Statistical Validation of the Split-Time Method with Volume-Based Numerical Simulation

Kenneth Weems, Vadim Belenky, Bradley Campbell, and Vladas Pipiras

Abstract The application of a statistical validation procedure for estimating the probability of capsizing with the split-time method is described. The method is a numerical-extrapolation scheme incorporating motion perturbation simulations to evaluate a critical roll rate leading to capsizing following an up-crossing event. Fast volume-based numerical simulations create a sample of capsizing events in realistic irregular wave conditions that serves as a "true" value. Subsets of this data are used with the split-time method to estimate the capsizing probability. The split-time estimates are compared to the "true" value to judge the validity of the estimate. A short description of volume-based numerical simulation, review of the essence of the split-time methods, and the statistical validation and performance assessment of the estimation of the probability of capsizing is contained in the chapter.

Keywords Probability of capsizing · Validation · Split-time method

1 Introduction

The application of advanced hydrodynamic codes in the probabilistic assessment of capsizing in irregular waves inevitably leads to the solution of an extrapolation problem. The Monte-Carlo application cannot be applied effectively with advanced numerical methods, as capsizing in realistic sea conditions is too rare to be directly observed within a limited simulation time and the computation cost of such codes

K. Weems · V. Belenky (✉) · B. Campbell
David Taylor Model Basin (NSWCCD), West Bethesda, Maryland, USA
e-mail: vadim.belenky@navy.mil

K. Weems
e-mail: kenneth.weems@navy.mil

B. Campbell
e-mail: bradley.campbell@navy.mil

V. Pipiras
University of North Carolina, Chapel Hill, NC, USA
e-mail: pipiras@email.unc.edu

K. J. Spyrou et al. (eds.), *Contemporary Ideas on Ship Stability*, Fluid Mechanics and Its Applications 134, https://doi.org/10.1007/978-3-031-16329-6_14

prohibits the time and cost of obtaining a sufficiently large sample size. At the same time, the complexity of the problem's physics precludes the application of overly simplified simulations. This conundrum has led to development of extrapolation methods that attempt to characterize the probability of rare events from limited simulation data (for example, [1, 3, 6, 11, 16, 25, 34]). These methods are typically performed with hybrid numerical seakeeping codes such as LAMP [23, 27] and TEMPEST [12], which can practically generate hundreds or even thousands of hours of quantitatively relevant responses in random irregular wave fields.

The validation of the extrapolation methods, however, presents a challenge, as the data set must be extremely large in order to be able to observe the "true" extreme value and yet capture the principal physics of the large amplitude motion in order to be relevant [28]. Moreover, the result of simulation-based extrapolation is a random number that is estimated with uncertainty quantified as a confidence interval. If the "true" value is known, the extrapolation can be regarded as successful if this "true" value falls within the confidence interval. However, due to the very same random nature, a single successful extrapolation result is hardly convincing. How would one know if this was not just a coincidence?

To ensure that the result is representative relative to the environmental conditions, [28] introduced a multi-tier concept of statistical validation. The first tier is elemental: it is successful if the extrapolation result contains a "true" value within its confidence interval (the methodology of obtaining the true value is considered in the next section). The extrapolation procedure is then repeated several times for exactly the same condition but with independent data sets, this is second tier. A successful validation for a given condition produces a certain percentage of successes, referred to as a "passing rate"; [28] proposed 90% as a level for acceptance for 100 extrapolations. The third tier of statistical validation includes consideration of several conditions reflecting the expected operational conditions. How many of those conditions need to be successful for an extrapolation method to pass is not clear. Examples of the application of the procedure for the EPOT (Envelope Peak over Threshold) method [15] are considered in [28] as well as in [16] . This chapter describes the application of this multi-tiered procedure to the evaluation of the probability of capsizing in irregular waves with the split-time method.

2 Estimation of "True Value"

The extrapolation validation procedure reviewed in the sect. 1 requires a priori knowledge of the probability of capsizing. Theoretical solutions for the probability of capsizing are available for piecewise linear models [5], but while these models do describe capsizing qualitatively, i.e. as a transition between two stable equilibria, they are too simplistic to be considered as quantitative ship motion models. In particular, they cannot describe the realistic change of stability in waves as well as the fact that the restoring is inseparable from wave excitation for large-amplitude ship motions.

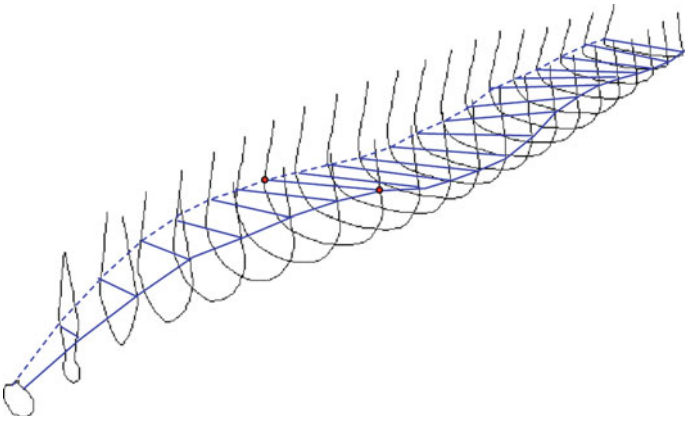


Fig. 1 Station/incident wave intersection for volume based hydrostatic and Froude-Krylov forces for the ONR Tumblehome hull in stern oblique seas [30]

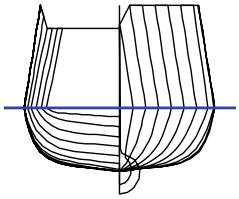
A solution was proposed by [30]. The main idea is to compute the inseparable nonlinear hydrostatic and Froude-Krylov forces from the distribution of the instantaneous submerged volume along the hull, implemented as a sectional-based calculation to preserve the variation of relative motion along the ship's hull, as illustrated in Fig. 1.

In a typical hybrid numerical method, hydrostatic and Froude-Krylov forces are computed by pressure integration over the instantaneous wetted surface:

$$\vec{F}_{FK+HS}(t) = -\rho \iint\limits_{S_B(t)} \left(\frac{\partial \varphi_0(x, y, z, t)}{\partial t} + gz \right) \vec{n} ds \tag{1}$$

where ρ is density, g is gravity acceleration t is time, x, y, z are spatial coordinates, $\overrightarrow{n}$ is a unit vector, normal to a time-variant surface of submerged portion of ship hull $S_B(t)$, and $\varphi_0(x, y, z, t)$ is the incident wave velocity potential, whose time derivative is the pressure distribution of the undisturbed wave field.

While straightforward to evaluate in a standard spectrum-based wave field, formula (1) can be very expensive to calculate for an irregular seaway with many components. If the incident wave pressure can be approximated by constant gradient over each section, Gauss theorem relates the integration of pressures to the instantaneously submerged volume, while the moment can be expressed through the coordinate of the centroid of this volume. This idea has evolved into a very fast algorithm, comparable in performance with calm-water *GZ* for restoring and effective slope for excitation, but with a much more complete model of nonlinear forcing and stability variation in waves. A known limitation of the volume-based technique is related to short wave lengths that are comparable to or shorter than the ship's beam. Derivation of the formulae, a detailed description of the algorithm, and cross-validation with LAMP can be found in [33] and [32]. Additional hydrodynamic



Principal Dimensions and Calculation Conditions

Length BP, m	154
Breadth molded, m	18.8
Draft amidships, m	5.5
GM, m	2.2
Speed, knots	6

Fig. 2 Lines, principal dimensions, and flotation of the ONR tumblehome topside configuration

forces including added mass, damping and maneuvering forces are approximated by ordinary differential equation (ODE) style models.

Weems and Belenky [31] reported that 10 h of data could be generated in 7 s on a single processor of a laptop computer, allowing millions of hours of simulation data to be computed practically on a standard workstation or modest sized cluster.

The subject ship for the validation exercise is the tumblehome configuration from the ONR topside series [14]. The ship lines, principal dimensions, and flotation are in Fig. 2. The statistical validation campaign included four different sea states and various relative wave headings, which are summarized in Table 1.

To avoid a self-repeating effect (e.g. [8]), the simulations for each sea state consisted of a large number of 30-min records. 240 frequency components provided

Table 1 Summary of validation conditions and "true" value estimates

Significant wave height, m	Modal Period, s	Heading, degrees	Total simulation time, hours	Number of capsizes	Estimate of rate 1/s	Low boundary of rate	Upper boundary of rate
8.5	14	45	200,000	8	1.13 E-08	4.24 E-09	1.98 E-08
8.5	14	60	200,000	31	4.38 E-08	2.97 E-08	5.93 E-08
9	14	35	720,000	12	4.71 E-09	2.04 E-09	7.37 E-09
9	14	40	200,000	12	1.70 E-08	8.48 E-09	2.68 E-08
9	14	45	200,000	51	7.20 E-08	5.37 E-08	9.18 E-08
9	14	50	20,000	7	9.89 E-08	2.83 E-08	1.84 E-07
9	14	55	60,000	69	3.25 E-07	2.50 E-07	4.05 E-07
9	14	60	200,000	176	2.49 E-07	2.12 E-07	2.85 E-07
9	14	65	200,000	80	1.13 E-07	8.90 E-08	1.38 E-07
9	14	70	200,000	6	8.48 E-09	2.83 E-09	1.55 E-08
9	15	45	345,000	10	8.19 E-09	3.11 E-09	1.33 E-08
9	15	60	300,000	11	1.04 E-08	4.71 E-09	1.70 E-08
9.5	15	45	1,000,000	157	4.44 E-08	3.74 E-08	5.13 E-08
9.5	15	60	1,000,000	242	6.84 E-08	5.98 E-08	7.70 E-08

a statistically valid model of irregular waves for 30 min duration. The total simulation time and number of observed capsizes are reported in Table 1.

The rate of the capsizing events, $\widehat{\lambda}_T$, based of these observations is estimated as

$$\hat{\lambda}_T = \frac{N_T}{T_T} = \frac{N_T}{N_R T_R - \sum_{i=1}^{N_T}(T_R - t_{Ci})} \tag{2}$$

where N_T is a number of capsize events observed and T_T is the total simulation time, N_R total number of records in the simulation campaign, T_R duration of a record, t_{Ci} time of ith recorded capsizing. The observed number of capsizes N_T is assumed to follow a binomial distribution as capsizings are rare and can be treated as Bernoulli trials. The binomial distribution has two parameters: the total number of trials N_R (which is a total number independent records) and the probability p of an event's occurring during a particular record.

$$p \approx \hat{p} = N_T / N_R \tag{3}$$

Boundaries of the confidence interval for the estimate $\widehat{\lambda}_T$ are computed by a binomial distribution (e.g. [5])

$$\hat{\lambda}_T^{Up,Low} = \frac{1}{T_T} Q_B\left(\frac{1 \pm P_\beta}{2}\right); \tag{4}$$

where Q_B is a quantile (inverse cumulative distribution function) for the binomial distribution with parameters (3) and P_β is a confidence probability. The calculation of this quantile encounters numerical error for the total time of 720,000 h and above (too large to compute a factorial in double precision), so a normal approximation for the estimate distribution was employed for those cases, with the mean value and variance ($\widehat{p}$ is small compared to 1.0) equal to the estimate itself:

$$E\left(\hat{\lambda}_T\right) = \frac{1}{T_T} p N_S \approx \frac{\hat{p} N_S}{T_T} = \hat{\lambda}_T; \tag{5}$$

$$Var\left(\hat{\lambda}_T\right) = \frac{1}{T_T} p N_S (1 - p) \approx \frac{\hat{p} N_S}{T_T}\left(1 - \hat{p}\right) \approx \hat{\lambda}_T \tag{6}$$

The boundaries of the normal-approximation-based confidence interval are:

$$\hat{\lambda}_T^{Up,Low} = \hat{\lambda}_T \pm Q_N\left(\frac{1 + P_\beta}{2}\right)\sqrt{\hat{\lambda}_T} \tag{7}$$

Q_N is the standard normal (with zero mean and unity variance) quantile. The boundaries of the confidence interval for the capsizing rate estimates, computed with a confidence probability of 0.95, are listed in Table 1.

3 Essence of the Split-Time Method

The objective of the split-time method is to provide a way to use an advanced numerical code for estimating the probability of a rare event without actually having to observe it in simulations. Its principal idea is to separate the estimation procedure into an observable or "non-rare" problem and a non-observable or "rare" problem. The "non-rare" problem is an estimation of the crossing rate of an intermediate roll threshold. The threshold roll angle must be low enough to observe a statistically significant number of up-crossing events in, say, 100 h, but high enough so that most of these up-crossings can be treated as independent events.

The "rare" problem is solved for each up-crossing with a motion perturbation scheme in Fig. 3. The roll rate is perturbed at the instant of up-crossing until capsizing is observed. The minimum value of roll rate perturbation leading to capsizing is computed by a metric of the danger of capsizing at the instant of up-crossing

$$y_i = c + \dot{\phi}_{U,i} - \dot{\phi}_{C,i}; c = 1 \text{ rad/s}; i = 1, \ldots, N_U \tag{8}$$

$\dot{\phi}_{Ci}$ is the critical roll rate calculated for the i-th up-crossing defined as the minimum perturbed roll rate leading to capsizing (corresponds to capsizing time history in Fig. 3), $\dot{\phi}_{Ui}$ is the roll rate observed at the i-th up-crossing, and N_U number of observed up-crossings. The constant $c = 1$ rad/s is introduced for convenience in working with the metric.

A rate of capsizing events λ_C (a number of events per unit of time) is expressed as

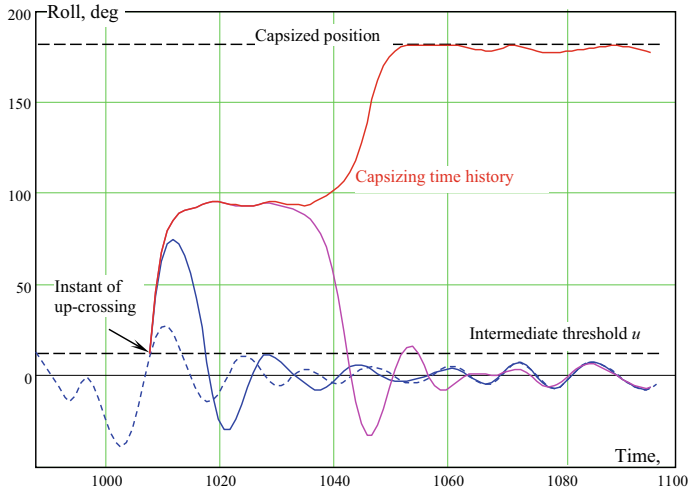


Fig. 3 Illustration of motion perturbations

$$\lambda_C = \lambda_U P(y \geq c | \phi = u \cap \dot{\phi} > 0) \tag{9}$$

where λ_U is a rate of up-crossings of the intermediate threshold u; $P(y \geq c | \phi = u \cap \dot{\phi} > 0)$ is a conditional probability that the capsizing occurs after an up-crossing of the intermediate threshold u (i.e. the capsizing metric y exceeds the constant $c =$ 1 rad/s). Following standard definition (e.g. [18]), an up-crossing event is defined when the roll angle crosses the intermediate threshold $\phi = u$ with a positive roll rate ($\dot{\phi} > 0$).

To find the conditional probability $P(y \geq c | \phi = u \cap \dot{\phi} > 0)$, modeling of the entire distribution of the capsizing metric y is not necessary (as was done by [9] for a time-variant piecewise linear model, Eq. 61 of the cited reference). As the capsizing event is rare, to fit the tail of the distribution of the capsizing metric is sufficient.

Following the second extreme value theorem (e.g. [17]), the tail of any distribution can be approximated with a Generalized Pareto Distribution (GPD), whose probability density function is described as

$$\text{pdf}(y) = \begin{cases} \frac{1}{\sigma} \exp\left(-\frac{y-w}{\sigma}\right) & \text{for } \xi = 0 \\ \frac{1}{\sigma}\left(1 + \xi \frac{y-w}{\sigma}\right)^{-(1+1/\xi)} & \text{for } \xi \neq 0 \text{ and } \xi \frac{y-w}{\sigma} > -1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where ξ is a shape parameter, σ is a scale parameters, and w is a threshold for the capsizing metric (secondary threshold, in angular velocity units), defining the beginning of the distribution tail. A brief overview on extreme value theorems is available from [7].

Fitting the GPD for the tail of a capsizing metric, which is described in detail in [10], consists of the following steps:

- Define a set of "candidate" secondary thresholds
- Estimate shape and scale parameters of GPD for each secondary threshold value
- Search for the secondary threshold
- Evaluate the confidence intervals for estimates of the conditional probability (with the distribution of the extrapolated estimate, derived from a bivariate normal distribution of $\hat{\xi}$ and $ln\hat{\sigma}$) and capsizing rate in Eq. (9).

This fitting procedure is completely data-driven and does not account for any physical considerations that may be available for the problem at hand. Adding physical considerations to a data-driven model may reduce statistical uncertainty (for example, Fig. 8 of [20]).

Nonlinearity of roll rate is usually considered to be weak as it is related to roll damping, which is a weakly nonlinear function of roll rate. As a result, a roll rate process is assumed normal. The capsizing metric contains a value of roll rate at an instant of up-crossing $\dot{\phi}_U$. If both roll and roll rate are normal, the value $\dot{\phi}_U$ follows a Rayleigh distribution (for example, p. 201 of [21]). Similar to the normal distribution, the Rayleigh distribution has an exponential tail (a proof is in Example 1.1.7 in [19]).

For the roll rate at the instance of up-crossing of an actual nonlinear roll process, the exponential tail is a plausible assumption.

The second random component of the capsizing metric (8) is the value of the critical roll rate $\dot{\phi}_C$. . Its randomness reflects the variation of stability in waves. The variability of the roll rate at up-crossing is assumed larger than the variability due to the changing stability in waves. Finally, the assumption of exponential tail is adopted for the entire capsizing metric (8). The exponential tail is a particular case of GPD (Eq. 10) when the shape parameter $\xi = 0$.

Applying the exponential tail, the conditional probability of capsizing after up-crossing of the intermediate threshold u is expressed as

$$P(y \geq c|\phi = u \cap \dot{\phi} > 0) = P(y \geq w)\exp(-(c - w)/\gamma_w) \tag{11}$$

where γ_w is the parameter of the exponential tail and w is the secondary threshold.

Fitting the exponential tail follows the same steps as fitting the GPD. Given a sufficient number of up-crossings of the intermediate threshold u, the parameter for the tail of the distribution can be estimated as

$$\hat{\gamma}_w = \frac{1}{N_w}\sum_{i=1}^{N_w}(y_i - w) \tag{12}$$

where N_w is the number of data points remaining above the secondary threshold w.

The value of the secondary threshold w is found by testing a number of "threshold candidates" and finding one that provides the best fit for the tail. Two methods were selected in Belenky et al. [6]: a prediction error criterion developed by Mager [24] and a goodness-of-fit test, modified for exponential distribution by [29].

The rate of up-crossing of the intermediate threshold u and the probability of exceedance of the secondary threshold w are estimated as

$$\hat{\lambda}_U = \frac{N_U}{T}; \quad \hat{P}(y \geq w) = \frac{N_w}{N_U} \tag{13}$$

where T is the total simulation time. The final expression for the capsizing rate estimate is

$$\hat{\lambda}_C = \frac{N_w}{T}\exp\left(-(c - w)/\hat{\gamma}_w\right) = \hat{\lambda}_w\exp\left(-(c - w)/\hat{\gamma}_w\right) \tag{14}$$

where $\hat{\lambda}_w = N_w/T$ is an estimate of exceedance rate of the secondary threshold w.

The estimate of the capsizing rate Eq. (14) is a function of two other estimates, $\hat{\lambda}_w$ and $\hat{\gamma}_w$, which are random numbers. To evaluate a confidence interval for the capsizing rate estimate, distributions are needed for the estimates $\hat{\lambda}_w$ and $\hat{\gamma}_w$.

Similarly to the capsizings, the exceedance events of the secondary threshold w can be considered rare enough to be treated as Bernoulli trials (independence assumed). The number of events observed within simulation time T then follows binomial

distribution. The binomial distribution has two parameters: number of trials N and probability $\hat{p}$ of an exceedance event at any instant of time so that

$$N = \frac{T}{\Delta t};\ \hat{p} = N_w/N \tag{15}$$

where Δt is the time increment in the simulations. The estimate of the exponential tail parameter (5) is essentially a mean value. Its distribution is approximately normal with the standard deviation

$$\hat{\sigma}_\gamma = \frac{1}{N_w}\sqrt{\widehat{\mathrm{Var}}(y-w)} = \frac{1}{N_w}\sqrt{\frac{1}{N_w}\sum_{i=1}^{N_w}(y_i - w)^2 - \hat{\gamma}_w^2} \tag{16}$$

where $\widehat{\mathrm{Var}}(y-w)$ is an estimate of the variance of the capsizing metric values on the tail.

Boundaries of confidence interval for the estimates $\hat{\lambda}_w$ and $\hat{\gamma}_w$ can be found as follows:

$$\begin{aligned}\hat{\lambda}_w^{Up,Low} &= \frac{1}{T}Q_B\left(\frac{1 \pm P_{\beta 1}}{2}\right);\\ \hat{\gamma}_w^{Up,Low} &= \hat{\gamma}_w \pm Q_N\left(\frac{1+P_{\beta 1}}{2}\right)\hat{\sigma}_\gamma\end{aligned} \tag{17}$$

where Q_B is a quantile (inverse cumulative distribution function) for binomial distribution with parameters (15), Q_N is standard normal (with zero mean and unity variance) quantile, and $P_{\beta 1}$ is confidence probability for the estimates $\hat{\lambda}_w$ and $\hat{\gamma}_w$.

The confidence probability of the estimates $\hat{\lambda}_w$ and $\hat{\gamma}_w$ is related to the confidence probability for the complete capsizing as estimate P_β as

$$P_\beta = \sqrt{P_{\beta 1}} \tag{18}$$

under an assumption of mutual independence of the estimates $\hat{\lambda}_w$ and $\hat{\gamma}_w$. The boundaries of the confidence interval for capsizing rate estimate $\hat{\lambda}_c^{Up,Low}$ can be obtained through the boundaries of the confidence intervals of the estimates $\hat{\lambda}_w^{Up,Low}$ and $\hat{\gamma}_w^{Up,Low}$:

$$\hat{\lambda}_c^{Up,Low} = \hat{\lambda}_w^{Up,Low}\exp\left(-(\mathrm{c}-w)/\hat{\gamma}_w^{Up,Low}\right) \tag{19}$$

Justification for Eq. (17), sometimes referred as "boundary method", can be found in Sect. 4.4 of [13].

4 Results of Statistical Validation

Examples of the tier-two validation are in Fig. 4 (GPD tail fit) and Fig. 5 (exponential tail fit). A seaway derived from a Bretschneider wave spectrum [22] with a significant wave height of 9.0 m and a modal period of 14 s is used in both examples. The tier-two validation data set consists of 50 independent extrapolations. Each extrapolation estimate uses 100 h of volume-based simulations, with no capsizing cases observed during those times. The extrapolation result is presented with a confidence interval for the 0.95 confidence probability. Besides these boundaries, each extrapolation has the most probable value (identified by red x-marks in Fig. 4) and the mean value (indicated as circles in Fig. 4). The calculation of the mean and most probable values is discussed in detail in [10]. The tier-one validation is successful if the confidence interval contains the "true" value. The case in Fig. 4 has 45 individual extrapolations that contain the "true" value within their confidence interval. The tier-two validation is successful when a percentage of the underlining tier-one validation successes ("passing rate") is close to the accepted confidence probability. This number is 0.90 for the case in Fig. 4, which would be considered a successful passing rate by [28].

The vertical scale of Fig. 4 is logarithmic. To indicate zero, a small value of 10^{-15} s^{-1} was applied. A total of 37 values of lower boundary of the confidence interval extends below 10^{-15} s^{-1}, and 11 most probable extrapolated estimates and even 1 value of upper boundary are also very small. The reason is an apparent light tail and associated right bound of the estimated distribution of the metric. It is one of the known issues of practical application of GPD [2, 4, 26].

Figure 5 has results for the exponential tail, inferred from weak nonlinearity of the roll rate and the assumption that the variability of roll rate at up-crossings is larger than the variability of critical roll rate caused by changing stability in waves see the sect. 3 of this chapter. This inference is essentially a choice of statistical model

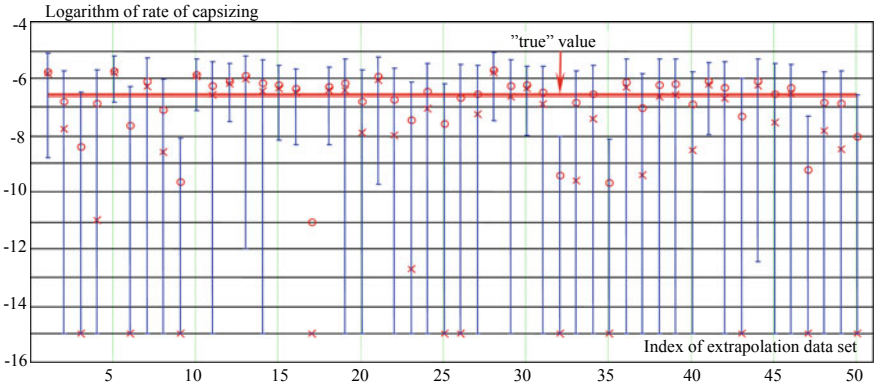


Fig. 4 Example of tier-two validation with the GPD tail fit; significant wave height 9.0, modal period 14 s, heading 60°,, passing rate 0.90; circles indicate mean value of extrapolated estimates, x-marks are most probable extrapolated estimates

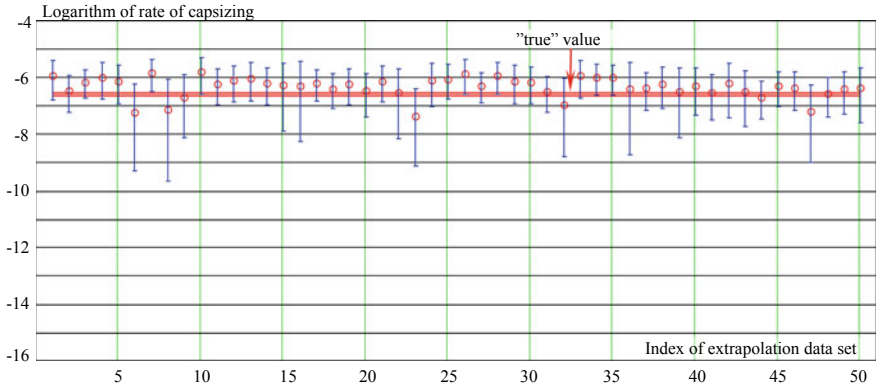


Fig. 5 Example of tier-two validation with the exponential tail fit (prediction error criterion); significant wave height 9.0, modal period 14 s, heading 60°, passing rate 0.98

(exponential tail) based on physical considerations. Including physical information reduces uncertainty, which is reflected in the decreased width of the confidence intervals in Fig. 5 as compared to Fig. 4. Similar results were reported previously by [20]. A mathematical aspect of the decreased uncertainty is a transition from the GPD tail with two estimated parameters to the exponential tail with a single estimated parameter. Comparing Figs. 4 and 5, the difference in the upper boundary is not that dramatic. The practical advantage of this physics-informed approach is improved reliability of prediction.

Besides the passing rate, assessing the performance of the different approaches and elements of an extrapolation is done with two other indicators: "conservative distance" *CD* and "relative bias" *RB*. These are defined as:

$$CD = \log\left(\frac{E\left(\hat{\lambda}_c^{Up}\right)}{\hat{\lambda}_T}\right); \quad RB = \frac{E\left(\hat{\lambda}_c\right) - \hat{\lambda}_T}{\hat{\lambda}_T} \tag{20}$$

where $E\left(\hat{\lambda}_c^{Up}\right)$ is the upper boundary of extrapolated estimates averaged over all the considered extrapolation data sets, $E\left(\hat{\lambda}_c\right)$ is the extrapolated estimate (most probable estimate is used for GPD) averaged over all the considered data sets, and $\hat{\lambda}_T$ is the true value estimated from capsizing observations with Eq. (2).

In a sense, the *CD*-value expresses the practicality of the extrapolation. The upper boundary of the extrapolated estimate is likely to be utilized for the final answer, to keep the whole procedure conservative. If the upper boundary is too far from the "true" value, the result may be too conservative to be practical. The *CD*-value shows, on average, by how many orders of magnitude the upper boundary exceeds the true value. The *RB*-value may be helpful for comparing the accuracy of the fit, including most probable estimate vs. mean value estimate of GPD and two different techniques

of the exponential fit. It also indicates if a method is conservative (when its sign is positive).

As mentioned above, the third tier of the [28] validation procedure is carried out over a number of environmental conditions. Table 1 lists the conditions considered in the present validation campaign. Table 2 summarizes the results with the GPD fit (meaning of different font colors are explained in the next section). The tier-two validation procedure was repeated three times on independent data to check the variability of the results. Each data set included 50 records with a duration of 100 h. The passing rate for each individual data set is indicated as PR_1, PR_2, and PR_3, while PR_A stands for the passing rated averaged over all three data sets. Conservative distance and relative bias were also averaged over all three data sets. The symbol RB_M is for the relative bias, computed over the mean value of the extrapolated estimate, while RB_{MP} means relative bias of the most probable value of the extrapolated estimate. Two more values were included in Table 2 to indicate the ability to complete the extrapolation with a given data set. The value NF_{MP} shows how many times the calculations did not yield the most probable extrapolated value over 150 data sets, e.g. data set #3 in Fig. 4. The value NF_U indicates how many times over 150 data sets the upper boundary of the extrapolated estimate was not provided, e.g. data set #17 in Fig. 4. Finally, averaged quantities for all performance indicators are included in Table 2.

Table 2 Summary of validation results with GPD tail

Hs, m	Tm, s	β deg	PR_1	PR_2	PR_3	PR_A	CD	RB_M	RB_{MP}	NF_{MP}	NF_{UB}
8.5	14	45	1.00	0.98	0.90	0.96	2.31	32.21	21.77	10	0
8.5	14	60	0.92	0.96	0.94	0.94	1.85	10.96	8.43	25	4
9	14	35	1.00	0.98	0.98	0.99	2.53	45.39	25.15	7	0
9	14	40	1.00	0.98	1.00	0.99	2.20	22.61	13.15	6	0
9	14	45	0.98	0.98	0.96	0.97	1.68	7.62	5.66	18	1
9	14	50	0.98	0.92	0.94	0.95	1.55	5.32	4.31	30	2
9	14	55	0.90	0.80	0.92	0.87	0.89	0.38	0.20	42	3
9	14	60	0.90	0.86	0.94	0.90	1.02	0.81	0.57	42	3
9	14	65	0.94	0.92	0.94	0.93	1.33	2.47	1.75	35	3
9	14	70	0.92	1.00	0.90	0.94	2.20	16.89	9.01	46	3
9	15	45	0.98	0.96	0.96	0.97	2.53	50.58	32.68	14	1
9	15	60	0.96	0.98	0.98	0.97	2.40	40.27	27.94	13	0
9.5	15	45	0.96	0.94	0.96	0.95	1.80	8.71	5.35	14	1
9.5	15	60	0.98	0.94	0.96	0.96	1.64	6.70	5.25	25	1
Averaged quantities						0.95	1.85	17.92	11.52	23.36	1.57

Table 3 summarizes the results of the extrapolations with the exponential tail fit. Both methods of fit are included: prediction error criterion and goodness-of-fit test. Any justification for setting a level of significance α for the goodness-of-fit test is not apparent, the level of significance was varied from 0.1 to 0.5. Averaged quantities for all performance indicators are also included in Table 3.

Passing rate is the main criterion in tier-two validation. [28] considers the tier-two validation successful if the passing rate does not fall below a standard value that depends on the number of extrapolation data sets, and it equals to 0.9 for 100

Table 3 Summary of validation result with exponential tail fit

Hs, m	Tm, s	β deg	Prediction error criterion			Goodness-of-fit α = 0.1			Goodness-of-fit α = 0.2		
			PR	*CD*	*RB*	*PR*	*CD*	*RB*	*PR*	*CD*	*RB*
8.5	14	45	0.94	1.54	2.53	0.94	1.32	0.99	0.94	1.50	1.78
8.5	14	60	0.96	1.56	5.96	0.94	1.43	4.28	0.96	1.64	6.11
9	14	35	0.94	1.41	1.20	1.00	1.42	0.32	1.00	1.75	1.03
9	14	40	0.98	1.71	5.64	0.92	1.33	2.88	0.98	1.65	3.93
9	14	45	0.98	1.35	3.20	0.98	0.95	0.66	0.98	1.19	1.12
9	14	50	1.00	1.23	2.64	0.98	1.14	1.79	1.00	1.27	1.63
9	14	55	1.00	0.75	0.77	0.98	0.76	0.61	0.98	0.84	0.30
9	14	60	0.98	0.88	1.31	0.92	0.89	1.11	0.98	0.97	0.69
9	14	65	0.90	1.12	2.82	0.92	0.99	1.54	0.98	1.16	1.56
9	14	70	0.98	1.75	7.96	0.86	1.67	7.81	0.96	1.82	7.07
9	15	45	0.92	2.01	13.55	0.74	1.82	8.09	0.90	1.97	7.32
9	15	60	0.98	1.76	6.76	0.96	1.54	3.31	1.00	1.79	4.76
9.5	15	45	0.98	1.26	1.73	0.94	1.06	0.71	0.96	1.22	0.56
9.5	15	60	0.92	1.32	3.41	0.84	1.05	1.72	0.96	1.28	2.06
Averaged quantities			0.96	1.40	4.25	0.92	1.24	2.56	0.97	1.43	2.85

Hs, m	Tm, s	β deg	Goodness-of-fit α = 0.3			Goodness-of-fit α = 0.4			Goodness-of-fit α = 0.5		
			PR	*CD*	*RB*	*PR*	*CD*	*RB*	*PR*	*CD*	*RB*
8.5	14	45	0.98	1.65	2.77	0.98	1.75	3.49	0.96	1.78	4.10
8.5	14	60	0.96	1.69	6.35	0.96	1.71	5.96	0.96	1.71	5.55
9	14	35	1.00	1.84	1.58	1.00	1.90	1.87	1.00	1.98	2.92
9	14	40	1.00	1.75	4.16	1.00	1.80	4.68	0.98	1.81	5.11
9	14	45	0.96	1.28	1.47	0.98	1.34	1.57	0.98	1.36	1.87
9	14	50	1.00	1.32	1.78	1.00	1.36	1.73	1.00	1.38	1.76
9	14	55	0.98	0.88	0.31	0.98	0.90	0.30	0.98	0.91	0.29
9	14	60	0.96	1.01	0.72	0.96	1.03	0.70	0.96	1.04	0.69
9	14	65	1.00	1.21	1.38	0.96	1.22	1.46	0.96	1.24	1.26
9	14	70	0.98	1.85	5.45	1.00	1.88	5.37	0.98	1.89	4.52
9	15	45	1.00	2.07	8.09	1.00	2.13	9.76	1.00	2.19	11.61
9	15	60	1.00	1.79	3.34	1.00	1.88	4.20	1.00	1.96	5.73
9.5	15	45	0.96	1.32	0.71	0.96	1.39	1.02	0.96	1.45	1.37
9.5	15	60	1.00	1.39	2.30	1.00	1.43	2.32	1.00	1.44	2.25
Averaged quantities			0.98	1.50	2.89	0.98	1.55	3.17	0.98	1.58	3.50

Table 4 Upper and lower acceptable passing rates

N_e	50	100	150	700	2100
Lower	0.88	0.90	0.91	0.93	0.94
Upper	1.00	0.99	0.98	0.97	0.96

extrapolations. The standard should be lower for 50 extrapolations in this validation campaign, as the random variability is expected to be larger. To adjudicate a tier-two validation with an arbitrary number of data sets, consider each extrapolation (*i.e.* tier-one validation) as a Bernoulli trial. If an extrapolation procedure works perfectly, the probability of covering a true value with the confidence interval is equal to the accepted confidence probability P_β. Then the number of successful extrapolations N_S of a total of N_e extrapolations is a random number following binomial distribution with parameters N_e and P_β.

Using the same confidence probability P_β, the expected boundaries of the passing rate can be computed as

$$PT^{Up,Low} = \frac{1}{N_e} Q_B\left(\frac{1 \pm P_\beta}{2}; N_e, P_\beta\right) \tag{22}$$

The upper and lower acceptable passing rates for different number of extrapolations are listed in Table 4.

5 Discussion

Only the lower boundary for the passing rate (Table 4) is proposed for validation use by [28]. An apparent reason is that exceeding the upper boundary of the passing rate indicates that the width of the confidence interval was likely overestimated. The results are likely to be conservative, but the extrapolation method is still usable.

If the passing rate falls below the lower boundary from Table 4, the extrapolation result may be questionable. A likelihood that its confidence interval does not contain a true value may be too high and cannot be explained by natural variability. These cases were encountered during the described validation study for both GPD and exponential tail fits. They are indicated by the red font in Tables 2 and 3.

Two tier-two validation failures were observed for GPD tail fit: in date set 2 for the heading 55 and 60° at significant wave height of 9 m and modal period 14 s. Two other data sets for these conditions did not indicate a failure. The reason for failure is likely that the shape parameter was significantly underestimated, leading to a very light tail and to one of the "pitfalls" of GPD tail fitting described by [26]. Table 2 also contains the passing rate averaged over three data sets in the column marked PR_A. As the total number of "trials" for this column is 150, the acceptable passing rate is between 0.91 and 0.98 (Table 4). The cases when the passing rate exceeded the upper boundary of 0.98 are in the blue font. The observed number of failures for

the most probable estimate and upper boundary are given for 150 data sets as well as values of the conservative distance and relative bias.

Two failures were observed with the exponential tail: for a heading of 70° with a modal period 14 s and at 45° with a model period 15 s and significant wave 9 m. Both failures were observed when applying goodness-of-fit with the level of significance 0.1. As no failures were recorded for a level of significance exceeding 0.1, the reason is likely to be the fitting method. Based on these observations, the significance level must be 0.2 or more for successful use of the goodness-of-fit test. No failures were observed for the prediction error criterion.

As mentioned above, the requirements for tier three are not yet clear. One approach described by [28] is to fail a validation if one of the conditions at tier two did not pass. Following this approach, the GPD tail fit and exponential tail fit with the goodness-of-test and significance level of 0.1 should be limited in application to those conditions that passed tier two.

Alternatively, the passing rates averaged over all the considered conditions could be reviewed. That would correspond to 2100 extrapolation data sets for the GPD tail and 700 for the exponential tail. Acceptable boundaries for the passing rates are available in Table 4. This approach finds the GPD tail acceptable with a perfect passing rate of 0.95, indicating that the light tail "pitfall" still can be overcome by a large-volume sample. This is also a possible indication of slow convergence of GPD. The "averaged" approach still fails the exponential tail with a significance level goodness-of-fit of 0.1. Significance levels of 0.3 and above may be seen as too conservative with a passing rate of 0.98, exceeding an acceptable level of 0.97 from Table 4. The exponential tail estimated with 0.2 significance level for goodness-of-fit and prediction error criterion are found acceptable by both tier-three approaches.

The conservative distance, *CD*, as follows from its name and definition in Eq. (20), is an indicator of how conservative the extrapolated estimate could be, expressed as an order of magnitude. The *CD* values are evaluated for all the extrapolation data sets individually and averaged over all conditions. The latter is a convenient metric to compare the performance of different tail fits. The exponential tail reduces the *CD* value to 1.4–1.5 from the 1.85 evaluated for GPD. This conclusion is consistent with the visual observation in Figs. 4 and 5, showing a more significant decrease for the lower boundary of the confidence interval.

The relative bias, *RB*, is defined in Eq. (20) and is similar to the *CD*-value. It measures the conservativeness of the extrapolated estimate but uses the most probable value (and mean value for the GPD tail) rather than the upper boundary of the confidence interval. Since the upper boundary is expected to be of practical use, the *RB* value can be observed as an auxiliary performance indicator. Similar to the conservative distance, *RB* values are evaluated for all the extrapolation data sets individually and averaged over all the conditions. The *RB* value is formulated as a factor rather than an order of magnitude, so the most probable value can be expected to converge to the true value if the extrapolation is perfect.

The relative bias values reveals that for the GPD tail, the most probable value is a better estimate, as $RB_{MP} < RB_M$ in Table 2. However, the estimation of the most probable extrapolated value fails on average in about 15% (22.36/150, the column

identified NF_{MP} in Table 2) of extrapolation attempts, while the mean value estimate always can be computed (Eq. 27 in [10]). The difference between *RB* values for GPD and exponential tail is also a good illustration of improvement made by the physics-informed approach: on average 11 for GPD in Table 2 versus 2.9–4.25 for the exponential tail in Table 3. This difference is believed to be caused by slower convergence of GPD vs. exponential tail.

The last column in Table 2 (identified as NF_U) is the number of failures for the calculation of the upper boundary of the confidence interval for the GPD extrapolated estimate. The percentage of failures is about 1% (1.57/150), which is smaller than the percentage of failures for the most probable estimate, $NF_U < NF_{MP}$; therefore, so even if the calculation of the most probable GPD value fails, the upper boundary of the confidence interval still may be available.

Concluding the overall performance assessment, the best method was found to be an exponential fitted with goodness-of-fit test with significance level of 0.2 with $CD = 1.43$ and $RB = 2.85$, while the fitting with the error prediction criterion having a similar $CD = 1.4$ but $RB = 4.25$. However, since no theoretical background exists for the choice of the significance level in the goodness-of-fit test, the recommendation is to use the extrapolation with exponential tail fitted with the error prediction criterion.

6 Summary and Conclusions

This chapter describes the statistical validation of the split-time method for estimating the probability of capsizing in irregular waves. The main feature of the split-time method is to compute a metric of the likelihood of capsizing as a difference between the observed roll rate at a roll threshold up-crossing and a critical roll rate leading to capsizing at a particular instant of time. Statistics for the metric values can be collected without actual observation of capsizing and extrapolated to estimate the probability of capsizing.

Extrapolation is performed with Generalized Pareto distribution (GPD) following the second extreme value theorem. Accounting for weak nonlinearity of roll rate and assuming lesser influence of stability variation in waves, an exponential tail can be applied instead of GPD. Including physical information into extrapolation scheme (i.e. the physics-informed approach) allows a significant decrease in the statistical uncertainty and improvement of the reliability of the prediction.

Validation of extrapolation is determined with a fast numerical simulation algorithm, capable of qualitatively reproducing the most principle nonlinearity of roll motion by computing the instantaneous submerged volume and its centroid. These calculations were carried out for sufficiently long times to observe capsizing in realistic conditions. The validation is considered to be successful if a small subset of this data can predict the capsizing probability without observing capsize.

The statistical validation considered 14 conditions for the ONR tumblehome top configuration. A three-tiered validation procedure was employed for GPD and exponential tail extrapolation. Two tail fitting techniques were applied for the exponential tail: prediction error criterion and goodness-of-fit test, with the series level of significance varying from 0.1 to 0.5.

If the successful multi-condition validation requires that all the conditions to be validated individually, only extrapolation with exponential tail fitted with error prediction criterion or goodness-of-fit test with the significance level 0.2 and above can pass. If adjudication of success is based on the averaged outcomes, GPD extrapolation also passes.

In addition to validation, performance of the extrapolation methods was assessed with criteria for conservativeness and accuracy. The best performing methods were extrapolations by exponential tail fitted with error prediction criterion and goodness-of-fit test with a significance level of 0.2. The final recommendation is application of the split-time method with exponential tail fitted with the error prediction criterion.

Acknowledgements The work described in this chapter has been funded by the Office of Naval Research (ONR) under Dr. Woei-Min Lin. This work was also supported by the NSWCCD Independent Applied Research (IAR) program under Dr. Jack Price. The participation of Prof. Pipiras was facilitated by the NSWCCD Summer Faculty and Sabbatical Programs, both of which were also managed by Dr. Jack Price. The authors are very grateful for the support that made this work possible.

References

1. Anastopoulos PA, Spyrou KJ (2019) Evaluation of the critical wave groups method in calculating the probability of ship capsize in beam seas. Ocean Eng 187:106213
2. Anastopoulos P, Spyrou KJ (2019) Can the generalized Pareto distribution be useful towards developing ship stability criteria? In: Proceedings of the 17th international ship stability workshop (ISSW 2019), 10–12 June, Helsinki
3. Anastopoulos PA, Spyrou KJ (2023) An efficient formulation of the critical wave groups method for the assessment of ship stability in beam seas. In: Spyrou K, Belenky V, Katayama T, Bačkalov I, Francescutto A (eds) Contemporary ideas on ship stability—from dynamics to criteria, Chapter 10. Springer, Berlin, pp 157–174. ISBN 978-3-031-16328-9
4. Anastopoulos PA, Spyrou KJ (2023) Effectiveness of the generalized Pareto distribution for characterizing ship tendency for capsize. In: Spyrou K, Belenky V, Katayama T, Bačkalov I, Francescutto A (eds) Contemporary ideas on ship stability—from dynamics to criteria, Chapter 15. Springer, Berlin, pp 245–263. ISBN 978-3-031-16328-9
5. Belenky V, Weems K, Lin WM (2016) Split-time method for estimation of probability of capsizing caused by pure loss of stability. Ocean Eng 122:333–343
6. Belenky V, Weems K, Pipiras V, Glotzer D (2018) Extreme-value properties of the split-time metric. In: Proceedings of 13th international conference on stability of ships and ocean vehicles (STAB 2018), Kobe, Japan
7. Belenky V, Weems KM, Spyrou K, Pipiras V, Sapsis T (2023) Modeling broaching-to and capsizing with extreme value theory. In: Spyrou K, Belenky V, Katayama T, Bačkalov I, Francescutto A (eds) Contemporary ideas on ship stability—from dynamics to criteria, Springer, Chapter 26. Berlin, pp 435–457. ISBN 978-3-031-16328-9

8. Belenky V (2011) On self-repeating effect in reconstruction of irregular waves. In: Neves MAS, Belenky V, de Kat JO, Spyrou K, Umeda N (eds) Contemporary ideas on ship stability, Chapter 33. Springer, Berlin, pp 589–598. ISBN 978-94-007-1481-6
9. Belenky V, Reed AM, Weems KM (2011) Probability of capsizing in beam seas with piecewise linear stochastic GZ curve. In: Neves MAS, Belenky V, de Kat JO, Spyrou K, Umeda N (eds) Contemporary ideas on ship stability, Chapter 30. Springer, Berlin, pp 531–554. ISBN 978-94-007-1481-6
10. Belenky V, Weems K, Campbell B, Pipiras V (2014) Extrapolation and validation aspects of the split-time method. In: Proceedings of 30th symposium on naval hydrodynamics, Hobart, Tasmania, Australia
11. Belenky V, Weems K, Pipiras V, Glotzer D, Sapsis T (2018) Tail structure of roll and metric of capsizing in irregular waves. In: Proceedings of 32nd symposium on naval hydrodynamics, Hamburg, Germany
12. Belknap WF, Reed AM (2019) TEMPEST: a new computationally efficient dynamic stability prediction tool. In: Belenky V, Spyrou K, van Walree F, Neves MAS, Umeda N (eds) Contemporary ideas on ship stability. Risk of capsizing, Chapter 1. Springer, Berlin, pp 3–21. ISBN 978-3-030-00514-6
13. Bickel PJ, Doksum KA (2001) Mathematical statistics: basic ideas and selected topics, vol 1, Prentice Hall. ISBN 978-0138503635
14. Bishop RC, Belknap W, Turner C, Simon B, Kim JH (2005) Parametric investigation on the influence of GM, roll damping, and above-water form on the roll response of model 5613, hydromechanics department report, Naval Surface Warfare Center Carderock Division, West Bethesda, Maryland, USA, NSWCCD-50-TR-2005/027
15. Campbell B, Belenky V, Pipiras V (2016) Application of the envelope peaks over threshold (EPOT) method for probabilistic assessment of dynamic stability. Ocean Eng 120:298–304
16. Campbell, B, Weems KM, Belenky V, Pipiras V, Sapsis T (2023) Envelope peaks over threshold (EPOT) application and verification. In: Spyrou K, Belenky V, Katayama T, Bačkalov I, Francescutto A (eds) Contemporary ideas on ship stability—from dynamics to criteria, Springer, Chapter 16. Berlin, ISBN 978-3-031-16328-9
17. Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London. ISBN 978-1849968744
18. Cramér H, Leadbetter MR (2004) Stationary and related processes. Dover, Sample function properties and their application
19. De Haan L, Ferreira A (2007) Extreme value theory: an introduction. Springer Science & Business Media. ISBN 978-0387239460
20. Glotzer D, Pipiras V, Belenky V, Campbell B, Smith T (2017) Confidence interval for exceedance probabilities with application to extreme ship motions. REVSTAT Statistical Journal 15(4):537–563
21. Leadbetter MR, Lindgren G, Rootzen H (1983) Extremes and related properties of random sequences and processes, Springer series in statistics. Springer, Berlin. ISBN 978-1461254492
22. Lewis EV (ed) (1989) Principles of naval architecture. Vol. 3: motions in waves and controllability. SNAME, Jersey City, 429 p. ISBN 0-939773-02-3
23. Lin WM, Yue DKP (1990) Numerical solutions for large amplitude ship motions in the time-domain. In: Proceedings of the 18th symposium on naval hydrodynamics, Ann Arbor, Michigan, USA, pp. 41–66
24. Mager J (2015) Automatic threshold selection of the peaks over threshold method. Master's thesis, Technische Universitat Munchen
25. Mohamad MA, Sapsis T (2018) Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. Proceedings of the national academic of sciences of United States of America (PNAS) 115:11138–11143
26. Pipiras V (2020) Pitfalls of data-driven peaks-over-threshold analysis: perspectives from extreme ship motions. Probab Eng Mech 60:103053
27. Shin YS, Belenky VL, Lin WM, Weems KM, Engle AH (2003) Nonlinear time domain simulation technology for seakeeping and wave-load analysis for modern ship design. SNAME Trans 111:557–578

28. Smith TC (2019) Validation approach for statistical extrapolation. In: Belenky V, Neves M, Spyrou K, Umeda N, van Walree F (eds) Contemporary ideas on ship stability. Risk of capsizing, Chapter 33. Springer, Berlin, pp 573–589. ISBN 978-3-030-00514-6
29. Stephens MA (1974) EDF statistics for goodness of fit and some comparisons. J Am Stat Assoc 69(347):730–737
30. Weems K, 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 forces. In: Proceedings of 13th international ship stability workshop, Brest, France
31. Weems K, Belenky V (2015) Fast time-domain simulation in irregular waves with volume-based calculations for Froude-Krylov and hydrostatic force. In: Proceedings of 12th international conference on stability of ships and ocean vehicles (STAB 2015), Glasgow, UK
32. Weems K, Belenky V (2018) Extended fast ship motion simulations for stability failures in irregular seas. In: Proceedings of 13th international conference on stability of ships and ocean vehicles (STAB 2018), Kobe, Japan
33. Weems K, Belenky V, Spyrou K (2018) Numerical simulations for validating models of extreme ship motions in irregular waves. In: Proceedings of 32nd symposium on naval hydrodynamics, Hamburg, Germany
34. Weems K, Belenky V, Campbell B, Pipiras V, Sapsis T (2019) Envelope peaks over threshold (EPOT) application and verification. In: Proceedings of 17th international ship stability workshop, Helsinki, Finland, pp 71–79