Journal of Marine Technology and Environment
DOI:10.53464/JMTE.02.2026.04

ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www. jmte.eu

SERIOUS HUMAN CONSEQUENCES IN SHIPPING ACCIDENTS: TEMPORAL DISCRIMINATION AND CALIBRATION USING TSB MARSIS DATA, 1995-2026

Dr. David O. KREMELBERG[1]

[1]Adjunct Instructor, Maine Maritime Academy, 1 Pleasant Street, Castine, ME 04420, USA
ORCID: 0009-0003-1260-5779; e-mail address: david.kremelberg@mma.edu

Abstract: Maritime risk management requires separating accident occurrence from consequence severity conditional on an accident. This study used the Transportation Safety Board of Canada public Marine Safety Information System (MARSIS) to examine serious human consequences among 9,491 reportable shipping accidents from 1995 through July 2026; 293 (3.09%) involved death, a missing person, or serious injury. Among accident types, the all-period occurrence-level logistic model showed the largest adjusted association for capsize (odds ratio, OR 10.36, 95% confidence interval, CI 6.76-15.88), followed by sinking/loss/abandonment (OR 2.74, 95% CI 1.79-4.19); moderate/adverse/ice sea state had OR 2.25 (95% CI 1.65-3.08). A historical model developed on 6,923 accidents from 1995-2014 retained discrimination in 2,568 later accidents (area under the receiver operating characteristic curve (AUC) 0.826; precision-recall AUC (PR-AUC) 0.237) but under predicted absolute risk: 93 events were observed versus 41.9 expected (observed/expected ratio 2.22). Connected-component resampling supported these conclusions. Post hoc time-specification checks predicted 61.3 events without calendar year and 67.7 with a linear year term, indicating that extrapolation of the original spline contributed to under prediction while not fully explaining it. The full historical model also materially outperformed an accident-type-only benchmark for AUC, PR-AUC, and Brier score. Expanding-origin refits using more contemporary data improved average calibration in later forward blocks. The results support transparent retrospective consequence-risk stratification, but not a fixed operational probability threshold: contemporary refitting, prospective evaluation, and a defined prediction time are required before deployment.
Key words : Accident severity; calibration; consequence risk; maritime safety; MARSIS; temporal validation.

1. INTRODUCTION

Maritime safety risk is not a single probability. Risk-management frameworks separate hazard identification, event occurrence, consequence development, controls, and monitoring [1]. For accident databases, an additional distinction is essential: the probability that a vessel experiences an accident is different from the probability that an accident already recorded by an authority produces death or serious injury. Commercial-fishing research has made this distinction explicitly when comparable data on safe trips were unavailable [2]. The public Marine Safety Information System (MARSIS) is likewise occurrence-centred rather than a denominator of safe voyages or vessel-time exposure [3]. The estimand here is therefore P(serious human consequence | TSB-reportable shipping accident), not accident incidence.

Human-consequence modelling has substantial precedent. Weng and Yang analysed 24,301 worldwide ship-accident records from 2001 through February 2011, using logistic regression for fatal versus nonfatal accidents and a zero-truncated negative-binomial model for mortality among fatal accidents; their design used a random training/test split and found especially strong consequences for sinking as well as associations with accident type, adverse weather, darkness, distance from shore, and cruise-vessel status [4]. Talley, Jin, and Kite-Powell modelled nonfatal injuries, deaths, and missing crew in U.S. Coast Guard vessel accidents with count methods [5]. More recently, Wang et al. modelled four ordered injury-severity levels using 1,128 investigation reports from seven national agencies, including the TSB, and a zero-inflated ordered probate model; collision vessels were represented separately, producing 1,294 ship-level observations [6]. These studies establish that regression-based and human-specific maritime consequence models are not themselves novel.

A broader literature has modelled accident severity and accident mechanisms using ordered regression, Bayesian networks, and data-driven methods [7,8,9,10,11]. Recent studies add text mining, explainable ensemble methods, and operationally richer Bayesian or machine-learning frameworks [12,13,14,15]. Reviews emphasize that

Journal of Marine Technology and Environment
DOI:10.53464/JMTE.02.2026.04
ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www. jmte.eu

maritime risk analyses are strongly shaped by the characteristics and availability of underlying data [1,16].

Canadian fishing studies further show that environmental conditions can affect both incident occurrence and consequences, while policy translation requires attention to stability, weather information, training, communication, and search-and-rescue planning [17,18].

The present contribution is therefore methodological and data-structural rather than algorithmic. It uses the full public TSB structured occurrence record over approximately three decades, defines an occurrence-level human endpoint so that a single accident outcome is not duplicated across its vessels, excludes obvious post-outcome response variables, and distinguishes relative discrimination from absolute probability calibration. Calibration is important because a model may rank cases well while systematically misstating risk levels [19]. Temporal structure also matters: random cross-validation can understate predictive error when nearby observations share temporal or hierarchical dependence, so out-of-time evaluation is more appropriate when the intended question concerns later periods [20]. Reporting and model-identity terminology were additionally informed by TRIPOD+AI guidance for regression and machine-learning prediction studies, adapted here to a nonclinical maritime setting [21].

The objectives were to identify accident, vessel, operational, geographic, temporal, and environmental characteristics associated with serious human consequences among TSB-reportable shipping accidents; to determine how well a transparent multivariable model stratified later accidents; and, after cross-era under prediction was observed, to conduct a limited set of explicitly post hoc checks addressing calendar-time specification, incremental value, recurrent-vessel dependence, missingness, multi-vessel aggregation, and a specific gross-tonnage data-quality anomaly. Because the source contains finalized occurrence fields rather than time-stamped initial reports, the analysis is described as retrospective consequence profiling.

Operational triage or pre-accident deployment is not claimed.

2. DATA AND METHODS

2.1 Data source, cohort, and outcome:

Data were obtained from the TSB public MARSIS release, comprising occurrence, vessel, navigation-equipment, lifesaving-equipment, recording-equipment, injury, and data-dictionary files [3]. The study window was 1 January 1995 through 31 July 2026. Starting from 88,054 raw occurrence rows representing 48,820 unique occurrence identifiers, 27,035 identifiers fell within the study window, 11,228 were classified as accidents, and 9,491 met the predefined shipping-accident classifications.

All 9,491 linked to occurrence-level vessel aggregates. Reportable incidents were excluded. Accidents aboard ship were also excluded because the TSB definition incorporates a person being killed or seriously injured, which would make the analytic denominator partly outcome-defined [22].

The primary endpoint was occurrence-level: at least one death, missing person, or serious injury. TSB regulations define serious injury using specified clinical and hospitalization criteria [23]. Of the 9,491 accidents, 293 met the primary endpoint, 9,198 did not, 199 involved death or a missing person, and 94 involved serious injury without death or missing persons. Death, missing-person, and injury fields can overlap and were not summed as mutually exclusive person counts. Injury records were de-duplicated by victim-count identifier before reconciliation, and reconstructed July 2026 shipping-accident counts agreed with the TSB published table [24]. A prespecified secondary endpoint included death or missing person only.

The TSB introduced a new marine occurrence database in November 2013 and revised reporting requirements effective 1 July 2014 [22]. The primary all-period association model used all eligible accidents. For temporal evaluation, 1995-2014 accidents formed the development sample and 2015-July 2026 accidents formed the later evaluation sample, thereby beginning evaluation with the first full calendar year after the reporting change. The later era was subsequently inspected during model revision, so no post hoc comparison is described as a new untouched holdout.

2.2 Predictors and pre-processing:

Accident type was coded as collision/allision (reference), grounding, fire/explosion, capsize, sinking/loss/abandonment, or unseaworthy/unfit. Vessel composition was represented by overlapping occurrence indicators for fishing, passenger/ferry, cargo/tanker, and tug/barge/service vessels. Maximum gross tonnage (GT) across vessels represented size and was $\log_2$-transformed so that one coefficient represented a doubling of GT. Operational phase used overlapping indicators for underway, secured/stationary, and not under own power. Region was harmonized to Pacific, Central/Arctic, Atlantic, or Foreign waters. Calendar year, light condition, weather, and sea state completed the core set. Post-accident response, damage, pollution, investigation-class, evacuation, and lifesaving-use variables were excluded from the primary model.

Weather was grouped as clear/overcast, adverse, or unknown/unrecorded; sea state as calm-to-slight, moderate/adverse/ice, or unknown/unrecorded. Positive GT was available for 8,689 accidents (91.5%). Missing $\log_2$(GT) was replaced by the median observed $\log_2$(GT)

Journal of Marine Technology and Environment
DOI:10.53464/JMTE.02.2026.04
ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www. jmte.eu

within each training sample and accompanied by a missing-GT indicator. Unknown categorical states were retained because recording completeness varied strongly by era. This approach preserves a recording-state signal but does not imply that missing-indicator methods recover unbiased physical-variable associations in observational data; that limitation is well established in the missing-data literature [25]. A post hoc deterministic, outcome-blind conditional GT-imputation sensitivity therefore assessed whether the principal associations and temporal calibration were dependent on median imputation. The unusable people-aboard field was excluded before outcome-association fitting after audit showed 3,795 accidents recorded zero aboard, including 75 primary events.

2.3 *Statistical model and validation:*

The primary model was occurrence-level multivariable logistic regression with 26 slope parameters plus an intercept. Calendar year used a natural cubic spline with three degrees of freedom, with knots and boundaries determined only from the corresponding development sample. All core predictors were entered simultaneously; stepwise selection, univariate screening, and outcome-driven deletion were not used. Results are reported as adjusted odds ratios (ORs) with 95% model-based Wald confidence intervals (CIs). Average standardized probabilities and risk differences were obtained by marginal standardization over the empirical all-period covariate distribution; their CIs used the delta method. These are descriptive model translations, not contemporary calibrated predictions and not causal intervention effects. A Firth-type logistic model with intercept correction was prespecified only as a numerical fall-back for sparse sensitivity fits [26].

Internal validation used 500 bootstrap resamples with training-dependent pre-processing repeated inside each resample. Five repeats of stratified 10-fold cross-validation were retained as a secondary internal check. The primary transport assessment applied the 1995-2014 historical equation unchanged to 2015-July 2026. Discrimination was measured with ROC AUC and precision-recall AUC (PR-AUC); overall probability error used Brier score and log loss. Calibration was summarized by calibration intercept, calibration slope, observed-to-expected (O/E) event ratio, and a flexible calibration curve. The calibration intercept used the predicted logit as an offset with slope fixed at one; the calibration slope was estimated in a separate logistic model with an intercept. In the convention used here, a positive calibration intercept denotes systematic under prediction. Because only 93 events occurred in the later sample, the flexible high-risk tail was interpreted cautiously, consistent with recommendations that calibration curves need substantial event counts [19].

Recurrent vessel identifiers created dependence across accidents. Vessel-linked connected components were therefore constructed across the full sample. Later-period performance uncertainty used 2,000 resamples of whole connected components while holding each historical fitted equation fixed. The later sample contained 1,830 components; the largest had 71 accidents. These intervals quantify evaluation-sample dependence conditional on the fitted historical model, not model-development uncertainty. Random cross-validation was not treated as independent temporal evidence because structured random splits can produce optimistic error estimates [20].

Model complexity was interpreted without a mechanical events-per-parameter (EPP) rule. Modern prediction-methodology work shows that required sample size depends on anticipated model performance, outcome risk, and parameterization rather than a universal threshold [27]. Accordingly, the 26-slope historical model and smaller modern updating samples were accompanied by internal validation, sensitivity analyses, and a limited ridge experiment rather than a retrospective claim that a simple EPP threshold proved adequacy.

2.4 *Post hoc revision analyses*

After the original temporal result had been examined, three calendar-time diagnostics were specified using the same historical development sample and later evaluation sample: refitting without calendar year, refitting with a linear year term, and holding the original spline input at the 2014 boundary while retaining the original coefficients. An accident-type-only logistic model provided a simple benchmark. Fixed historical models were also evaluated in 2015-2018, 2019-2022, and 2023-July 2026 blocks; expanding-origin comparisons refit each specification using only preceding eligible accidents. All of these comparisons are labelled post hoc; none restores the independence of the already inspected later outcomes. Paired component resampling compared the original full historical equation with the accident-type-only benchmark.

Additional focused checks addressed data quality and aggregation. Two records for the 42-m ferry R.G.L. Fair weather contained GT=301,287. The design firm lists gross registered tonnage of 304 for the sister-ferry family [28], which flags implausibility but does not verify the individual vessel’s exact GT. Therefore, the raw value was retained in the baseline and treated as missing only in sensitivity analyses; neither 304 nor 301.287 was substituted. A single-vessel sensitivity removed 1,663 multi vessel accidents. Conditional GT imputation also treated the two flagged ferry values as missing and used training-only vessel and era information without the outcome. Performance was also

Journal of Marine Technology and Environment
DOI:10.53464/JMTE.02.2026.04
ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www.jmte.eu

examined separately for later accidents involving vessel identifiers seen or unseen historically. Finally, a limited ridge model was tuned within 2015-2020 using three forward folds and evaluated on the already examined 2021-July 2026 period. These analyses were designed to test stability, not to search until a preferred result appeared.

The font for first and last name of the authors is Times New Roman 11.

3. RESULTS

3.1 Cohort and primary associations

Table 1 describes the cohort. Among 9,491 eligible accidents, 293 (3.09%) met the primary endpoint. Collision/allision (30.7%) and grounding (29.1%) were the most common classifications, but serious outcomes were concentrated in capsize and sinking/loss/abandonment: 90 of 315 capsizes and 81 of 830 sinking/loss/abandonment accidents met the endpoint.

The outcome rate was 28.57% for capsize, 9.76% for sinking/loss/abandonment, 2.30% for fire/explosion, 2.03% for collision / allision, 0.87% for grounding, and 0.19% for unseaworthy/unfit accidents.

The all-period model converged normally (Table 2). Capsize had the largest adjusted association (OR 10.36, 95% CI 6.76-15.88). Its all-period model-standardized descriptive risk was 19.95% (95% CI 15.73%-24.16%), compared with 2.73% for collision / allision. Sinking/loss/abandonment had OR 2.74 (95% CI 1.79-4.19), whereas grounding had OR 0.26 (95% CI 0.16-0.43). Moderate/adverse/ice sea state had OR 2.25 (95% CI 1.65-3.08). The broad adverse-weather indicator did not show a clear additional association conditional on the specified variables (OR 1.04, 95% CI 0.73-1.46); this comparison does not establish that weather has little total causal importance. Passenger/ferry involvement had OR 2.18 (95% CI 1.23-3.85). Each doubling of recorded maximum GT had OR 0.87 (95% CI 0.83-0.92), while secured/stationary status had OR 0.37 (95% CI 0.22-0.63).

Table 1. Selected characteristics of reportable shipping accidents, overall and by serious human-consequence status

Characteristic	**Level**	**Overall *N*=9,491**	**No event *N*=9,198**	**Event *N*=293**
Sample size	*N*	9491	9198	293
Accident type	Collision/allision	2913 (30.7%)	2854 (31.0%)	59 (20.1 %)
Accident type	Grounding	2763 (29.1%)	2739 (29.8%)	24 (8.2%)
Accident type	Fire/explosion	1611 (17.0%)	1574 (17.1%)	37 (12.6 %)
Accident type	Capsize	315 (3.3%)	225 (2.4%)	90 (30.7 %)
Accident type	Sinking/loss/abandonment	830 (8.7%)	749 (8.1%)	81 (27.6 %)
Accident type	Unseaworthy/unfit	1059 (11.2%)	1057 (11.5%)	2 (0.7%)
Any fishing vessel involved	Yes	3795 (40.0%)	3633 (39.5%)	162 (55.3 %)
Any passenger/ferry vessel involved	Yes	1126 (11.9%)	1085 (11.8%)	41 (14.0 %)
Any tug/barge /service vessel involved	Yes	2318 (24.4%)	2269 (24.7%)	49 (16.7 %)
Maximum gross tonnage	Median [Q1, Q3]	184.0 [24.0, 9304.1]	204.1 [25.1, 9611.0]	25.0 [10.0, 145.6]
Maximum vessel length (m)	Median [Q1, Q3]	23.4 [12.0, 126.9]	25.0 [12.2, 129.9]	11.4 [8.0, 19.1]
Any vessel secured/stationary	Yes	1723 (18.2%)	1696 (18.4%)	27 (9.2%)
Weather condition	Adverse	1422 (15.0%)	1366 (14.9%)	56 (19.1 %)

Note: Percentages are column percentages. Vessel-type flags can overlap and need not sum to 100%. Continuous summaries use observed positive values, with length additionally limited to 500 m. Available-value denominators and unknown categories are in Supplement Table S14.

3.2 Temporal discrimination, calibration, and incremental value

DOI:10.53464/JMTE.02.2026.04
ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www. jmte.eu

The historical development sample contained 6,923 accidents and 200 events; the later evaluation sample contained 2,568 accidents and 93 events. The original historical equation retained AUC 0.826 and PR-AUC 0.237. Connected-component percentile intervals were 0.780-0.869 for AUC and 0.164-0.334 for PR-AUC. The Brier score was 0.03178 (component interval 0.02546-0.03832). Relative ranking therefore transported reasonably well. Table 3 distinguishes the all-period model from the historical fitted equations.

Absolute probabilities did not. The original historical equation predicted 41.9 events where 93 occurred, for O/E 2.22 (component interval 1.80-2.63), calibration intercept 0.938 (0.686-1.148), and calibration slope 0.905 (0.764-1.062). The flexible calibration curve (Figure 1) was above the identity line through most of the supported probability range, with wide uncertainty in the sparse high-risk tail. Thus, discrimination and calibration gave different conclusions, consistent with the distinction emphasized in prediction-methodology literature [19,29].

The post hoc accident-type-only benchmark had lower AUC (0.741) and PR-AUC (0.141) and a higher Brier score (0.03268) but better average calibration (O/E 1.25). Relative to that benchmark, the original full historical equation improved AUC by 0.0846 (paired component interval 0.0517-0.1186), PR-AUC by 0.0955 (0.0580-0.1582), and had a Brier-score difference of -0.00090 (-0.00168 to -0.00016). The fuller model therefore added meaningful ranking information beyond accident type alone, even though its fixed historical probability scale transported less well.

Table 2. Selected all-period adjusted associations and standardized descriptive risk translations.

Predictor / contrast	**Adjusted OR (95% CI)**	**Standardized risk (95% CI)**	**Risk difference (95% CI)**
Grounding vs collision/allision	0.26 (0.16-0.43)	0.74% (0.44-1.03)	-1.99 pp (-2.81 to -1.17)
Fire/explosion vs collision/allision	0.96 (0.61-1.51)	2.62% (1.78-3.46)	-0.10 pp (-1.25 to +1.05)
Capsize vs collision/allision	10.36 (6.76-15.88)	19.95% (15.73-24.16)	+17.22 pp (+12.85 to +21.59)
Sinking/loss/abandonment vs collision/allision	2.74 (1.79-4.19)	6.87% (5.25-8.48)	+4.14 pp (+2.24 to +6.04)
Passenger/ferry vessel involved	2.18 (1.23-3.85)	5.32% (3.21-7.43)	+2.45 pp (+0.25 to +4.65)
Tug/barge/service vessel involved	0.58 (0.35-0.97)	2.15% (1.38-2.93)	-1.24 pp (-2.30 to -0.18)
Maximum GT, per doubling	0.87 (0.83-0.92)	-	-
Vessel secured/stationary	0.37 (0.22-0.63)	1.49% (0.86-2.13)	-2.00 pp (-2.86 to -1.15)
Adverse weather vs clear/overcast	1.04 (0.73-1.46)	3.59% (2.70-4.48)	+0.11 pp (-0.89 to +1.11)
Moderate/adverse/ice sea state	2.25 (1.65-3.08)	4.88% (3.93-5.84)	+2.40 pp (+1.35 to +3.45)

Note: Model-based Wald CIs are shown for ORs; delta-method CIs for standardized probabilities and differences. Standardization uses all 9,491 accidents. Risk units are percent; difference units are percentage points (pp). References are collision / allision for accident type, clear/overcast for weather, calm-to-slight for sea state, and absence for binary vessel/phase flags. Reference risks are in Supplement Table S15. These are descriptive associations, not contemporary calibrated probabilities or causal effects.

3.3. *What the time diagnostics changed*

Calendar-time specification materially affected the magnitude of under prediction without materially changing discrimination. Using the same 1995-2014 development and 2015-July 2026 evaluation samples, removing year predicted 61.3 events (O/E 1.52; AUC 0.830), while a linear year term predicted 67.7 (O/E 1.37; AUC 0.831). Holding the original spline input at 2014 predicted 52.8 (O/E 1.76). Under prediction therefore persisted, but extrapolation of the original natural-spline time term contributed substantially to its magnitude. Because these alternatives were evaluated after the later outcomes were known, none is presented as a newly validated replacement model.

Shorter forward blocks showed the same pattern. With the fixed 1995-2014 original equation, O/E increased from 1.66 in 2015-2018 to 2.31 in 2019-2022 and 3.32 in 2023-July 2026. The fixed linear-year sensitivity reduced those values to 1.19, 1.38, and 1.65. However, expanding-origin refits added an important qualification: when the original specification was refit through 2018, O/E was 1.02 for 2019-2022, and when

Journal of Marine Technology and Environment
DOI:10.53464/JMTE.02.2026.04
ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www.jmte.eu

refit through 2022, O/E was 0.96 for 2023-July 2026. This pattern suggests that updating with contemporary data can improve average calibration more effectively than treating any one historical time parameterization as permanent. It does not identify a single causal source of the temporal shift.

3.4 Risk concentration and robustness

Development-defined risk strata separated later accidents strongly. The highest-risk 302 accidents, 11.8% of the later sample, contained 52 of 93 events (55.9%); the highest two strata contained 71.0% of events in 24.3% of accidents. This demonstrates concentration of retrospective consequence risk, not an optimal operational threshold or decision benefit.

The new robustness checks did not materially alter the central conclusions. Treating the two unresolved ferry GT values as missing changed temporal AUC from 0.82604 to 0.82591 and O/E from 2.2206 to 2.2159; all-period capsize OR changed from 10.357 to approximately 10.38. Train-only conditional GT imputation yielded temporal AUC 0.827 and O/E 2.19. The single-vessel subset contained 7,828 accidents and 248 events; its historical-to-later AUC was 0.833 and the same cross-era under prediction remained (O/E 2.39). In later data, 863 accidents involved at least one vessel identifier seen historically but contained only eight events; 1,705 accidents with no historically seen vessel contained 85 events, so the subgroup contrast is exploratory rather than a stable transport estimate. The limited ridge diagnostic in the 2021-July 2026 reused era produced AUC 0.839 and O/E 1.26, essentially similar to the modern reduced-model refit.

Table 3 Predictive performance, model identity, and post hoc temporal comparisons

Model / evaluation	AUC	PR-AUC	Brier	O/E	Cal. int.	Cal. slope
All-period apparent (1995-Jul 2026)	0.851	0.275	0.02519	1.00	0.000	1.000
All-period bootstrap corrected	0.838	0.250	0.02580	-	0.005	0.945
Historical original: 1995-2014 → 2015-Jul 2026	0.826	0.237	0.03178	2.22	0.938	0.905
Post hoc: historical model without year	0.830	0.242	0.03084	1.52	0.499	0.929
Post hoc: historical model with linear year	0.831	0.243	0.03066	1.37	0.383	0.929
All-period apparent (1995-Jul 2026)	0.851	0.275	0.02519	1.00	0.000	1.000
Post hoc: accident-type-only benchmark	0.741	0.141	0.03268	1.25	0.259	0.803

Note: The historical original row records the preexisting historical equation applied to the later era. All no-year, linear-year, and benchmark comparisons were added after the later outcomes had been examined and are therefore exploratory. Connected-component 95% intervals for the historical original model were AUC 0.780-0.869, PR-AUC 0.164-0.334, O/E 1.80-2.63, calibration intercept 0.686-1.148, and calibration slope 0.764-1.062. The all-period model uses 9,491 accidents/293 events for development and internal evaluation; historical comparisons use 6,923/200 for development and 2,568/93 for evaluation. O/E = observed/expected events; Cal. int. = calibration-in-the-large; PR-AUC = precision-recall AUC.

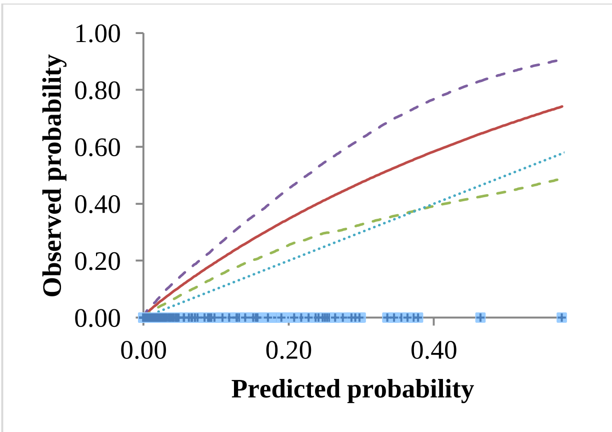

Figure 1 Calibration of the original 1995-2014 model in 2015-July 2026.

Journal of Marine Technology and Environment
DOI:10.53464/JMTE.02.2026.04
ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www. jmte.eu

The solid curve is a three-degree-of-freedom natural-spline logistic smoother of the predicted logit; dashed boundaries are pointwise 95% limits from 500 component resamples. The dotted diagonal denotes equality and baseline marks show predicted probabilities. The sparse high-risk tail is uncertain. The historical model is held fixed.

4. DISSCUTIONS

The principal substantive finding is a clear hierarchy of serious human-consequence risk conditional on a reportable shipping accident. Capsize and sinking/loss/abandonment were the strongest major accident-type signals, and moderate/adverse/ice sea state remained an important environmental marker after adjustment. These patterns agree with earlier evidence linking sinking or stability loss to fatalities or higher injury severity [2,4,5,6]. The present occurrence-level design differs from ship-level analyses that represent the vessels in a collision separately: that choice prevents one accident outcome from being counted repeatedly merely because multiple vessels were involved.

The environmental result should be interpreted specifically. Sea state showed a clear adjusted association, whereas the broad weather composite did not show a clear additional association conditional on the model's other variables. Canadian fishing research has linked wind, temperature, ice, storm intensity, and other environmental factors to incident severity, and policy reviews emphasize weather information, stability, and response planning [17,18]. The current findings therefore do not imply that weather is unimportant; they show that, within this coarse administrative representation, sea state carried a more clearly estimated conditional association than the general adverse-weather flag.

Vessel findings require similar restraint. Passenger/ferry involvement was associated with higher consequence risk, while greater recorded maximum GT was associated with lower conditional odds. Wang et al. also found that GT can act differently across injury-propensity and conditional-severity processes [6]. Because the MARSIS people-aboard field could not be used reliably, passenger/ferry involvement may partly proxy the number of persons exposed to harm within an accident. Neither that association nor the inverse GT association can be translated into fleet-wide accident incidence without appropriate traffic or vessel-time denominators; passenger-ship safety studies using exposure concepts address a different estimand [30].

The methodological result is at least as important as the association estimates. The fixed historical equation retained good discrimination yet substantially under predicted later events. Post hoc time diagnostics showed that the original spline's extrapolation amplified this problem, but they did not reduce the issue to a single modelling mistake: a linear term still under predicted, and regular expanding-origin refitting improved average calibration more effectively in later blocks. This is consistent with the general prediction literature's warning that population drift, measurement changes, and model specification can all affect calibration, and that model updating may be required [19]. The TSB's 2013-2014 system and reporting transition provides a plausible administrative discontinuity, but the data do not identify how much of the temporal pattern reflects reporting, fleet composition, safety practice, case mix, or unmeasured variables.

The accident-type-only benchmark clarifies what the full model contributes. Accident type alone was better calibrated on average in the later era, but the full model materially improved AUC, PR-AUC, and Brier score under paired component resampling. Thus, the choice is not between a 'good' simple model and a 'bad' full model. The fuller score carries additional information for ranking accidents, while its absolute probability scale requires updating. This is precisely the situation in which reporting only discrimination would be incomplete [19,21].

Finally, the dependence analyses support a more qualified interpretation of validation. Chronological splitting is preferable to random splitting for later-period transport, but recurrent vessel identifiers mean that time alone does not remove every hierarchical connection. Whole-component resampling accounted for recurrent-vessel dependence in performance uncertainty while leaving the core conclusion intact. This aligns with structured-validation work showing that validation design should reflect the dependence structure relevant to the intended prediction target [20].

5. PRACTICAL MARITIME RISK-MANAGEMENT IMPLICATIONS

The results are most defensible as a retrospective consequence-risk layer for safety review, scenario planning, and research prioritization. They are not a pre-accident incidence model and do not establish that all retained fields are available at an operational decision time. In the later sample, risk was nevertheless strongly concentrated: 55.9% of events fell within the highest-risk 11.8% of accidents. That concentration may help identify scenarios deserving deeper review, but no decision-curve, cost function, or resource-allocation threshold was evaluated.

The substantive signals support attention to stability and loss-of-buoyancy scenarios, sea-state conditions, and the response implications of passenger/ferry involvement, while not estimating the effectiveness of any particular intervention. Canadian fishing-policy work illustrates how environmental evidence can be translated into stability, weather-information, training,

Journal of Marine Technology and Environment
DOI:10.53464/JMTE.02.2026.04
ISSN (Print): 1844-6116
ISSN (Online): 2501-8795
https://www. jmte.eu

communication, and search-and-rescue planning questions [18]. Within this journal, Demirci and Cicek similarly used data-driven risk-area prioritization to focus port-state-control inspections by ship characteristics [31]. The present model can inform that kind of prioritization logic only as a research signal: operational use would require a clearly defined prediction time, a predictor-availability audit, contemporary fitting, prospective evaluation, and an explicit utility framework.

6. LIMITATIONS

Several limitations constrain the inference. First, MARSIS is occurrence-centred and provides no corresponding population of safe voyages or vessel-time exposure [3]. The study cannot estimate accident incidence. Accidents aboard ship were excluded because their definition incorporates human harm, and the binary composite endpoint gives equal occurrence-level weight to one serious injury and multiple fatalities. The fatal/missing sensitivity partially addresses, but does not eliminate, that loss of severity detail.

Second, the public records span more than three decades of changes in reporting, taxonomy, fleet composition, safety practice, and field completeness. Explicit unknown categories preserve information but are not substitutes for measured physical conditions. The post hoc conditional GT sensitivity is assumption-dependent deterministic imputation, not multiple imputation, and the missing-data literature cautions against interpreting missing indicators as unbiased confounder adjustment in observational settings [25]. The GT=301,287 ferry value remains unresolved; retaining the raw value in the baseline and isolating it in sensitivity avoids inventing an unsupported correction.

Third, the structured public fields omit many human, organizational, management, and response-context factors available in detailed investigation reports [10,12,16]. Residual confounding and measurement error are therefore plausible, and adjusted associations are not causal effects. The all-period standardized probabilities describe the observed 1995-July 2026 covariate distribution; they are not calibrated contemporary risk estimates.

Fourth, the 2015-July 2026 period was originally held out for the historical equation but was subsequently inspected during revision. The original temporal result remains a valid record of how that pre-existing historical equation performed when first applied, but post hoc no-year, linear-year, benchmark, expanding-origin, updating, and ridge comparisons are exploratory analyses within already examined data. No language in this paper treats those comparisons as a new independent validation cohort.

Finally, this is temporal evaluation within the same national administrative system, not external-country or prospective validation. Connected-component performance intervals condition on the fitted historical equations and do not incorporate development-sample uncertainty. No decision-analytic net benefit was estimated. These constraints preclude operational probability thresholds while leaving the observed association patterns and relative risk concentration scientifically informative.

7. CONCLUSIONS

Across 9,491 TSB-reportable shipping accidents from 1995 through July 2026, serious human consequences were uncommon overall but concentrated in capsize, sinking/loss/abandonment, and moderate or adverse sea-state contexts. Passenger/ferry involvement was also associated with higher conditional consequence risk, whereas grounding and secured/stationary status were associated with lower risk. These are observational, occurrence-conditional associations rather than estimates of accident incidence or causal intervention effects.

A historical multivariable model retained useful discrimination in later accidents and materially outperformed an accident-type-only benchmark for ranking, but its absolute probabilities under predicted later events. Post hoc analyses showed that calendar-time extrapolation contributed to the discrepancy and that refitting with more contemporary data improved average calibration in later forward blocks. The strongest interpretation is therefore not that one fixed equation has been operationally validated, but that transparent consequence profiling can identify stable relative risk concentration while absolute probability models require ongoing temporal evaluation and updating.

8. AVAILABILITY AND DECLARATIONS

8.1 Data and code availability

The MARSIS public marine occurrence data and data dictionary are available from the Transportation Safety Board of Canada [3]. The analysis code and supporting numerical outputs are available from the author upon reasonable request. Analyses used R 4.2.2 with data.table 1.17.6, ggplot2 4.0.0, pROC 1.19.0.1, PRROC 1.4, and logistf 1.26.1 where required.

8.2 Ethics statement

The study used publicly released administrative occurrence records from which the TSB excludes third-party, personal, privileged, and other protected information [3]. No interaction with human participants occurred; institutional review board approval and informed consent were not applicable.

8.3 Funding

This research received no external funding.

8.4 Declaration of interest

The author declares no competing financial or personal interests that could have influenced the work reported in this paper.

9. CREdit authors statement DATA

Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Visualization; Writing - original draft; Writing - review and editing; Project administration - David O. Kremelberg.

10. REFERENCES

[1] Huang, X., Wen, Y., Zhang, F., Han, H., Huang, Y., & Sui, Z., 2023, *A review on risk assessment methods for maritime transport*, Ocean Engineering, 279, 114577. https://doi.org/10.1016/j.oceaneng.2023.114577

[2] Jin, D., Kite-Powell, H., & Talley, W. K., 2001, *The safety of commercial fishing: Determinants of vessel total losses and injuries*, Journal of Safety Research, 32(2), 209-228. https://doi.org/10.1016/S0022-4375(01)00047-0

[3] Transportation Safety Board of Canada, 2026, Marine transportation occurrence data from January 1995: MARSIS public dataset and data dictionary. https://www.tsb.gc.ca/eng/stats/marine/data-6.html (accessed 3 September 2026).

[4] Weng, J., & Yang, D., 2015, *Investigation of shipping accident injury severity and mortality*, Accident Analysis & Prevention, 76, 92-101. https://doi.org/10.1016/j.aap.2015.01.002

[5] Talley, W. K., Jin, D., & Kite-Powell, H., 2005, *Determinants of crew injuries in vessel accidents*, Maritime Policy & Management, 32(3), 263-278. https://doi.org/10.1080/03088830500139760

[6] Wang, H., Liu, Z., Wang, X., Huang, D., Cao, L., & Wang, J., 2022, *Analysis of the injury-severity outcomes of maritime accidents using a zero-inflated ordered probit model*, Ocean Engineering, 258, 111796. https://doi.org/10.1016/j.oceaneng.2022.111796

[7] Wang, H., Liu, Z., Wang, X., Graham, T., & Wang, J., 2021, *An analysis of factors affecting the severity of marine accidents*, Reliability Engineering & System Safety, 210, 107513. https://doi.org/10.1016/j.ress.2021.107513

[8] Wang, L., & Yang, Z., 2018, *Bayesian network modelling and analysis of accident severity in waterborne transportation: A case study in China*, Reliability Engineering & System Safety, 180, 277-289. https://doi.org/10.1016/j.ress.2018.07.021

[9] Li, B., Lu, J., & Li, J., 2021, *Investigation of accident severity in sea lanes from an emergency response perspective based on data mining technology*, Ocean Engineering, 239, 109920. https://doi.org/10.1016/j.oceaneng.2021.109920

[10] Fan, S., Blanco-Davis, E., Yang, Z., Zhang, J., & Yan, X., 2020, *Incorporation of human factors into maritime accident analysis using a data-driven Bayesian network*, Reliability Engineering & System Safety, 203, 107070. https://doi.org/10.1016/j.ress.2020.107070

[11] Cao, Y., Wang, X., Wang, Y., Fan, S., Wang, H., Yang, Z., Liu, Z., Wang, J., & Shi, R., 2023, *Analysis of factors affecting the severity of marine accidents using a data-driven Bayesian network*, Ocean Engineering, 269, 113563. https://doi.org/10.1016/j.oceaneng.2022.113563

[12] Liu, D., Cheng, L., Ai, B., Xie, R., Wang, X., & Shang, H., 2026, *Determinants of casualty severity in maritime accidents: Evidence from text mining of 581 investigation reports*, 2014-2023, Ocean Engineering, 351, 124456. https://doi.org/10.1016/j.oceaneng.2026.124456

[13] Khan, R. U., & Yang, Z., 2026, *Maritime accident severity modelling with operationally realistic multi-factor data-driven Bayesian networks*, Reliability Engineering & System Safety, 275, 112823. https://doi.org/10.1016/j.ress.2026.112823

[14] Duan, N., Guan, W., Wang, X., & Li, L., 2026, *An explainable ensemble learning framework for maritime accident severity and type identification: Multi-method feature selection and SHAP*

interpretation, Ocean Engineering, 363, 126776. https://doi.org/10.1016/j.oceaneng.2026.126776

[15] Cao, W., Wang, X., Feng, Y., Zhou, J., & Yang, Z., 2026, *Improving maritime accident severity prediction accuracy: A holistic machine learning framework with data balancing and explainability techniques*, Reliability Engineering & System Safety, 266, 111648. https://doi.org/10.1016/j.ress.2025.111648

[16] Shi, X., Zhuang, H., & Xu, D., 2021, *Structured survey of human factor-related maritime accident research*, Ocean Engineering, 237, 109561. https://doi.org/10.1016/j.oceaneng.2021.109561

[17] Rezaee, S., Pelot, R., & Finnis, J., 2016, *The effect of extratropical cyclone weather conditions on fishing vessel incidents' severity level in Atlantic Canada*, Safety Science, 85, 33-40. https://doi.org/10.1016/j.ssci.2015.12.006

[18] Rezaee, S., Brooks, M. R., & Pelot, R., 2017, *Review of fishing safety policies in Canada with respect to extreme environmental conditions and climate change effects*, WMU Journal of Maritime Affairs, 16(1), 1-17. https://doi.org/10.1007/s13437-016-0110-z

[19] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W., 2019, *Calibration: The Achilles heel of predictive analytics*, BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7

[20] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al., 2017, *Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure*, Ecography, 40(8), 913-929. https://doi.org/10.1111/ecog.02881

[21] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al., 2024, *TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods*, BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

[22] Transportation Safety Board of Canada, 2015, Statistical summary - Marine transportation occurrences in 2014. https://www.tsb.gc.ca/eng/stats/marine/2014/ssem-ssmo-2014.html (accessed 3 September 2026).

[23] Transportation Safety Board Regulations, SOR/2014-37 (Canada), current consolidated regulations. https://laws-lois.justice.gc.ca/eng/regulations/SOR-2014-37/ (accessed 3 September 2026).

[24] Transportation Safety Board of Canada, 2026, Table 1. Marine occurrences (types, vessels and losses involved), July 2026. https://www.tsb.gc.ca/eng/stats/marine/2026/07/2026-07-t1.html (accessed 3 September 2026).

[25] Groenwold, R. H. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. M., 2012, *Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis*, CMAJ, 184(11), 1265-1269. https://doi.org/10.1503/cmaj.110977

[26] Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A., 2017, *Firth's logistic regression with rare events: Accurate effect estimates and predictions?*, Statistics in Medicine, 36(14), 2302-2317. https://doi.org/10.1002/sim.7273

[27] Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G. M., & Collins, G. S., 2019, *Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcome*s, Statistics in Medicine, 38(7), 1276-1296. https://doi.org/10.1002/sim.7992

[28] E.Y.E. Marine Consultants, n.d., Henry Nace / William Pitt II / RGL Fairweather, vessel particulars. https://www.eyemarine.com/vessel/henry-nace-william-pitt-ii-rgl-fairweather/ (accessed 5 September 2026).

[29] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W., 2010, *Assessing the performance of prediction models: A framework for traditional and novel measures*, Epidemiology, 21(1), 128-138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

[30] Eliopoulou, E., Alissafaki, A., & Papanikolaou, A., 2023, *Statistical analysis of accidents and review of safety level of passenger ships*, *Journal of Marine Science and Engineering*, 11, 410. https://doi.org/10.3390/jmse11020410

[31] Demirci, S. M. E., & Cicek, K., 2022, *Deficiency analysis identified in PSC inspections using event tree analysis*, Journal of Marine Technology and Environment, 2022(1), 40-46. https://doi.org/10.53464/JMTE.01.2022.05