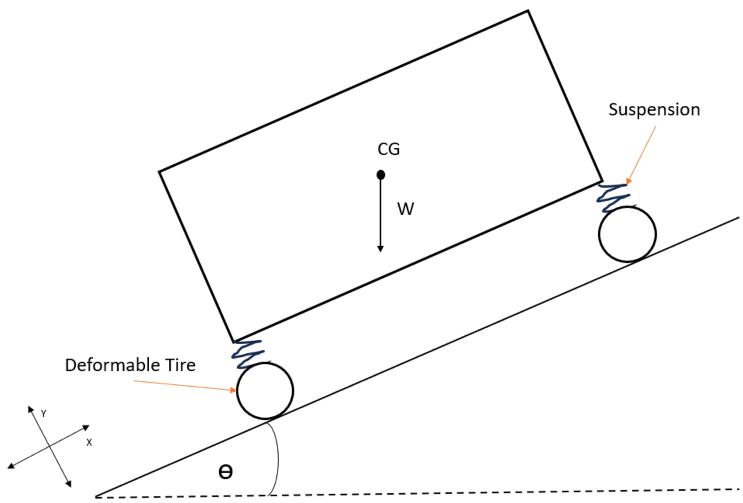


Figure 7. Referent Model Configuration

The different models were then evaluated over a range of CG locations and weights. The weights and CG locations were varied by 20% from their original specified values to simulate different cargo loading scenarios, and their effects on the CG location and total vehicle weight.

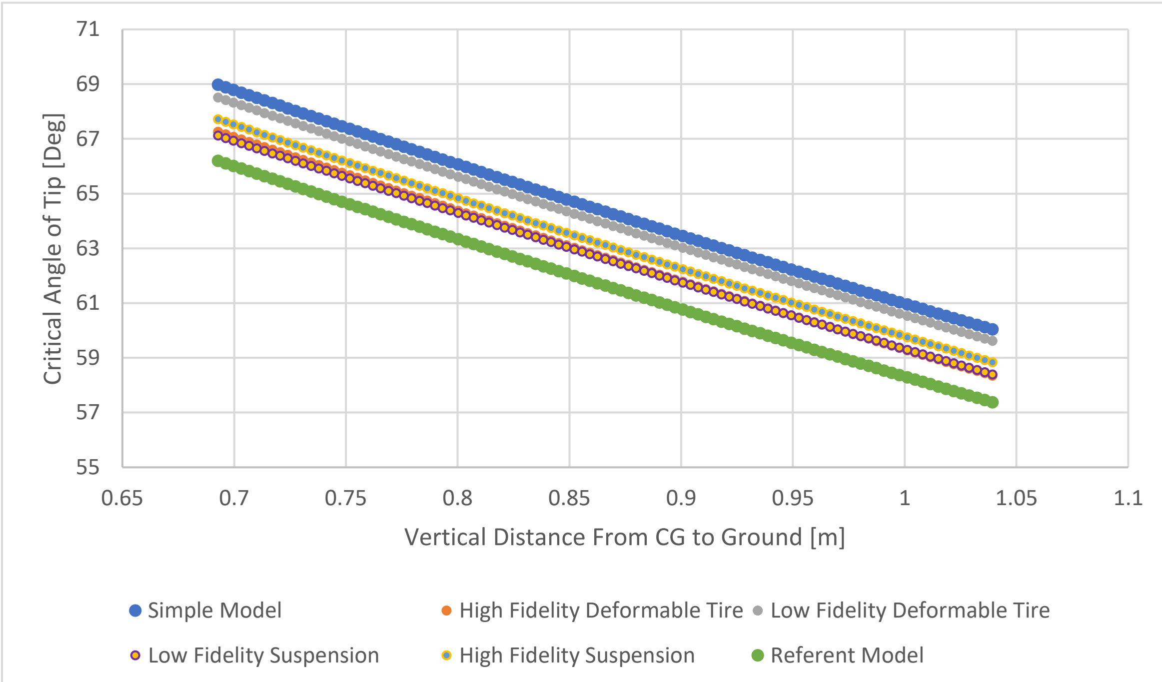


Figure 8. Critical Angle of Tip Response to Varied CG Horizontal Location

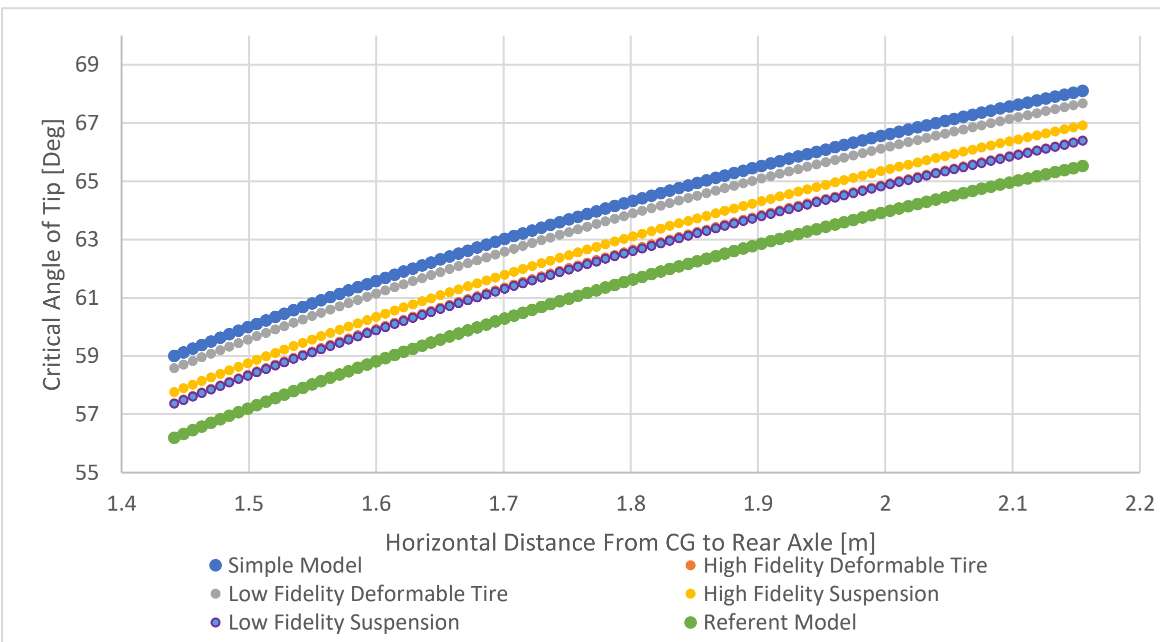


Figure 9. Critical Angle of Tip Response to Differing CG Vertical Location

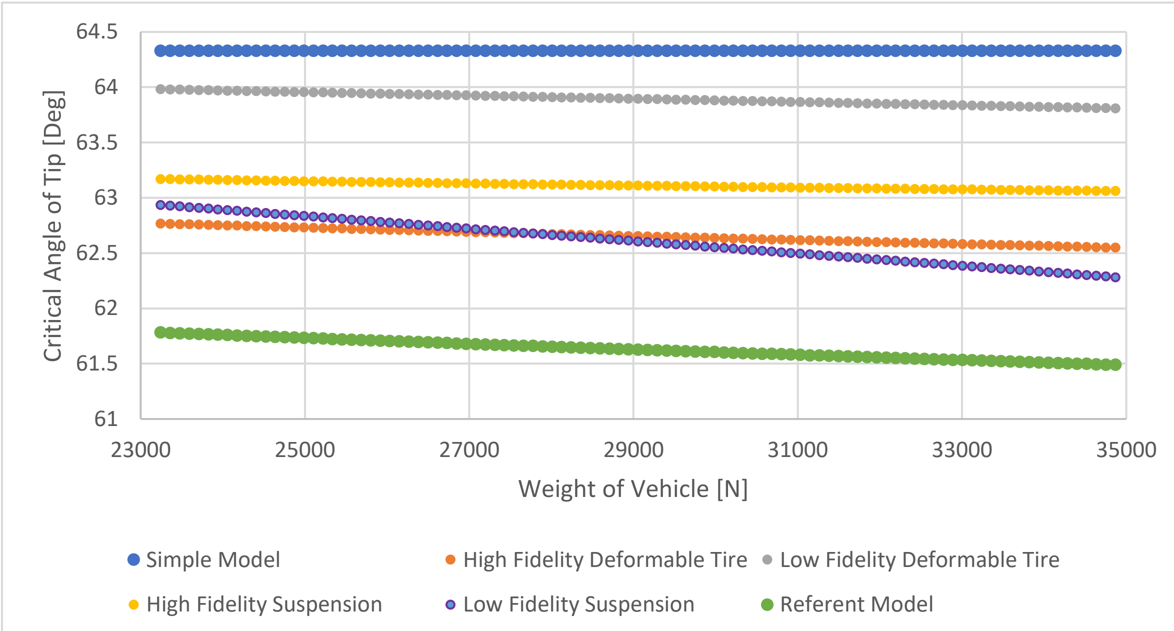


Figure 10. Critical Angle of Tip Response to Differing Vehicle Weight

The results of the simple model, referent model, and model variants were graphed to show the response of each model to the varied input parameter (see Figures 8-10). It is important to isolate variables in evaluation so that the individual response to the varied parameter can be seen. This importance is illustrated in figure 10 where a change to the vehicle weight has no impact on the simple model's output, whereas it does have an impact on the variant cases with tire and suspension features. Simple model fidelity, feature importance rates, and feature fidelity metrics were then calculated using the methods outlined above. The results are tabulated below.

	Varied Horizontal Location of CG	Varied Vertical Location of CG	Varied Weight of Vehicle
F_{SM}	0.955	0.956	0.956
w_i Suspension Relaxation	0.019	0.018	0.018
w_i Deformable Tire	0.026	0.026	0.026
f_i Low Fidelity Suspension	1.402	1.413	1.413
f_i Low Fidelity Deformable Tire	0.258	0.259	0.259

Table 1. Calculated Fidelity Parameters

From the above metrics calculated a few conclusions can be made. Firstly, a deformable tire feature is a more important phenomenon to include in a critical angle of tip simulation for an M1152 HMMWV. Secondly it can be concluded that the lower fidelity representations of the phenomenon lose a significant amount of information from the higher fidelity variants. It should be noted that the f_i value for the suspension does not mean that the lower fidelity feature has a fidelity 1.4 times that of the higher fidelity feature. It should be interpreted as having a representation that retains 40% of the information of the higher fidelity feature representation. It can also be interpreted as an overestimation of the effect of a particular phenomenon.

Another important item of note from the above results is that feature inclusion only accounts for approximately 4.5% of this model's fidelity. This gives a modeler a much more concrete understanding of the loss of information that can be expected from neglecting to include specific phenomenon in a model. This information, when combined with the computational time associated with a particular phenomenon and its complexity in representation, can be helpful in determining the requirements of a model for a particular application. The computational time of the models and corresponding fidelity is presented in the table 2 below (note: fidelity scores were averaged across all input variables).

	Computational Time [s]	Fidelity
Simple Model	0.009975	0.956037
Simple Model + Low Fidelity Suspension	0.028187	0.984011
Simple Model + High Fidelity Suspension	59.201132	0.963097
Simple Model + Low Fidelity Tire Deformation	0.022711	0.975865
Simple Model + High Fidelity Tire Deformation	25.89011	0.983298
Referent Model	84.519834	1

Table 2. Fidelity Score and Computational Time

It can more clearly be seen in Table 2 that as features are included, computational time increases. It can also be seen that as feature fidelity increases, so does computational time.

From Table 1 theoretical fidelity scores can then be calculated using Equation 1 for different configurations of models. Below the predicted fidelity scores are shown for a model using all low fidelity features and for a model using only high-fidelity features. These calculations were done based on the input parameter of interest. Additionally, these models were constructed in MATLAB and evaluated over the same range of input parameters to show how the theoretically calculated value compares to the actual.

	Low Fidelity Features	High Fidelity Features
Predicted	0.989298204	1.001026
Observed	0.992725448	1

Table 3. Calculated Fidelity Scores for Changing Horizontal CG Location

	Low Fidelity Features	High Fidelity Features
Predicted	0.989774	1.00123
Observed	0.993002	1

Table 4. Calculated Fidelity Scores for Changing Vertical CG Location

	Low Fidelity Features	High Fidelity Features
Predicted	0.989705	1.001163
Observed	0.992916	1

Table 5. Calculated Fidelity Scores for Changing Vehicle Weight

When combining the information from Tables 2 through 4 decisions can be made on which features to include and at what level for a particular application. If a modeler wanted to conduct a design sweep of variables, running hundreds or thousands of different test cases, it may be advantageous to use a model with both features represented at a lower fidelity. From Table 2 one could expect approximately 99% accuracy with respect to the referent model. Additionally, they could expect a much lower computational cost. If a modeler was using a model for a single case, and had a high need for accuracy, they may select the higher fidelity feature model. This would trade a high computational cost for a higher level of accuracy and subsequent fidelity.

5 Metric Limitations

Although the proposed metric does offer, in a non-subjective way, a method to compare similar models' fidelity, it is very time intensive to implement. A modeler must configure models with individual features and evaluate them over the ranges of input variables that they intend to use the model for. This may not be feasible for more complex models with long computational times, wide use cases, and uneasily reconfigured features. Additionally, a model evaluated using a limited range of use cases may have fidelity scores with limited applicability. If the model is needed to be used for a wider range of input values, later in its life, the fidelity scores will need to be recalculated over its intended input value range to ensure accuracy. Additionally, dissimilar models measuring different KPI's cannot be compared using this metric. Models designed to measure differing KPI's must go through an evaluation process with isolated KPI's, which can also add to the fidelity evaluation time.

Another drawback is the assumptions that this metric uses. One large assumption is that all features are implemented correctly. If a modeler makes a mistake in feature implementation it will carry over to the referent model leading to fidelity scores which may be misleading. It is also assumed that the modeler can identify the highest fidelity feature out of several variant features; in some instances, this may not be the case. In these instances, referring to the experimental frame analysis fidelity metric or the SAE level system may be helpful in identifying the highest fidelity feature variant.

This metric also does not produce 100% accurate predictions in generating fidelity scores for hypothetical models. It was shown that when predicting theoretical fidelity scores for the high fidelity and low fidelity feature cases, presented in tables 3-4, an error of approximately .32% was observed in predicted vs actual results. Although the level of error was low for the example presented, it may be higher for cases where phenomenon play a larger role in a model's accuracy of results.

The largest issue with the proposed metric is that it only relies on model accuracy, to a model referent, to determine a fidelity score. Fidelity is a concept which describes how well a model can represent its referent case. Soley relying on accuracy to determine this has the potential to misrepresent this value. Although accuracy and fidelity are generally related, this is not always the case. Determining a model's actual fidelity will likely require a more qualitative metric to be used in conjunction with a quantitative one. The methods outlined in this paper in determining feature weight could be used in tandem with a qualitative metric (like the SAE levels of refinement) to get a more accurate fidelity score.

6 Future Work

Looking to the future it would be advantageous to see how well this metric applies to different simulations. The example outlined in this document is a somewhat simple quasi-static model, meant to simulate the theoretical critical angle at which a vehicle would experience rollover. Applying this metric to more complex simulations, both static and dynamic, and evaluating the feasibility of this metric's implementation would be helpful in determining its true scope of application. Additionally, the results of a study on a broader implementation would be advantageous in assessing its true ability to aid in model selection.

Another potential area of future work would be implementing a qualitative metric, for fidelity, in conjunction with the methods outlined above to determine feature weights. This could be done by using a level system, like the SAE level of refinement model, in conjunction with feature weights to get a more accurate measurement fidelity. This could improve the proposed metrics dependents on accuracy of a model to quantify fidelity.

7 Conclusion

The proposed metric allows for modelers to evaluate the fidelity of a model based on individual feature evaluation. This gives insight into the importance of individual feature inclusion in a model without the use of subjective metrics. By comparing individual model features to a referent model, a better understanding can be developed of the degree individual phenomena contribute to a specific KPI. Additionally, the proposed metric allows the comparison of differing fidelity features to a referent feature to better quantify the loss of information associated with them. When combined with the computational time (cost) associated with the addition of individual features, decisions can be made in determining the ideal model configuration for a given application.

8 References

[1] Kim, H., McGinnis, L. F., & Zhou, C. (2010). On fidelity and model selection for discrete event simulation. *SIMULATION*, *88*(1), 97–109. https://doi.org/10.1177/0037549710371217

[2] SAE International Surface Vehicle Recommended Practice, "Model Specification Process Standard," SAE Standard J2546, Rev. Feb. 2002.

[3] Schow, G., Freeman, R., Lashley, T., Meier, H., & Swauger, J. (1997). DERIVATION AND PUBLICATION OF SIMULATION FIDELITY METRICS . *TASC*.