Validation metric based on relative error

Cor-Jacques Kat and Pieter Schalk Els

Department of Mechanical and Aeronautical Engineering, University of Pretoria, co Lynnwood Road and Roper Street, Pretoria 0002, South Africa

Abstract – Engineers and scientists are often faced with the problem of objectively comparing time histories of measured and/or simulated data. This paper presents a reliable and intuitive validation metric for use in the validation process. The proposed validation metric is able to quantify the agreement/disagreement between deterministic system response quantities of interest obtained from measurements on a physical system and predictions from a mathematical model. The validation metric is based on the relative error and the challenges concerning the use of the relative error on periodic signals are addressed. The validation metric is compared to similar metrics and their advantages and limitations are discussed. The results show that the proposed validation metric gives a comprehensive error that is able to quantify the agreement between two periodic signals and is easily interpretable.

Keywords: validation, validation metric, relative error, quantitative measure, quantitative validation, error metric, error measure

MSC2010: 00A06, 00B99

1. Introduction

Mathematical and computer modelling have been playing an increasingly important role in the computer aided engineering (CAE) process of many products in the last 60 years. Simulation offers great advantages in the development and analysis phase of products and offers a faster, better and more cost effective way than using physical prototypes alone. Engineers develop mathematical models of varying complexity to emulate various physical systems. The engineer needs to evaluate the mathematical model and decide whether the model does indeed represent the physical system to an acceptable level of accuracy. Therefore, in order to obtain meaningful simulation models it is necessary to verify and validate them. The need for a formal validation method for quantifying the accuracy of simulation models emulating physical systems has become increasingly important with the greater reliance on the CAE process during product development. The drive for a formal validation method is fuelled by the need for obtaining simulation models which satisfy accuracy requirements, and can be used with confidence to base key engineering and business decisions on.

The verification and validation (V&V) process is an important part of any model that is created to emulate physical events and engineering systems. Reference [1] states that “the terms verification and validation have a wide variety of meanings in the various technical disciplines”. Similarly [2] states that “the broad interest in V&V in many different scientific areas has led to a diverse and often incompatible list of definitions and concepts as it pertains to different disciplines. Moreover, despite the fact that modern views of the subject have been under development for nearly a decade, much remains to be done towards developing concrete approaches for implementing V&V procedures for particular applications”. Reference [1] refers to work that played a major role in attempting to standardize the terminology within the engineering community. Similarly, a committee was formed known as the ASME Committee for Verification and Validation in Computational Solid Mechanics whose purpose is to develop standards for assessing the correctness and credibility of modelling and simulation in computational solid mechanics. This committee released a guide for the verification and validation in computational solid mechanics [3]. They give the following definitions for the verification and validation process:

Verification - The process of determining that a computational (or simulation) model accurately represents the underlying mathematical model and its solution.

Validation - The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.

Figure 1 gives an overview of the V&V process and the tasks associated with the process. The two primary elements of a V&V process are 1) the physical system of interest and 2) the mathematical (or simulation) model that is created to emulate the physical system. An experimental setup of the physical system is constructed in task *(a)* from which the experimental data is obtained. In task *(b)* measurements or analysis (e.g. Computer Aided Design (CAD)) are made on the physical system in order to obtain the properties and parameters of the physical system, such as mass, mass moments of inertia, etc., which are required as inputs to the mathematical model. The path from the conceptual model to the simulation model is shown as well as the stages where the code and calculation verification is performed. With the simulation model verified it can be used to generate the simulation data. From the experimental data and the simulation data the system response quantity (SRQ) of interest can be obtained. The measured system response quantity (SRQ^m) is obtained from the measurements on the physical system and the predicted system response quantity (SRQ^p) is obtained from the predictions of the model. The measured and predicted SRQs are the required inputs into the validation process.

Various uncertainties exist that will affect both the measured and predicted SRQs. Reference [4] categorizes the sources of uncertainty in the simulation model broadly into uncertainty occurring in the model inputs, in the numerical approximations or in the model form. Similarly, uncertainty may exist in the measurements taken during the experiment due to measurement errors. These measurement errors may arise from various elements such as for example the individual measuring instruments. The characterization of the numerical approximation errors associated with a simulation is called verification. Verification, as described in reference [3], is divided into code verification and calculation verification and is indicated in Figure 1. Reference [4] states that the characterization of the model form uncertainty is estimated during the validation process. The uncertainty quantification in the experimental measurements and in the simulation model is outside the scope of this paper. Therefore, both the measured and predicted SRQs considered in this study are deterministic. The reader is referred to [1], [4] and [5] for more detail on uncertainties.

With the SRQs from the experimental and simulation model obtained, the validation process can commence. The validation process can be divided into two steps [1]. The first step is the quantitative comparison of the measured and predicted SRQs. The measured and predicted data can however also be compared qualitatively by superimposing them on graphs but the subjective conclusions on the correlation of bad, good or excellent makes quantifying the accuracy very difficult. Qualitative validation may be useful in certain scenarios, especially in identifying possible causes of errors in the model, but its inability to give a quantitative measure of the agreement/disagreement between the experimental and simulated data makes it difficult to use in determining whether the accuracy requirements are satisfied (2^{nd} step of the validation process). Quantitative comparisons attempt to circumvent the limitations of qualitative comparisons. Quantitative comparisons consist of comparing defined error measures or error metrics (validation metrics). Reference [6] makes the following distinction between an error measure and an error metric: "An error measure provides a quantitative value associated with differences in a particular feature of time series. An error metric provides an overall quantitative value of the discrepancy between time series; it can be a single error measure or a combination of error measures". The error measures to be used are chosen by the engineer and will vary depending on the data. Examples of error measures are steady state gains, response times, peak response times, percent overshoot for time domain data and peak frequency, peak amplitude ratio and phase angle for frequency domain data [7]. Reference [7] states however that certain data will not lend itself to the identification of such error measures. Instead of defining error measures of certain features of the data, the measured and predicted data can be compared by using error metrics (or validation metrics) which do not require the extraction of specific features in the data. The validation metric (or measure of comparison) attempts to give an overall measure of the comparison between the data being compared. Validation metrics will be discussed in section 2. It is the authors' opinion that both quantitative and qualitative comparisons of measured and predicted responses are useful to employ. During model refinement and fault-finding, qualitative comparisons can supply the modeller with valuable information and may give much more insight into the possible causes for the deviation than a validation metric. However, in determining whether the model is valid or not, the qualitative comparisons should be substituted with a quantitative comparison method.

The second step of the validation process shown in Figure 1 is concerned with determining whether the results obtained from the quantitative validation metric satisfies the accuracy requirements. When the result of the validation metric satisfies the accuracy requirements the model can be considered to be valid. Alternatively, it may be that the validation metric gives results that do not satisfy the accuracy requirements. Depending on the reason for the accuracy requirements not being met one of the two dash-line paths can be taken. Either better/more experimental data may be required or the model needs to be refined.

Although validation is essential in assuring that the model is valid, validation does have some shortfalls and the engineer should be aware of them and should try to avoid them. Various studies ([8-10]) validated models against certain parameters and then used them to predict others. For example, a vehicle model is developed for durability analysis but is only validated against accelerations. This approach may have certain risks involved such as stated in reference [11]. They use the example of a vehicle doing a severe J-turn with the assumption that the measured yaw rate and lateral acceleration are available from vehicle tests, but measured normal loads on the tyres are not. They compare the simulated yaw rate and lateral acceleration of two models of the same vehicle with the difference being that the centre of gravity height of one of the models is 10% higher. Comparing the yaw rate and lateral acceleration the models seem to give similar results, but comparing the lateral load transfer it becomes clear that there is some discrepancy between the two models. The importance of validating the model for the correct parameters is also shown in reference [12].

The focus in this study will be on the validation process and more specifically on the first step of the validation process concerned with the validation metric. The reader is referred to reference [2] and [13] for further details on the complete V&V process. The rest of this paper will be concerned with the development and evaluation of a quantitative validation metric based on relative error for use in the first step of the validation process. Of primary interest will be quantifying the agreement/disagreement between SRQs that are periodic in nature with a combination of many frequencies that may or may not oscillate around zero. An example of a SRQ that exhibits behaviour as described above is an acceleration measurement on a vehicle driving over a discrete bump or the accelerometer measurements on a vibrating beam. In this study deterministic SRQs with time as the independent variable will be compared.

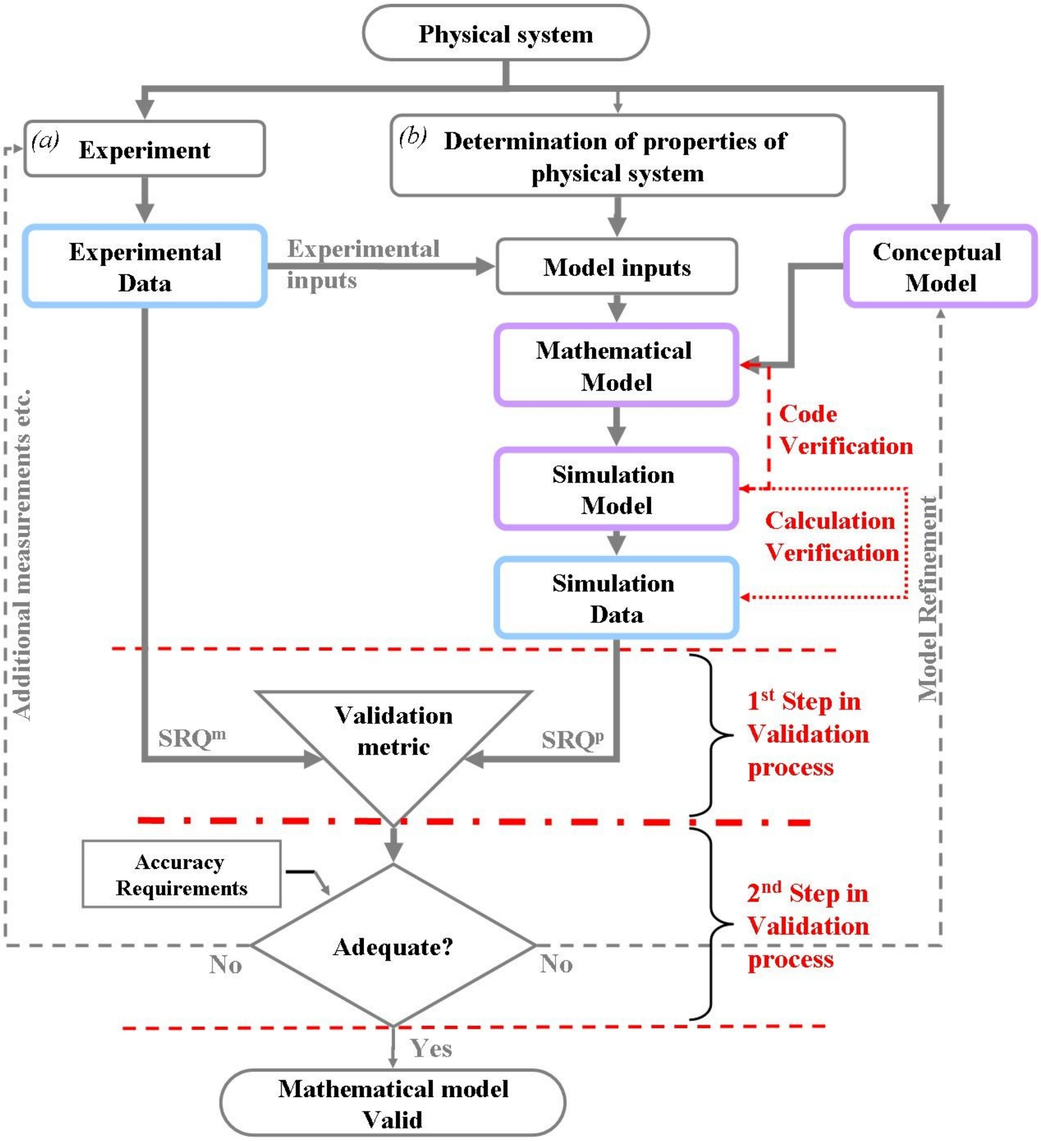


Figure 1. Overview of Verification and Validation processes

2. Quantitative validation metrics

A quantitative validation metric should be able to provide a measure that quantifies the overall error (or agreement/disagreement) between two sets of data, for example between measured and predicted data. In the context of the validation process we would like the validation metric to quantify the level of agreement/disagreement of the model with respect to the physical system in order to conclude whether the model satisfies the accuracy requirements and can be considered valid for the intended use. The validation metric's result should be an easily interpretable value that can be used to determine whether the agreement between the physical system and model satisfies the accuracy requirements.

Although many different error measures and validation metrics can be found in the literature for quantitatively comparing SRQs with time as the independent variable, not many studies concerning validation of simulation models make use of them. Rather the validation is done qualitatively with subjective conclusions such as the correlation is good, excellent or fair. This begs the question: Why are these measures not used? Do they give engineers physically meaningful and interpretable results? In an attempt to answer these questions, a literature survey was conducted in order to form an idea of the measures and metrics available, their capabilities, limitations and whether they give physically meaningful and easily interpretable results in order to determine whether or not the model satisfies the accuracy requirements.

2.1. Literature survey

Reference [1] divides traditional quantitative comparison approaches into three categories:

i) Techniques developed by structural dynamists for assessing agreement between computational and experimental results as well as techniques for improving agreement. These techniques are known as parameter estimation, model parameter updating or system identification.

ii) Hypothesis testing or significance testing.

iii) Bayesian analysis or Bayesian statistical inference.

Reference [1] mentions the following on the approaches used in the three categories:

i) "Although these techniques are used to compare computational and experimental results, their primary goal is to improve agreement based on newly obtained experimental data".

ii) "A validation metric is not specifically computed as a stand-alone measure that indicates the level of agreement or disagreement between computational and experimental results. The results of a hypothesis test is focused, instead, on obtaining a yes-no statement of computational-experimental consistency for a pre-specified level of significance"

iii) "Much of the theoretical development in Bayesian estimation has been directed towards optimum methods for updating statistical models of uncertain parameters in the computational model. In validation metrics, however, the emphasis is on methods for assessing the fidelity of the physics of the *existing* computational model".

They state that the primary goal of both parameter estimation and Bayesian inference is model updating and model calibration. This may be the goal in many situations but is different from the aim of the validation metric in the validation process. The purpose of a validation metric is to be able to assess the predictive capability of the mathematical model and not to optimize the agreement between the mathematical model and the experimental measurements. The functionality of the parameter estimation and Bayesian inference to optimize the agreement between the mathematical model and the physical system can be useful in the model refinement stage shown in Figure 1.

Reference [1] presents an approach that evaluates the accuracy of the model based on comparing deterministic computational results with the estimated mean of the experimental measurements.

The primary differences between their approach from the three traditional quantitative comparison approaches they mention are that: (a) “a stand-alone validation metric is constructed to provide a compact, statistical measure of quantitative disagreement between computational and experimental results”, and (b) “a statistical confidence interval is computed that reflects the confidence in the accuracy of the experimental data“. They state however that their validation metric is applicable to SRQs that do not have a periodic character and do not have a complex mixture of many frequencies. They state that the types of SRQs that are periodic and contain many frequencies require sophisticated time-series analysis and/or transformation into the frequency domain. They suggest using validation metrics constructed by Geers [14], Russell [15] and Sprague and Geers [16] for periodic systems or system responses with many frequencies.

Along with the three validation metrics (Geers, Russell and Sprague and Geers) many other error measures and error metrics exist that can be used to quantify the agreement between two time histories. Table 1 attempts to summarize the various error measures and error metrics found in literature. For a detailed discussion on each error measure/metric the reader is referred to the study that treats them in detail.

Table 1. Summary of Error Measures and Metrics

Error measure/ Metric	Advantages	Disadvantages
Discussed in [6][1]		
Vector norms		Norm choice leads to different conclusions. Not capable of distinguishing error due to phase from error due to magnitude.
Average Residual and its standard deviation		Positive and negative differences at various points may cancel out. Results of Average Residual and its standard deviation are conflicting.
Coefficient of correlation		Sensitive to phase difference and cannot distinguish between error due to phase and error due to magnitude.
Cross-correlation		Can only measure difference in phase
Sprague & Geers Metric	Gives error due to magnitude and phase separately which is useful when more detailed investigation of the error source is necessary.	Not symmetric. Cannot consider shape of the time histories.
Russell’s Error Measure	Symmetric	Same problem with respect to magnitude error as Sprague & Geers Metric.
Normalized Integral Square Error (NISE)		Magnitude error can be negative, which can decrease the combined error erroneously.
Dynamic Time Warping (DTW)	Effect of phase deviation on magnitude error can be minimized by using DTW.	
Discussed in [17]		
Sprague & Geers Metric	**Magnitude error** – Insensitive to phase discrepancies	

[1] The comments made in reference [6] regarding the advantages and disadvantages are made with respect to their application to vehicle safety applications.

Error measure/ Metric	**Advantages**	**Disadvantages**
	Phase error – Uses error proposed by Russell. Insensitive to magnitude differences. Defines a **Comprehensive error**	
Discussed in [18]		
Russell's error measure	Magnitude error is unbiased and signed.	
Geers' error measure		May only be an appropriate choice when a high level of confidence exists in the test data
Whang's inequality		No means for evaluating phase and magnitude errors
Theil's inequality		No means for evaluating phase and magnitude errors
Zilliacus' error		Incorrectly identifies the degree of error
RSS error factor		Incorrectly identifies the degree of error
Regression coefficient		Incorrectly identifies the degree of error
Johansen's magnitude		Should not be used in current state.
Johansen's energy		Should not be used in current state.

Reference [18] evaluated various measures in Table 1 and concluded that some error measures are very similar and others incorrectly identify the error. He recommends using Geers', Whang's or Russell's error measure [15], but state that Geers' error may only be an appropriate choice when a high level of confidence exists in the test data, and that Whang's inequality is very sensitive to phase errors. Russell [15] developed a set of magnitude, phase and comprehensive error measures that can be used to evaluate the deviation between two general functions or test and analytical data. Russell's error measures address some of the issues associated with some of the existing measures given in Table 1. He states the following five deficiencies with existing error measures, which he claims his proposed error measure resolves:

i. The value may not be well bounded and therefore may make it difficult to evaluate and compare results,
ii. the physical interpretation of the results may not be intuitive,
iii. the degree of error may not be correctly identified,
iv. the results can not be used to identify the cause of the error
v. the basis of the error factor may not be understood, which can lead to false interpretations of the results.

Reference [6] proposes three error measures describing the error in magnitude, the error in phase and the error in slope by combining existing measures. The three measures are then combined into a single validation metric based on linear regression using Subject Matter Expert (SME) ratings. Much of the objectivity of the proposed validation metric is lost as the metric is based on the subjective opinions of SMEs. Before the validation metric can be used to validate a model, the validation metric has to be created by training it, in order for it to be able to evaluate the model. This training is done by fitting the regression model to the SME ratings of the comparison between different data. This makes it highly dependent on the SMEs and it will therefore not be possible to compare the quantitative results of comparisons between two different models to a single set of "true" data (or test data) made using two different sets of SMEs, unless the SMEs' assessment is the same and given that the SME exists. The error measure proposed by [6] is not used further in this paper as the metric is heavily dependent on SMEs. This causes it to lose a lot of the required objectivity of a quantitative validation metric. This metric may however be useful in certain applications.

From the above mentioned studies ([6], [18]) in which various error measures/metrics were evaluated, it would seem that the two most likely error measures/metrics to give the most reliable validation results are Russell's error measure and Sprague & Geers' metric. These two metrics will now be discussed in more detail.

2.1.1. Russell's error measure

The following equations are used to calculate the magnitude, phase and comprehensive error measures as presented in reference [15].

For the magnitude error the following equation is used:

$$M_R = sign(rme) Log_{10}(1+|rme|)$$

With the relative magnitude error (*rme*) computed by,

$$rme = \frac{\sum_{i=1}^{N} p_i^2 - \sum_{i=1}^{N} m_i^2}{\sqrt{\sum_{i=1}^{N} p_i^2 \sum_{i=1}^{N} m_i^2}}$$

p and *m* represent the two signals that are being compared. *p* Represents the predicted data obtained from the simulation model and *m* is the measured data obtained from the experiment. *N* equals the number of data points in the measured (*m*) and predicted (*p*) data. The length of *p* and *m* should be the same.

For the phase error the following equation is used:

$$P_R = \frac{1}{\pi}\cos^{-1}\left(\frac{\sum_{i=1}^{N} p_i m_i}{\sqrt{\sum_{i=1}^{N} p_i{}^2 \sum_{i=1}^{N} m_i{}^2}}\right)$$

The magnitude and phase error are combined into a comprehensive error, C_R :

$$C_R = \sqrt{\frac{\pi}{4}(M_R^2 - P_R^2)}$$

2.1.2. Sprague & Geers' metric

The most recent version of Geers' error measure [14], presented in reference [19], will be used. In this version the equation for the phase error has been updated.

The Sprague & Geers' (S&G) magnitude error is calculated by:

$$M_{S\&G} = \sqrt{\frac{\sum_{i=1}^{N} p_i^2}{\sum_{i=1}^{N} m_i^2}} - 1$$

The phase error is calculated by:

$$P_{S\&G} = \frac{1}{\pi} \cos^{-1} \left(\frac{\sum_{i=1}^{N} p_i m_i}{\sqrt{\sum_{i=1}^{N} p_i^2 \sum_{i=1}^{N} m_i^2}} \right)$$

The Sprague and Geers' comprehensive error measure is given by:

$$C_{S\&G} = \sqrt{M_{S\&G}^2 + P_{S\&G}^2}$$

From the above equations it can be observed that the phase error of both metrics is calculated in the same way. The calculation of the magnitude and comprehensive error differs between S&G's metric and Russell's error measure. Whether these two validation metrics are able to 1) give results for which the physical interpretation is intuitive and 2) identify the degree of error correctly, is not clear. A validation metric will be proposed in the next section that will address these two aspects directly. This proposed validation metric will then be compared to Russell's error measure and to the Sprague & Geers (S&G) metric in section 3.

2.2. Validation metric based on relative error

The validation metric that is proposed will use the simple and commonly used relative error to quantify the agreement/disagreement between two data sets. The data sets may be SRQs obtained from a physical system and a model. The use of the relative error as a validation metric has been employed in previous studies ([1], [13]). Reference [1] states that, as long as the measured (*m*) data is not near zero, the relative error metric is a useful quantity. A similar remark is made by [17] stating that "a simple metric such as relative error works well for point-to-point comparisons, e.g. maximum deflection of a cantilever beam. However, when comparisons involve time or spatial variations, e.g. velocity history at a point or deflection along a beam, the application of a simple metric like relative error becomes sensitive to inaccuracies in time and space dimensions as well as the system response quantity (SRQ)". As mentioned, this paper will consider the comparison of SRQs with a periodic nature and which may have values at or near zero, which according to reference [1] and [17] will cause difficulties in using the relative error as a validation metric. Before discarding the use of the relative error as a validation metric on periodic systems where the measured (*m*) data might be near or equal to zero, we'll investigate the characteristics of the relative error, its various challenges and suggest ways to circumvent them.

2.2.1. Relative error (*RE*)

The equation for the relative error between two values is given in eq.{1}. Consider the two values as one being the measured (*m*) and the other the predicted (*p*) value, with the measured value taken as the true (or reference) value.

$$RE = \left| \frac{p-m}{m} \right| \qquad \{1\}$$

The calculation of the relative error between a measured (*m*) and predicted (*p*) value is simple and when expressed as a percentage (see eq.{2}) easy to interpret.

$$\%RE = \left| \frac{p-m}{m} \right| \times 100 \qquad \{2\}$$

The relationship between the *RE* and the ratio *p/m*, which represents the respective over or under prediction of the measured value, is shown in Figure 2.

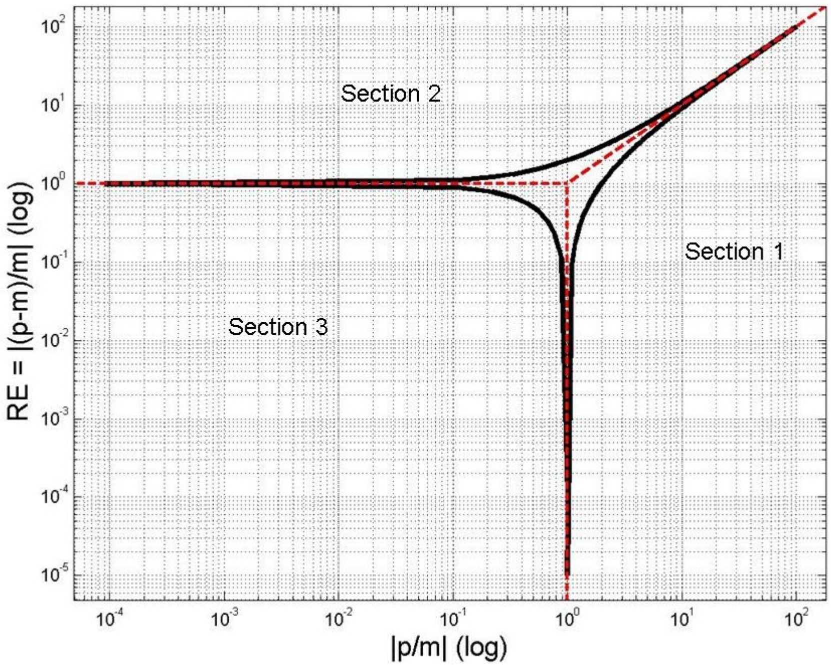


Figure 2. Relationship between the *RE* and the ratio *p/m*

From Figure 2 the following observations can be made:

i. The first obvious observation is that in the limit of *p* approaching *m*, *RE* goes to zero ($\lim_{p \to m} RE = 0$).

ii. When $\frac{p}{m} >> 1$ the relative error goes to positive infinity (along the line in section 1),

iii. Similarly, when $\frac{p}{m} << 1$ the relative error goes to negative infinity. However, because we plot the absolute values of the ratio *p/m* these large negative values instead go to positive infinity (along the line in section 2)

The ratio of *p/m* indicates whether the predicted value *p* is an over or under prediction of the measured value *m*. The predicted value *p* is said to be an over prediction of *m* if *p* is a larger positive value when *m* is positive, or when *p* is a larger negative value when *m* is negative. Similarly, *p* is classified as an under prediction when *p* is a smaller positive value (or any negative value) when *m* is positive, or when p is a larger negative value (or any positive value) when *m* is negative. With this convention relative errors that fall in section 1 are over predictions and relative errors in either section 2 or 3 are under predictions.

From Figure 2 and eq.{1} it is obvious that the relative error may result in infinite values and NaNs (Not-a-Number) due to the operations of 0/0 and 1/0, which may make further calculations on the relative error difficult. These challenges are discussed in the following paragraph.

2.2.2. Challenges in using the *%RE* as validation metric

The challenges concerning the use of the percentage relative error as a validation metric mainly arise when data has to be compared that have been obtained from a periodic system and the measured and/or predicted data has values equal to, or near, zero. Figure 3 shows an example of measured and predicted data obtained from a periodic system.

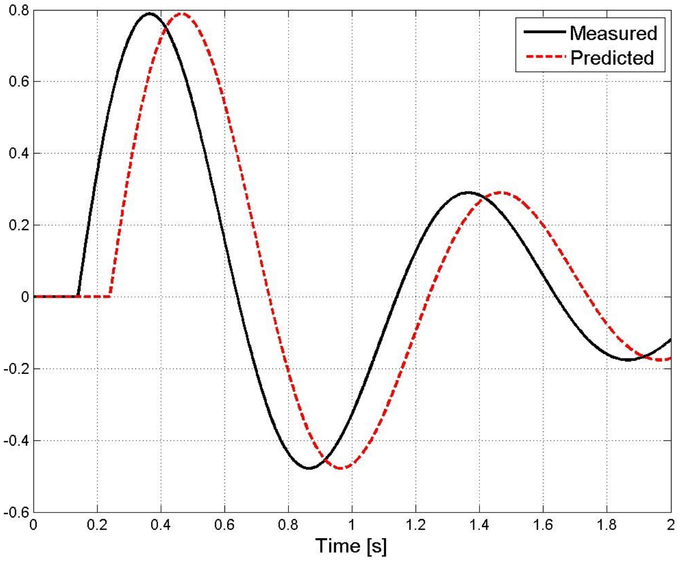


Figure 3. SRQs from experimental measurements and model simulations of a periodic system

The challenges associated with the *%RE* when periodic data, as shown in Figure 3, is compared are:

i. Non-constant *%RE* over the independent variable (time in the case of Figure 3),
ii. NaNs (Not-a-Number) present in the *%RE*s,
iii. Inf (infinite) values present in the *%RE*s.

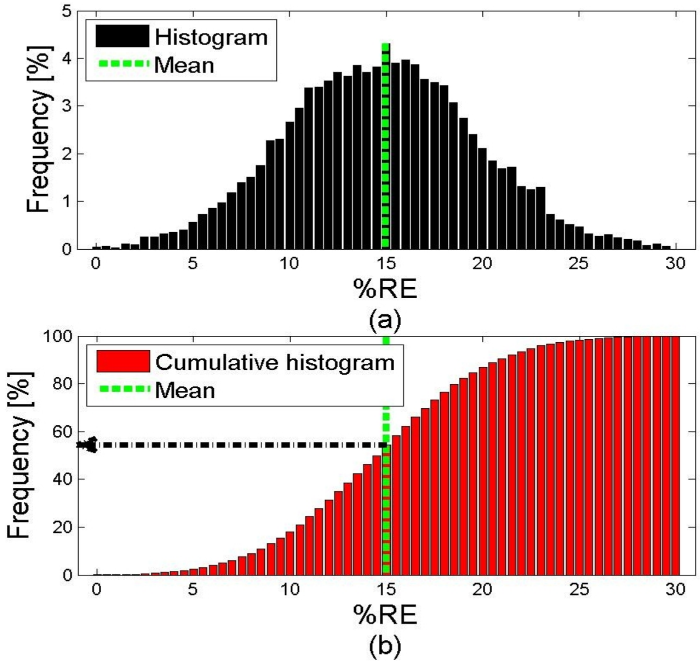
The first challenge faced when using the *%RE* in comparing two periodic SRQs, is that the *%RE* at each data point may not be the same. This makes it difficult to report a single representative result indicating the overall agreement/disagreement. Further challenges that are associated with using the *%RE* arise from comparing periodic SRQs that are near zero. When calculating the

%RE of periodic SRQs near zero, NaNs and Inf values may be present in the *%RE*. These values result from the operations 0/0 and 1/0, respectively, and make further calculations on the *%RE* difficult. These challenges are discussed in further detail in the following paragraphs and methods to overcome them, are proposed.

Non-constant *%RE* over the independent variable

The *%RE* may not have a constant value over the entire range of the data. In other words, the *%RE* may have different values for each data point. This makes it difficult to report on the agreement between the measured and predicted data using the *%RE*. When the *%RE* does not have a constant value, one of two methods can be used to report a single representative value for the *%RE*. In the two methods the non-constant *%RE* will be represented by a modified *%RE* defined either by the mean of the *%RE*s or by a specific *%RE*. In both methods a probability will be given that represents the percentage of *%RE*s that are below, or equal to, either the mean of the *%RE*s or the specific *%RE* that was chosen. When the mean of the *%RE*s is used to define the modified *%RE*, it will be denoted as $m\%RE^{m}$ and by $m\%RE^{s}$ when it is defined by a specific *%RE*.

In order to define the $m\%RE^{m}$ the mean and cumulative histogram of the *%RE*s are calculated. Using the cumulative histogram and the mean, the probability is calculated that the *%RE*s are at or below the mean *%RE*. Figure 4(a) shows the histogram and the mean of the *%RE*s and Figure 4(b) the cumulative histogram and the mean of the *%RE*s for an arbitrary set of *%RE*s. Figure 4 illustrates that when we take the y-intercept (representing the frequency of a specific *%RE*) of the cumulative histogram where the mean value intersects the cumulative histogram, we can obtain the probability that the data is at, or below the mean value. Therefore, for the data in Figure 4 the result will be that 55% of the *%RE*s are at or below the mean *%RE* of 15% ($m\%RE^{m}$ = 15% P(55%)).



(a)

(b)

Figure 4. (a) Superimposed mean and histogram, and (b) superimposed mean and cumulative histogram of *%RE* with a normal distribution

Even if the *%RE* does not have a normal distribution (see Figure 5(a)) the mean *%RE* can still be used to define the $m\%RE^{m}$. Figure 5(a) shows the non-normal distribution of the *%RE* and Figure 5(b) presents the cumulative histogram with the mean *%RE* superimposed on it. For the example in Figure 5 we obtain that 57.6% of the *%RE*s are at or below the mean *%RE* of 48.5% ($m\%RE^{m}$ = 48.5% P(57.6%)).

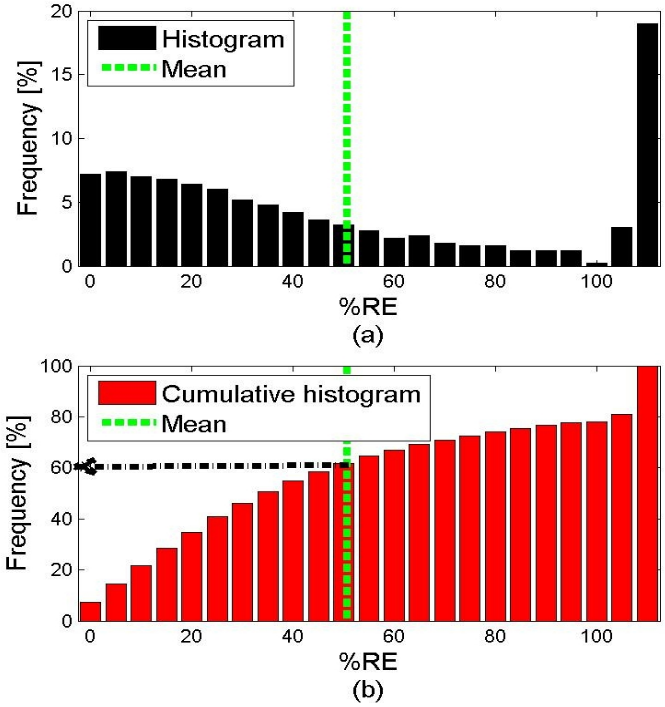


Figure 5. (a) Superimposed mean and histogram, and (b) superimposed mean and cumulative histogram of *%RE* with a non-normal distribution

In the second method, instead of defining the modified *%RE* by using the mean *%RE*, a specific *%RE* can be chosen and the probability that the *%RE*s are below, or equal to, this specific *%RE* can be calculated. However it may be that there is no *%RE*s below this chosen *%RE* and the result will be that there is zero probability that the *%RE*s are below this chosen *%RE* ($m\%RE^{s}$ = x% P(0%)). This result will make comparing and selecting the best model from a group of models impossible if this is the result for all the models. In this situation it may be better to use the mean *%RE* to define the modified *%RE*. However, in the situation where accuracy requirements are set for the model the result from the $m\%RE^{s}$ can easily be used to check whether the requirements are satisfied or not. It will therefore depend on the intended use of the validation metric whether the modified *%RE* is defined by a specific *%RE* or the mean *%RE*.

NaNs present in the *%RE*s

The calculation of the *%RE* is subject to the operation 0/0. This is one of the major problems encountered when using the relative error to quantify the agreement/disagreement between the measured and predicted SRQs obtained from periodic systems near zero. This occurs when the *%RE* is calculated at a point where the measured and predicted values equal zero. The IEEE floating point representation of 0/0 is NaN (Not-a-Number). The presence of NaNs in the *%RE* makes it difficult to perform further calculations on it. It is proposed that any NaN is set equal to

0 as the operation 0/0 implies that the *%RE* is equal to zero. The next step is now to use either the *m%RE*^{m} or the *m%RE*^{s} to report a single value for the *%RE*s in order to quantify the overall agreement between two SRQs.

The following considerations should be kept in mind when using the *m%RE*^{m}. Zeros in the *%RE* results in problems with the representation of the *m%RE*^{m}. Consider the *%RE* between measured and predicted data having 10 data points. Nine of them are equal to 8% and one is a NaN. Assigning a zero to the NaN the mean of the *%RE* is 7.2%. Using the mean of 7.2% we will obtain a probability that 0.1% of the *%RE*s are lower than 7.2% (*m%RE*^{m} = 7.2% P(0.1%)). If we calculate the mean of the *%RE* but now ignoring any zero value we will obtain a mean of 8%. This will give us the result that 100% of the *%REs* are at, or below, 8% (*m%RE*^{m} = 8% P(100%)). Ignoring the zeros gives a result that represents the agreement better. Consider the example shown in Figure 6 that reiterates this. We have a true time response and an approximation to the true response. We know that the amplitude of the approximate response deviates from the true response by 10%. If we calculate the modified *%RE* defined by the mean *%RE*, including the zero values, the result is that there is a 0.2% probability that the *%RE*s are smaller or equal to 9.98% (*m%RE*^{m} = 9.98% P(0.2%)). Even though this is true, the results without including the zeros in the calculation of the mean *%RE*, are considered more meaningful. Excluding the zeros when the mean of the *%RE* is calculated, the result is obtained that a 100% of the *%RE*s are equal to, or below 10% (*m%RE*^{m} = 10% P(100%)), which we know to be true. Therefore the mean of the *%RE*s, defining the *m%RE*^{m}, will be calculated neglecting all the zero values in the *%RE*s. Zeros in the *%RE*s do not have the same affect on the *m%RE*^{s} and does not need any special consideration.

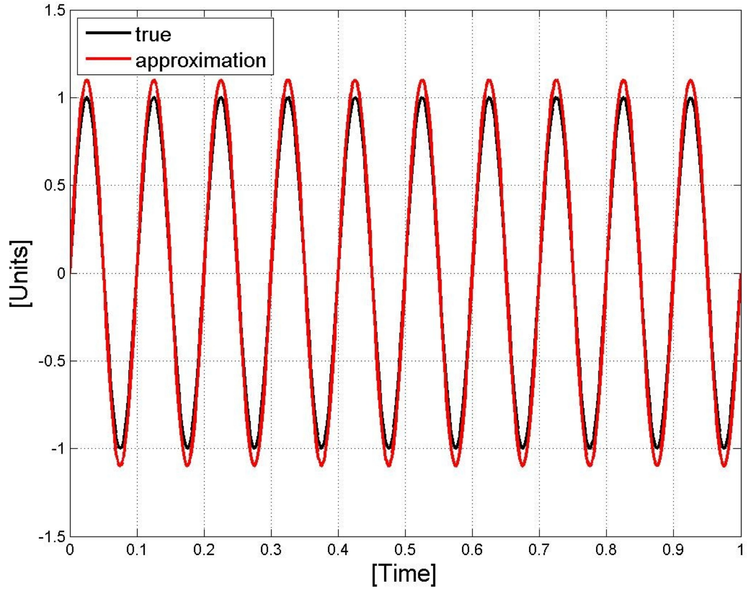


Figure 6. Example of time response histories of SRQs for the physical system (true) and computation model (approximation)

Inf values present in the *%RE*s

We already looked at how to handle operations involving 0/0 which the IEEE represents as NaNs. Another problem with using the relative error in comparing periodic signals is introduced by operations involving 1/0. This occurs when the measured value is zero and the predicted value has a non-zero value. The IEEE uses Inf to represent these operations The effects of operations involving 1/0 on the modified *%RE* are discussed at the hand of an example.

In Figure 7, two mathematical models (approximation 1 and 2) are compared to the physical system (true). Approximation 1 has a 10% deviation and approximation 2 a 30% deviation from the true value. In Figure 7(a) the data of both approximations are perfectly in-phase with the true data, whereas in Figure 7(b) the approximations and true data have some phase difference. Table 2 shows the results for the modified *%RE*. For the in-phase case we obtain the expected $m\%RE^m$ of 10% P(100%) and 30% P(100%) respectively, however for the out-of-phase case we obtain the result that there is a 100% probability that the *%RE*s are smaller than infinity ($m\%RE^m = \infty$ P(100%)), which has no meaning even though it is correct. The $m\%RE^s$ gives meaningful results for both the in-phase and out-of phase example.

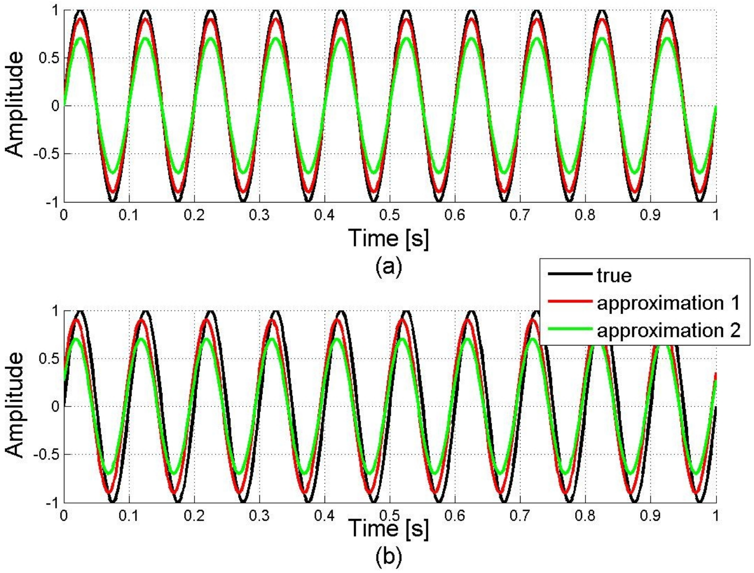


Figure 7. (a) Approximation 1 and 2 in-phase with true data. (b) Approximation 1 and 2 out-of-phase with true data

Table 2. Effect of *%RE* not being bounded on the results of the *m%RE* (Not bounded)

	(a) In-phase		**(b) Out-of-phase**	
	Approximation 1	**Approximation 2**	**Approximation 1**	**Approximation 2**
$m\%RE^m$	**10%** P(100%)	**30%** P(100%)	Inf P(100%)	Inf P(100%)
$m\%RE^s$	**15%** P(100%)	**15%** P(0.2%)	**15%** P(21.6%)	**15%** P(13.6%)

From the results in Table 2 it is clear that the presence of Inf values do not affect the results of $m\%RE^s$. However, the presence of Inf values in the *%RE*s makes it difficult to compare the models using the $m\%RE^m$. There is one of two ways to deal with Inf values in the *%RE*s when the $m\%RE^m$ is used namely:

i. If $\%RE_i$ > Inf, then remove Inf value from *%RE* data, or
ii. Bound the *%RE*s.

Completely removing the Inf values, as proposed in method (i), will imply that the *%RE*s at these points are ignored. The implication of this is that when the *m%RE^m* is used, the mean of the *%RE* will be lower than it really is. In the case were both models have the same amount of values above the specified *%RE threshold* method (i) will not influence the results negatively. However, if only one model has values above the threshold that is removed this may lead to the incorrect model being chosen as the more accurate model. Bounding the *%RE* as proposed in method (ii) will be less likely to make an erroneous model choice. Reference [13] presents a validation metric that uses the relative error combined with the hyperbolic tangent function which results in the relative error being bounded. Their equation is changed and presented here as equation {3}. *V* is the bounded *RE*.

$$V = \tanh\left|\frac{p-m}{m}\right| \qquad \{3\}$$

Plotting this equation on Figure 8 shows that the implementation of *tanh* bounds the *RE*, for all ratios of *p/m*, to 1. However, using *tanh* the "true" relative error is distorted. As can be seen from Figure 8, eq.{3} deviates from the true relative error as it moves away from *p/m* = 1. Therefore, for a ratio of p/m = 2 equation {3} results in a relative error of 0.7616 instead of 1. This results in a 24% lower error than which truly exists. The use of *tanh* in combination with the *RE* bounds the *RE* to 1, but has the implication that the true relative error is lost.

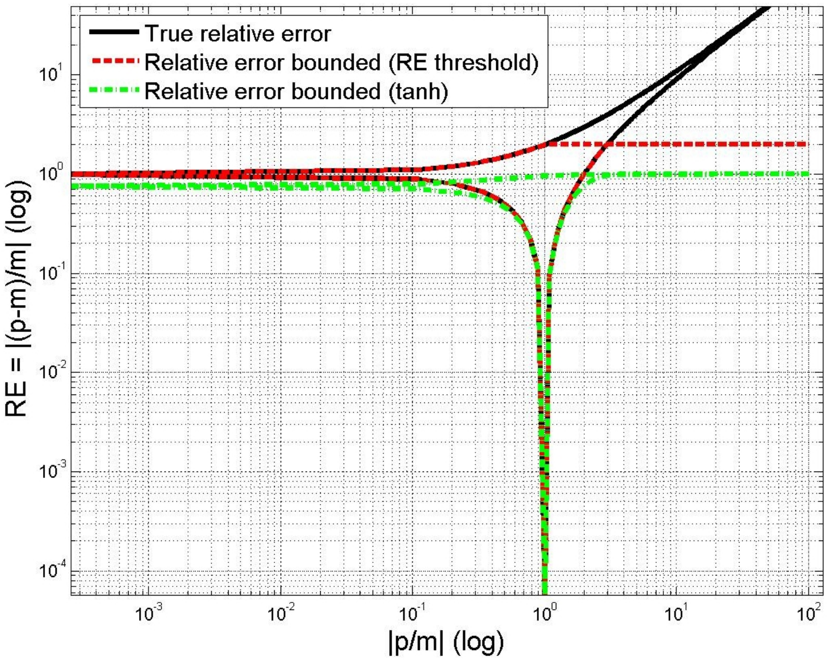


Figure 8. Relationship between the *RE* and the ratio *p/m* (relative error bounded)

The relative error can be bounded without distorting the true relative error by setting any *RE* that is greater than a chosen *RE threshold* equal to the *RE threshold*. This implies that the *RE* is now bounded but unlike eq.{3} all the *RE*s below the *RE threshold* value are the true relative errors. The implementation of this and its effect on the relationship between the *RE* and the ratio *p/m* is also shown on Figure 8 as the graph *Relative error bounded* (*RE threshold*). From the figure it can be observed that the true relative error is obtained until the *RE threshold* is reached. Above the *RE threshold* the true relative error is set equal to the *RE threshold*. In figure one the RE threshold was set equal to 2.

We again calculate the %*RE* between the measured and predicted data in Figure 7 but now using the bounded *%RE*. The %RE threshold is set equal to a 100%. Using the bounded *%RE* we obtain results for the $m\%RE^{m}$ that can actually be interpreted (see Table 3). It is now possible to evaluate approximation 1 and approximation 2, using the $m\%RE^{m}$, in order to conclude that approximation 1 is more accurate than approximation 2. This is similar to the results obtained from the $m\%RE^{s}$ which also indicates that approximation 1 is better than approximation 2.

Table 3. Effect of *%RE* not being bounded on the results of the *m%RE* (Bounded)

	(a) **In-phase**		**(b)** **Out-of-phase**	
	Approximation 1	**Approximation 2**	**Approximation 1**	**Approximation 2**
$m\%RE^{m}$	**10%** P(100%)	**30%** P(100%)	**48.5%** P(57.6%)	**51.3%** P(56.8%)
$m\%RE^{s}$	**15%** P(100%)	**15%** P(0.2%)	**15%** P(21.6%)	**15%** P(13.6%)

From the results in Table 2 and Table 3 it is clear that whether the *%RE* is bounded or not the $m\%RE^{s}$ is unaffected. Therefore unlike the $m\%RE^{m}$ the $m\%RE^{s}$ will not be affected by the choice of the *%RE threshold*. It is important that when the $m\%RE^{s}$ is used, that the specific *%RE* that is chosen to define the modified *%RE*, is never above the *%RE threshold*. The effect of the choice of the *%RE threshold* on the $m\%RE^{m}$ will be discussed by considering three scenarios:

i.) All *%RE*s < *%RE threshold*,
ii.) Some *%RE*s > *%RE threshold*, and
iii.) All *%RE*s > *%RE threshold.*

It is obvious that for scenario 1 the choice of the *%RE threshold* is irrelevant. With scenario 2 having some *%RE* greater than the *%RE threshold* the choice of the *%RE threshold* will affect the result of the $m\%RE^{m}$. Consider the example given in Table 4. We have a set of true values and their associated *%RE* between the true and approximate data. Table 5 presents the results for the $m\%RE^{m}$ for two *%RE threshold* values.

Table 4. Known *%RE* between true and approximate data

Data point	True	*%RE*
1	0	0
2	0.5	90
3	0.8	80
4	1	60
5	1.2	50
6	0.8	4
7	0.7	6
8	0.6	8
9	0.5	10
10	0	1
11	-0.5	-20
12	-0.8	-30

Data point	True	*%RE*
13	-1	-55
14	-1.2	-35
15	-0.8	-25
16	-0.7	-20
17	-0.6	-15
18	-0.5	-10
19	1	200
20	1	200
21	1	200

Table 5. Results for the $m\%RE^m$ using different *%RE threshold* values

%RE threshold	110%	250%
$m\%RE^m$	44.6% P(61.9%)	58.8% P(71.4%)

Figure 9 shows the histogram, cumulative histogram and the mean of the *%RE* for both the *%RE threshold* equal to 110% and 250%. From this figure it can be observed that the *%RE*s smaller than the *%RE threshold* are not affected by the choice of the *%RE threshold*. Therefore the histogram and cumulative histograms will be identical up until the *%RE threshold* after which they will differ. This implies that both results in Table 5 are correct and it can be concluded that it does not matter what value is chosen for the *%RE threshold*, as long as the same *%RE threshold* is used when two models are compared.

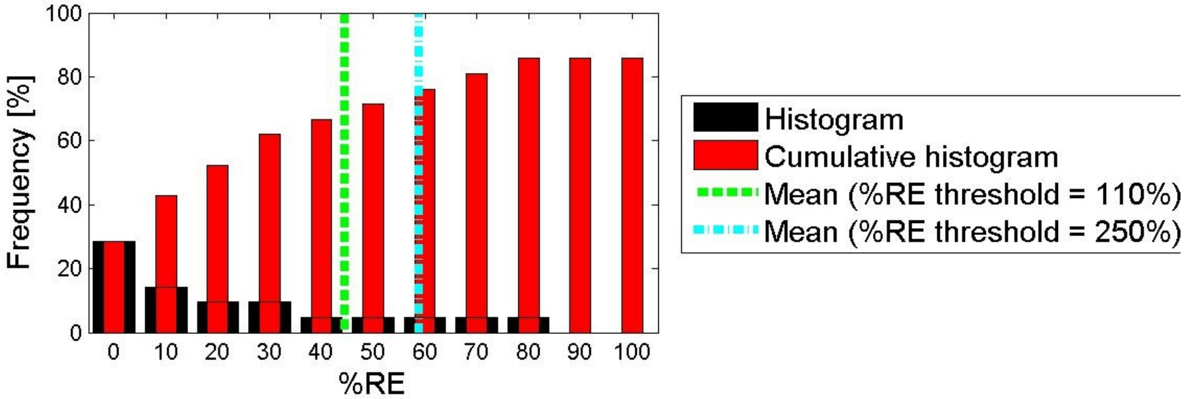


Figure 9. Superimposed histograms, cumulative histograms and means of *%RE* using different γ values

This brings us to the third scenario. If all the *%RE*s are above the *%RE threshold*, all the *%RE*s will be set equal to the *%RE threshold*. This implies that if the *%RE threshold* = 100% the results for both model 1 with a constant *%RE* = 200% and model 2 with a constant *%RE* = 150% will be that 100% of the *%RE*s are below a 100% ($m\%RE^m$ = 100% P(100%)). This result is obtained because the percentage relative errors that are above the *%RE threshold* are bounded by the *%RE threshold*. Having two models with an accuracy worse than 100% may already make them invalid models, however if they need to be analysed, the threshold value for the *%RE* can be adjusted. By adjusting the *%RE threshold* to 200% we get that 100% of the *%RE*s are below 200% for model 1 and that 100% of the *%RE*s are below 150% for model 2. Therefore model 2, although bad, is a better approximation to the measurements than model 1.

It is important to remember that because the *%RE*s greater than the *%RE threshold* is bounded to the *%RE threshold*, the *%RE*s above the *%RE threshold* are not the true *%RE*s. This is important especially when a specific *%RE* is chosen to define the modified *%RE*. The specific *%RE* should always be below the *%RE threshold*. When the specific *%RE* is chosen below the *%RE threshold* both the $m\%RE^s$ and the $m\%RE^m$ will give the true relative error.

2.2.3. Summary of the modified *%RE* validation metric

The modified percentage relative error validation metric and its two formulations ($m\%RE^m$ and $m\%RE^s$) were presented. It should be realized that both formulations can be used to either compare (and rank) a set of models or to evaluate the model against accuracy requirements using either formulation of the modified *%RE*. It is however suggested that the $m\%RE^m$ should be used for comparing a set of models and the $m\%RE^s$ used to evaluate the model against the accuracy requirements. Table 6 summarizes the two formulations of the modified percentage relative error validation metric.

Table 6. Summary of the two formulations of the modified *%RE* validation metric

	$m\%RE^m$	$m\%RE^s$
Defined by	The percentage of *%RE*s (given as a probability) that are equal to, or below, the mean of the *%RE*s: $m\%RE^m = \text{mean}(\%RE\text{s})\ P(\%)$ **Note**: the mean of the *%RE*s is calculated neglecting all zero values in the *%RE*s	The percentage of *%RE*s (given as a probability) that are equal to, or below, the specified *%RE* (for example x%): $m\%RE^s = x\%\ P(\%)$ **Note**: The specified *%RE* must always be below the *%RE threshold*
Preprocessing of the *%RE*s	NaNs (Not-a-Number)	
	Set NaNs = 0	Set NaNs = 0
	Inf (infinite) values	
	Bound *%RE*s with *%RE threshold* **Note**: Choice of *%RE* threshold influences result of the $m\%RE^m$. *%RE* threshold must be the same when comparing $m\%RE^m$ results	None required **Note**: If the *%RE*s were bounded, x% must be below the *%RE* threshold
Suggested uses	Primary use	
	Comparing and selecting the best model from a group of models	Evaluation of model against accuracy requirements
	Secondary use	
	Evaluation of model against accuracy requirements	Comparing and selecting the best model from a group of models

In the previous paragraph the use of the relative error as basis for a validation metric between two data sets was investigated. The relative error gives intuitive results, but has certain challenges. We discussed how these challenges can be overcome to still get useful intuitive results from the *%RE* when it is presented in the modified form (either $m\%RE^m$ or $m\%RE^s$). Because the modified *%RE* includes both the error due to a magnitude difference, as well as the error due to a phase difference, it is considered to be a comprehensive error. The modified *%RE* validation metric will now be compared to the validation metrics of Russell [15] and Sprague & Geers [19].

3. Comparison of validation metrics

The modified *%RE* validation metric will now be compared to the Sprague & Geers metric [19] and Russell's metric [15] that were presented in section 2.1. It should be noted that the magnitude, phase and comprehensive error measures of both S&G's and Russell's metric are multiplied by a hundred in order to present them as a percentage. This is done to compare it directly to the modified *%RE* metric. The *%RE threshold* value that is used throughout this study is a 100%.

From the comparison of the validation metrics we would like to conclude two things. Firstly, and most importantly, we would like to establish whether the validation metrics give a useful and reliable measure that quantify the agreement between the experimental and simulated data. It is important that the validation metrics give a reliable and easily interpretable metric which can be used to determine whether the model satisfies the accuracy requirements. Secondly we would like to evaluate the ability of the validation metrics to rank models and select the best model from a group of models. Analytical functions will firstly be used to compare the capabilities of the validation metrics to rank models. The analytical functions will also aid in determining whether the validation metrics can indeed quantify the agreement of the model and give a useful and reliable metric which will aid the engineer in deciding whether the model is valid or not. Case studies will then be used to further show the advantages and limitations of the different metrics.

3.1. Analytical functions

The analytical functions that we will use include the functions used in previous studies ([17], [18]) which are based on, and extensions of, the functions given in reference [14]. The analytical functions that are used are given in Table 7.

The analytical functions 1 to 15, listed in Table 7, were used in reference [18]. Functions 1 to 8 represent the predicted data and are compared to the measured data given by $m(t) = e^{-t}\sin(2\pi t)$. Similarly, functions 9 to 15 represent the predicted data that uses the measured data given by $m(t) = 1 - e^{-t/0.1} - 0.6e^{-t/0.4}\sin(5\pi t) + 0.01\sin(200\pi t)$. Functions 21(a), 22(a) and 22(b) are three additional functions used by [17] that were not considered by either [18] or [14]. The reference function for function 21(a), 21(b), 22(a) and 22(b) is given by $m(t) = e^{-(t-0.14)}\sin 2\pi(t-0.14)$.

Table 7. Equation for the various analytical functions

Function	Equation
Reference function for 1 to 8	$m(t) = e^{-t}\sin(2\pi t)$
1	$p(t) = 0.8e^{-t/0.8}\sin(2\pi t)$
2	$p(t) = e^{-t}\sin(1.6\pi t)$
3	$p(t) = 1.2e^{-t/1.2}\sin(1.6\pi t)$
4	$p(t) = 0.4e^{-t/1.2}\sin(1.6\pi t)$
5	$p(t) = 0.5e^{-t}\sin(1.6\pi t)$

Function	Equation
6	$p(t) = 0.6e^{-t/1.2}\sin(1.6\pi t)$
7	$p(t) = e^{-t}\sin(2\pi t) + 0.1e^{-t}\sin(30\pi t)$
8	$p(t) = e^{-t}\sin(2\pi t) + 0.3e^{-t}\sin(30\pi t)$
Reference function for 9 to 15	$m(t) = 1 - e^{-t/0.1} - 0.6e^{-t/0.4}\sin(5\pi t) + 0.01\sin(200\pi t)$
9	$p(t) = 1 - e^{-t/0.1} - 0.6e^{-t/0.3}\sin(5\pi t)$
10	$p(t) = 1 - e^{-t/0.1} - 0.6e^{-t/0.4}\sin(4\pi t)$
11	$p(t) = 0.6 - 0.6e^{-t/0.1} - 0.8e^{-t/0.3}\sin(4\pi .t)$
12	$p(t) = 0.6 - 0.6e^{-t/0.3} - 0.3e^{-t/0.5}\sin(3\pi .t)$
13	$p(t) = 0.3 - 0.3e^{-t/0.3} - 0.2e^{-t/0.5}\sin(3\pi .t)$
14	$p(t) = 1 - e^{-t/0.1} - 0.5e^{-t/0.4}\sin(5\pi .t) - 0.25t$
15	$p(t) = 1 - e^{-t/0.1} - 0.6e^{-t/0.4}\sin(4\pi .t) - 0.5t$
Reference function for 21 to 22	$m(t) = e^{-(t-0.14)}\sin 2\pi(t - 0.14)$
21(a)	$p(t) = 1.2e^{-(t-0.14)}\sin 2\pi(t - 0.14)$
21(b)	$p(t) = 0.8e^{-(t-0.14)}\sin 2\pi(t - 0.14)$
22(a)	$p(t) = e^{-(t-0.24)}\sin 2\pi(t - 0.24)$
22(b)	$p(t) = e^{-(t-0.04)}\sin 2\pi(t - 0.04)$

3.1.1. Ability to rank models and identify the best model

We start by comparing the validation metrics using functions 1 to 8 given in Table 7. The analytical functions 1 to 8 are compared to the same reference function. Table 8 shows how each validation metric ranks the 8 examples. All the functions are ranked the same by the three validation metrics except for Function 1 and 2. This gives an average agreement of 91.7%. The $m\%RE^m$ ranks Function 8 in 2nd and Function 1 as 3rd. The $m\%RE^m$ gives a lower mean for Function 1 than for Function 8 but Function 8 has a higher amount of *%RE*s below the mean *%RE* and therefore Function 8 is ranked higher than Function 1. The ranking of the functions in this order is confirmed when the results for the $m\%RE^s$ is considered. The $m\%RE^s$ for Function 8 is 37% P(63.85) and Function 8 therefore has a higher amount of *%RE*s below the same *%RE* than Function 1. The same functions were given to subject matter experts (SMEs) and asked to rank the comparisons of the eight functions to the reference function. The SMEs ranked all the functions the same as the three validation metrics except for Function 2 and 3 (see Table 9). The overall average agreement between the seven SMEs is 64.3%, which is a lot lower than between the three validation metrics.

Table 8. Ranking of comparisons by different validation metrics (Functions 1 to 8)

Function	S&G	Rank	Russell	Rank	$m\%RE^m$	Rank	Overall rank
1	28.68	4	20.2	3	37 P(47.9)	3	3 (66.6%)

Function	**S&G**	**Rank**	**Russell**	**Rank**	***m%RE^m***	**Rank**	**Overall rank**
2	23.4	3	20.7	4	68.9 P(40.1)	4	4 (66.6%)
3	38.6	5	27.3	5	71.5 P(36.9)	5	5 (100%)
4	61.9	8	46.2	8	76.8 P(44.2)	8	8 (100%)
5	55.6	7	41.2	7	75 P(43.6)	7	7 (100%)
6	42.9	6	32.9	6	72.4 P(43)	6	6 (100%)
7	3.3	1	2.87	1	18.8 P(76.8)	1	1 (100%)
8	10.4	2	8.9	2	39.2 P(66.2)	2	2 (100%)

Table 9. Ranking of comparisons by SME's (Functions 1 to 8)

Function	**SME #1**	**SME #2**	**SME #3**	**SME #4**	**SME #5**	**SME #6**	**SME #7**	**Overall rank**
1	2	3	5	3	3	3	3	3 (71.4%)
2	3	5	3	2	5	5	4	5 (42.8%)
3	4	6	4	4	4	4	5	4 (71.4%)
4	8	8	7	6	8	8	8	8 (71.4%)
5	6	4	8	7	7	6	7	7 (42.8%)
6	7	7	6	8	6	7	6	6 (42.8%)
7	1	1	1	1	1	1	1	1 (100%)
8	5	2	2	5	2	2	2	2 (71.4%)

The validation metrics were compared using another seven functions (Functions 9 to 15 given in Table 7). The results for the validation metrics and the SMEs are given in Table 10 and Table 11, respectively. The ranking by the three validation metrics and the seven SMEs are again the same for all but two Functions. For Functions 13 and 15 the SMEs rank these two functions as either 6th or 7th. The three validation metrics again have a higher average agreement of 90.5% against the 71.4% of the SMEs.

Table 10. Ranking of comparisons by different validation metrics (Functions 9 to 15)

Example	**S&G**	**Rank**	**Russell**	**Rank**	**m%RE**	**Rank**	**Overall rank**
9	0.91	1	0.8	1	5 P(85.5)	1	1 (100)
10	3.37	2	2.98	2	11.3 P(67.7)	2	2 (100)
11	40.1	4	28.2	4	45 P(77.6)	4	4 (100)
12	45.7	5	32.3	6	49.9 P(73.1)	5	5 (66.6)

Example	S&G	Rank	Russell	Rank	m%RE	Rank	Overall rank
13	72.89	7	57	7	75 P(72.5)	7	7 (100)
14	26.4	3	18.91	3	29.52 P(54.2)	3	3 (100)
15	50.1	6	36.4	5	54.6 P(48.5)	6	6 (66.6)

Table 11. Ranking of comparisons by SMEs (Functions 9 to 15)

Function	SME #1	SME #2	SME #3	SME #4	SME #5	SME #6	SME #7	Overall rank
9	1	1	1	1	1	1	1	1 (100)
10	2	2	2	2	2	2	2	2 (100)
11	4	4	4	3	4	4	3	4 (71.4)
12	5	6	5	5	5	5	4	5 (71.4)
13	6	7	6	6	7	7	5	6 (42.8) 7 (42.8)
14	3	3	3	4	3	3	6	3 (71.4)
15	7	5	7	7	6	6	7	6 (42.8) 7 (42.8)

From the above results it was observed that the three validation metrics and the SMEs tend to rank models similarly. The ranking of the functions by the validation metrics were done with more coherence than the ranking by the SMEs. The results may be influenced having more SMEs or using different groups of SMEs. Having additional validation metrics may also influence the results of the overall ranking of the models. These effects are outside the scope of this study. The results obtained seem to indicate that the validation metrics are able to rank the models and indicate which model is the best model from a group of models. The question is now whether all the metrics are able to give a reliable and useful measure of the level of agreement between the experimental and the simulated data.

3.1.2. Reliability and usefulness of validation metrics

The following two examples, indicated in Figure 10 and Figure 11, discuss the reliability and usefulness of the quantitative measure of the agreement/disagreement between two SRQs given by the various validation metrics. The example we consider in Figure 10 uses function 21(a) and 21(b) given in Table 7. There is no phase difference between the function representing the measured data and the two sets of predicted data represented by Function 21(a) and 21(b). The magnitude of Function 21(a) is 20% larger than the magnitude of the measured response and Function 21(b) is 20% smaller. We therefore know the error in magnitude between the measured response and the two models. This makes it possible to evaluate which of the metrics can indeed give the agreement between the two data sets correctly. Table 12 shows the results for the various metrics. Only S&G and the $m\%RE^{m}$ give the correct percentage relative error between the two signals. S&G is also capable of stating whether the magnitude is smaller or larger than the measured magnitude.

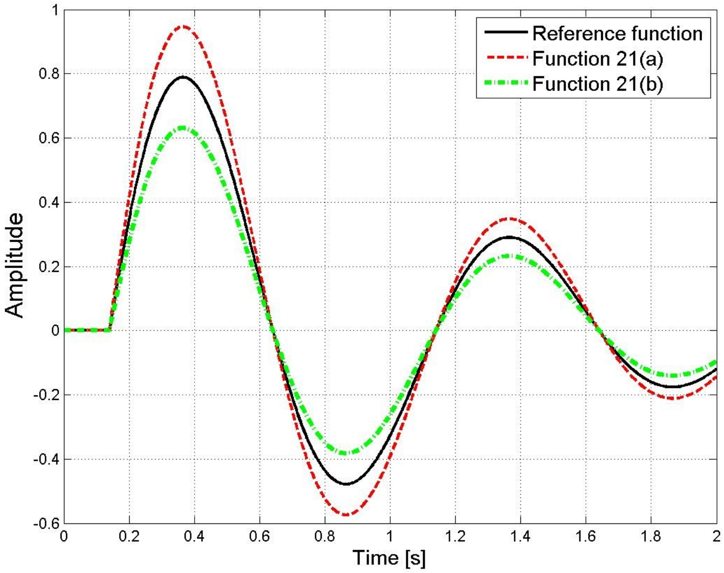


Figure 10. Comparison of Functions 21(a) and 21(b) to the reference function

Table 12. Comparison between the error measures' ability to quantify the accuracy (Functions 21(a) and 21(b))

	Function 21(a)			**Function 21(b)**		
	S&G	**Russell**	***m%RE^m***	**S&G**	**Russell**	***m%RE^m***
Magnitude	20	13.57		-20	-16.14	
Phase	0	0		0	0	
Comprehensive	20	12	20 P(100)	20	14.3	20 P(100)

The comprehensive error of S&G and *m%RE^m*, in the example where Functions 21(a) and 21(b) were used, is easy to interpret and captures the agreement of the two models. The magnitude and phase error of S&G provide additional information indicating that the error is due to a difference in the magnitude. However, when we consider two models with only a phase difference and no magnitude difference, as in Figure 11 for Functions 22(a) and 22(b), the results of the validation metrics need more consideration to understand what they actually mean. Considering the magnitude and phase error obtained from S&G for Function 22(a) and 22(b), it is clear that there is little difference in the magnitude compared to the reference function and that there exists a phase difference of almost 20% (see Table 13). However, the meaning of the comprehensive errors is not as clear. The comprehensive error of S&G in the comparisons of Functions 21(a) and (b) and Functions 22(a) and (b) are effectively equal. However, at time 0.4s the value that Function 21(a) had to predict is 20% higher than the reference function's value. At the same time (0.4s) the value that Function 22(a) had to predict is 30% lower than the reference function's value.

The *m%RE^m* gives a comprehensive error that is easier to interpret. In comparing Function 21(a) to the reference function the *m%RE* indicates that all the predicted values deviate less than 20% from the measured value. In comparing Function 22(a) to the reference function, the *m%RE^m* indicates that 53.1% of the errors between the responses are smaller than 60.3%. The magnitude and phase error of S&G for Functions 21(a) and (b) and 22(a) and (b) is easily interpretable, whereas its comprehensive error is not, as discussed above. Combining the magnitude and phase errors of S&G with the comprehensive error of the *m%RE^m*, we obtain a validation metric that has a meaningful comprehensive error. Furthermore, this combination of S&G and the $m\%RE^m$ makes it possible to determine whether the error is in the magnitude and/or in the phase.

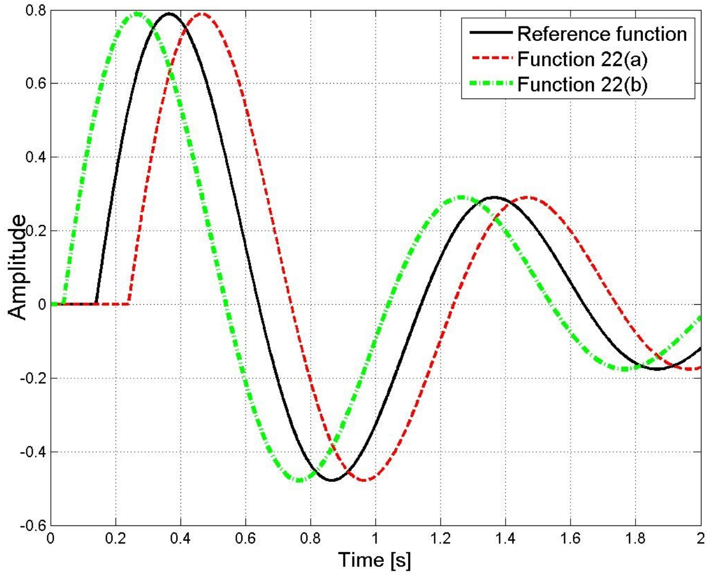


Figure 11. Comparison of Functions 22(a) and 22(b) to the reference function

Table 13. Comparison between the error measures ability to quantify the accuracy (Functions 22(a) and 22(b))

	Function 22(a)			Function 22(b)		
	S&G	**Russell**	$m\%RE^m$	**S&G**	**Russell**	$m\%RE^m$
Magnitude	-0.48	-0.41		0.14	0.122	
Phase	19.5	19.5		19.5	19.5	
Comprehensive	19.5	17.3	60.3 P(53.1)	19.5	17.3	59.4 P(51.9)

3.1.3. Combination of S&G and the modified *%RE*

Figure 12 shows two approximations obtained from Model 1 and Model 2 both having the same deviation in phase from the true value. The amplitude of Model 1 is 10% higher than the measured value and Model 2 is 10% lower. The results for the different validation metrics are shown in Table 14. Analyzing the results of the different validation metrics on their own are not as insightful as combining them. When we combine the magnitude and phase error of S&G with the comprehensive error of the $m\%RE^m$ we can form the following conclusion. The agreement of both Model 1 and Model 2 is approximately similar with roughly 58% of the *%RE* being below 51%. The deviation of both Model 1 and Model 2 is due to a difference in both phase and magnitude. Model 1 and Model 2 have the same difference in phase with the amplitude of model 1 being 10 % higher than the true signal's amplitude and Model 2, 10% lower. In the context of the validation procedure the magnitude error measure does not mean that Model 1 over predicts the true (measured) values and that Model 2 under predicts the values. If the phase difference between the two signals were zero then the magnitude error measured of S&G would have indicated that Model 1 over predicts the true data and Model 2 under predicts the data. In order to comment on Model 1 and Model 2 over or under predicting the values, the relationship between the relative error and the ratio of p/m, as discussed in paragraph 2.2.1, should be used to calculate whether the model is under or over predicting.

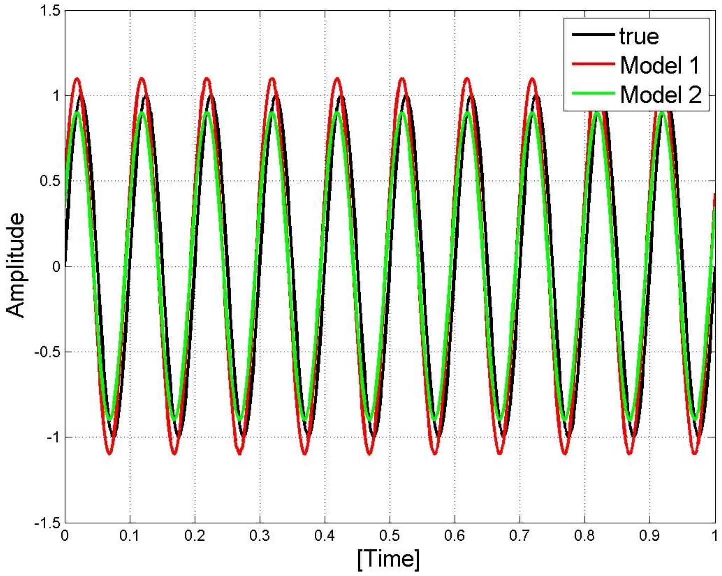


Figure 12. Model 1 and Model 2 with same phase shift but different magnitudes

Table 14. Comparison between error measures for models with same phase shift but different magnitudes

	Model 1			**Model 2**		
	S&G	**Russell**	***m%RE^m***	**S&G**	**Russell**	***m%RE^m***
Magnitude	10	7.6		-10	-8.3	
Phase	12.7	12.7		12.7	12.8	
Comprehensive	16.2	13.2	51.3 P(55.6)	16.2	13.5	48.5 P(57.6)

3.2. Case studies

Two case studies will now be used to further compare the validation metrics. The reliability and usefulness of the validation metric's results in quantifying the measure of agreement between the experimental and simulated data are further investigated using these case studies.

3.2.1. Case study 1: Known error between signals

Consider the two predicted SRQs obtained from Model 1 and Model 2 shown in Figure 13. The *%RE* between the two predicted SRQs and the measured SRQ are known and shown in Table 15. The results for the different metrics are shown in Table 16. Looking at the comprehensive errors of S&G and Russell, Model 1 seems to be a closer fit to the measured data than Model 2. However, when we consider the *%RE* between the models and the measured data, shown in Table 15, it is clear that Model 2 has the smaller *%RE* and is therefore closer to the measured data. The *m%RE^m* metric correctly shows that Model 2 is closer to the measured data stating that 60% of the errors are smaller than 35.2%. When the magnitude and phase errors of S&G are considered along with the results from the *m%RE^m* metric for Model 2 it can be seen that the difference in magnitude is the major contributor to the errors as the error in phase is small. For Model 1 the magnitude and phase errors of S&G give similar results and it is difficult to conclude whether the deviation is due to an error in the magnitude or an error in the phase. From Figure 13 it seems as if the deviation is largely due to an error in the magnitude.

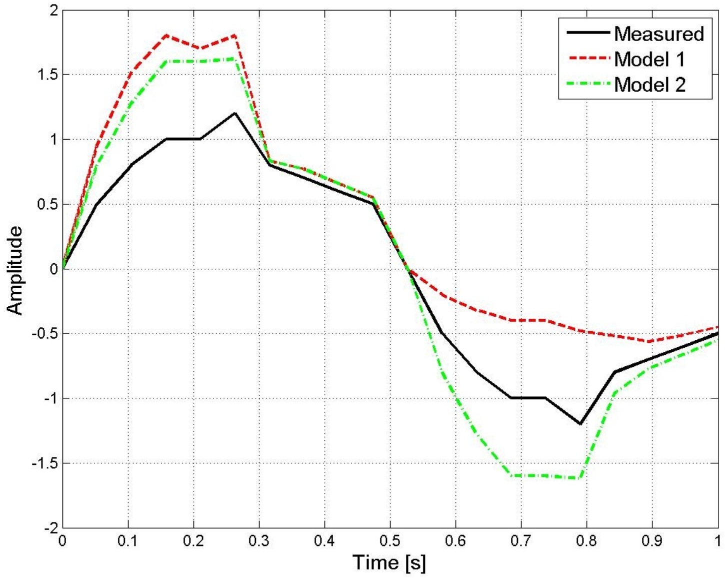


Figure 13. Two models with known *%RE* relative to the measured data

Table 15. Relative error between Model 1, Model 2 and the measured data

Data point	Model 1	Model 2
1	0	0
2	90	60
3	90	60
4	80	60
5	70	60
6	50	35
7	4	4
8	10	10
9	10	10
10	10	10
11	0	0
12	-60	60
13	-60	60
14	-60	60
15	-60	60
16	-60	35
17	-35	20
18	-20	10
19	-15	10
20	-10	10
mean(|*%RE*|)	39.7	31.7
mean(|*%RE*|) (without zero)	44.1	35.2

Table 16. Comparison between error measures for known *%RE*

	Model 1			**Model 2**		
	S&G	**Russell**	***m%RE^m***	**S&G**	**Russell**	***m%RE^m***
Magnitude	16.28	11.3		41.8	23.3	
Phase	16.27	16.2		4.75	4.8	
Comprehensive	23	17.5	44.1 P(50)	42.1	21.1	35.2 P(60)

3.2.2. Case study 2: Elasto-plastic leaf spring models

In this case study the force of a multi-leaf spring, induced by a certain displacement input, has been measured. We have two models which predict the behaviour of the multi-leaf spring. Model 1 uses a linear elasto-plastic formulation and Model 2 uses a nonlinear formulation to emulate the multi-leaf spring. For details of the two models, the reader is referred to reference [20]. The two models are given the same displacement input as was given to the physical spring. Figure 14 shows the qualitative comparison of the two models against the measured data and it seems as if Model 2 is the better model. Table 17 shows the quantitative results for the different validation metrics. All the metrics except the $m\%RE^m$ indicates that Model 2 is the better model. After closer inspection of the measured signal we see that there exists noise around zero on the measurement's reading which is shown in Figure 15.

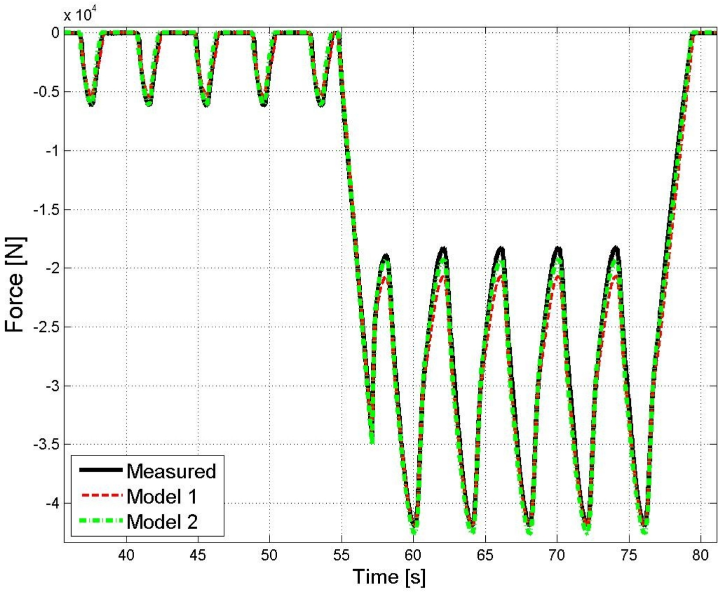


Figure 14. Qualitative comparison of predictions by leaf spring models and measured data

Table 17. Results with noise on measured data around zero

	Model 1				Model 2			
	S&G	**Russell**	$m\%RE^m$	$m\%RE^s$	**S&G**	**Russell**	$m\%RE^m$	$m\%RE^s$
Magnitude	2.0	1.7			1.3	1.1		
Phase	1.4	1.4			0.83	0.83		
Comprehensive	2.4	1.94	60.1 P(38.1)	10 P(21.9)	1.6	1.2	68.7 P(33.1)	10 P(28.9)

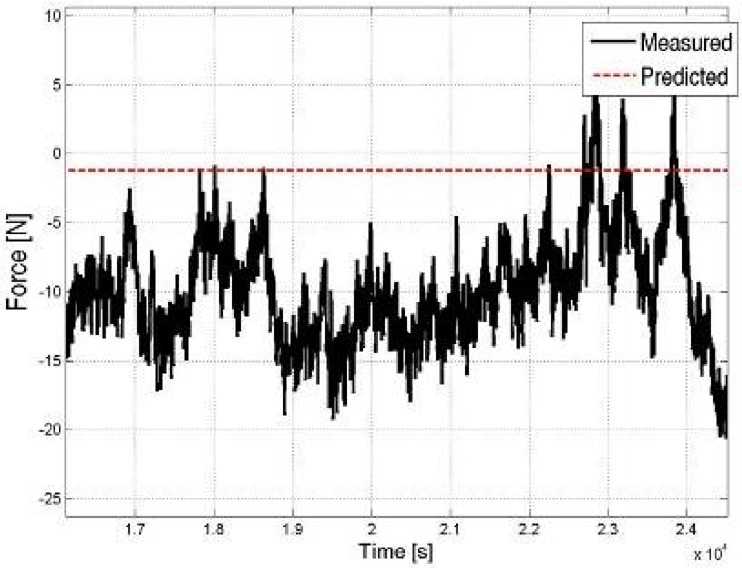


Figure 15. Noise on measurement signal around zero (zoomed in on Figure 14)

Removing the noise on the measurement error around zero by reassigning all measurements lower than 25N to 0N, the results shown in Table 18 are obtained. The results for S&G and Russell stay the same whereas the results from the modified *%RE* (the *$m\%RE^m$* and the *$m\%RE^s$*) changes and show that Model 2 is much better than Model 1. When Figure 14 is viewed it would be expected that the two models should give similar results. After closer inspection, it was found that Model 1 had an error in predicting the zero values correctly (see Figure 16). After the cause for the error in the prediction of Model 1 has been indentified and the model refined the results shown in Table 19 are obtained. S&G and Russell still gives the same results. The results from both the *$m\%RE^m$* and the *$m\%RE^s$* now show that Model 2 is the better model, but that there is not a big difference in agreement between Model 1 and Model 2.

Table 18. Results with noise on measured data around zero removed

	Model 1				**Model 2**			
	S&G	**Russell**	***$m\%RE^m$***	***$m\%RE^s$***	**S&G**	**Russell**	***$m\%RE^m$***	***$m\%RE^s$***
Magnitude	2.0	1.7			1.3	1.1		
Phase	1.4	1.4			0.83	0.83		
Comprehensive	2.4	1.94	64.4 P(46.9)	10 P(35.2)	1.6	1.2	13.5 P(94.5)	10 P(92.8)

Table 19. Results with noise on measured data around zero removed and Model 1 refined

	Model 1				**Model 2**			
	S&G	**Russell**	***$m\%RE^m$***	***$m\%RE^s$***	**S&G**	**Russell**	***$m\%RE^m$***	***$m\%RE^s$***
Magnitude	2.0	1.7			1.3	1.1		
Phase	1.4	1.4			0.83	0.83		
Comprehensive	2.4	1.94	19.3 P(92.3)	10 P(83.7)	1.6	1.2	13.5 P(94.5)	10 P(92.8)

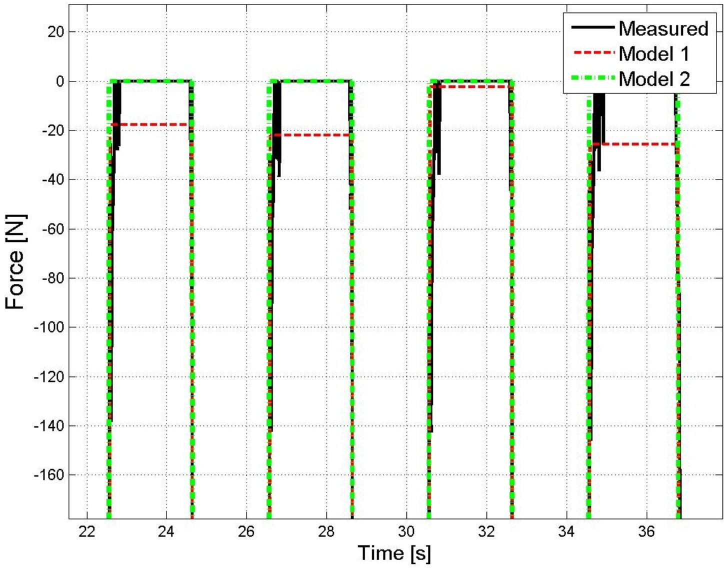


Figure 16. Error in predictions of Model 1

It is interesting to note that S&G, Russell and the *m%RE*^{s} ranked the two models correctly from the start. S&G and Russell indicated from the start that the difference between the models should not be far from each other but the use of the modified *%RE* metric showed that there were large errors between the SRQs and helped with identifying the error in Model 1. Both formulations of the modified *%RE* continually gave an accurate representation of the accuracy between the models. This example also shows that the modified *%RE* and especially the *m%RE*^{s} can easily be used to compare the validation measure's results to predefined accuracy requirements. An accuracy requirement of 10% or closer could have been defined and Model 2 having 92.8% of the model's predictions below 10% may indeed satisfy the requirements.

4. Conclusion

An overview of the V&V process was presented and briefly discussed. From literature two validation metrics were identified and compared to the proposed validation metric that is based on relative error. The challenges associated with using the *%RE* as a validation metric was discussed and techniques were presented to circumvent these challenges. From the comparisons of the three validation metrics it was found that the validation metrics give similar results when ranking models and in selecting the best model. It was shown that the comprehensive error of the modified *%RE* validation metric is the most reliable in providing a representative measure of the agreement/disagreement between two SRQs. Furthermore, when used in combination with the magnitude and phase errors of other measures such as S&G it gives information that enables the ranking of models, selecting the best model, fault finding and refinement, and ultimately validation of the model.

The modified *%RE* validation metric gives a comprehensive error but cannot distinguish between an error in phase or an error in magnitude. It is suggested that when comparing analytical functions that the modified *%RE* be used together with the magnitude and phase error measures such as presented by S&G. When SRQs are compared that are obtained from a simulation model and a physical system, the modified *%RE* should rather be used with qualitative comparison methods as this might give the analyst a holistic view and make the identification of the possible causes for the deviation more likely.

It was shown that the modified *%RE* validation metric gives a reliable and easily interpretable metric that will enable the quantification of the agreement of the simulation model's predictions against the measurements on the physical system and the comparison to the accuracy requirements. The modified %RE can also be used on analytical functions and on deterministic SRQs with an independent variable other than time.

Acknowledgements

This research has been made possible through the financial support from Afrit and from funding received from the Technology and Human Resources for Industry Program (THRIP) of the Department of Trade and Industry of South Africa.

References

[1] **Oberkampf, W.L. and Barone, M.F. (2006),** "Measures of agreement between computation and experiment: Validation metrics", *Journal of Computational Physics*, 217, pp.5-36

[2] **Babuska, I. and Oden, J.T. (2004)**, "Verification and validation in computational engineering and science: basic concepts.", *Comput. Methods Appl. Mech. Engrg.* 193:4057-4066

[3] **ASME standards (2006)**, "ASME V&V 10-2006 Guide for Verification and Validation in Computational Solid Mechanics", The American Society of Mechanical Engineers, New York.

[4] **Roy, C.J. and Oberkampf, W.L. (2011)**, "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing", *Computer methods in applied mechanics and engineering*, 200:2131-2144

[5] **Figliola, R.S. and Beasley, D.E. (2006)**, "Theory and Design for Mechanical Measurements", Fourth Edition, Wiley and Sons.

[6] **Sarin, H., Kokkolaras, M., Hulbert, G., Papalambros, P., Barbat, S. and Yang, R.-J. (2010)**, "Comparing Time Histories for Validation of Simulation Models: Error Measures and Metrics", *J. Dynamic Systems, Measurement, and Control*, Vol. 132

[7] **Heydinger G.J., Garrot W.R., Chrstos J.P. and Guenther D.A. (1990),** "A Methodology for Validating Vehicle Dynamics Simulations", *SAE Technical Paper 900128*

[8] **Ferry, W.B., Frise, P.R., Andrews, G.T. and Malik, M.A. (2002)**, "Combining virtual simulation and physical vehicle test data to optimize durability testing", *Fatigue & Fracture of Engineering Materials and Structures*, Vol. 25, Issue 12, pp.1127-1134

[9] **Edara, R. Shih, S., Tamini, N., Palmer, T. and Tang, A. (2005)**, "Heavy Vehicle Suspension Frame Durability Analysis Using Virtual Proving Ground", *SAE Transactions*, 2005-01-3609

[10] **Cosme, C., Ghasemi, A. and Gandevia, J. (1999)**, "Application of Computer Aided Engineering in the Design of Heavy-Duty Truck Frames", *SAE Transactions*, 1999-01-3760

[11] **Bernard, J.E. and Clover, C.L. (1994)**, "Validation of Computer Simulations of Vehicle Dynamics". *SAE Transactions*, 940231

[12] Kat, C. and Els, P.S. (2011a), “Importance of correct validation of simulation models”, *Proc. ASME 2011 International Design Engineering Technical Conferences & Computers and Information in Engineering Conferences*, Aug. 29-31, Washington, DC, USA

[13] Oberkampf, W.L. and Trucano, T.G. (2002), “Verification and validation in computational fluid dynamics”, *Progress in Aerospace Sciences*, Vol. 38 pp. 209-272

[14] Geers, T.L. (1984), “An objective error measure for the comparison of calculated and measured transient response histories”, *The shock and vibration bulletin*, 54:99-107

[15] Russell, D.M. (1997a), “Error Measures for Comparing Transient Data: Part I: Development of a Comprehensive Error Measure”, *Proceedings of the 68th Shock and Vibration Symposium*, *Hunt Valley, MD,* pp. 175 – 184

[16] Sprague, M.A. and Geers, T.L. (2003), “Spectral elements and field separation for an acoustic fluid subject to cavitation”, *J. Computational Physics,* 184:149-162

[17] Schwer, L.E. (2007), “Validation metrics for response histories: perspectives and case studies”, *Engineering with Computers,* 23:295-309

[18] Russell, D.M. (1997b), “Error Measures for Comparing Transient Data: Part II: Error Measures Case Study”, *68th Shock and Vibration Symposium*, *Hunt Valley, MD,* pp. 185 – 198

[19] Sprague, M.A. and Geers, T.L. (2006), “A spectral-element/finite-element analysis of a ship-like structure subjected to an underwater explosion”, *Computer methods in applied mechanics and engineering*, 195:2149-2167

[20] Kat, C. and Els, P.S. (2011b), “Elasto-plastic leaf spring model”, *International Journal of Engineering Systems Modelling and Simulation*, Vol. 3, Nos. 3/4, pp. 126 - 139