Applicability of the Difference between Population Statistics as an Acceptance Criteria Metric for Seakeeping Validation

Aurore V. Zuzick*, Arthur M. Reed, William F. Belknap, and Bradley L. Campbell
David Taylor Model Basin (NSWCCD), USA

Abstract: The difference between population statistics is proposed as a primary acceptance criteria metric for the direct quantitative validation of ship simulation tools in support of accreditation for uses related to ship motions in irregular seas. The discussion is applicable to comparisons of statistical quantities calculated from ship motion time histories generated by simulations and benchmark data such as scale-model test results. The difference between population statistics provides several of the key characteristics desirable in acceptance criteria, including quantifiable measures of accuracy, completeness, and self-consistency. Further, this metric can be applied to a variety of statistical quantities of interest, provides an opportunity to extend parameter-level comparison results to a broader measure of overall accuracy, and allows for straightforward application of engineering margins traceable to simulation tool performance requirements. Use of the difference between values (often called the error) as the foundation of comparison metrics is not a new concept in the field of validation, but its use is not frequently associated with acceptance criteria for simulations of stochastic processes. Much work has been completed to characterize the total uncertainties from various sources associated with each data set in a comparison of this type. Extension of that body of work to the uncertainty associated with the comparison itself provides a robust measure of parameter accuracy and a flexible and adaptable acceptance criteria foundation.

Key words: Validation; Simulation; Seakeeping

1. Introduction

Validation and accreditation of simulation tools for modeling ship motions in irregular seas is a challenging endeavor for which no single straightforward methodology has been proven universally applicable. Rather, several key comparison approaches are typically employed through a multifaceted comparison of simulation results to benchmark data.

Belknap, *et al.* (2011) describes two categories of validation techniques: qualitative and quantitative. Qualitative validation methods examine trends and expected behaviors to provide confidence that the underlying assumptions within the code lead to reasonable, physical results. Quantitative validation methods establish the simulation tool's ability to meet specific performance requirements associated with the Specific Intended Uses (SIUs) for which accreditation is sought.

Acceptance criteria are typically implemented as part of quantitative validation to provide non-subjective assessment of desired simulation tool performance. Of course, subjectivity is inherently present in the development of acceptance criteria, themselves, but ideally the quantitative criteria are directly traceable to performance capabilities defined by the simulation tool user for each SIU.

Establishment of appropriate acceptance criteria for SIUs related to ship motions in irregular waves is not straightforward. Smith (2012) describes some of the complexities of this task, including the development of acceptance criteria structure. Validation methods should extend single parameter comparisons to overall assessment of the code through examination of multiple degrees of freedom and conditions. Smith (2012) proposes a three-tier structure of parameter

* **Corresponding author:** Aurore V. Zuzick, research fields: surface ship hydromechanics, computational hydrodynamics. E-mail: aurore.zuzick@navy.mil

criteria, condition criteria, and set criteria. Parameter criteria are applied to a single degree of freedom and a single condition; results from the parameter criteria form the inputs to condition criteria, and so forth.

The challenge of establishing appropriate parameter criteria is complicated by uncertainty associated with the data values compared. Statistical values calculated from ship motion time histories are not known exactly. Uncertainty associated with calculated time history statistics comes from several sources including stochastic process uncertainty, instrumentation uncertainty (if model results are used as benchmark data), and uncertainty in simulation results due to uncertainty in simulation input parameters (input sensitivity). Known uncertainties should be quantified and incorporated into acceptance criteria for robust comparison assessment.

Characteristics of good acceptance criteria for validation and accreditation of computer models have been identified by previous efforts within and beyond the field of ship dynamic stability. Oberkampf & Barone (2006) outline features of good validation metrics within their discussion of criteria development. Smith (2012) discusses the importance of many of these characteristics to the development of acceptance criteria for irregular seas ship motion prediction validation. Perhaps most significant among these characteristics is the ability to provide quantifiable measures accuracy through comparisons. Also notable are the importance of self-consistency (non-contradictory assessment outcomes) and completeness (consideration given for all relevant sources of uncertainty associated with validation data sets).

The quantitative acceptance criteria metric proposed in this paper is intended to be applied on a parameter level for direct validation through comparison with benchmark (model-test) data.

2. Definitions

Ship motion response in irregular seas can be characterized in many ways, the most common of which is the standard deviation (or square root of variance) of a particular ship motion parameter time history. The discussion below will be presented in the context of standard deviation comparisons, but the concepts are applicable for other statistical quantities which may be applicable to the SIUs (*e.g.* exceedence rate, percentile of peak amplitudes). Belenky, *et al.* (2013) provides discussion on the calculation of mean ensemble variance values from typical irregular seas model-test time histories.

2.1 Difference Between Data Points

The foundation of the proposed metric for this application is the difference between statistical quantities calculated from simulation and model test data sets.

$$\Delta = \sigma_{simulation} - \sigma_{benchmark} \quad (1)$$

A positive value is associated with simulation over-prediction, and a negative value denotes simulation under-prediction. This concept is certainly not new to the field of validation, but its use is often associated with largely deterministic processes. Both Oberkampf & Barone (2006) and ASME (2009) refer to this quantity as the error between model and experimental results, noting that the experimental results are only an estimated measure of the "true" parameter value.

2.2 Confidence Intervals

The confidence interval is an conventional mathematical quantity which NIST (2014) defines as a range of values which is likely to contain the population parameter of interest. Its purpose is to account for the possible difference between a discreet value derived from limited population samples from the underlying population value. The level of confidence associated with the interval defines its length and corresponds to the probability that the sampled value and intervals encompass the true population value.

When defined relative to a mean value and assuming a large sample size, the confidence interval is defined as

$$CI_{\mu} = z_{1-\alpha/2}\frac{s}{\sqrt{N}}$$

where s is the sample standard deviation, N is the number of samples, α is the desired significance level (corresponds to confidence level), and z is the two-tailed Gaussian distribution factor with significance level, α. The upper and lower bounds of the confidence intervals applied to the sample mean are defined as

$$\mu_{sample} \pm CI_{\mu} \tag{2}$$

where μ_{sample} is the sample mean. Belenky, *et al.* (2013) provides an extension of this theory to calculate the confidence interval on the ensemble mean standard deviation value from a set of time histories of ship motions for one parameter and one condition.

When comparing samples from two populations, the confidence interval on the difference between mean values is of interest. The confidence interval on the difference between mean values is defined as

$$CI_{\Delta\mu} = z_{1-\alpha/2}\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \tag{3}$$

where the subscripts 1 and 2 distinguish between data sets.

2.3 Combined Uncertainty

Additional sources of uncertainty may be applicable to the sample value, including uncertainty due to instrumentation limitations and uncertainty due to variability of the conditions under which the data was generated. Combined uncertainty intervals constructed from multiple sources of uncertainty are typically the root sum of squared intervals calculated separately for each source. While confidence intervals (based only on sampling characteristics) are symmetric, combined uncertainty intervals may be asymmetric.

To compare two data sets with equal number of samples (*i.e.* $N_1 = N_2$) and symmetric confidence intervals, (3) can be rearranged and described in terms of the confidence intervals associated with each data set value as

$$CI_{\Delta\mu} = z_{1-\alpha/2}\sqrt{\left(\frac{CI_{\mu\_1}}{z_{1-\alpha^*/2}}\right)^2 + \left(\frac{CI_{\mu\_2}}{z_{1-\alpha^*/2}}\right)^2}$$

where α^* refers to the level of significance associated with the sample intervals and α refers to the level of significance associated with the uncertainty in the difference.

Equation (3) lends itself to a definition of the combined uncertainty (*e.g.* statistical, instrument, etc.) in the difference between samples which is agnostic to the methods used to define the combined uncertainty intervals associated with each data set, assuming the uncertainties of each set are Gaussian distributed. Further, (3) can be adapted to account for asymmetric intervals by distinguishing between the upper and lower intervals associated with each set.

For validation purposes, consider the definition of the difference provided in (1) to compare two ensemble mean standard deviation quantities. Given combined uncertainty intervals associated with each data set of significance level α^*, the upper and lower combined uncertainty intervals on the difference can be calculated as

$$CI_{\Delta}^{+} = z_{1-\alpha/2}\sqrt{\left(\frac{CI_{bench}^{-}}{z_{1-\alpha^*/2}}\right)^2 + \left(\frac{CI_{sim}^{+}}{z_{1-\alpha^*/2}}\right)^2}$$

and

$$CI_{\Delta}^{-} = z_{1-\alpha/2}\sqrt{\left(\frac{CI_{bench}^{+}}{z_{1-\alpha^*/2}}\right)^2 + \left(\frac{CI_{sim}^{-}}{z_{1-\alpha^*/2}}\right)^2}$$

where the subscripts "bench" and "sim" refer to the benchmark and simulation data sets, respectively. Fig. 1 below illustrates the relationships between the uncertainty intervals on both data sets and the uncertainty interval on the difference.

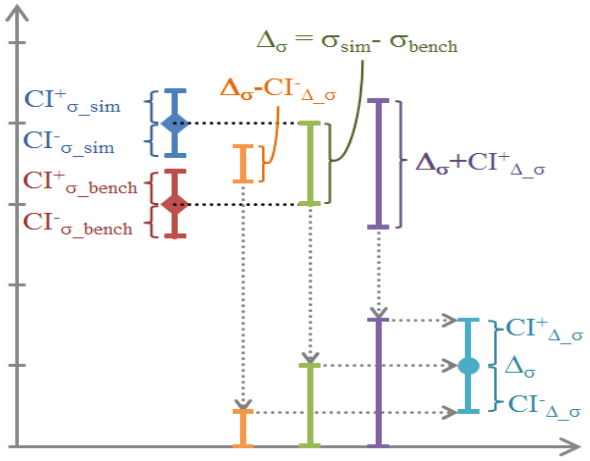

Fig. 1: Uncertainty Intervals On Two Data Sets and On the Difference Between Data Sets

3. Validation Utility

Quantitative validation for stochastic processes is often centered around traditional methods of statistical inference. Hypothesis testing and interval overlap examination address the question: "Could the underlying populations (described by the sample data sets) be the same?" Smith (2011) presents techniques to extend these types of statistical methods to continuous and non-independent samples, such as ship motion time histories. The level of significance associated with these statistical testing methods can satisfy a requirement for quantified accuracy. However, for engineering purposes, statistical similitude does not necessarily constitute a required level of correlation.

Some quantifiable differences between the simulation and the benchmark data may be acceptable for a given SIU. Further, the intended use of a simulation tool may allow for application of a margin to simulation results. If so, the validation effort may be most effective if it provides quantifiable measures of the demonstrated accuracy of the tool in lieu of a binary accreditation outcome (*i.e.* accredited or not accredited).

3.1 Interpretation of Results

The combined uncertainty intervals surrounding a difference between simulation and benchmark statistics enclose the region within which the "true" difference between populations is found. The level of confidence associated with interval calculations corresponds to the probability that the true difference is within the interval limit. For a 90% level of confidence, there is a 90% probability that the difference between the simulation and benchmark results is between the lower and the upper interval
extents.

Positive values denote a simulation value which is greater than the benchmark (over-prediction) while negative values denote under-prediction. A zero-crossing of an interval denotes the possibility that there is no difference between the underlying populations (similar to the objectives achieved through statistical inference tests). It should be noted, however, that the confidence level associated with the interval does not equal the probability that the difference is zero. In fact, there is equal likelihood that the true difference falls anywhere else within the interval extents.

For some purposes, a relative metric may be more suitable for comparison purposes. By dividing the difference (and interval limits) by the benchmark statistic value, a %-difference (and associated uncertainty) is generated. Both the difference and %-difference comparison measures can be used to apply parameter-level acceptance criteria for a single condition or can be used to examine simulation parameter accuracy trends over a range of conditions.

3.2 Parameter Acceptance Criteria

As noted above, when the uncertainty interval on the difference crosses zero, there may be no difference between the two populations. As a potential parameter-level acceptance criteria, a zero-crossing of difference intervals is most analogous to an overlap of

uncertainty intervals associated with two data sets. Note, however, that zero-crossing is a more "strict" measure of similitude than interval overlap. For same level of significance, it is mathematically possible for the intervals to slightly overlap without the corresponding interval on the difference crossing through zero. Comparison A in Fig. 2 below illustrates a case of the uncertainty interval on percent difference crossing through zero.

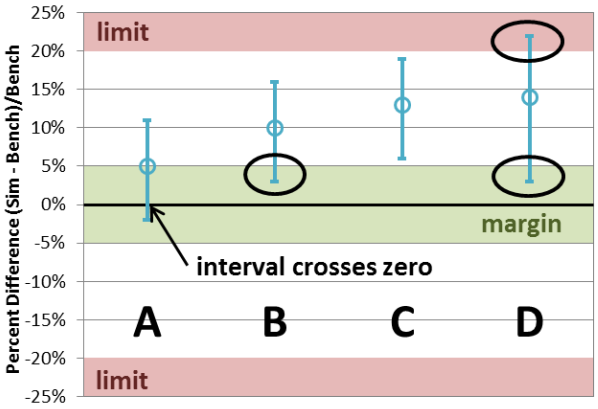

Fig. 2: Options for parameter-level criteria application based on the difference between population statistics

While zero difference is generally a goal for most validation efforts, a region around zero can be defined to capture a broader and more requirement-specific measure of successful agreement. A criterion which would require some part of the difference interval to fall within this region (*i.e.* margin) can be an effective way to link specific requirements to validation comparisons. For example, the acceptance criteria may state that simulation roll standard deviation values must potentially agree with benchmark values by 5%. For a specified significance level, any interval falling at least partially within –5% and +5% could satisfy the criterion. Comparisons A, B, and D in Fig. 2 illustrate cases in which the uncertainty intervals on the percent difference extend into a margin region defined about zero.

In additional to providing a test for statistical equivalence or good correlation, the difference can be used to investigate other aspects of comparison results. Based on the requirements associated with the SIUs, the application of a criterion to bound the differences (*i.e.* limit) may be a more applicable approach. This may be the case if the SIU is related to safety and potential differences are more important than potential similarities. For example, acceptance criteria may require that simulation results not differ from benchmark data by more than 20%. For a specified significance level, any interval falling entirely within the limits of –20% and +20% would satisfy the requirement. Further, if conservatism is a goal, the bounds on the difference could be asymmetric, such as requiring all parts of the interval to fall between –10% and +20%. Cases A, B, and C in Fig. 2 satisfy this type of limit criterion, while the uncertainty associated with case D suggests the differences could be greater than the limiting value.

If validation requirements are well-defined, a combination of limits and margins can be employed to form a multi-faceted parameter-level criterion. For example, the requirements associated with the SIUs may identify good correlation as standard deviation values within ±5% of benchmark results and unacceptable differences greater than ±20%. Cases A and B in Fig. 2 would satisfy both the margin and limit criteria. Case C passes the limit criterion but fails to demonstrate sufficient performance by not extending into the margin region.

Case D demonstrates sufficient correlation (interval extends into the margin region) while also suggesting the possibility of excessively large differences (interval extends into limit region). In this case, the uncertainty in the comparison is too large to adequately determine a successful comparison outcome. For validation purposes, it is important to distinguish between the outcomes of cases C and D. While case C demonstrates a lack of correlation, case D provides evidence of good correlation, but the comparison is hindered by the large uncertainty in the validation data sets.

3.3 Broad View of Simulation Accuracy

The acceptance criteria methodologies based on the differences between benchmark and simulation statistics described in the previous section are most useful as part of a multi-level acceptance criteria structure. Typically, for a given condition, several key parameters (*e.g.* roll, pitch, etc.) must pass parameter criteria in order to consider the overall comparison of the condition a success. The purpose of this requirement is to provide evidence that the overall physics underlying a given condition are adequately captured by the simulation tool. While this methodology is sound, the information provided as a result of multi-level criteria application is limited in value. A condition, or some percentage of conditions, is determined to pass or fail the criteria. The question answered through accreditation is, "Is the tool accurate enough to be used for the SIUs?". The question which cannot be answered by this approach is, "How accurate is the simulation tool?".

A particularly useful attribute of the difference between statistics is its ability convey information about a simulation's accuracy for a given parameter across a range of conditions. The following section provides an example of this utility using notional comparison data.

4. Sample Data

Fig. 3 below presents notional data similar to results which may be used for simulation tool validation. The values shown have been generated using typical ship response in large seas, but are not attributable to any particular ship. Standard deviation values are plotted with 90% confidence intervals for ten different environmental conditions. Benchmark results are shown in red and simulation results are shown in blue. The data presented are assumed to share the same operational condition (*e.g.* head seas, 10kts full-scale ordered speed) with varying wave characteristics.

Application of an interval overlap criterion would result in the passing of all conditions except for B and D. However, the differences between standard deviation values in those cases may still be small enough to satisfy the requirements of the SIUs. Fig. 4 shows the difference (and 95% confidence interval on the difference) for the same ten cases.

Examination of the comparison difference results provides additional quantifiable comparison results. For example, there is at least a 95% probability that the difference in all conditions is less than +1.5 deg (over-prediction) and –1.0 deg (under-prediction). Also, all ten comparisons show evidence of possible correlation within ±0.5 deg. Note that cases B and D, whose individual data set intervals do not overlap, are the only conditions whose interval on the difference does not pass through zero.

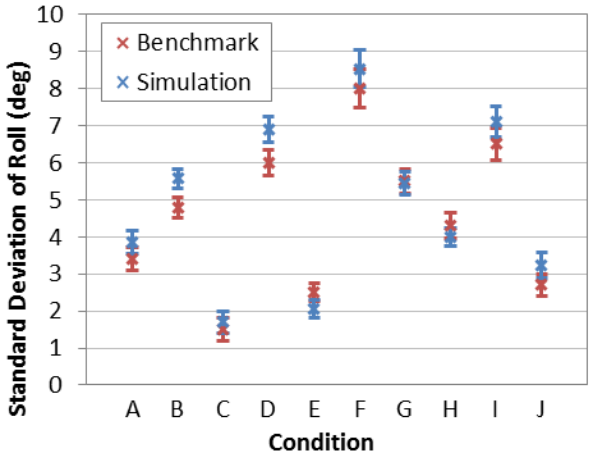

Fig. 3: Notional Validation Data Sets for Roll Standard Deviation of Ship In Heavy Seas

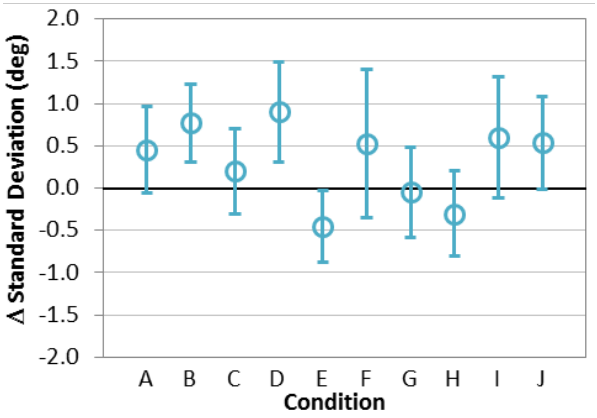

Fig. 4: Difference in Roll Standard Deviation Values (Notional Validation Data)

Fig. 4 provides the viewer the opportunity to focus on the quantified accuracy of each comparison. However, presented this way, the context of each comparison is not apparent. Specifically, Fig. 4 does not allow the viewer to draw inferences about the simulation accuracy as it relates to ship response. By plotting the difference values for all conditions (A–J) as a function of the benchmark standard deviation value, this context is returned to the comparison. Fig. 5 illustrates this approach.

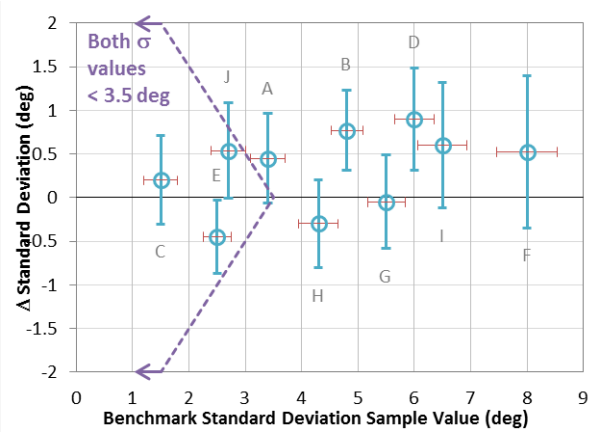

Fig. 5: Difference in Roll Standard Deviation Values as a Function of Benchmark Value (Notional Data)

Horizontal red error bars are added to the difference values plotted in Fig. 5 to indicate the 90% uncertainty associated with the benchmark standard deviation value (*x*-axis). The overall quantified accuracy is unchanged from that described in discussion of Fig. 4, but additional insight is provided by the plotting technique of Fig. 5. For example, the simulation tends to over-predict when the ship is most excited (*i.e.* differences are more positive than negative at higher *x*-values).

The ability to distinguish between conditions as a function of expected ship motion response may be important if, for example, the SIU is directly tied to safety. In this case, simulation accuracy may be most important for conditions in which the roll standard deviation values are large. As such, it may be determined that conditions with "small" expected responses are not desired for validation. The purple dashed line in Fig. 5 identifies the threshold between values (both simulation and benchmark) which are less than and greater than 3.5 deg. This line is analogous to a horizontal line on Fig. 3 at y = 3.5 deg. Using this metric, conditions C and E are denoted as small response conditions, while conditions J and A may or may not be considered small motions.

The comparisons presented in Fig. 4 and Fig. 5 can be modified to reflect the percent difference between the simulation and benchmark results, as shown in Fig. 6 and Fig. 7. The purple dashed line in Fig. 7 again identifies the threshold between values less than and greater than 3.5 deg.

As indicated in Fig. 6, the simulation correlates with benchmark data to within ±30% for most cases, but conditions C, E, and J display larger percent differences. Fig. 7 shows that these conditions coincide with the smallest ship motion response cases. If these cases are not of interest for accreditation, the simulation can be said to agree with benchmark results within –20% and +30%.

5. Conclusions

The difference between population statistics is proposed as a primary acceptance criteria metric for the direct quantitative validation of ship simulation tools in support of accreditation for uses related to ship motions in irregular seas. This metric can be used to achieve traditional goals of this type of validation, including investigation of statistical similitude. It provides a quantifiable measure of comparison accuracy, which incorporates a level of significance associated with the uncertainty in the data as well as a quantified measure of agreement between benchmark and simulation statistical results. This metric can be used to generate acceptance criteria linked to the requirements of the SIUs by incorporating margins and limits on simulation accuracy. Finally, this metric can be used for parameter level criteria application (*i.e.* resulting in pass/fail conclusions) as well as across multiple conditions simultaneously to provide a broad view of simulation accuracy.

Use of the difference between population statistics (including uncertainty) as an acceptance criteria metric builds upon established validation techniques typically reserved for deterministic processes while also utilizing the body of work associated with the quantification of uncertainty of ship motion responses in irregular seas.

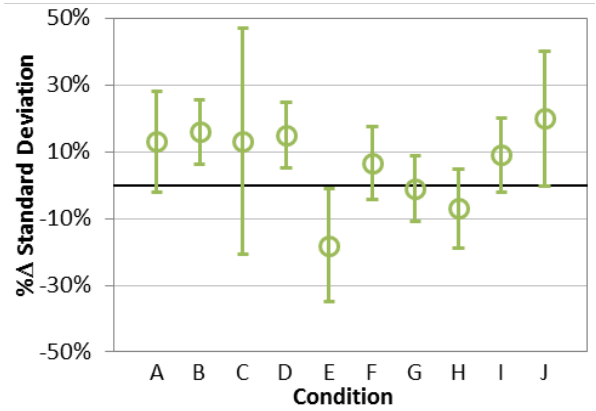

Fig. 6: Percent Difference in Roll Standard Deviation Values (Notional Validation Data)

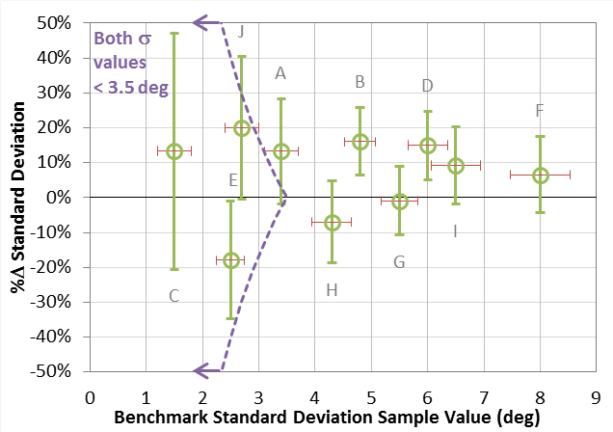

Fig. 7: Percent Difference in Roll Standard Deviation Values as a Function of Benchmark Value (Notional Data)

Acknowledgments

The authors would like to thank Timothy Smith and Vadim Belenky and for their encouragement throughout the development of this paper and their efforts to facilitate attendance of this workshop.

References

ASME (2009) "Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer." American Society of Mechanical Engineers, New York.

Belenky, V., V. Pipiras, C. P. Kent, M. Hughes, B. L. Campbell & T. C. Smith (2013) "On the Statistical Uncertainties of Time-domain-based Assessment of Stability Failures: Confidence Interval for the Mean and Variance of a Time Series." *Proc. 13th Int'l. Ship Stability Workshop*, Brest, France.

Belknap, W. F., T. C. Smith & B. L. Campbell (2011) "Addressing Challenges in the Validation of Dynamic Stability Simulation Tools." *Proc. 12th Int'l. Ship Stability Workshop*, Washington, DC

NIST/SEMATECH (2014) e-*Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 20 June 2014.

Oberkampf, W. L. & M. F. Barone (2006) "Measures of Agreement Between Computation and Experiment: Validation Metrics." *J. Comp. Physics*, 217: 5–36

Smith, T. C. (2011) "Statistical Data Set Comparison for Continuous, Dependent Data," *Proc. 12th Int'l. Ship Stability Workshop*, Washington DC.

Smith, T. C. (2012) "Approaches to Ship Motion Simulation Acceptance Criteria." *Proc. 11th Int'l. Conf. Stability of Ships & Ocean Vehicles*, Athens, Greece.