Validation of Statistical Extrapolation Methods for Large Motion Prediction

Timothy Smith, *NSWCCD (Naval Surface Warfare Center, Carderock Division)*
timothy.c.smith1@navy.mil
Aurore Zuzick, *NSWCCD (Naval Surface Warfare Center, Carderock Division)*
aurore.zuzick@navy.mil

ABSTRACT

Large amplitude ship motion often results from nonlinear aspects of hull geometry and wave excitation. As a result, large amplitude ship motion occurs infrequently, making direct simulation problematic. Statistical extrapolation is a methodology to assess the probability of large amplitude ship motion from shorter duration simulations or model test data that may not contain such large motion. The validation process involves the fitting of an extreme value distribution to data, generation and identification of a "true value," and formulation of a comparison such that a definitive answer can be made.

This paper presents a worked numerical example of statistical extrapolation considering large amplitude roll, pitch, vertical acceleration, and lateral acceleration. Examining different motions addresses different types and levels of nonlinearity. Data are fit with the Generalized Pareto Distribution to formulate a statistical extrapolation. The generation of a "true value" for comparison is discussed. Lastly, the formulation of three-tier acceptance criteria is demonstrated to fully answer the question of statistical extrapolation accuracy. The paper will stress the desired traits and interactions between the main parts of extrapolation, true value, and acceptance criteria.

Keywords: *statistical extrapolation, validation, non-linear motion*

1. INTRODUCTION

The validation of numerical simulations is addressed by various professional societies and governmental bodies for many engineering disciplines. There are established verification and validation outlines, guides, and processes to follow when performing numerical simulation verification and validation (AIAA, 1998; ASME, 2009; ITTC, 2011). These processes and guides are often generalized with details left to the engineers actually performing the verification and validation. Validation at its core consists of a comparison between the simulation and the "true value," and becomes the basis for a validation decision. The true value comes from scale model testing or higher fidelity simulations and implies enough physical understanding to recognize it as the true value.

Focusing on the phenomena of large motion and capsizing, the true value is at once both non-linear and rare. The simulation of these phenomena requires advanced, hydrodynamic blended method prediction tools due to the non-linearity involved (de Kat and Pauling, 1989; Lin and Yue, 1990; Shin *et al*., 2003). Furthermore, the ITTC parametric roll study (Reed, 2011; ITTC Stability in Waves, 2011) showed the uncertainty can be quite large due to practical non-ergodicity. This further increases the difficulty in understanding the result of the validation effort and achieving a definitive result.

This paper continues Smith (2014) and Smith and Campbell (2013) by providing a complete worked example to demonstrate a

multi-tiered validation approach with large amplitude motions and accelerations including statistical extrapolation of rare events.

2. TEST CASE

This test case considers ship roll and pitch motion and lateral and vertical acceleration for a range of relative wave heading in a high sea state. Extrapolations are made based on a sub-set of time history data and compared to a directly counted true value at a motion level not necessarily seen in the data sub-set.

2.1 Extrapolation Method

Following (Smith, 2014), this paper uses the extrapolation technique based on Generalized Pareto Distribution (GPD) as implemented by Campbell, *et al*, (2014). GPD can be used to approximate the tail of any distribution that makes use of a scale and shape parameter to fit the data. There are various implementation details in terms of selecting a threshold and determining the scale and shape parameter.

The confidence intervals for the extrapolated estimate were calculated using two approaches assuming a normal distribution of the GPD parameters. The first is based on the confidence interval of the GPD parameters and generation of a boundary based on the upper and lower extremes of the possible combinations. The second follows the confidence interval method from Campbell, *et al*, (2014) except the logarithm of the scale parameter was used instead of the scale parameters itself. The use of the logarithm of the scale parameter ensures its positive value. These are referred to as the boundary CI and logarithm CI in this paper.

2.2 True Value

The true value was determined by calculating hundreds of thousands of hours of ship motion simulation using a fully coupled, 3 degree of freedom (DOF) simulation tool based on volume calculation (Weems and Wundrow, 2013, Weems and Belenky, 2015). This model assumes constant radiation and diffraction forces with non-linear hydrostatics on 2D strip hull representation. As such it captures essential hydrostatic non-linearity and maintains very fast computation time. Note that in the case of validation against model test data, the true value is typically characterized by some non-negligible amount of uncertainty related to instrumentation and sampling limitations. By simulating against large amounts of simulated data, this uncertainty can be reduced such that a single true value may be identified.

The appropriateness of the 3DOF simplified simulation tool was checked by comparing various instantaneous roll and pitch parameters to those same parameters as calculated with a higher fidelity, 6 DOF blended simulation tool (Lin and Yue, 1990). The parameters compared dealt with the roll and pitch restoring force such as metacentric height, area under GZ curve, and peak of the GZ curve. The most useful comparison was plotting instantaneous roll angle and GZ value. Due to the difference in degree of freedom, the simulation tools could not be compared time step by time step for the same wave realization. The determination of appropriate and adequate physical representation was based on general agreement between the 3DOF and 6DOF simulations (Weems and Belenky 2015).

The peaks were extracted using an envelope approach (Belenky and Campbell, 2012). This method ensures independent data samples as required to apply GPD. The true value of the exceedance rate is found using a direct counting procedure studied in detail in Belenky and Campbell (2012).

3. VALIDATION APPROACH

This example expands the multi-tier validation approach consisting of a parameter, condition and set criteria to include vertical and lateral acceleration (Smith, 2012). The three tiered structure reflects typical scale model data structures of individual motion channels, a run condition of speed-heading-seaway, and a test consisting of many conditions. Criteria are set to determine an acceptable parameter comparison, and what constitutes an acceptable condition and overall set.

A parameter comparison, Tier I, is the elemental comparison between the simulation and true value. It refers to a single motion or response. Choosing an appropriate parameter-level comparison metric depends largely on the problem under examination. Metric options typically share underlying principles and utilize similar properties of the sample data to draw conclusions. They tend to differ in terms of the specific information they provide about each comparison. Some metrics produce binary outcomes (pass or fail) while others provide quantified measures of correlation. Smith (2012) discusses possible comparisons for motions. Further discussion of comparison methods appears later in this paper.

Tier II is a condition comparison. Typically, a condition is the environment, speed and heading used to define the simulation and the associated motion response. So a set of environmental descriptors (e.g. significant wave height, period, etc.), speed and heading and four motions would be four conditions due to the four motions. Thus, a condition can be defined as a deterministic vector:

$$\vec{S} = \left(H_S, T_m, V_S, \beta, i_{dx}\right) \qquad (1)$$

where H_S is a significant wave height, T_m modal frequency, V_S, forward speed, β-heading, i_{dx} –motion index (say, i_{dx}=4 corresponds to roll). This considers motions or parameters independently and Tier II mirrors Tier I.

Alternatively, motions (parameters) can be considered collectively with all or multiple motions (parameters) being included in the condition definition. Then the number of passing motion (parameter) responses becomes a criterion for a condition passing. This is a more stringent criterion to pass with slightly different bookkeeping. The Tier II criterion defines what constitutes a passing condition in terms of number or combinations of passing Tier I comparisons.

Tier III, the set comparison, defines how many conditions have to pass for the simulation to pass the validation criteria. The condition definition needs to be considered in setting the Tier III validation criteria to avoid an impossible criterion.

There is an inter-relationship between the parameter comparisons, second tier condition definition and third tier acceptance levels. Other parameter comparisons besides confidence interval capture of the true value may be used depending on what is important to the application. For instance, the amount of conservatism or absolute difference may be used as a metric. A change of the parameter comparison could change the condition criteria. The multi-tier validation definition used in this study provides a check on both the extrapolation and the confidence interval formulation as both are included in the parameter comparison.

For this example, the parameter comparison is the comparison of a statistical extrapolation to the true value at a specified critical motion level. The parameter comparison passes the test if the extrapolation confidence interval captures the true value. Multiple extrapolations are made from different data sets all representing the same condition, that is speed-heading-seaway-motion combination. A condition passes if the true value is captured by the confidence interval at a percentage roughly equal to the confidence probability. This is repeated for a number of different conditions.

The extrapolation method is considered valid if a high percentage of conditions pass.

These acceptance criteria assume a valid confidence interval formulation. As two confidence interval calculation methods are assessed, this example also serves as a validation of the confidence interval calculation method.

4. FURTHER PARAMETRIC COMPARISONS

The validation approach described above examines the effectiveness of the method used to calculate the confidence interval on the extrapolated value. If the acceptance criteria are passed, we have demonstrated that the 95% confidence interval for any given population sample set does indeed capture the true population value 95% of the time. Once we have confidence that our methods can accurately calculate the uncertainty associated with an extrapolated value, the extrapolated values and their associated uncertainty can be used to validate the simulation tool's ability to model real ship motions. This section discusses parameter comparisons appropriate for non-rare and rare comparisons between simulation data and model test data as an expansion of Smith (2012, 2014). The comparisons discussed are: confidence interval overlap, hypothesis testing, and quantiles (percentiles).

4.1 Confidence Interval Overlap

Confidence interval overlap is a straightforward comparison metric which provides an unambiguous outcome applicable to both non-rare and rare comparisons. Note that the application of the confidence interval overlap metric described earlier for validation of the confidence interval formulation is distinctly different from its use as validation metric for comparisons between simulation and model test data. When evaluating the confidence interval formulation, a true value is known and there are many sample sets of the population from which to draw conclusions. For model test comparisons, the "true values" from two populations (model and simulation) are being compared, and only one sample set from each population is available. When validating simulation results against model test data, confidence intervals are calculated at a specified confidence probability for both sets of sample data; if the intervals from both sets overlap one another, the comparison is considered successful. However, the existence of overlapping 95% confidence intervals does not necessarily signify a 95% chance that both samples share the same underlying population characteristics. The combined probability that the true population values of both data sets lie within the overlapped interval range is significantly less than 95% and depends on the lengths and relative position of both confidence intervals.

The interval overlap metric also provides no information about the probability of differences between the populations. For example, significant overlap of relatively long intervals (high uncertainty) suggests that both populations lie on the same interval; but if the interval is large, the populations could be very different. Alternatively, if intervals are very small, this metric can reject comparisons when population differences are very (perhaps acceptably) small. A perhaps undesirable characteristic of the interval overlap metric is that comparisons are inherently less likely to pass the criteria as uncertainty is reduced.

The confidence interval on the difference between parameter populations is an extension of the interval overlap method. The level of significance is associated with the comparison (i.e. a form of combined probability), rather than with each individual data set. The extents of the confidence interval on the difference provide information about how similar and how different the populations are likely to be. A 95% confidence interval on difference provides the range of values for which the following is true: there is a 95% probability

that the true difference between the two populations lies within that range. Setting limits on the allowable difference (including uncertainty) forms a pass/fail application of this comparison metric. Unlike the interval overlap method, this criterion does not become more difficult to pass when the data are more well-known (less uncertainty). In addition, the difference (including uncertainty) can be used to quantify the simulation's overall level of accuracy. For example, basic statistics (minimum, mean, etc.) of the observed lower uncertainty limits across Tier I comparisons quantify the amount of under-prediction and provide information on safety margin for simulation results.

Another metric making use of the confidence interval is the maximum conservative distance (MCD) which is the difference between the extrapolation upper confidence level and the true value. The upper confidence level is often the "not to exceed" limit. Maximum conservative distance then provides a direct measurement of accuracy of the important parameter.

4.2 Hypothesis Testing

Hypothesis testing on the difference between population statistics is another way to associate a level of significance with a quantified measure of correlation agreement based on two sets of sample data. Formulation of an appropriate null hypothesis is integral to applying this metric. When attempting to identify evidence of good correlation that cannot initially be assumed to exist, the null hypothesis should be contrary to the outcome desired. Formulated this way, strong evidence must be present in the sample data to reject the null hypothesis (acceptance of desired outcome).

For example, the following null hypothesis for a one-tailed Student's t-test may be well-suited as parameter-level criteria metric: the difference between the population mean significant single amplitude (SSA) values is greater than a specified amount. The level of significance used in the test dictates the probability of wrongly rejecting the null hypothesis. For the given null hypothesis, it is possible to quantify (and set to an acceptably low level) the probability of incorrectly identifying good correlation. The probability of failing to correctly identify good correlation is not associated with a specific probability (often known as the Type II or beta error); while this value is of interest, it is typically of less concern than incorrectly identifying poor correlation as good correlation. By defining both a specific limit on the allowable population difference and a level of significance for the test, the Student's t-test can provide a pass/fail outcome for the comparison.

Alternatively, by specifying the level of significance and solving for the critical value of the limit on the difference allowed to pass the test, a quantified measure of correlation agreement is produced. Similarly, by specifying the limit on the difference and solving for the critical level of significance to pass the test, the probability associated with success of the comparison is produced; see Appendix A. Both the quantified measure of agreement and probability of test success are comparison outcomes which can be used to develop measures of correlation across multiple comparisons. The beta error is not explicitly calculated in this process.

Hypothesis tests rely upon measures of the variance in both populations; for comparison of extrapolated values, the confidence intervals on the extrapolated values can be used to calculate the parameters necessary for hypothesis test calculations.

4.3 Percentiles – Rare Comparisons

To make comparisons farther out on the tail of the motion distribution, a cumulative distribution of the measured peaks is useful. The cumulative distributions or quantiles can

be compared at specific percentiles with uncertainty bands. Uncertainty bands at any percentile can be calculated using the normal approximation to the binomial distribution (Belenky and Campbell, 2012).

Percentiles comparisons across the range of available data (shown as a Quantile-Quantile (Q-Q) plot) provide strong visual indications about model and simulation correlation across the distribution of ship response. However, establishing parameter criteria metrics to quantify correlation (including uncertainty) across a range of percentiles is challenging; applying comparison metrics at one or more discrete percentile of peak values is recommended.

One should be cautious when deciding at which percentile to apply parameter criteria comparison metrics. Data at the highest percentiles are prone to large sampling uncertainties; repetition of an ensemble of runs often leads to very different values of the 99th percentile of roll peaks due to the small number of samples associated with extreme motions. The 90th percentile of ship motion peak responses has been observed to be a stable level at which quasi-rare behavior can be observed without being obscured by very large uncertainty. Note that high percentile of peak values are related to threshold exceedance rate. Both high percentiles of peaks and exceedance rates are useful parameters for which non-rare motion channel criteria may be applied, though percentiles tend to lend themselves more easily to the definition of margins and limits. The authors have not yet attempted to apply this to extrapolations.

5. RESULTS

Ten of thousands of hours simulated ship motions were calculated using the 3DOF simplified simulation tool in large sea states for a range of heading. The seaway was a 9.5 m significant wave height and 15 sec modal period with a Bretschneider spectral shape. The headings ranged from near following (15 deg) to bow seas (135 deg). The headings were 15, 30, 45, 60, 90, and 135 deg. The speed was 12 knots for all cases.

The total exposure time was accumulated by ensembling many half-hour simulations. Each simulation had a unique set of random phases to generate a unique and independent wave realization. Time histories of roll, pitch, vertical acceleration, and lateral acceleration were analyzed to extract peaks. This distribution of peaks is the true value for each heading-motion combination. The length of exposure time varies between headings as a matter of convenience. The difference in exposure time does not affect the validation results beyond potentially limiting the maximum comparison value.

Around 100 extrapolations were made with sub-sets of the total exposure time for each heading-motion combination. The use of multiple extrapolations allows for a direct check on confidence interval formulation and gives understanding of data sub-set dependency. The data sub-set exposure time was either 50 hours or 100 hours depending on the number of peaks extracted, with 50 hours being used for cases with more peaks. This was due to a memory limitation on the analysis software.

The extrapolations were compared to the true value at an evaluation level corresponding to a high motion level in the true data set. The comparison was based on overlap of the 95% confidence interval with the true value. The evaluation level was selected as the highest level in the true data set that had more than 30 data samples. Thirty samples are enough to have meaningful uncertainty. With less than 30 samples, the uncertainty becomes very large and the true value has not stabilized.

Figure 1 shows an example of the roll parameter comparisons for stern seas, 30 deg heading. In this figure, the true value is represented by a solid line (1.1111^{-8}). Each extrapolation confidence interval is represented

by a vertical line and represents a single Tier I comparison. The extrapolation captures the true value if the vertical line crosses the horizontal true value line. The estimate of mean value of crossing rate is denoted by a circle and the most probable crossing rate is denoted by a cross. The confidence intervals are asymmetric relative to the mean or most probable crossing rate. This is a property of GPD and different than the symmetric confidence intervals more commonly seen with the normal distribution.

The entirety of Figure 1 represents a Tier II comparison comprised of 100 Tier I comparisons. The expected passing rate percentage is the same as the confidence interval due to use of confidence interval overlap for the parameter comparison. Due to the finite number of data sets, the passing rate can vary from 90 to 100% and still be acceptable; though the mean rate across all the conditions should be close to the confidence level. Table 1 shows the passing rates for all the conditions for the two different confidence formulations. Conditions that pass are bold; failing conditions are italicized. Smith (2014) indicated both CI approaches were acceptable based on roll and pitch. The addition of lateral and vertical acceleration shows a difference between the two CI approaches. The logarithm CI has many instances where the true value capture rate is less than 90% and fail the Tier I comparison (parameter).

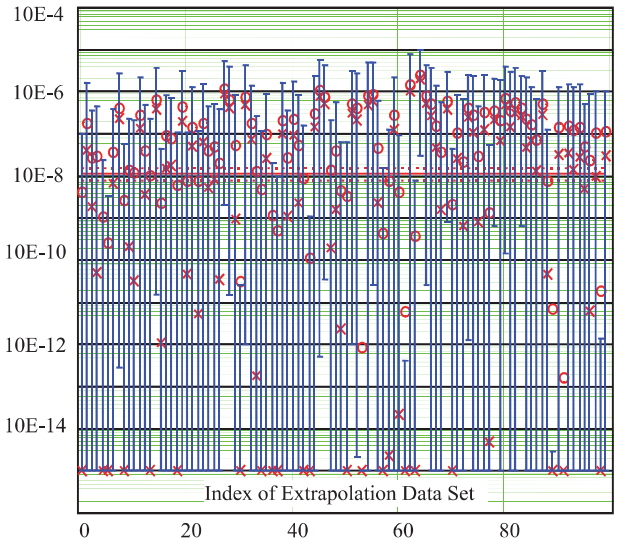

Figure 1 Confidence interval overlap of true value for roll at 30 deg heading at 30 deg comparison level using logarithm CI (91%) (true value 1.111E-08)

As noted in Smith (2014), the GPD can have zero probability which results in asymmetric confidence intervals that have very small lower confidence limits. This can result in automatically capturing the true value if the extrapolation is at all conservative; larger than the true value.

Figure 1 shows roll comparison at 30 deg wave heading for the logarithm boundary CI. The estimates of the mean value are distributed about the true value, while the most probable values are mainly less than the true value. This is a property of the mean value averaging over the entire confidence interval region, whereas the most probable value is selected at a particular point. This difference is discussed in Campbell *et al*. Figure 2 shows the pitch comparison at the same condition where both the mean value estimates and most probable values are greater than the true value. This is indicative of a conservative bias.

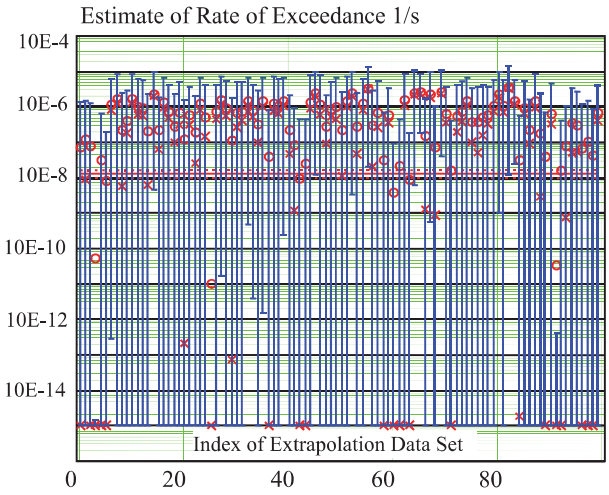

Figure 2 Confidence interval overlap of true value for pitch at 30 deg at 11.5 deg using logarithm CI (97%)

Figures 3 and 4 show the difference in confidence interval method for pitch at 45 deg heading at 11.5 deg comparison level. The

boundary CI method has a higher average upper boundary and a lower boundary always at the lowest value considered (10^{-15}). The logarithm CI method produces smaller confidence intervals.

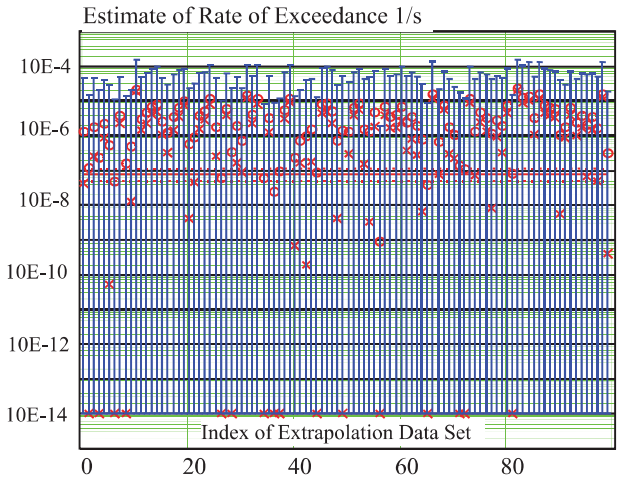

Figure 3 Confidence interval overlap of true value for pitch at 45 deg heading at 11.5 deg comparison level using boundary CI (100%)

Table 1 Confidence interval overlap results

Heading	Motion	Avg GPD Threshold	Comparison Level	# Points at Threshold	Exposure Hours	Boundary CI pass %	Boundary CI MCD	Logarithm CI pass %	Logarithm CI MCD	Shooting Distance
15	roll	6.947	15	53	230000	**96**	119.210	*84*	60.867	1.159
15	pitch	7.359	12	69	230000	**99**	278.512	**94**	166.585	0.653
15	lat accel	No data > 0.2g								
15	vert accel	0.125	0.2	468	230000	**100**	89.878	*86*	49.892	0.607
30	roll	12.877	30	40	100000	**96**	178.616	**91**	101.038	1.330
30	pitch	7.296	11.5	46	100000	**99**	420.751	**97**	244.296	0.576
30	lat accel	0.091	0.2	16	100000	**100**	1061.000	**98**	591.799	1.186
30	vert accel	0.133	0.25	13	100000	**100**	955.188	**96**	484.258	0.877
45	roll	17.094	60	30	230000	**99**	193.110	**94**	112.282	2.510
45	pitch	7.012	11.5	28	230000	**100**	503.342	**98**	269.962	0.640
45	lat accel	0.135	0.3	37	230000	**98**	237.340	**94**	118.744	1.221
45	vert accel	0.158	0.25	518	230000	**99**	74.673	**90**	41.558	0.578
60	roll	18.754	50	49	100000	**100**	276.284	**100**	169.168	1.666
60	pitch	6.257	9.5	71	100000	**100**	330.874	**91**	179.786	0.518
60	lat accel	0.157	0.35	19	100000	**98**	415.630	**97**	213.923	1.227
60	vert accel	0.194	0.3	169	100000	**100**	193.528	*86*	98.556	0.547
90	roll	16.055	32.5	41	230000	**99**	329.773	**99**	192.832	1.024
90	pitch	1.517	2.5	170	230000	**100**	136.716	**94**	68.382	0.648
90	lat accel	0.154	0.25	43	230000	**94**	136.448	*86*	74.551	0.621
90	vert accel	0.270	0.4	287	230000	**100**	98.103	**96**	52.179	0.480
135	roll	12.350	17.5	186	230000	**100**	92.851	**92**	60.694	0.417
135	pitch	4.909	7	172	230000	**100**	134.019	**94**	70.853	0.426
135	lat accel	0.137	0.25	25	230000	**96**	424.676	*88*	232.106	0.827
135	vert accel	0.283	0.4	192	230000	**98**	150.486	**96**	85.572	0.416
Average Value						99		93		

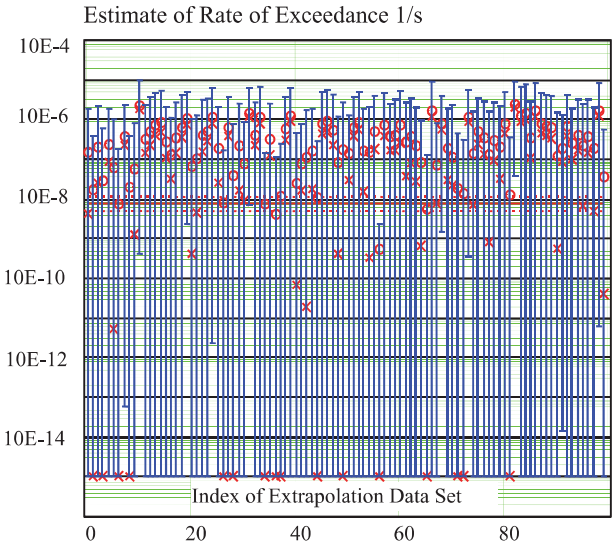

Figure 4 Confidence interval overlap of true value for pitch at 45 deg heading at 11.5 deg comparison level using logarithm CI (98%)

Figure 5 shows asymmetric confidence intervals for the 45 deg heading, lateral acceleration logarithm CI at 0.3 g comparison. At 0.2g comparison level for lateral acceleration at 45 deg heading, Figure 6 shows much smaller confidence intervals and a conservative bias. Although the true value capture rate is very low, the actual difference is small.

Figures 7 and 8 compare the boundary and logarithm CI methods for vertical acceleration at 30 deg heading and 0.2g comparison level. For comparisons at higher levels, both methods have 100% capture rate. The boundary method has larger confidence intervals; both a higher upper bound and many more instances of near zero lower bound. The boundary CI higher upper bound is indicated by the higher MCD; 99.5 vice 54.2. These are similar to the trends seen for pitch in Figures 3 and 4.

Pipiras, *et al.* (2015) compares the boundary and logarithm (lognormal) confidence interval approaches with a preference for lognormal as anti-conservative.

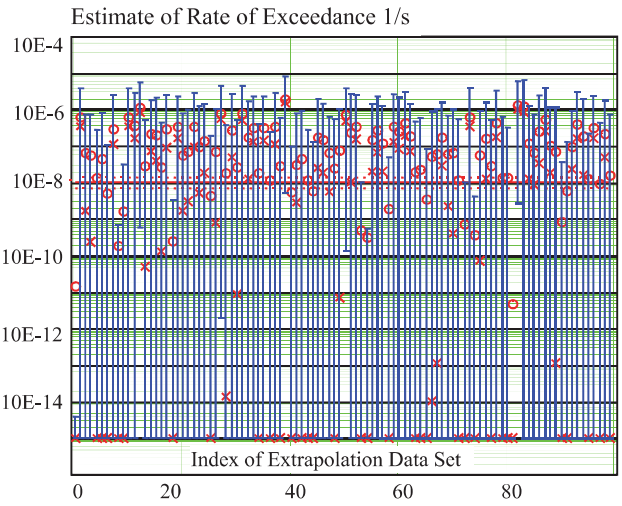

Figure 5 Confidence interval overlap of true value for lateral acceleration at 45 deg heading at 0.3g comparison level using logarithm CI

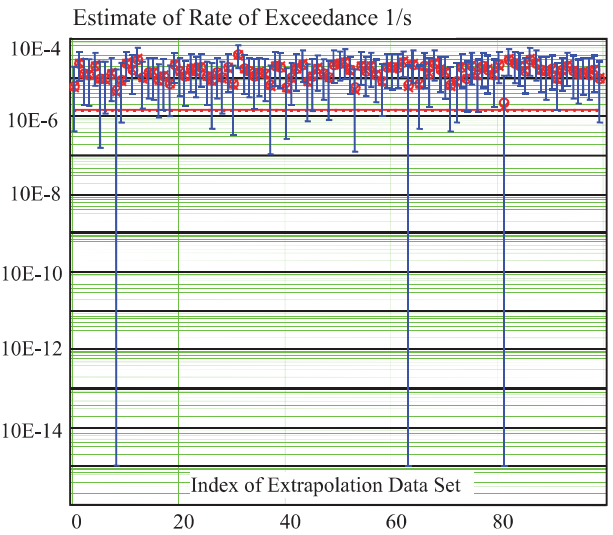

Figure 6 Confidence interval overlap of true value for lateral acceleration at 45 deg heading at 0.2g comparison level using logarithm CI

In terms of acceptance criteria, the boundary CI approach had all the parameter comparisons pass. Therefore, all the Tier II conditions pass and overall acceptance, Tier III, automatically passes if all condition comparisons, Tier II, are acceptable. In this case, even the alternate Tier II definition requiring all the motions to pass for a condition to pass results in overall acceptance.

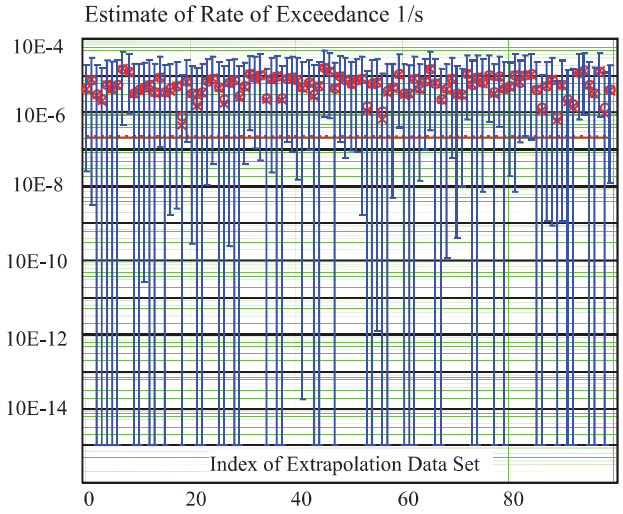

Figure 7 Confidence interval overlap of true value for vertical acceleration at 30 deg heading at 0.2g comparison level using boundary CI

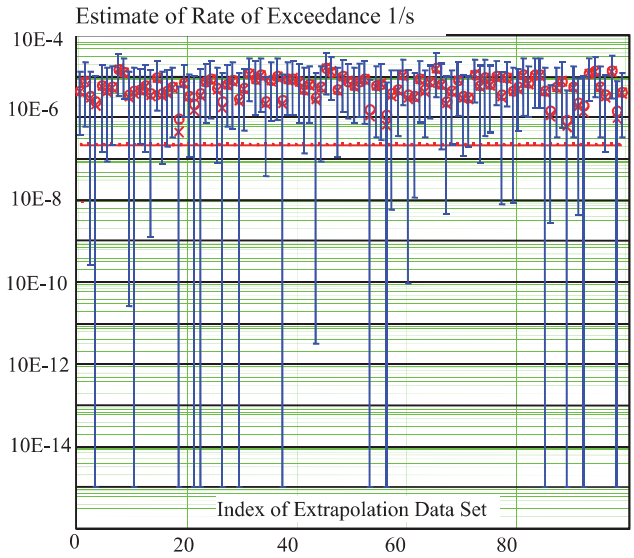

Figure 8 Confidence interval overlap of true value for vertical acceleration at 30 deg heading at 0.2g comparison level using logarithm CI

It is perhaps more instructive to look at the logarithm CI results. Here some of the acceleration parameter comparisons are not acceptable; true value capture rates less than 90%. As a result, some conditions do not pass but the condition pass rate is acceptable for overall acceptance, 18/23 (78%). However, a condition criterion requiring the passing of all motion comparisons is failed for two-thirds the conditions defined by unique speed-heading-seaway combination. There are six such conditions in this example. Overall acceptance fails as well.

However, this does show there is a heading range that is acceptable; aft of beam seas. There could be a limited acceptance with the restricted range of headings. This may also highlight a difference in performance due to behavior of the distribution tail, that is, heavy or light.

As an alternative, the mean conservative distance is a metric which uses the upper confidence limit on an extrapolated sample value. This metric estimates how much conservatism (or over-prediction) is present in the simulation results. For validation of a simulation tool against model data for ship guidance, this quantity may be more important than overall total confidence interval. The difficulty is agreeing to what is an acceptable value. In this example, only one case was over 3 orders of magnitude and almost all were over 2 orders of magnitude. At first glance, this appears to be completely unacceptable as a 100% difference is usually considered unacceptable. However, for exceedance rates of extreme values the uncertainty is inherently high and 1 in a billion is essentially the same as 10 in a billion. The acceptable MCD can be determined by the level at which the over conservatism produces an undesired operational restriction or life time risk level.

The MCD is calculated from the upper confidence limit; the upper confident limit suggests that 95% of the time, the true value is smaller than the limit value. Re-analysis of the existing data would show how successfully both methodologies estimate this quantity. Because this investigation would be focused on only the upper interval limit, the overall methodology validation conclusions may differ from those related to formulation of the entire confidence interval.

This example demonstrates the many factors influencing the comparison: CI method, comparison level, and comparison metric.

6. CONCLUSIONS

This paper demonstrated the applicability of a multi-tier validation approach to the validation of an extrapolated value confidence interval calculation method based on the Generalized Pareto Distribution. The first tier, parameter comparison, was made by comparing the 95% confidence interval from a GPD extrapolation to the true value. This was done 50 to 100 times to determine a passing rate for the Tier II, condition, comparison. Lastly, most of the conditions passed the second tier criterion, which passes the Tier III, or set, comparison. The extrapolation method would be considered validated. A rigorous validation effort would specify passing percentages at Tiers I and II.

Discussion was extended from parameter comparisons of a confidence interval calculation methodology against a known population to comparisons of extrapolated data between simulation and model data. The confidence interval overlap validation approach for these Tier I comparisons requires a validated confidence interval formulation. Other comparison metrics such as maximum conservative distance, mean difference, hypothesis testing, and percentiles were discussed as alternatives to confidence interval overlap. The boundary CI method proved marginally better than the logarithm CI method due to more acceptable conditions and mean value closer to confidence level. Still the boundary CI has many instances of 100 percent capture. This could be indicative of an overly large CI. The logarithm CI method had many failing conditions. Both approaches showed a conservative bias. Extrapolation of acceleration peaks tended to have lower capture rates than for roll or pitch for both CI methods. In terms of the acceptance criteria, the boundary CI method passed when accelerations are included and the logarithm CI method did not. It was shown that the acceptance criteria can indicate the presence of different behavior of distribution tail based on which parameters/motions pass.

The ratio of the GPD threshold and the evaluation level provides a metric for practical use. This ratio was less for accelerations than motions. The conditions with low motions can either have more data added, in the hope of increasing the GPD threshold level, or ignoring the condition as having negligible motions.

7. ACKNOWLEDGEMENT

This work was funded by the Office of Naval Research under Dr. Ki-Han Kim. The authors are grateful to Dr. V. Belenky for providing the data and analysis tools for the numerical example and thorough editing of the manuscript.

8. REFERENCES

AIAA, 1998, Guide for the Verification and Validation of Computational Fluid Dynamics Simulations. American Institute for Aeronautics and Astronautics, Reston, Virginia.

ASME, *2009, Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer*, American Society of Mechanical Engineers, NY.

Belenky, V. and Campbell, B. 2012 “Statistical Extrapolation for Direct Stability Assessment”, Proc. 11th Intl. Conf. on Stability of Ships and Ocean Vehicles STAB 2012, Athens, Greece, pp. 243-256.

Belenky, V., Pipiras, V., Kent, C., Hughes, M., Campbell, B. and Smith, T. 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 Intl.

Ship Stability Workshop, Brest, France, pp. 251-258.

Belenky, V., Pipiras, V. and Weems, K. 2015 "Statistical Uncertainty of Ship Motion Data" Proc. of 12th Intl Conference of the Stability of Ships and Ocean Vehicles STAB 2015, Glasgow, UK

Campbell, B., Belenky, V. and Pipiras, V. 2014, "On the Application of the Generalized Pareto Distribution for Statistical Extrapolation in the Assessment of Dynamic Stability in Irregular Waves," Proc. 14th Intl. Ship Stability Workshop, Kuala Lumpur, Malaysia.

de Kat, J. O. and Paulling, J. R., 1989, "The Simulation of Ship Motions and Capsizing in Severe Seas," Transactions of The Society of Naval Architects and Marine Engineers, vol. 97.

ITTC Loads and Responses Seakeeping, 2011, "Verification and Validation of Linear and Weakly Nonlinear Seakeeping Computer Code," Procedure 7.5-02-07-02.4 Rev 1.

Lin, W.M., & D.K.P. Yue, 1990, "Numerical Solutions for Large Amplitude Ship Motions in the Time-Domain." Proc. 18th Symp. of Naval Hydrodynamics, Ann Arbor, Michigan, pp. 41–66.

Pipiras, V., Glotzer, D., Belenky, V., Campbell, B., Smith, T. 2015 "Confidence Interval for Exceedance Probabilities with Application to Extreme Ship Motions", Journal of Statistical Planning and Inference (Submitted).

Reed, A. M., 2011, "26th ITTC Parametric Roll Benchmark Study," Proc. 11th Intl. Ship Stability Workshop, Washington, DC, USA.

Shin, Y.S., Belenky, V. Lin, W.M. Weems, K.M. and Engle, A.H. 2003, "Nonlinear Time Domain Simulation Technology for Seakeeping and Wave-Load Analysis for Modern Ship Design," SNAME Transactions, Vol. 111.

Smith, T. C., 2012, "Approaches to Ship Motion Simulation Acceptance Criteria" Proc. of 11th Intl Conference of the Stability of Ships and Ocean Vehicles STAB 2012, Athens, Greece pp 101-114.

Smith, T. C., 2014, "Example of Validation of Statistical Extrapolation Example of Validation of Statistical Extrapolation," Proc. of the 14th Intl. Ship Stability Workshop, Kuala Lumpur, Malaysia.

Smith, T. and Campbell, B. 2013, "On the Validation of Statistical Extrapolation for Stability Failure Rate," Proc. 13th Intl. Ship Stability Workshop, Brest, France.

The Specialist Committee on Stability in Waves, 2011, Final Report and Recommendations to the 26th ITTC, Proc. 26th International Towing Tank Conference, Vol II, Rio de Janeiro, Brazil.

Weems, K. and Wundrow, D. 2013, "Hybrid Models for Fast Time-Domain Simulation of Stability Failures in Irregular Waves with Volume-Based Calculations for Froude-Krylov and Hydrostatic Force", Proc. 13th Intl. Ship Stability Workshop, Brest, France.

Weems, K. and Belenky, V. 2015, "Fast Time-Domain Simulation in Irregular Waves With Volume-Based Calculations for Froude-Krylov and Hydrostatic Force" Proc. of 12th Intl Conference of the Stability of Ships and Ocean Vehicles STAB *2015,* Glasgow, UK

9. APPENDIX A: DISCUSSION OF HYPOTHESIS TESTING ERROR

Hypothesis testing involves the formulation of a null hypothesis and has two errors that need to be controlled. The Type I error is associated with rejecting a null hypothesis that should be accepted - false negative. It is characterized with probability α (also known as the level of significance). On the probability

density function (PDF), this is the area in the tails of the distribution. The confidence interval is the area between the tails, 1-α (probability that null hypothesis is accepted correctly).

Type II error is the error associated with accepting a null hypothesis that should be rejected - false positive. It is characterized with probability β. Indeed, 1- β is the probability that the null hypothesis is rejected correctly. Type II error is calculated based on the difference in means between the data sets relative to an allowable or desired difference. The Type II error is overlap area of the original PDF and a PDF shifted by the difference in means.

If the mean shift is large, then it is relatively easy to detect false positives as the means being compared are far apart and β is small. Conversely, if the mean shift is small, then it is difficult to detect false positives and they are more likely to occur. For this case, β would be larger and even greater than 50%. A normal distribution is often used to describe the data probability density function for the Type II comparison.

This example deals with motion data sets with equal number of records and the comparison metric is the standard deviation. The null hypothesis is that the two variances are assumed to come from the same data set or the variances are statistically the same. Probability of Type II error is arbitrarily desired to be less than 50% using 90% confidence probability.

It is possible to check if $\beta < 0.5$ without actually calculating β by taking advantage of the fact that once normalized, the mean difference must be greater than 1.645 (90% quintile of a normal distribution with zero mean and unity variance) based on the shift to achieve less than 50% probability with 90% confidence and two-tailed Normal distribution. This critical difference, δ_{cr}, is applicable to all comparisons using the same confidence and β. Using other values of β results in different mean critical differences.

As the comparison metric is related to standard deviation, the starting point is normalizing the allowable mean difference D_A between the variance estimates of the two data samples:

$$\delta_{cr} = \frac{D_A}{\sqrt{Var(\hat{D})}} \tag{A1}$$

$\hat{D}$ is the difference between variance estimates of the two samples and $Var(\hat{D})$ is the variance of this difference.

$$\hat{D} = \hat{V}_1 - \hat{V}_2 \tag{A2}$$

The samples are independent, thus:

$$Var(\hat{D}) = Var(\hat{V}_1) + Var(\hat{V}_2) \tag{A3}$$

Each sample consists of a number of records obtained from simulation or model experiment. Variance estimates of each sample are the result of averaging the variance estimates of each record. Thus, variance of the variance estimate is expressed as:

$$Var(\hat{V}_1) = \frac{Var(\hat{V}_{R1})}{N_1}; Var(\hat{V}_2) = \frac{Var(\hat{V}_{R2})}{N_2} \tag{A4}$$

$Var(\hat{V}_{R1})$ and $Var(\hat{V}_{R2})$ are the variances of the variance of a single record.

The validity of the code is the hypothesis being tested. Thus, the code is expected to recover the theoretical variance reflected in a model test. It also means that the variance of the variance estimate of a single record is the same between the code and model test:

$$Var(\hat{V}_{R1}) = Var(\hat{V}_{R2}) = Var(\hat{V}_R) \tag{A5}$$

Thus, the theoretical value of the variance of the differences is

$$\begin{aligned} \hat{Var}(D) &= \frac{Var(\hat{V}_{R1})}{N_1} + \frac{Var(\hat{V}_{R2})}{N_2} \\ &= Var(\hat{V}_R)\left(\frac{1}{N_1} + \frac{1}{N_2}\right) \end{aligned} \quad \text{(A6)}$$

The problem is that the theoretical value $Var(\hat{V}_R)$ is not known. Instead, the estimates $\hat{Var}(\hat{V}_{R1})$ and $\hat{Var}(\hat{V}_{R2})$ can be computed using an option for a large number of records, see Belenky, *et al.* (2013, 2015). The best estimate of $Var(\hat{V}_R)$ can be obtained by pooling together the two available estimates:

$$\begin{aligned} &\hat{Var}(\hat{V}_R) \\ &= \frac{(N_1 - 1)\hat{Var}(\hat{V}_{R1}) + (N_2 - 1)\hat{Var}(\hat{V}_{R2})}{N_1 + N_2 - 2} \end{aligned} \quad \text{(A7)}$$

Finally:

$$\begin{aligned} &\hat{Var}(\hat{D}) = \\ &= \frac{(N_1 - 1)\hat{Var}(\hat{V}_{R1}) + (N_2 - 1)\hat{Var}(\hat{V}_{R2})}{N_1 + N_2 - 2} \\ &\times \left(\frac{1}{N_1} + \frac{1}{N_2}\right) \end{aligned} \quad \text{(A8)}$$

Setting δ_{cr} to 1.645 and substituting (A8) into equation (A1)

$$1.645\sqrt{\hat{Var}(\hat{D})} < D_A \quad \text{(A9)}$$

For the case when number of records in each sample is the same: $N_1=N_2=N$:

$$Var(\hat{D}) = \frac{\hat{Var}(\hat{V}_{R1}) + \hat{Var}(\hat{V}_{R2})}{N} \quad \text{(A10)}$$

Equal number of records allows producing a simple final formula to assess Type II error:

$$1.645\sqrt{\frac{\hat{Var}(\hat{V}_{R1}) + \hat{Var}(\hat{V}_{R2})}{N}} < D_A \quad \text{(A11)}$$