

Article

Uncertainty-Aware Short-Horizon Warning of Large-Inclination Exceedance in Small Fishing Vessels: A Simulation-Based Multi-Model Benchmark

Hyungtak Joo ¹, Byoungchul Song ², Kyoungwon Park ², Kiwon Kwon ² and Taeho Im ^{1,*} 

¹ Division of Information and Telecommunication Engineering, Hoseo University, 20 Hoseo-ro 79beon-gil, Baebang-eup, Asan-si 31499, Chungcheongnam-do, Republic of Korea; 375jdw@gmail.com

² Korea Electronics Technology Institute, 43 Changeop-ro, Sujeong-gu, Seongnam-si 13449, Gyeonggi-do, Republic of Korea; songbc@keti.re.kr (B.S.); kwpark@keti.re.kr (K.P.); kwonkw@keti.re.kr (K.K.)

* Correspondence: taehoim@hoseo.edu; Tel.: +82-41-540-9642

Abstract

A large heel can develop within seconds on a small fishing vessel, so a short-horizon forecast of inclination is useful for safety only if it reports both the predicted angle and its uncertainty and is turned into an explicit warning decision. We present an uncertainty-aware early-warning benchmark and decision layer for the total inclination angle, framing the task as warning of *large-inclination exceedance*—that the heel will cross a fixed operational threshold ($15^\circ/20^\circ/25^\circ$)—rather than predicting a vessel-specific capsizing or dynamic-stability limit. Models are trained and evaluated on 90 time series generated from 6-DOF simulations spanning five tonnages, three sea states, and six independent wave-phase realizations under a leakage-safe, multi-seed protocol. The strongest in-distribution forecaster (an LSTM) reaches $R^2 = 0.677$ (RMSE 3.51°) pooled over the 1–5 s horizon, and converting its predictive distribution into a probabilistic exceedance alarm lowers the event-level false-alert burden at matched recall; the receiver-operating advantage is threshold-specific—clear at 15° , marginal at 20° , and reversed at the rarest 25° —so the alarm is not uniformly better than a point alarm. Leave-one-realization-out folds confirm that forecasting is robust to wave phase, whereas the smallest (10 t) vessel remains the dominant generalization failure mode. Because nominal interval coverage drops in the large-inclination tail, a conformal recalibration layer is required at deployment. The study is a simulation benchmark: operational value remains conditional on validation against measured seakeeping data and on mapping the thresholds to vessel-specific stability criteria.

Keywords: fishing vessel safety; large-inclination exceedance; inclination-threshold early warning; uncertainty quantification; calibration; time-series forecasting; deep learning; generalization



Academic Editors: Wengang Mao,
Lianzhong Huang and Jinlu Sheng

Received: 9 June 2026

Revised: 25 June 2026

Accepted: 25 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Capsizing remains a leading cause of loss of life on small fishing vessels [1,2]. Unlike large ships, small vessels respond rapidly to wave forcing: a large heel can develop within a few seconds, leaving the crew little time to react. Short-horizon forecasting of the vessel's dynamic inclination, together with a timely *warning* when the heel is about to cross a large-inclination threshold, could eventually support operational decision making, after validation on measured data and the mapping of those thresholds to vessel-specific

stability criteria. We treat the fixed thresholds used here ($15^\circ/20^\circ/25^\circ$) as *operational* large-inclination exceedance levels, not as vessel-specific capsizes or dynamic-stability limits; the capsizes motivation sets the safety context, but the prediction target throughout is threshold exceedance of the measured heel.

However, three gaps limit the engineering value of point forecasts for small-vessel inclination. First, they produce predictions with no measure of confidence; a safety system that cannot say *how certain* a forecast is cannot rationally trade off missed warnings against false alarms. Second, the forecast is rarely turned into an explicit *decision*: the link from a predicted angle to a calibrated, evaluated early-warning alarm (with detection rate, false-alarm rate, and warning lead time) is missing. Third, evaluations typically rely on a single model and a single operating condition, so it is unclear how the approach compares to standard baselines or how well it transfers to unseen sea states and vessels.

We address these gaps with an uncertainty-aware early-warning framework for the total inclination angle $\theta_{\text{TIA}} = \sqrt{\phi^2 + \theta_p^2}$ of small fishing vessels. The novelty is not a new forecasting architecture—the component models and uncertainty methods are standard—but the path from a calibrated *predictive distribution* to an evaluated *warning decision*: we convert short-horizon predictive distributions into threshold-specific exceedance probabilities and score the resulting alarm at the operational event level, with explicit false-alert and lead-time accounting, under leakage-safe, conditional-coverage, and out-of-distribution protocols. The framework is trained and evaluated on 90 series of 6-DOF simulation spanning five tonnages (10–50 t), three sea states ($H_s \in \{3, 5, 7\}$ m), and six independent wave-phase realizations, so that wave-phase variability is sampled and tested rather than assumed away. Specifically:

- **C1 (Probabilistic exceedance alarm with event-level evaluation).** We convert the predictive distribution into a probabilistic exceedance alarm $P(\theta_{\text{TIA}}(t+k) > \theta_d) \geq \alpha$ and score it against the conventional point alarm by ROC/PR and, at the operational event level, false-alert episodes and warning lead time with a leakage-free operating point—reporting the alarm’s advantage threshold by threshold rather than as a blanket claim.
- **C2 (Predictive uncertainty, calibration, and recalibration).** On the selected backbone we obtain MC-Dropout [3], deep-ensemble [4], and quantile-regression [5] distributions, assess them with calibration and sharpness metrics, and show that nominal coverage fails in the large-inclination tail, requiring a conformal recalibration layer.
- **C3 (Leakage-safe multi-model benchmark).** We compare seven forecasters (persistence, AR, Kalman, LSTM [6], GRU [7], TCN [8], and a light Transformer [9]) plus two window-regression baselines under one leakage-safe, multi-seed protocol, isolating the value of temporal modeling.
- **C4 (Generalization and failure diagnosis).** We quantify transfer with leave-one- H_s -out, leave-one-vessel-out, and leave-one-realization-out folds—finding forecasting robust to wave phase but failing on the smallest vessel—and trace part of that failure to a hydrostatic (GM -floor) artifact.
- **C5 (Target representation).** We compare predicting θ_{TIA} directly against reconstructing it from (ϕ, θ_p) , and give a closed-form, Monte-Carlo-validated error-propagation analysis explaining the reconstruction penalty.

The remainder of the paper is organized as follows. Section 2 reviews related work; Section 3 formulates the problem and details the materials and methods (data, curation, models, uncertainty quantification, and the early-warning decision layer); Section 4 reports results; Section 5 discusses findings and limitations; and Section 6 concludes.

2. Related Work

We organize prior work into three threads that the present study draws together: (i) ship-stability assessment and operational fishing-vessel warning, which sets the safety context but does not forecast the operating inclination; (ii) data-driven ship-motion forecasting, which supplies the forecasting backbones but stops short of a calibrated warning decision; and (iii) uncertainty quantification and distribution-free recalibration in safety-critical forecasting, which provides the tools needed to turn a forecast into a trustworthy exceedance alarm.

2.1. Ship Stability and Operational Fishing-Vessel Warning

Intact-stability assessment is classically governed by hydrostatic righting-arm (GZ) curves and regulatory criteria such as the IMO weather criterion [10], while time-domain vessel motion is simulated with 6-DOF hydrodynamic models [11]. These tools characterize stability but do not provide short-horizon *predictions* of the instantaneous inclination during operation. Machine-learning surrogates have recently been used to accelerate the dynamic-stability calculation itself [12], but they target the stability assessment rather than short-horizon forecasting of the operating inclination. This motivates a data-driven sequence-forecasting layer that can be benchmarked, calibrated, and converted into an explicit early-warning decision. The closest operational fishing-vessel warning study in this journal is the integrated monitoring system of Yang et al. [13], which combines a nine-axis attitude sensor, environmental sensors, GPS, camera data, wireless communication, and a fuzzy-logic danger score to trigger alerts during field trials. That study provides measured-system validation; the present work is narrower in deployment readiness but more explicit about forecast uncertainty, false-alarm operating points, and lead-time accounting.

2.2. Data-Driven Ship-Motion Forecasting

Ship-motion forecasting is an active area rather than an unexplored one. Chen et al. [14] used LSTM networks to predict measured heave, pitch, and roll over an 8 s horizon and infer resting periods for shipborne-helicopter operations. Li et al. [15] studied short-term roll prediction with an encoder–decoder Bi-LSTM and examined the effects of training data and sea state. Lee and Ahn [16] compared RNN-family models on simulated 6-DOF ship motions and reported the effect of input dimensionality, downsampling, model scale, and peak-focused metrics. Convolutional and attention-based architectures have likewise been applied to short-term roll [17] and pitch [18] prediction, and quaternion-based attitude representations have been studied with a Transformer–LSTM model to avoid the limitations of Euler angles [19]. The underlying sequence models are standard: LSTM [6] and GRU [7] recurrent networks for motion time series, temporal convolutional networks [8] for long causal receptive fields without recurrence, and Transformer encoders [9] for long-range dependence via attention. We include all four—together with persistence and classical AR/Kalman baselines—so that the marine-forecasting gains are reported against a common, transparent reference rather than a single architecture. What these studies share, and what the present work adds, is the step *after* the forecast: none converts a calibrated predictive distribution into a threshold-specific probabilistic exceedance alarm scored at the event level.

2.3. Uncertainty Quantification and Recalibration for Safety-Critical Forecasting

Predictive uncertainty in deep learning is commonly obtained from MC-Dropout [3], deep ensembles [4], or direct quantile regression [5], with a well-established epistemic/aleatoric decomposition [20]; quality is judged with proper scoring rules and calibration diagnostics—PICP, interval width, reliability diagrams, and the CRPS [21]—and with

classification-style calibration error [22]. Uncertainty has begun to appear in ship-motion prediction specifically: Wei et al. [23] used an improved quantile recurrent network for probabilistic pitch prediction from sea-trial data; Kim et al. [24] applied MC-Dropout to LSTM and Transformer roll predictors and validated it on ferry voyage data; and Li et al. [25] reported bootstrap uncertainty for a WT-A-LSTM-KAN roll/pitch model at 5, 8, and 10 s horizons. A recurring gap is that marginally calibrated intervals need not hold *conditionally*—least of all in the rare large-inclination tail that an exceedance alarm depends on. Distribution-free conformal prediction [26] and conformalized quantile regression [27] restore finite-sample coverage guarantees without distributional assumptions, which makes them the natural recalibration layer for a deployed warning system; we adopt them for exactly this role. Table 1 summarizes how the present manuscript is positioned relative to the closest studies.

Table 1. Positioning against closely related ship-motion forecasting and fishing-vessel warning studies. The table is intended to delimit the present contribution, not to claim that the component methods themselves are new.

Study	Data and Target	Warning/UQ Element	Position of the Present Study
Yang et al. [13]	Fishing-vessel field system with attitude/environment sensing.	Fuzzy-logic danger score and real-time alerts.	Closest operational analogue in the target journal; our work instead studies short-horizon probabilistic forecasts, ROC/PR operating points, and lead-time accounting but lacks field validation.
Chen et al. [14]	Measured heave, pitch, and roll for an 8 s LSTM resting-period forecast.	Deterministic real-time forecast for shipborne-helicopter operations.	Demonstrates measured-data feasibility for multi-DOF forecasting; our target is total inclination and calibrated exceedance probability on a small-vessel simulation grid.
Li et al. [15]	Short-term roll prediction with an encoder–decoder Bi-LSTM.	Deterministic roll forecast; analyzes training-data and sea-state effects.	Provides a strong roll-forecasting reference; our study adds multi-model baselines, UQ, and alarm metrics.
Lee and Ahn [16]	Simulated 6-DOF ship motions in multiple environments.	Deterministic RNN-family comparison and peak-oriented metrics.	Most similar benchmark setting; our focus is smaller vessels, probabilistic warning, transfer folds, and target-representation error propagation.
Kim et al. [24]	Roll prediction from HydroD–Wasim simulations with ferry voyage validation.	MC-Dropout uncertainty for LSTM and Transformer models.	Directly overlaps with UQ; our added layer is threshold-specific exceedance alarming and comparison of multiple uncertainty estimators.
Li et al. [25]	Moored-vessel roll/pitch prediction at 5, 8, and 10 s.	Bootstrap uncertainty for a WT-A-LSTM-KAN hybrid model.	Shows recent short-horizon UQ progress; our contribution is a calibrated decision layer and reproducible benchmark.
Yan et al. [19]	Short-term ship-motion forecasting with quaternion attitude representation.	Representation study with a Transformer–LSTM hybrid.	Motivates future quaternion vs Euler tests; our current C5 is narrower, treating GZ as a deterministic consistency check and analyzing reconstruction error.

3. Materials and Methods

3.1. Target Definition and Simulation Dataset

We forecast the *total inclination angle*, defined from the roll (ϕ) and pitch (θ_p) angles as

$$\theta_{\text{TIA}} = \sqrt{\phi^2 + \theta_p^2}, \quad (1)$$

which we verified holds on the dataset to within 5×10^{-7} rad. We treat θ_{TIA} as a *simulator-defined inclination magnitude*: it is the quantity the simulator records and the target we forecast. It is not identical to the geometric tilt of the vessel's vertical axis from vertical, $\arccos(\cos \phi \cos \theta_p)$; the two agree to within 0.6° for inclinations up to 25° and differ by about 2.5° near 40° , so the small-angle quadrature form is an accurate proxy in the operational regime studied here, while mapping to physical deck tilt at very large angles is left to future work. The simulator's righting-arm proxy is

$$GZ = GM \cdot \sin(\theta_{\text{TIA}}), \quad (2)$$

where the metacentric height GM is constant per vessel. Equation (2) is therefore a *deterministic, monotone* transform of θ_{TIA} (it holds in the data to machine precision); we treat GZ as a re-parameterization of the target throughout, never as an independent signal (Section 3.9).

The data comprise 90 time series from a 6-DOF time-domain simulator: the five vessel tonnages (10, 20, 30, 40, 50 t) and three significant wave heights ($H_s \in \{3, 5, 7\}$ m) give 15 conditions, and each condition is generated for *six independent wave-phase realizations* ($r_0, \dots, r_5$). Realization r_0 carries the original published random-phase seed; r_1 – r_5 draw new, independent phase sets at the same spectrum, so the corpus samples the wave-phase variability of each condition rather than a single deterministic draw. Each series runs at 1 Hz for $N = 5001$ steps (eighty-seven of the ninety series are full length; see below), giving 441,562 rows in total (per-realization row counts are released in `T2_realization_counts.csv`). Every series provides the eight kinematic variables $\{u, v, w, p, q, r, \phi, \theta_p\}$ (surge/sway/heave velocities, roll/pitch/yaw rates, and roll/pitch angles) from the 6-DOF state, together with θ_{TIA} and GZ . Table 2 lists the vessel dimensions and metacentric heights, and Table 3 summarizes the per-condition inclination statistics and exceedance-threshold frequencies (averaged over the six realizations). Figure 1 shows the inclination distribution. Exceedance is dominated by sea state: exceedances above 15° rise from below 2% at $H_s = 3$ m to roughly 25% at $H_s = 7$ m, giving the class imbalance that makes precision–recall analysis necessary alongside ROC.

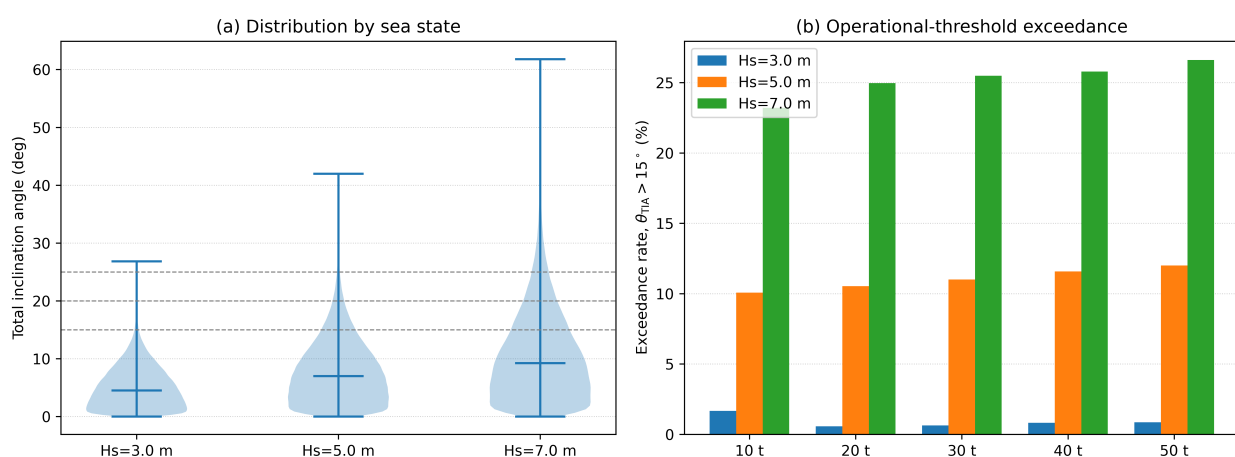


Figure 1. Distribution of the total inclination angle (degrees). (a) by sea state, with the $15^\circ/20^\circ/25^\circ$ exceedance thresholds; (b) exceedance rate above 15° by tonnage and sea state.

Table 2. Principal dimensions of the five vessels and the metacentric height GM recomputed from the dynamics-consistent dimensions (Munro–Smith hydrostatics; $C_B = 0.68$, $C_W = 0.80$, $KG = 0.60$ T). Only the 10 t vessel’s unclamped GM (0.258 m) falls below the simulator’s $GM_{\min} = 0.30$ m floor and is clamped.

Tonnage (t)	L (m)	B (m)	T (m)	GM (m)	Note
10	7.93	1.90	0.95	0.300	clamped ($GM_{\text{raw}} = 0.258$)
20	9.99	2.40	1.20	0.326	—
30	11.43	2.74	1.37	0.372	—
40	12.58	3.02	1.51	0.410	—
50	13.56	3.25	1.63	0.440	—

Table 3. Dataset summary by tonnage and significant wave height H_s , averaged over the six wave-phase realizations (r_0 – r_5) of each condition (90 series in total). Each series has $N = 5001$ steps at 1 Hz, except the three 10 t/ $H_s = 7$ m series truncated by integrator divergence in the large-roll regime (87/90 series full); statistics use all valid steps. θ_{TIA} statistics are in degrees and the last three columns give the strict exceedance frequency $P(\theta_{\text{TIA}} > \theta_d)$. Values are mean $\pm$ std across realizations (the std on mean/std is $\leq 0.1^\circ$ and omitted). Exceedance is driven by sea state ($<2\%$ above 15° at $H_s = 3$ m, rising to $\approx 25\%$ at $H_s = 7$ m).

Ton (t)	H_s (m)	Mean ($^\circ$)	Std ($^\circ$)	Max ($^\circ$)	$>15^\circ$ (%)	$>20^\circ$ (%)	$>25^\circ$ (%)
10	3	7.36	3.34	21.01 ± 0.57	1.66 ± 0.20	0.04 ± 0.01	0.00 ± 0.00
10	5	9.08	4.47	31.96 ± 5.16	10.06 ± 0.34	1.82 ± 0.13	0.21 ± 0.04
10	7	11.13	5.88	43.79 ± 8.98	23.35 ± 0.37	8.06 ± 0.43	2.40 ± 0.29
20	3	4.53	3.25	19.68 ± 1.40	0.57 ± 0.13	0.01 ± 0.02	0.00 ± 0.00
20	5	7.46	5.45	34.56 ± 2.75	10.53 ± 0.15	2.92 ± 0.17	0.64 ± 0.07
20	7	10.44	7.60	48.87 ± 4.26	24.97 ± 0.75	12.10 ± 0.18	5.13 ± 0.21
30	3	4.48	3.35	20.33 ± 0.99	0.63 ± 0.08	0.04 ± 0.03	0.00 ± 0.00
30	5	7.49	5.58	36.13 ± 2.74	11.00 ± 0.17	3.12 ± 0.26	0.61 ± 0.11
30	7	10.52	7.85	49.28 ± 1.60	25.49 ± 0.34	13.07 ± 0.23	5.74 ± 0.29
40	3	4.55	3.40	21.52 ± 0.73	0.82 ± 0.13	0.04 ± 0.02	0.00 ± 0.00
40	5	7.61	5.68	33.76 ± 1.01	11.57 ± 0.26	3.48 ± 0.27	0.73 ± 0.15
40	7	10.57	8.00	50.27 ± 2.75	25.79 ± 0.54	13.05 ± 0.23	5.85 ± 0.12
50	3	4.60	3.47	20.95 ± 2.90	0.86 ± 0.12	0.02 ± 0.03	0.00 ± 0.01
50	5	7.75	5.74	35.83 ± 2.94	12.00 ± 0.18	3.64 ± 0.25	0.79 ± 0.04
50	7	10.79	8.13	52.51 ± 3.62	26.61 ± 0.53	13.88 ± 0.32	6.53 ± 0.27

Three of the ninety series—all in the most violent cell (10 t at $H_s = 7$ m)—terminated early when the time-domain integrator diverged in the large-roll regime (r_1 retained 4182 rows, r_3 643 rows, and r_4 1650 rows; the remaining 87 series are full 5001-row records). We retained the truncated portions rather than discarding them: the divergence is genuine simulator behavior at extreme inclination, not a data defect, and it is consistent with the 10 t cell being the hardest case for forecasting (Section 4.4). Windowing (Section 3.4) was applied per series, so the truncations only reduced the number of windows drawn from those three series.

3.2. Operational Exceedance Thresholds Versus Stability Limits

The forecast target is exceedance of the fixed inclination levels $\theta_d \in \{15^\circ, 20^\circ, 25^\circ\}$, evaluated with the strict comparator $\theta_{\text{TIA}}(t+k) > \theta_d$ (the same comparator behind the frequencies in Table 3). These are *operational*, *percentile-based* levels for this dataset, chosen to span the rare-event tail, not vessel-specific capsizes or dynamic-stability-loss criteria. They are distinct from IMO intact-stability criteria, the downflooding angle, and the true hydrostatic GZ curve—none of which the single-angle, constant- GM simulator resolves.

We therefore report a *large-inclination exceedance warning*, not a capsize warning, throughout; mapping these thresholds to vessel-specific stability limits requires full hydrostatics and is left to future work. The strong class imbalance is why precision–recall is reported alongside ROC.

3.3. Simulation Configuration

The series were generated with the Marine Systems Simulator (MSS) fishing-vessel model under irregular wave forcing; the released code regenerated the dataset end-to-end. Waves follow a modified Pierson–Moskowitz spectrum (100 frequency components up to 3 rad/s) at the three significant wave heights $H_s \in \{3, 5, 7\}$ m, with a *fixed* modal period $T_0 = 8$ s (peak frequency $\omega_0 = 2\pi/T_0$) and a single wave heading of 140° without directional spreading. Each (tonnage, H_s) condition was generated for six independent random-phase realizations $r_0, \dots, r_5$ of this spectrum: r_0 reproduced the original published seed, and r_1 – r_5 drew new, statistically independent phase sets at the same significant wave height and modal period. The realizations therefore share the encounter geometry and spectral shape while differing only in the random phases of the wave components, isolating wave-phase variability as a controlled axis. The vessel was simulated at zero forward speed from rest, so the output is the wave-induced 6-DOF response at a fixed encounter geometry; angle wrapping and divergence guards were enabled for numerical stability at the 1 Hz step. Because vessel-specific response-amplitude operators were not available for these hulls, a simplified wave-force model driven by the wave elevation was used; consequently absolute motion magnitudes are simulator-specific and should be read as relative rather than field-calibrated values. The metacentric height was computed from the precise dimensions with the simulator’s Munro–Smith hydrostatics and floored at $GM_{\min} = 0.30$ m (reached only by the 10 t hull), after which θ_{TIA} and the proxy GZ follow from Equations (1) and (2).

This configuration fixes the spectral peak period, heading, speed, and loading condition, so the 90 series sample tonnage, H_s , and wave phase but not encounter angle, speed, or loading condition. Wave-phase variability was then sampled across the six realizations and explicitly tested with a leave-one-realization-out protocol (Section 3.8), under which forecasting skill is essentially unchanged from the in-distribution baseline (Section 4.4); the generalization difficulty that remains is driven by vessel and sea state, not by wave phase. We accordingly read the present results as a rigorous simulation benchmark over this simulator population and identify varied headings/speeds/loading, RAO-based forcing, and validation against measured sea-trial data as the priority extensions (Section 5).

3.4. Data Curation, Windowing, and Splits

Two issues were corrected during curation. First, the metacentric height was recomputed from the dynamics-consistent (precise) vessel dimensions using the simulator’s Munro–Smith hydrostatics ($C_B = 0.68$, $C_W = 0.80$, $KG = 0.60$ T); the only vessel affected by the simulator’s $GM_{\min} = 0.30$ m floor is the 10 t case (unclamped $GM = 0.258$ m), which we report transparently in Table 2. The original GZ column had been computed from rounded filename dimensions, introducing a $\sim 7\%$ inconsistency at 30 t; we adopted the dynamics-consistent GM and recomputed GZ from Equation (2). Second, we found the 1 Hz inclination signal to have low short-lag autocorrelation ($\text{lag-1} \approx 0.16$ – 0.24 , rising with sea state), so a persistence baseline attains a *negative* overall R^2 (reported in full in Section 4.1). A signal-diagnostics analysis (Supplementary Figure S1) confirms this is a genuine property of the target rather than a sampling or target-construction artifact: the raw roll ϕ peaks near the 7 s wave period (the modal period is $T_0 = 8$ s), well below the 0.5 Hz Nyquist limit, so the 1 Hz rate does not alias the carrier, but the rectified magnitude

$\theta_{\text{TIA}} = \sqrt{\phi^2 + \theta_p^2}$ has a broadband spectrum and large step-to-step changes that produce the weak lag-1 correlation. The task is therefore non-trivial, and the learned gains are not attributable to signal smoothness.

Inputs are sliding windows of length $L = 20$ s; targets are the next $H = 5$ s, predicted jointly (direct multi-step, $t + 1, \dots, t + 5$). Windows never straddle a series boundary, so each window lies within a single tonnage- H_s -realization series. For the in-distribution (ID) protocol, each series is split chronologically 70/15/15 with an embargo of $L + H - 1$ steps so that train, validation, and test windows share no timestep; pooling the per series test splits yields 63,880 test windows. Feature standardization is fit on the training split only. For out-of-distribution evaluation (Section 3.8) the held-out group (one sea state, one vessel, or one wave-phase realization) forms the test set in full and is entirely absent from training and validation.

3.5. Forecasting Models and Training Protocol (C3)

All models shared the inputs, windows, target scaling, optimizer, and early-stopping protocol of Table 4; only the component under study varied. Learned models were trained with the Adam optimizer and mean-squared error on per-horizon standardized targets, with early stopping on the validation loss; predictions were inverse-scaled to degrees for reporting.

Table 4. Shared experimental configuration and per-model hyperparameters. All learned models use the same windowing, target scaling, optimizer, and early-stopping protocol; only the architecture differs. Classical baselines (AR, Kalman) are closed-form and deterministic.

Setting	Value
Inputs (8 variables)	$\{u, v, w, p, q, r, \phi, \theta_p\}$ (from the 6-DOF state)
Lookback L /horizon H	20 s/5 s (predict $t + 1, \dots, t + 5$, direct multi-step)
Split (ID)	per-record chronological 70/15/15, embargo $L + H - 1$
Split (OOD)	leave-one- H_s -out, leave-one-vessel-out, leave-one-realization-out
Target scaling	per-horizon standardization, fit on train only
Optimizer/loss	Adam, MSE
Early stopping	patience 12 on validation loss, max 100 epochs
Seeds	5 ($\{0, 1, 2, 3, 4\}$) for every stochastic experiment (benchmark, UQ, early warning, generalization, OOD); mean $\pm$ std reported
Operational exceedance thresholds	$\{15^\circ, 20^\circ, 25^\circ\}$; alarm false-alarm budget $\text{FPR} \leq 0.10$
LSTM/GRU	hidden 128, 2 layers, dropout 0.3, lr 10^{-3} , batch 128
TCN	hidden 64, 4 dilated residual blocks (dilation 1, 2, 4, 8), dropout 0.1
Transformer	d_{model} 64, 2 encoder layers, 4 heads, dropout 0.1, lr 5×10^{-4}
AR	order $p = 8$, OLS fit, recursive multi-step
Kalman	local-level filter, q, R from training first differences

The analysis pipeline was implemented in Python 3.12.3 with PyTorch 2.6.0+cu124, NumPy 1.26.4, Pandas 2.3.3, SciPy 1.13.1, scikit-learn 1.8.0, PyYAML 6.0.1, Matplotlib 3.10.8, and PyArrow 23.0.1. All experiment settings were stored in YAML configuration files, and the entry point that generates each result artifact is listed in the accompanying code repository.

We benchmarked seven primary forecasters under the identical protocol of Section 3.4: (i) *persistence*, repeating the last observed value; (ii) $AR(p)$, an order-8 autoregression fit by ordinary least squares and rolled out recursively; (iii) a *local-level Kalman* filter whose process/observation variances are estimated from the training first differences; (iv) the reference *LSTM* [6] (two layers, 128 hidden units, a two-layer MLP head); (v) a *GRU* [7] of matched capacity; (vi) a *TCN* [8] of dilated causal residual blocks whose receptive field covers the lookback; and (vii) a light *Transformer* encoder [9] with sinusoidal positional

encoding. The classical AR and Kalman baselines forecast recursively from the target's own history and are deterministic (reported once, not over seeds). We additionally included two non-neural window-regression baselines, ridge regression and gradient boosting, fit on the same flattened input window as the neural models, to test whether the gains arise from temporal modeling rather than from access to the full window; together with the seven primary models these give the nine forecasters evaluated in the benchmark of Section 4.1.

3.6. Uncertainty Quantification Methods (C2)

On the shared LSTM backbone we obtained three predictive distributions. *MC-Dropout* [3] keeps dropout active at inference and aggregates T stochastic passes into a Gaussian $(\hat{\mu}, \hat{\sigma})$. *Deep ensembles* [4] use $M=5$ independently seeded models and take the across-member mean and spread. *Quantile regression* [5] trains a multi-quantile head ($\tau \in \{0.05, 0.5, 0.95\}$) with the pinball loss, repairing any quantile crossing by sorting. For the Gaussian methods the across-model (epistemic) spread captures only part of the predictive uncertainty; we therefore added an aleatoric variance term (a horizon-specific homoscedastic validation-residual variance, used as a pragmatic residual-variance proxy rather than a pure estimate of aleatoric uncertainty) and combined the two in quadrature, $\hat{\sigma}_k^2 = \hat{\sigma}_{\text{epi},k}^2 + \hat{\sigma}_{\text{alea},k}^2$ [20]. The aleatoric term was estimated *per forecast horizon*: for each horizon $k \in \{1, \dots, 5\}$ we set $\hat{\sigma}_{\text{alea},k}^2$ to the mean-squared residual of the point forecast on the validation split (a single homoscedastic scalar per horizon, in standardized target units) so that longer horizons—which are intrinsically less predictable—receive a larger residual-variance term. This horizon-resolved residual component is necessary for the Gaussian methods to be calibrated, as the epistemic spread alone substantially under-covers (Section 4.2).

Calibration and sharpness are assessed with the prediction-interval coverage probability (PICP@90, target 0.90), the mean prediction-interval width (MPIW), reliability curves (empirical vs. nominal central coverage), and the continuous ranked probability score (CRPS) [21]. For a Gaussian forecast the CRPS has the closed form

$$\text{CRPS}(\mathcal{N}(\mu, \sigma), y) = \sigma \left[z(2\Phi(z) - 1) + 2\varphi(z) - \frac{1}{\sqrt{\pi}} \right], \quad z = \frac{y - \mu}{\sigma}, \quad (3)$$

which we validated against the sample (energy) form. Quantile forecasts are available only at $\tau = \{0.05, 0.5, 0.95\}$, so their CRPS was computed as a coarse pinball-integral approximation and is used as a secondary score; interval coverage, interval width, reliability, and point RMSE are the primary criteria for comparing the three-quantile model. The quantile model feeds the early-warning layer (it is the sharpest interval producer), with the deep ensemble reported as the better-calibrated alternative.

Beyond the marginal scores we assessed three further calibration properties (Section 4.2.1). *Conditional coverage* reports PICP@90 sliced by forecast horizon, sea state, tonnage, and inclination regime (below vs. above the 15° threshold) since a marginally calibrated interval can still under-cover the operationally critical large-inclination tail. *Distribution-free recalibration* wraps the forecasters with split conformal prediction [26] (absolute-residual nonconformity on the ensemble mean) and conformalized quantile regression (CQR) [27] (conformity score $E = \max(q_{\text{lo}} - y, y - q_{\text{hi}})$ on the quantile band); the finite-sample $\lceil (n+1)(1-\alpha) \rceil / n$ nonconformity quantile is calibrated on the earlier temporal half of each test group and applied to the later half. Under exchangeability conditions this split-conformal construction yields finite-sample marginal coverage; because the present windows overlap and are serially dependent, strict exchangeability does not hold, so we interpret the result as empirical calibration on a held-out temporal split rather than as a formal finite-sample guarantee. *Exceedance-probability calibration* scores the alarm's probability $P(\theta_{\text{TIA}}(t+k) > \theta_d)$ directly with the Brier score and the expected calibration error (ECE, 10 equal-width bins) [22].

3.7. Early-Warning Decision Layer (C1)

An exceedance event at horizon k is the strict exceedance $\theta_{\text{TIA}}(t+k) > \theta_d$ for a threshold $\theta_d \in \{15^\circ, 20^\circ, 25^\circ\}$; these are operational, not vessel-specific stability, levels (Section 3.2). We compared two alarms scored against these labels: the *point* alarm uses the predicted value as score (equivalently, fires when $\hat{y} > \theta_d$), whereas the *probabilistic* alarm uses the exceedance probability $P(\theta_{\text{TIA}}(t+k) > \theta_d)$, obtained from the Gaussian survival function or, for the quantile model, from a piecewise-linear predictive CDF. The two scores differ only through the per-window uncertainty. A *decision cell* is one scored window-horizon-threshold instance; with N test windows and H forecast horizons there are NH decision cells per threshold.

Detection quality is summarized by ROC-AUC (computed via the Mann-Whitney statistic and cross-checked against a reference implementation), PR-AUC, and the best F_1 . We define the *effective horizon* within the tested range as the longest contiguous horizon, up to the designed $H = 5$ s forecast horizon, at which $\text{AUC} \geq 0.8$. Warning lead time is measured within each record's contiguous test timeline: a future timestep that exceeds θ_d is *detected* if any window/horizon pair that forecasts it fires, and its lead time is the earliest such firing; events are counted once per future timestep (deduplicated across the windows that see them), while false alarms are counted per decision cell. Alarm thresholds are selected under a common false-alarm budget ($\text{FPR} \leq 0.10$); the achieved FPR is reported because discrete thresholds can make the realized operating point differ from the budget.

We additionally evaluated the alarm at the operational *event* level with a leakage-free operating point (Section 4.3.1). The alarm threshold α is selected on the earlier temporal half of each record (the calibration partition) to meet the FPR budget, frozen, and applied only to the disjoint later half on which all event metrics are computed, so the operating point is never tuned on the evaluation data. On a record's reconstructed 1 Hz heel timeline, an *exceedance event* is a maximal run of timesteps with $\theta_{\text{TIA}} > \theta_d$ and a *false-alert episode* is a maximal run of consecutive alarm firings not associated with any event onset within the H -second lead window; counting maximal runs means one sustained alarm registers as a single episode, which is the operator-facing alert count. We report the event detection rate (recall) and its complement the missed-event rate, false-alert episodes per hour, event-level precision (true-alert episodes over all episodes), and lead-time quantiles, pooled across records. We additionally swept the false-alarm budget and an optional alert-suppression (refractory) period as a sensitivity analysis, reported in Table S7.

3.8. Generalization Protocols (C4)

We measured transfer to unseen conditions with three leave-one-out protocols. Leave-one- H_s -out (LOHO, three folds) trained on two sea states and was tested on the third across all vessels; leave-one-vessel-out (LOVO, five folds) trained on four tonnages and was tested on the fifth across all sea states; and leave-one-realization-out (LORO, six folds) held out one of the six wave-phase realizations $r_0, \dots, r_5$ at a time, training on the other five across all tonnages and sea states. LORO isolates wave-phase generalization—the held-out fold shares every vessel and sea state with training and differs only in the random phases of the wave forcing—whereas LOHO and LOVO probe extrapolation to unseen sea states and vessels. In every fold the held-out group was entirely absent from training and validation, which we verified programmatically. Degradation is reported as ΔR^2 relative to the ID baseline.

3.9. Target Representation and Error-Propagation Methods (C5)

We compared three ways to obtain a θ_{TIA} forecast, all scored in the same θ_{TIA} space: predicting it *directly*; predicting (ϕ, θ_p) and *reconstructing* $\theta_{\text{TIA}} = \sqrt{\phi^2 + \theta_p^2}$; and predicting GZ as a consistency check and inverting $\theta_{\text{TIA}} = \arcsin(GZ/GM)$. Here GZ denotes the *simulator-specific* righting-arm proxy $GZ = GM \sin(\theta_{\text{TIA}})$ (a single-angle, constant- GM surrogate), not a full vessel-specific hydrostatic GZ curve at large angles. The inverse is defined only for $|GZ/GM| \leq 1$; because the data satisfy this identity it holds for the targets, and predicted ratios are clipped to $[-1, 1]$ before inversion. Because $g(\phi, \theta_p) = \sqrt{\phi^2 + \theta_p^2}$ is nonlinear, the reconstruction path propagates the angle-prediction error. A second-order Taylor (delta-method) expansion, with angle-error variances $\sigma_\phi^2, \sigma_{\theta_p}^2$, covariance $\sigma_{\phi\theta_p}$, and $r = \sqrt{\phi^2 + \theta_p^2}$, gives the propagated variance (to leading order) and the curvature bias,

$$\text{Var}[\hat{g}] \approx \frac{\phi^2}{r^2} \sigma_\phi^2 + \frac{\theta_p^2}{r^2} \sigma_{\theta_p}^2 + \frac{2\phi\theta_p}{r^2} \sigma_{\phi\theta_p}, \quad \mathbb{E}[\hat{g}] - g \approx \frac{1}{2r^3} (\theta_p^2 \sigma_\phi^2 + \phi^2 \sigma_{\theta_p}^2 - 2\phi\theta_p \sigma_{\phi\theta_p}), \quad (4)$$

where the second term is the Jensen (curvature) bias. Since g is convex this expected bias is non-negative for a positive-semidefinite error covariance matrix, and the penalty grows as $r \rightarrow 0$. We validated Equation (4) against Monte Carlo and, on the real test angles in the large-inclination regime, compared the predicted reconstruction error to its measured value.

3.10. Metrics, Seeds, and Uncertainty Summaries

Forecast quality is reported per horizon (RMSE, MAE, and R^2) and was computed strictly column-wise to avoid non-monotonic per-step aggregation artifacts and pooled overall. The benchmark, UQ, early-warning, representation, out-of-distribution UQ, GM-floor, and evidence-accumulation experiments each trained learned models over 5 seeds and summarized them as mean $\pm$ std, with seed-level 95% t intervals reported for the key overall comparisons; the combined corner-cell holdout study used three seeds. The seed axis captures only model-training stochasticity (weight initialization and minibatch order): each simulation condition is already represented by six wave-phase realizations in the data, so the corpus carries wave-phase variability separately from—and in addition to—the training-seed variability. Because stride-one windows overlap strongly, window-level paired errors were used for descriptive ranking and plots, not as independent observations for formal hypothesis testing. Transfer uncertainty is reported through held-out-fold ranges in addition to the LOHO/LOVO/LORO averages. The naive domain-rule alarm is learning-free and therefore deterministic (zero seed variance).

4. Experiments and Results

Reproducibility note. The pipeline is config-driven and fully scripted; the entry point that produces each table and figure is listed in the accompanying code repository.

4.1. Multi-Model Benchmark (C3)

Table 5 reports the in-distribution benchmark. Persistence attains a negative overall R^2 (-0.385), confirming that the 1 Hz signal is not trivially predictable from its last value; the LSTM attains the strongest overall skill, with $R_{t+1}^2 = 0.932$ decaying to $R_{t+5}^2 = 0.427$ (overall $R^2 = 0.677$, RMSE = 3.51°). (The headline benchmark R^2 values (Table 5) are pooled over the five benchmark seeds, and the in-distribution baselines reported later in the generalization, representation, and series-bootstrap analyses come from their own experiment runs and resampling units, so they differ from this value by ≤ 0.002 in R^2 .) GRU is close behind ($R_{\text{all}}^2 = 0.674$). Using only the five training seeds as the uncertainty unit, the

overall R^2 seed-level 95% intervals are 0.677 ± 0.003 for LSTM and 0.674 ± 0.001 for GRU (RMSE $3.506 \pm 0.014^\circ$ and $3.527 \pm 0.006^\circ$, respectively), tighter than the previous single-realization estimates because the multi-realization test set averages out wave-phase noise; the ranking should be read as a small but consistent benchmark advantage rather than a large practical separation. A series-level bootstrap confirms that the LSTM–GRU gap is statistically significant but small ($\Delta R^2 = +0.003$), while the LSTM’s edge over the strongest non-neural baseline is large and significant ($+0.191$; Table 6). The two baselines that see the *same* full input window as the neural models, a ridge linear regressor ($R^2 = 0.174$ overall) and gradient boosting (0.485), remain well below the recurrent networks, so the neural advantage reflects temporal modeling rather than richer inputs; gradient boosting captures much of the one-step structure ($R^2_{t+1} = 0.849$) but decays fastest across the horizon. TCN and Transformer retain lower long-horizon skill ($R^2_{t+5} = 0.363$ and 0.335 , respectively). Per-horizon skill curves are shown in Figure 2, and Figure 3 ranks the overall performance with seed variability and descriptive error bars. The models are small (72k–220k parameters) and infer in under 1 ms per window on a CPU, well within the 1 Hz real-time budget on commodity shipboard hardware (Table S4). *Operationally*, a compact recurrent forecaster delivers the best in-distribution skill at negligible inference cost, so the modeling choice for shipboard deployment is driven by generalization and calibration (below) rather than by raw in-distribution accuracy.

Table 5. Main forecasting benchmark on the ID test set: per-horizon R^2 at $t+1, t+3, t+5$ and the pooled overall RMSE (degrees) and R^2 . Models are grouped into classical sequence baselines, window regressors (which share the neural models’ $\{u, \dots, \theta_p\} \times L$ input window), and neural sequence models. Non-neural baselines are deterministic (reported once); learned models are means over 5 seeds. Persistence yields a *negative* overall R^2 (Supplementary Figure S1), so the task is non-trivial.

Group	Model	R^2_{t+1}	R^2_{t+3}	R^2_{t+5}	RMSE _{all} (°)	R^2_{all}
Classical	Persistence	−0.322	0.018	−0.726	7.27	−0.385
	AR(8)	0.181	0.294	0.015	5.67	0.155
	Kalman	0.119	0.180	−0.215	6.12	0.018
Window reg.	Ridge	0.397	0.111	0.091	5.61	0.174
	GBM	0.849	0.577	0.150	4.43	0.485
Neural	TCN	0.916	0.706	0.363	3.70	0.641
	Transformer	0.912	0.700	0.335	3.78	0.625
	GRU	0.929	0.715	0.423	3.53	0.674
	LSTM (ref.)	0.932	0.719	0.427	3.51	0.677

Bold marks the best value in each column.

Table 6. Significance of the two headline comparisons. (a) Probabilistic-minus-point ROC-AUC, with a record-level (cluster) bootstrap 95% CI (seed 0, 600 resamples) and the DeLong p (max over 5 seeds); “sig. seeds” counts seeds whose CI excludes zero. (b) The LSTM’s overall- R^2 advantage over the GRU and the strongest non-neural baseline (GBM), with a record bootstrap CI and p (5000 resamples).

(a) Probabilistic—point alarm, ROC-AUC				
θ_d	ΔAUC	95% CI	DeLong p	sig. seeds
15°	+0.019	[+0.016, +0.021]	$<10^{-60}$	5/5 (prob > point)
20°	+0.004	[−0.000, +0.006]	0.026	2/5 (mixed)
25°	−0.018	[−0.024, −0.015]	$<10^{-30}$	0/5 (point > prob)
(b) LSTM—competitor, overall R^2				
Comparison	ΔR^2	95% CI	p	
LSTM – GRU	+0.003	[+0.002, +0.005]	<0.001	significant
LSTM – GBM	+0.191	[+0.185, +0.198]	<0.001	significant

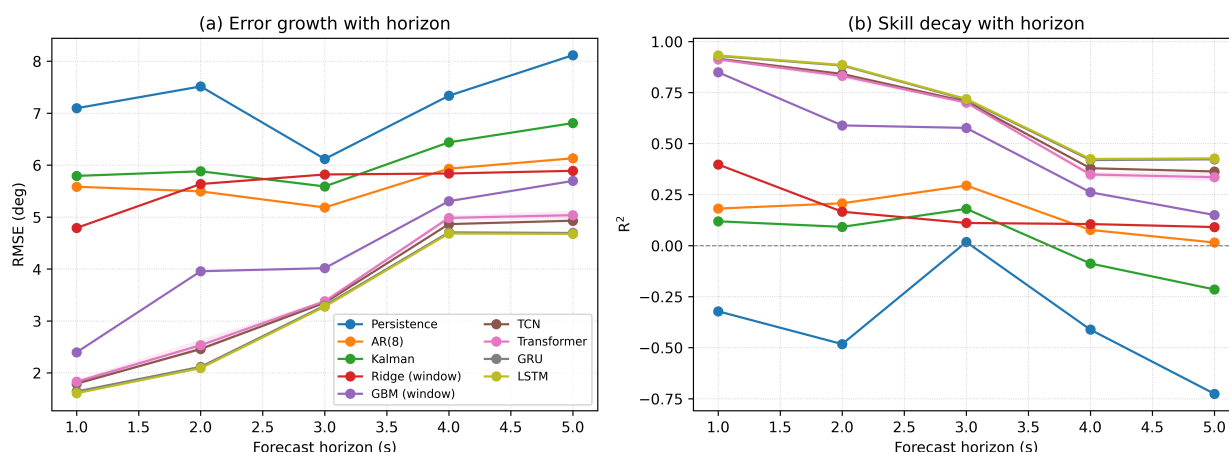


Figure 2. Per-horizon error growth and skill decay for the benchmark models (mean $\pm$ std over seeds).

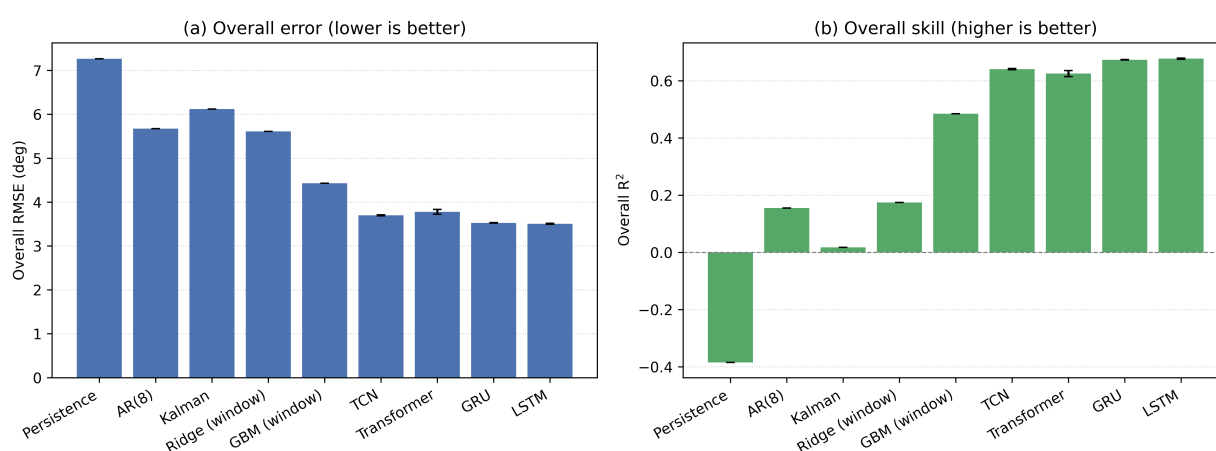


Figure 3. Overall ranking of the benchmark models with seed error bars.

4.2. Uncertainty Quantification (C2)

Table 7 compares the three predictive distributions. After combining epistemic spread with the validation-estimated aleatoric term, all three methods are close to nominal coverage ($\text{PICP@90} = 0.914\text{--}0.928$). The quantile model, which learns the conditional quantiles directly, is the sharpest ($\text{MPIW} = 10.36^\circ$). Its reported CRPS value (1.36°) is a three-quantile approximation, so we treat it as supporting evidence rather than a full-distribution ranking against the Gaussian methods. The deep ensemble gives the lowest point RMSE (3.45°). Figure 4 shows the reliability diagram and example prediction intervals. We note that without the aleatoric term the MC-Dropout and ensemble intervals are badly over-confident ($\text{PICP@90} \approx 0.45$ and 0.24 , respectively), so the post hoc aleatoric component should be viewed as part of the calibration procedure rather than a minor adjustment. *Operationally*, the quantile model is the natural alarm probability source for its sharpness, but its marginal calibration is necessary, not sufficient, and the conditional analysis below shows where it must be recalibrated before deployment.

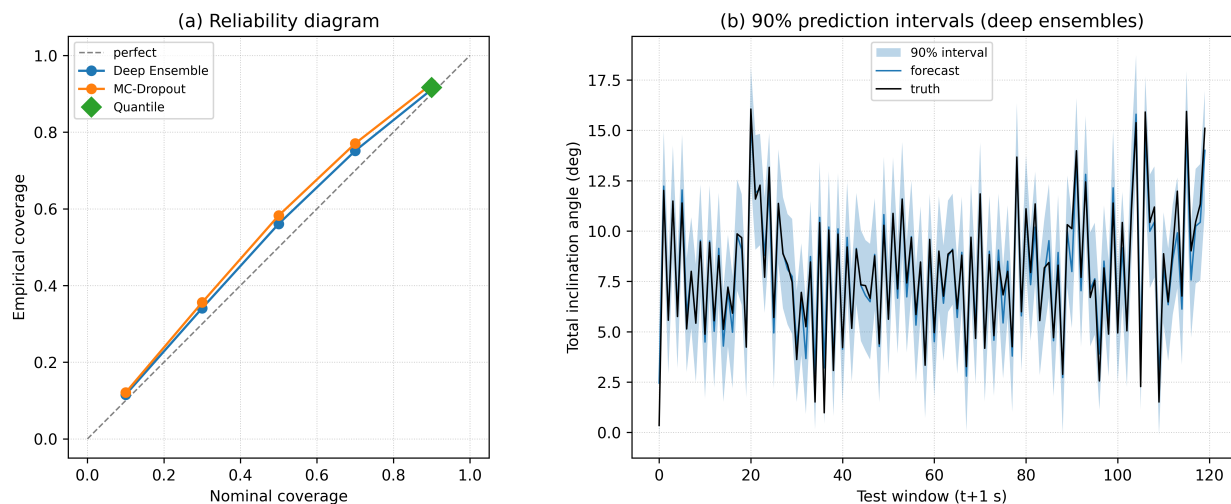


Figure 4. Calibration of the predictive distributions (one representative seed; Table 7 gives the seed mean $\pm$ std): (a) reliability diagram (empirical vs. nominal coverage); (b) 90% prediction intervals for the best-calibrated Gaussian method (deep ensembles); the early-warning alarm of Section 4.3 instead uses the quantile model.

Table 7. Uncertainty quantification on the ID test set: PICP@90 (coverage of the nominal 90% interval, target 0.90), MPIW (mean width), CRPS (a three-quantile proxy for the Quantile method), and point RMSE. The Gaussian methods add a validation-estimated aleatoric term, without which they under-cover strongly (PICP@90 $\approx$ 0.45 and 0.24). Mean $\pm$ std over 5 seeds.

Method	PICP@90	MPIW ($^{\circ}$)	CRPS/proxy ($^{\circ}$)	Point RMSE ($^{\circ}$)
MC-Dropout	0.928 $\pm$ 0.000	11.39 $\pm$ 0.06	1.79 $\pm$ 0.01	3.51 $\pm$ 0.01
Deep Ensemble	0.914 $\pm$ 0.001	10.71 $\pm$ 0.02	1.75 $\pm$ 0.00	3.45 $\pm$ 0.00
Quantile	0.923 $\pm$ 0.004	10.36 $\pm$ 0.11	1.36 $\pm$ 0.01	3.50 $\pm$ 0.03

Bold marks the best value in each column.

4.2.1. Conditional Coverage, Conformal Calibration, and Probability Scores

A marginal PICP near 0.90 is necessary but not sufficient: an interval can be calibrated on average while failing in the regimes that matter operationally. Table 8 slices the quantile band's coverage and shows exactly that. Coverage decays with horizon (0.946 at $t + 1$ to a minimum of 0.904 at $t + 4$ before partially recovering to 0.911 at $t + 5$), degrades at the most energetic sea state (0.899 at $H_s = 7$ m) and most consequentially drops to 0.832 in the large-inclination tail ($\theta_{TIA} > 15^{\circ}$) against 0.936 in the calm regime. The band is therefore optimistic precisely where an exceedance alarm reads it, which is a caveat any deployment must respect.

This tail miscoverage reflects the parametric interval rather than the model, and it can be corrected by distribution-free recalibration. Table 9 applies split conformal prediction (to the ensemble mean) and conformalized quantile regression (CQR, to the quantile band), calibrating the nonconformity quantile on the earlier temporal half of each test group and evaluating on the later half. Both wrappers pull empirical coverage onto the 90% target: CQR moves the quantile band from 0.923 to 0.899 while *tightening* its mean width from 10.34° to 9.95° , and split conformal moves the ensemble from 0.915 to 0.902. In this temporal calibration split, both wrappers reached near-nominal coverage without increasing the mean interval width. Because the tail miscoverage is exactly the regime an exceedance alarm reads, we treat this conformal/CQR recalibration as a *required deployment-time calibration layer*, not an optional refinement: the raw quantile band under-covers the tail at ~ 0.83 against the ~ 0.92 marginal level, and the conformal wrapper restores near-

nominal marginal coverage on the held-out temporal split (tail-specific post-recalibration coverage is not separately quantified here and would require dedicated verification). We report it separately from the headline in-distribution result because the calibration split reduces the evaluation set, but a fielded system must include it.

Table 8. Conditional coverage of the nominal 90% quantile band, sliced by horizon, sea state, tonnage, and inclination regime. The marginal PICP (0.923) hides the pattern: coverage decays with horizon and sea state and falls to 0.832 in the large-inclination tail ($\theta_{TIA} > 15^\circ$) versus 0.936 in the calm regime, motivating the recalibration in Table 9. Mean $\pm$ std over 5 seeds.

Slice	Level	Coverage	MPIW ($^\circ$)
Horizon (s)	$t + 1$	0.946 $\pm$ 0.003	5.78 $\pm$ 0.06
	$t + 2$	0.938 $\pm$ 0.004	7.12 $\pm$ 0.13
	$t + 3$	0.919 $\pm$ 0.007	10.52 $\pm$ 0.16
	$t + 4$	0.904 $\pm$ 0.007	14.10 $\pm$ 0.30
	$t + 5$	0.911 $\pm$ 0.006	14.25 $\pm$ 0.18
Sea state	$H_s = 3$ m	0.942 $\pm$ 0.004	6.89 $\pm$ 0.06
	$H_s = 5$ m	0.927 $\pm$ 0.004	10.66 $\pm$ 0.11
	$H_s = 7$ m	0.899 $\pm$ 0.006	13.72 $\pm$ 0.21
Regime	$\theta_{TIA} \leq 15^\circ$	0.936 $\pm$ 0.004	—
	$\theta_{TIA} > 15^\circ$	0.832 $\pm$ 0.014	—
Marginal	all cells	0.923 $\pm$ 0.004	10.36 $\pm$ 0.11

Bold highlights the large-inclination tail coverage (0.832), the safety-critical regime that falls furthest below the 0.90 target; the **Marginal** row reports the overall average.

Table 9. Distribution-free recalibration: split conformal prediction (on the ensemble mean) and conformalized quantile regression (CQR, on the quantile band) are calibrated on the earlier temporal half of each group and evaluated on the later half. Both pull empirical coverage onto the 90% target without assuming a parametric error distribution, and CQR also tightens the band. Mean $\pm$ std over 5 seeds.

Interval Method	Coverage (Eval Half)	MPIW ($^\circ$)
Quantile band (parametric)	0.923 $\pm$ 0.004	10.34 $\pm$ 0.11
+ CQR	0.899 $\pm$ 0.001	9.95 $\pm$ 0.04
Ensemble Gaussian (parametric)	0.915 $\pm$ 0.001	10.71 $\pm$ 0.02
+ split conformal	0.902 $\pm$ 0.000	10.26 $\pm$ 0.01

Bold marks the best value in each column (coverage closest to the 0.90 target and the smallest interval width).

Because the alarm thresholds the exceedance probability $P(\theta_{TIA}(t+k) > \theta_d)$ rather than the interval, we also score that probability directly: both the quantile-derived and ensemble-Gaussian probabilities are well calibrated ($ECE \leq 0.021$; Table S1). We use the quantile model as the alarm's probability source because it is the sharpest interval producer (Table 7); the ensemble is the better-calibrated alternative at the lowest threshold. Supplementary Figure S2 plots these conditional-coverage and conformal-calibration findings. *Operationally*, marginal calibration hides a tail under-coverage that conformal/CQR recalibration must repair as a mandatory deployment step before the band drives any exceedance alarm.

4.3. Early Warning (C1)

The operationally relevant result is the event-level evaluation of Section 4.3.1; the per-cell metrics here are diagnostic rather than operator-facing, and the $FPR \leq 0.10$ budget is a common comparison point, not a recommended operating point. Table 10 contrasts the probabilistic and point alarms at the three exceedance thresholds, and the comparison is *threshold-specific* rather than a blanket ordering. At 15° the probabilistic alarm is clearly

preferred: with the achieved false-alarm rates matched close to the $\text{FPR} \leq 0.10$ budget, it improves ROC-AUC (0.947 vs. 0.928), PR-AUC, best F1, and lead time, and the ROC-AUC gain is statistically significant across all five seeds (see Table 6). At 20° the two alarms are essentially tied on ROC-AUC (0.953 vs. 0.948; $\Delta = +0.004$): the gain is not robustly significant—the cluster bootstrap is non-significant ($p \approx 0.50$), while the DeLong test is borderline ($p \approx 0.03$) and only 2 of 5 seeds reach significance—so we do not claim a ROC-AUC advantage there, although the probabilistic alarm still leads on PR-AUC, best F1, and lead time. At the rarer 25° threshold the ROC-AUC ordering *reverses*: the point alarm has the higher ROC-AUC (0.964 vs. 0.946), while the probabilistic alarm improves PR-AUC (0.664 vs. 0.633) and best F1 (0.613 vs. 0.605)—the rare-event-appropriate metrics. Its mean lead time is longer (4.63 s vs. 4.52 s), but this is not a matched-cost comparison because the achieved FPR is lower for the probabilistic alarm (0.058 vs. 0.100). Reporting these metrics as mean $\pm$ std over five seeds shows that the probabilistic alarm's PR-AUC and lead-time advantages at 15° and 20° exceed the seed standard deviation by roughly an order of magnitude and are therefore robust, whereas the best-F1 gains (+0.008–0.009) are small but consistent in sign across seeds. All thresholds retain an effective horizon of 5 s only in the bounded sense that AUC remains above 0.8 for the full 1–5 s horizon range tested; longer horizons were not evaluated here. Both learned alarms also clearly dominate the non-learned trend-extrapolation baseline (*naïve*), which merely extrapolates the recent inclination trend: across thresholds its ROC-AUC is only 0.67–0.73, its PR-AUC 0.08–0.25, and its effective horizon is 0 s (AUC below 0.8 even at $t + 1$). This large margin shows that the large heel cannot be reliably anticipated by trend extrapolation alone, consistent with the low one-step autocorrelation of the 1 Hz signal, so the calibrated learned alarm rather than a re-expressed heuristic supplies the warning skill. Figure 5 shows the ROC and PR curves and Figure 6 the AUC decay with horizon (with the effective-horizon line) and the lead-time distribution. *Operationally*: switching from a point to a probabilistic alarm is the right default at the lower, more frequently exceeded $15^\circ/20^\circ$ thresholds, but at the rarest 25° threshold the choice should follow the cost model—the point alarm ranks better on ROC-AUC while the probabilistic alarm favors precision and lead time.

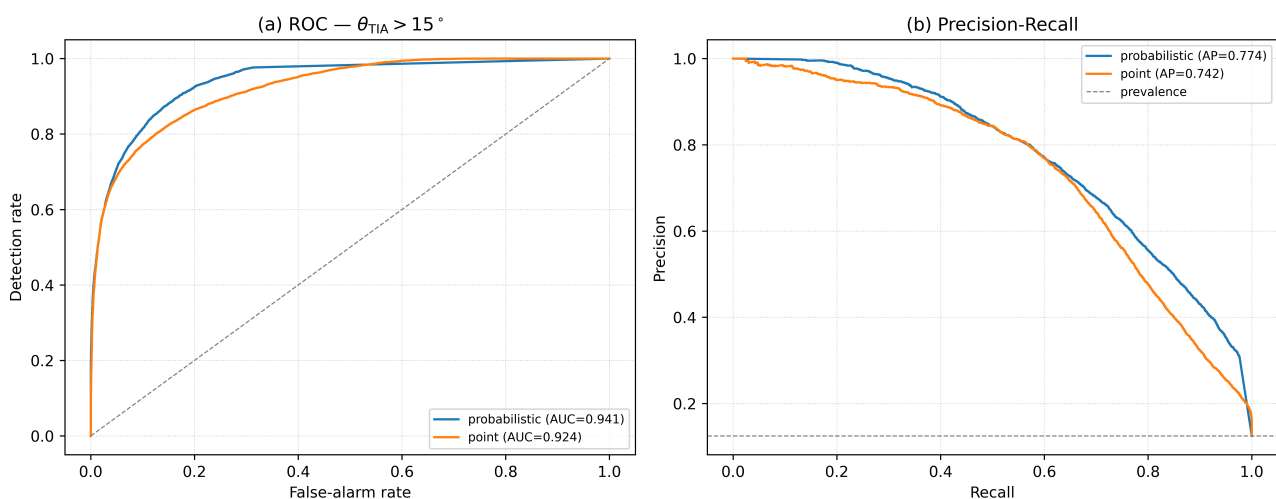


Figure 5. Detection performance at the 15° exceedance threshold (one representative seed; Table 10 gives the seed mean $\pm$ std): (a) ROC and (b) precision–recall, for the probabilistic and point alarms.

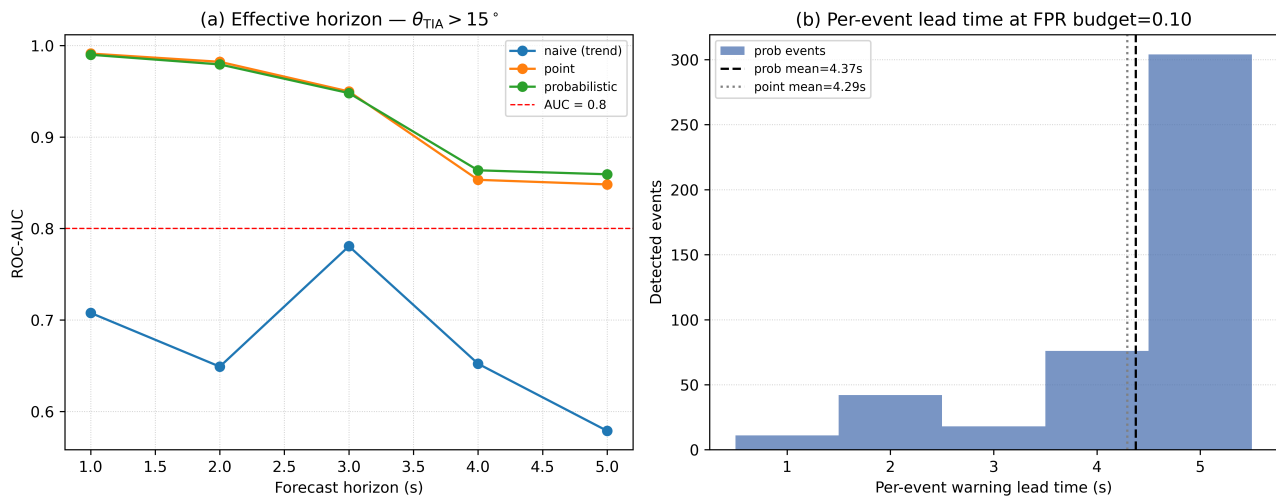


Figure 6. (a) ROC-AUC vs. horizon with the AUC = 0.8 reference line; (b) *per-event* warning lead-time distribution—one lead per detected event (its onset minus the earliest covering fire, hence bounded by the 5 s horizon)—under a leakage-free FPR ≤ 0.10 alarm-threshold budget (α fixed on the earlier temporal half of each group, leads measured on the disjoint later half; one representative seed). Aggregating one value per event, rather than per (window, horizon) cell, keeps the lead-time statistic symmetric with the one-episode-per-false-alarm-run counting so that long events do not dominate it.

Table 10. Per-cell early-warning detection at the three thresholds, comparing the point alarm ($\hat{y} > \theta_d$), the probabilistic alarm ($P(\theta_{TIA}(t+k) > \theta_d) \geq \alpha$), and a non-learned trend baseline (*naive*). Lead time is measured at the FPR ≤ 0.10 budget (realized point as achieved FPR); the effective horizon is bounded by the 1–5 s range. Learned metrics are mean $\pm$ std over 5 seeds (naive deterministic).

θ_d	Alarm	ROC-AUC	PR-AUC	Best F1	Ach. FPR	Eff. hor. (s)	Mean Lead (s)
15°	point	0.928 $\pm$ 0.002	0.756 $\pm$ 0.003	0.690 $\pm$ 0.004	0.100	5	4.03 $\pm$ 0.04
	prob	0.947 $\pm$ 0.001	0.788 $\pm$ 0.002	0.698 $\pm$ 0.002	0.100	5	4.36 $\pm$ 0.02
	naive	0.670	0.254	0.316	0.100	0	3.59
20°	point	0.948 $\pm$ 0.002	0.699 $\pm$ 0.004	0.652 $\pm$ 0.005	0.100	5	4.27 $\pm$ 0.04
	prob	0.953 $\pm$ 0.003	0.733 $\pm$ 0.004	0.661 $\pm$ 0.002	0.100	5	4.66 $\pm$ 0.02
	naive	0.703	0.150	0.233	0.100	0	3.60
25°	point	0.964 $\pm$ 0.002	0.633 $\pm$ 0.006	0.605 $\pm$ 0.005	0.100	5	4.52 $\pm$ 0.04
	prob	0.946 $\pm$ 0.005	0.664 $\pm$ 0.005	0.613 $\pm$ 0.002	0.058	5	4.63 $\pm$ 0.04
	naive	0.734	0.085	0.161	0.100	0	3.66

Bold marks the best value in each column within each threshold block; the naive baseline is shown for reference.

4.3.1. Event-Based Evaluation with a Leakage-Free Operating Point

The per-cell metrics above select the alarm threshold on the same data they are evaluated on, and they count detection at the (window, horizon) level rather than at the level of physical events, both of which flatter the alarm. Here we distinguish two notions of a false alarm. A *case-level* false alarm is a single misfiring (window, horizon) prediction, the quantity the per-cell FPR of Table 10 counts; with a sustained but spurious alarm this inflates the apparent nuisance rate because one continuous alert is counted once per window. An *event-level false-alert episode* instead collapses each maximal run of consecutive spurious firings into a single nuisance event, which is what an operator actually experiences. The event-level metrics report the latter. Table 11 removes the first issue by selecting α on the earlier temporal half of each test group (to meet the FPR ≤ 0.10 budget) and evaluating only on the disjoint later half and addresses the second by scoring at the event level: an *exceedance event* is a run of actual large heel, and a *false-alert episode* is a maximal run of

alarm firings not associated with any upcoming event (so one sustained alarm counts as a single episode). The lead time is aggregated symmetrically with this episode counting: each detected event contributes a single lead—its onset minus the earliest covering fire, hence at most the 5 s horizon—so that neither the false-alert burden nor the lead time is inflated by the duration of long events. (The per-cell lead times in Table 10, by contrast, weight every timestep of an event equally and are reported only as a finer diagnostic.)

Table 11. Event-based early-warning evaluation with a leakage-free operating point: α is selected on the earlier temporal half of each group ($FPR \leq 0.10$) and applied to the disjoint later half. An *event* is a run of actual exceedance and a *false-alert episode* a maximal run of alarm firings unassociated with any event (one sustained alarm counts once). Mean $\pm$ std over 5 seeds ($n_{\text{events}} = 2545, 1149, 491$).

θ_d	Alarm	Detection	Missed-Event	False ep./h	Precision	Lead p_{50} (s)
15°	point	0.996 $\pm$ 0.001	0.004	135.7 $\pm$ 5.0	0.510 $\pm$ 0.006	5.0
	prob	0.993 $\pm$ 0.001	0.007	112.9 $\pm$ 3.5	0.546 $\pm$ 0.009	5.0
	naive	0.651	0.349	125.1	0.597	4.0
20°	point	0.997 $\pm$ 0.000	0.003	184.8 $\pm$ 6.7	0.282 $\pm$ 0.006	5.0
	prob	0.997 $\pm$ 0.000	0.003	154.3 $\pm$ 5.8	0.312 $\pm$ 0.009	5.0
	naive	0.760	0.240	182.0	0.356	3.0
25°	point	0.996 $\pm$ 0.000	0.004	200.9 $\pm$ 9.9	0.149 $\pm$ 0.005	5.0
	prob	0.994 $\pm$ 0.001	0.006	135.0 $\pm$ 7.9	0.212 $\pm$ 0.012	5.0
	naive	0.823	0.177	213.0	0.183	3.0

Bold marks the more favorable of the point and probabilistic alarms in each column within each threshold block (lower false-alert rate, higher precision); the naive baseline is shown for reference.

At the $FPR \leq 0.10$ budget both learned alarms achieve near-complete event recall (≥ 0.99), so events are essentially never missed and the operational cost lies in the false-alert burden. Here the probabilistic alarm helps at *every* threshold, in contrast to the threshold-dependent per-cell ROC-AUC picture above: it cuts false-alert episodes per hour by 16–33% (112.9 vs. 135.7 at 15°, 154.3 vs. 184.8 at 20°, 135.0 vs. 200.9 at 25°) and raises event-level precision at every threshold (0.546 vs. 0.510, 0.312 vs. 0.282, 0.212 vs. 0.149) at equal recall, while the non-learned trend baseline detects only 0.65–0.82 of events. The naive rule’s higher event-level precision at the 15° and 20° thresholds must be read together with this much lower recall; it is not an equal-recall comparison. The burden remains high in absolute terms (even the probabilistic alarm raises ~ 113 –154 false-alert episodes per hour at this fixed 10% budget), so the budget is an operational knob rather than a fixed setting. The operating curve of Figure 7 shows this trade-off quantitatively: sweeping the budget, the probabilistic alarm achieves a more favorable trade-off than the point alarm (same detection at a lower alert rate, higher precision across the tested budgets), and at the operationally preferable $FPR = 0.05$ it raises only 79.5 false-alert episodes per hour at 0.98 recall and 0.66 precision, against 96.2/h for the point alarm. We therefore present the probabilistic alarm as a way to lower the false-alert burden at fixed recall, not as a deployment-ready system. Supplementary Figure S3 plots these threshold-wise event-level metrics—detection rate, false-alert episodes per hour, and event-level precision—for the point, probabilistic, and trend alarms. *Operationally*, at the leakage-free event level the probabilistic alarm lowers the operator’s nuisance-alert load at every threshold without sacrificing recall, and the false-alarm budget—not the choice of alarm—is the dominant lever on that load.

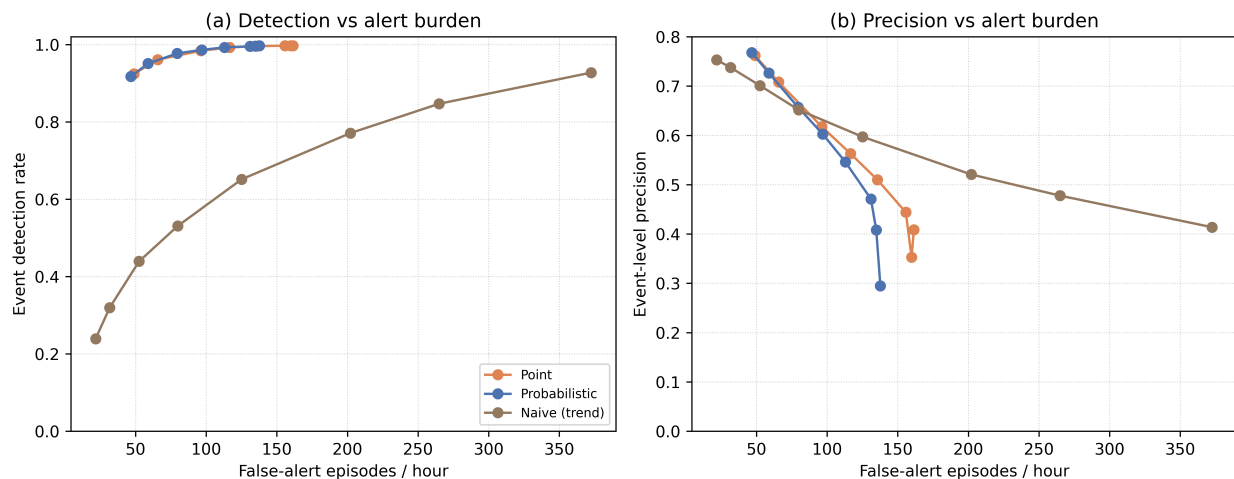


Figure 7. Event-level operating curve at 15° (mean over 5 seeds), sweeping the false-alarm budget: (a) detection rate and (b) precision versus false-alert episodes per hour. The probabilistic alarm lies above/left of the point alarm.

4.3.2. Evidence Accumulation Reduces the Nuisance-Alert Load

The false-alert burden above arises because a single threshold crossing fires the alarm. A cheap and widely used remedy is temporal hysteresis: require k of the last n horizon predictions to exceed the threshold before arming the alarm so that isolated spurious crossings are filtered while sustained evidence still fires promptly. Sweeping (k, n) on the probabilistic alarm at 15° (FPR-matched as above) trades event recall against the false-alert episode rate along a smooth curve. Firing on a single crossing ($k = 1, n = 1$) gives recall 0.993 at 112.9 false-alert episodes per hour, whereas a $k = 3$ -of- $n = 10$ rule cuts the load to 15.9 episodes per hour—a roughly seven-fold reduction—while retaining 0.941 event recall and raising event-level precision from 0.55 to 0.72, at the cost of a modest lead-time change (4.69 s vs. 4.34 s). Larger k/n ratios suppress alerts further but begin to drop recall (e.g., $k = 10$ -of- $n = 10$ falls to 0.30 recall). Evidence accumulation is therefore a practical second lever, complementary to the false-alarm budget, for tuning the alarm to an operator's tolerance for nuisance alerts. Operationally, a k -of- n accumulation rule converts most of the residual false-alert burden into filtered noise, cutting nuisance episodes about seven-fold at ~ 0.94 recall, and should be set alongside the FPR budget when configuring a deployed alarm.

4.4. Generalization (C4)

Table 12 and Figure 8 report transfer to unseen conditions along three axes: sea state (LOHO), vessel (LOVO), and wave-phase realization (LORO). Holding out a sea state (LOHO) reduces overall R^2 from 0.677 to 0.572 ($\Delta R^2 = -0.105$). Holding out a vessel (LOVO) is more severe on average ($R^2 = 0.360, \Delta R^2 = -0.317$), and the fold range reveals a clear failure mode: the held-out 10 t vessel has strongly negative overall skill ($R^2 = -0.911; R^2_{t+5} = -2.516$), whereas the 20–50 t held-out folds remain positive ($R^2 = 0.659$ – 0.693). The fold-wise heatmap therefore separates interpolation within the training envelope from the difficult smallest-vessel fold; per-fold values with seed variability are listed in Table S6, where the 10 t fold is also the only seed-unstable one (± 0.367).

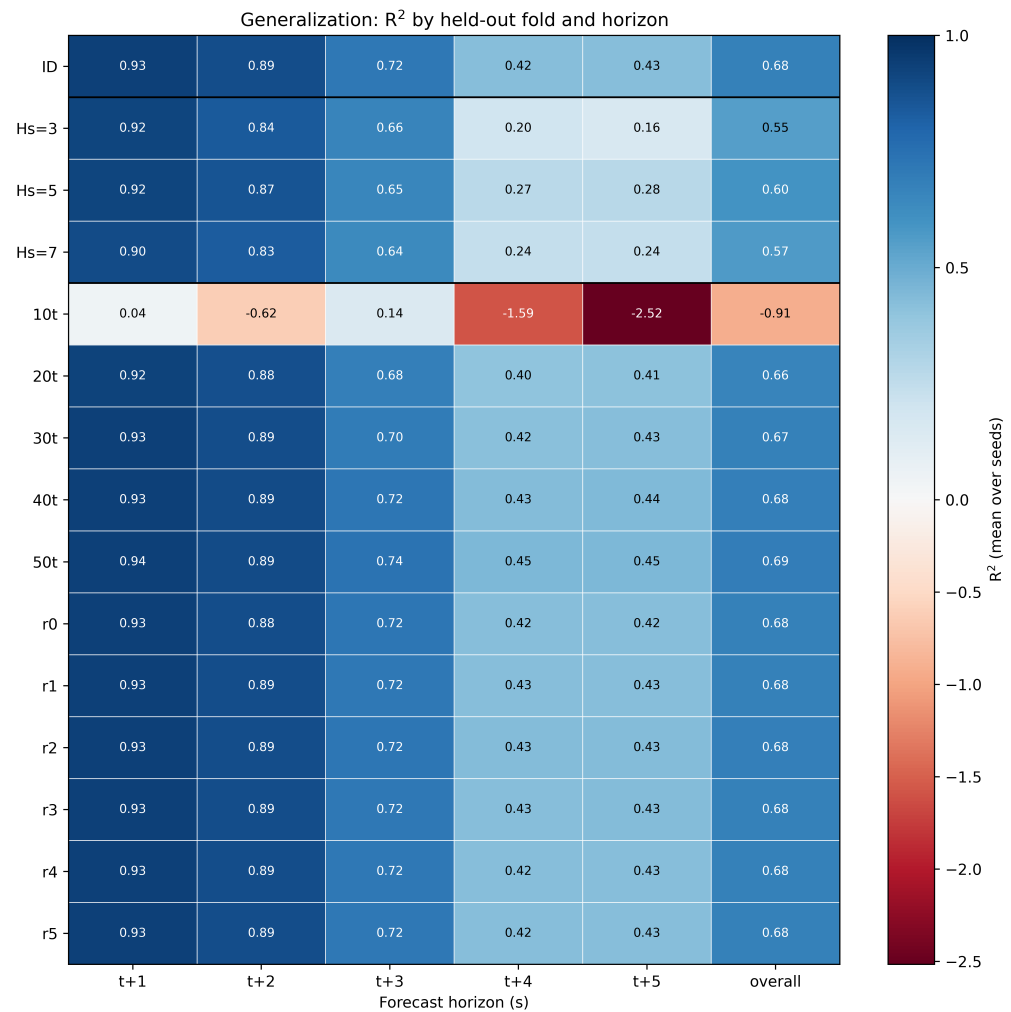


Figure 8. R^2 for each held-out fold (ID, then LOHO sea states, then LOVO tonnages, then LORO wave-phase realizations r0–r5) across forecast horizons.

Table 12. Generalization to unseen conditions. ID is the in-distribution baseline; LOHO (leave-one- H_s -out), LOVO (leave-one-vessel-out), and LORO (leave-one-realization-out) average over their held-out folds. ΔR^2 is the change relative to ID. The fold range reports the overall R^2 range across held-out groups. LORO matches ID ($\Delta R^2 = +0.001$), so forecasting skill is robust to the wave-phase realization; the generalization gap is driven by sea state and vessel size, not by wave phase.

Regime	Overall RMSE (°)	Overall R^2	ΔR^2 vs. ID	R^2 Fold Range	# Folds
ID	3.51	0.677	0	0.677–0.677	1
LOHO (sea state)	3.62	0.572	−0.105	0.553–0.595	3
LOVO (vessel size)	4.22	0.360	−0.317	−0.911–0.693	5
LORO (wave phase)	3.50	0.678	+0.001	0.676–0.681	6

In contrast, holding out an entire wave-phase realization (LORO; train on five of the six independent phase realizations r0–r5, test on the held-out one) leaves skill essentially unchanged: overall $R^2 = 0.678$ ($\Delta R^2 = +0.001$ relative to ID), with all six folds tightly clustered (0.676–0.681). The forecaster is thus robust to wave phase—the quantity it never sees at deployment time—and the difficulty of out-of-distribution transfer comes from the vessel and sea-state axes, not from the random realization of the wave field. This directly answers whether the earlier single-realization results were an artifact of one lucky phase seed: they were not.

Because the 10 t vessel is also the only case affected by the simulator's $GM_{\min} = 0.30$ m clamp, the LOVO fold cannot be read as pure hull-size extrapolation: it confounds extrapolation with a hydrostatic-parameter discontinuity introduced by the simulator. We quantified this confound directly by re-simulating the 10 t vessel with the floor removed (its true $GM = 0.258$ m) and repeating the LOVO fold (the clamped -0.911 baseline is the lower end of the LOVO fold range in Table 12). Unclamping improves the held-out 10 t overall R^2 by 0.527, from -0.911 (clamped) to -0.384 (unclamped)—recovering part of the collapse without restoring positive skill. A substantial part of the 10 t failure is therefore an artifact of the hydrostatic-floor discontinuity rather than genuine small-vessel difficulty—but the fold remains negative even after unclamping, so a residual genuine extrapolation difficulty persists for the smallest hull. *Operationally*, the model can be trusted to transfer across wave phase but not yet onto vessels smaller than the training envelope; the smallest-hull failure is part simulator artifact and part real, so extending the trusted operating envelope downward needs both a floor-free hydrostatic model and small-vessel training data.

A stricter test holds out a sea state and a vessel *jointly*: the model is trained on the eight cells that share neither the held sea state nor the held vessel and tested on the single joint-novel cell (Table 13). This separates joint interpolation from joint extrapolation cleanly. The center cell ($H_s = 5$ m $\times$ 30 t, interior on both axes) retains near-ID skill ($R^2 = 0.584$), confirming the model interpolates jointly when the test condition is bracketed. The two large-vessel corners stay positive ($R^2 \approx 0.51$ – 0.57), but both corners involving the 10 t vessel collapse to negative skill (-0.71 at $H_s = 7$ m, -1.07 at the calm $H_s = 3$ m), so the smallest vessel, not the sea state, is the dominant failure mode, consistent with its in-distribution being the weakest condition (Table S2) and its single-axis LOVO fold being the worst (Table 12). Because the 10 t vessel is the only one affected by the simulator's $GM_{\min} = 0.30$ m clamp, this joint-extrapolation failure again carries the same hydrostatic-discontinuity confound quantified by the unclamped LOVO study above, rather than isolating hull-size extrapolation alone. *Operationally*, joint interpolation inside the training grid is safe, but any deployment that would place the system on a vessel near or below the 10 t corner must first close the small-hull gap.

Table 13. Joint $H_s \times$ vessel holdout (stricter than LOHO/LOVO): training sees neither the held sea state nor the held vessel, and the single joint-novel cell is the test. The center cell (joint interpolation) retains near-ID skill and the large-vessel corners stay positive, but every corner with the 10 t vessel collapses, identifying the smallest vessel (the only one under the $GM_{\min} = 0.30$ m clamp) as the dominant failure mode. Mean $\pm$ std over 5 seeds.

Held Cell ($H_s \times$ tonnage)	Type	Overall R^2	RMSE ($^\circ$)
5 m $\times$ 30 t	interpolation	0.584 ± 0.003	3.60 ± 0.01
3 m $\times$ 50 t	corner	0.507 ± 0.042	2.44 ± 0.10
7 m $\times$ 50 t	corner	0.574 ± 0.006	5.31 ± 0.03
7 m $\times$ 10 t	corner	-0.712 ± 0.120	7.66 ± 0.27
3 m $\times$ 10 t	corner	-1.071 ± 0.674	4.76 ± 0.75

Bold marks the best (highest) overall R^2 , the joint-interpolation cell that is interior on both axes.

4.5. Target Representation and Error Propagation (C5)

Table 14 compares the three target representations in θ_{TIA} space. Direct prediction is strongest overall (RMSE = 3.50° , $R^2 = 0.679$). The GZ path is close behind (RMSE = 3.51° , $R^2 = 0.677$) because it is a deterministic monotone transform of θ_{TIA} at fixed GM , and is therefore best read as a consistency check rather than an independent representation gain. Reconstruction matches direct prediction at one step ($R^2_{t+1} = 0.932$ for both) but has weaker pooled skill ($R^2 = 0.635$), consistent with the propagated-error penalty that grows

with the horizon; the per-horizon values (Table S5) show reconstruction slightly ahead at $t+1$ – $t+2$, the two representations crossing between $t+2$ and $t+3$, and the gap then widening to 0.430 vs. 0.330 by $t+5$ in direct prediction's favor. Figure 9 shows the skill by representation and the agreement between the analytic error-propagation prediction (Equation (4)) and Monte Carlo on the true test angles. The ablations over the lookback, feature set, model size, and horizon are summarized in Supplementary Figure S4, with the underlying values in Table S3. *Operationally*, predicting θ_{TIA} directly is the right default for multi-step exceedance warning; the reconstruction path buys nothing beyond the first step and pays a growing propagation penalty at the longer horizons that matter for lead time.

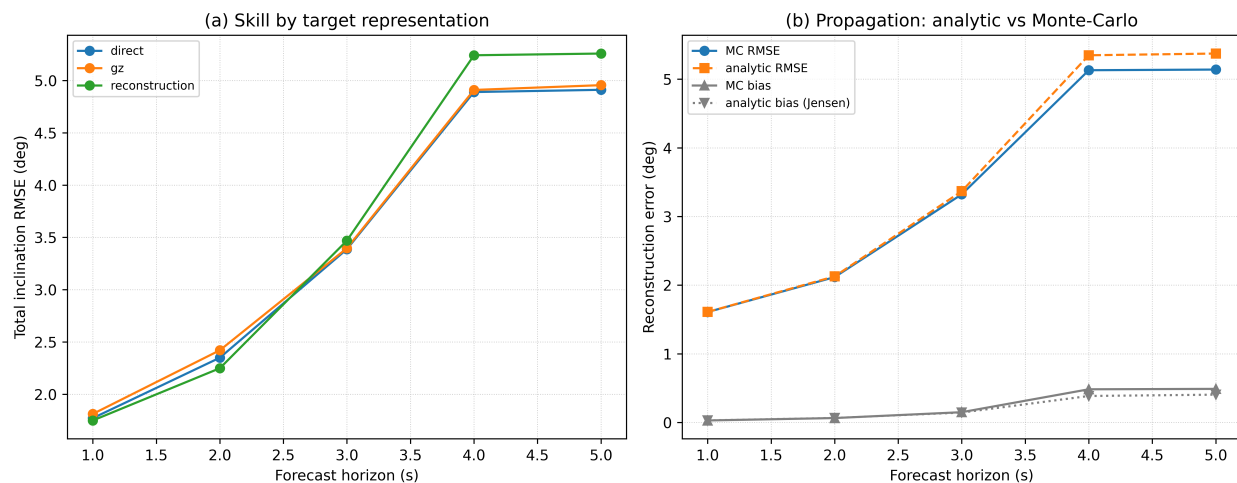


Figure 9. (a) θ_{TIA} RMSE by horizon for the direct/reconstruction/GZ representations (one representative seed; Table 14 gives the seed mean $\pm$ std); (b) analytic (delta-method + Jensen) vs. Monte-Carlo reconstruction error in the large-inclination regime.

Table 14. Target-representation comparison, scored in θ_{TIA} space (degrees): *direct* predicts θ_{TIA} , *reconstruction* predicts (ϕ, θ_p) and forms $\sqrt{\phi^2 + \theta_p^2}$, and GZ predicts and inverts the righting-arm proxy. As GZ is a deterministic monotone transform of θ_{TIA} , the GZ path is a consistency check, not independent skill. Mean $\pm$ std over 5 seeds.

Representation	RMSE _{all} (°)	R^2_{all}	R^2_{t+1}
Direct	3.50 $\pm$ 0.01	0.679 $\pm$ 0.001	0.932 $\pm$ 0.001
Reconstruction	3.73 $\pm$ 0.03	0.635 $\pm$ 0.005	0.932 $\pm$ 0.003
GZ [†]	3.51 $\pm$ 0.01	0.677 $\pm$ 0.002	0.931 $\pm$ 0.001

Bold marks the best value in each column (R^2_{t+1} is tied between Direct and Reconstruction). [†] Consistency check (deterministic transform of the target), not an independent representation.

4.6. Sampling Uncertainty and Condition-Resolved Skill

The seed standard deviations reported throughout capture training noise but not the sampling uncertainty of a test set whose windows are strongly dependent within each simulated series. Table 15 therefore reports series-level (cluster) bootstrap 95% intervals, resampling whole simulated series (vessel $\times$ sea-state $\times$ wave-phase realization) so that the within-series dependence is preserved. Two points follow. First, with the multi-realization dataset the bootstrap interval on overall R^2 (± 0.011 for the micro average) is now several times the seed standard deviation (± 0.002), so the dominant uncertainty is still which conditions are sampled rather than training randomness, though the gap is far smaller than in the previous single-realization design because the larger test set narrows the sampling interval; the LSTM–GRU separation (0.677 vs. 0.673), small relative to that interval, should

be read as a benchmark ordering rather than a practically meaningful gap. Second, the condition-balanced *macro* average ($R^2 = 0.617$ for the LSTM) is below the pooled *micro* value (0.677), showing that the abundant easy series inflate the headline number; we report both so the condition-balanced figure is visible. Resolving the in-distribution skill by condition (Table S2), it falls with sea state and rises with tonnage, and the weakest condition (10 t) is the same fold that fails out of distribution. *Operationally*, the headline *micro- R^2* overstates skill on the hard conditions, so deployment expectations should be anchored to the condition-balanced *macro* figure and its bootstrap interval.

Table 15. Series-level (cluster) bootstrap 95% CIs for the two headline models, resampling whole series (2000 resamples). The series-balanced *macro* R^2 is markedly below the pooled *micro* value, showing that the abundant easy series inflate the headline; the interval (± 0.011 on the *micro* R^2) still exceeds the seed std (± 0.002), so sampling over conditions dominates the seed variation.

Model	R^2 Micro (pooled)	R^2 Macro (record-bal.)	RMSE (°)
LSTM	0.677 [0.665, 0.687]	0.617 [0.606, 0.627]	3.51 [3.27, 3.73]
GRU	0.673 [0.661, 0.684]	0.612 [0.601, 0.622]	3.53 [3.29, 3.75]

Finally, we ask whether the calibrated predictive intervals survive distribution shift. Table 16 retrains the three UQ methods on the in-distribution conditions and scores them on the held-out $H_s = 7$ m sea state, with the aleatoric term estimated from in-distribution validation residuals, as a deployed system would. The intervals that cover near-nominally in distribution (PICP@90 ≈ 0.91 – 0.93) become over-confident out of distribution, with PICP collapsing to 0.70 – 0.81 as the point RMSE rises from $\sim 3.5^\circ$ to $\sim 5.0^\circ$. The quantile model degrades least in coverage and keeps the lowest CRPS, while the deep ensemble’s narrow intervals under-cover the most. Calibration thus does not transfer for free with extrapolation, which is the out-of-distribution counterpart to the conditional-coverage gap of Section 4.2.1 and a further argument for the required conformal recalibration layer in any deployment. *Operationally*, predictive intervals must not be trusted as-is once the sea state leaves the training envelope; recalibration on in-distribution data does not protect the tail under shift conditions.

Table 16. Out-of-distribution UQ calibration: the methods are trained on the in-distribution conditions and evaluated on the held-out $H_s = 7$ m sea state, with the aleatoric term estimated from in-distribution validation residuals. Intervals calibrated in-distribution (PICP@90 ≈ 0.91 – 0.93) become over-confident out of distribution (PICP 0.70 – 0.81 , RMSE $\sim 3.5^\circ \rightarrow 5.0^\circ$). Mean $\pm$ std over 5 seeds.

Method	PICP@90 (ID)	PICP@90 (OOD)	MPIW _{OOD} (°)	RMSE _{OOD} (°)
MC-Dropout	0.928	0.744 ± 0.003	9.75 ± 0.06	5.01 ± 0.04
Deep Ensemble	0.914	0.699 ± 0.002	8.78 ± 0.04	4.92 ± 0.01
Quantile	0.923	0.808 ± 0.006	11.15 ± 0.14	5.05 ± 0.02

Bold marks the best value in each column (highest out-of-distribution coverage, smallest interval width, smallest RMSE).

We close with formal significance tests for the two headline comparisons (Table 6), which the seed mean $\pm$ std cannot supply. For the alarm, the probabilistic-minus-point ROC-AUC difference was tested with both a series-level cluster bootstrap and the DeLong test, and the verdict is threshold-specific. The probabilistic gain is significant at 15° ($\Delta\text{AUC} = +0.019$, 95% CI $[+0.016, +0.021]$, all five seeds significant). At 20° the gain is small ($\Delta\text{AUC} = +0.004$) and not robustly significant—the cluster bootstrap is non-significant ($p \approx 0.50$) while DeLong is borderline ($p \approx 0.03$), and only 2 of 5 seeds reach significance—so we report it as mixed evidence rather than a clean win. At the rarest 25° threshold the difference is significantly

reversed ($\Delta\text{AUC} = -0.018$, CI $[-0.024, -0.015]$, DeLong $p \approx 10^{-33}$), the point alarm having the higher ROC-AUC; there (as Tables 10 and 11 show) the probabilistic alarm nonetheless leads on PR-AUC and event-level precision, the rare-event-appropriate metrics. For the models, the LSTM's overall- R^2 edge over the GRU is small but significant ($\Delta R^2 = +0.003$, CI $[+0.002, +0.005]$), and its edge over the strongest non-neural baseline (GBM) is large ($+0.191$, CI $[+0.185, +0.198]$). These model tests resample whole records (not individual windows) and bootstrap the *paired* difference between two forecasters scored on identical data: because the LSTM and GRU errors are strongly correlated, the standard error of their difference is much smaller than that of either model's own R^2 . The LSTM-GRU difference is therefore significant (Table 6) even though the two models' *marginal* bootstrap intervals overlap (Table 15)—overlapping marginal intervals do not imply a null difference for correlated estimators, and the record-level paired test is the appropriate one. The gap is nonetheless tiny ($\Delta R^2 = +0.003$), so we treat the LSTM and GRU as practically tied and read the ordering as a benchmark ranking only. An operating-point sweep (Table S7) shows the false-alarm budget is the effective lever (tightening from FPR = 0.10 to FPR = 0.05 cuts the false-alert rate by roughly 30%, from 112.9 to 79.5 episodes per hour, at > 0.97 recall), whereas an alert-suppression refractory mostly fragments sustained alarms and is not useful here. *Operationally*, the formal tests confirm that the probabilistic alarm's advantage is real only at 15° , mixed at 20° , and reversed (on ROC-AUC) at 25° , so the alarm choice must be made per threshold and per cost model, with the false-alarm budget as the primary tuning lever.

5. Discussion

5.1. When Probabilistic Early Warning Helps, and When It Does Not

The point and probabilistic alarms are derived from the same model and differ only in that the probabilistic score incorporates the per-window predictive uncertainty. When that uncertainty varies across windows, as it does for the quantile and ensemble predictors, the exceedance probability re-ranks borderline cases more informatively than the bare point forecast. The benefit, however, is *threshold-specific* rather than uniform. On receiver-operating discrimination the probabilistic alarm is clearly ahead only at 15° ($\Delta\text{ROC-AUC} = +0.019$, significant in 5/5 seeds); at 20° the gap is marginal ($\Delta = +0.004$, mixed evidence—bootstrap n.s. but DeLong $p \approx 0.03$ and 2/5 seeds significant); and at the rarest 25° it is *reversed*, with the point alarm scoring higher (0.964 vs. 0.946). The probabilistic alarm is consistently better is on precision–recall and on the event-level burden: across all three thresholds it improves PR-AUC and best F1 and lengthens warning lead time at nearly matched false-alarm rates. At 15° and 20° these gains are robust across the five seeds (the PR-AUC and lead-time advantages exceed the seed standard deviation by about an order of magnitude) but modest in size (PR-AUC $\approx +0.03$, best F1 $+0.008$ to $+0.009$, lead time $+0.3$ to $+0.4$ s). These per-cell lead-time gains are diagnostic: at the operator-facing event level both alarms reach the same median lead (5.0 s at every threshold; Table 11), so the probabilistic alarm's event-level benefit is its lower false-alert burden and higher precision rather than a longer warning. The larger separation is between either learned alarm and the trend-extrapolation baseline. Operationally, then, the probabilistic alarm should be preferred for its lower event-level false-alert burden and higher precision, but it is not a blanket improvement on discrimination, and its relevance—like the methodology as a whole—awaits confirmation on measured seakeeping data.

5.2. Scope of Novelty

The individual neural architectures and UQ techniques used here are established, and recent ship-motion studies already report LSTM, Transformer, MC-Dropout, bootstrap, and

quantile-style uncertainty results. The contribution is therefore an integrated evaluation package for small-vessel total inclination: a leakage-safe multi-model benchmark, calibrated predictive distributions, threshold-specific probabilistic alarms, explicit lead-time accounting, transfer folds, and a target-reconstruction error analysis under one scripted protocol. Relative to recent ship-motion benchmarks that compare RNN-family models on simulated 6-DOF motion [16], the distinguishing elements are the calibrated, threshold-specific alarm with explicit lead-time accounting and a domain-rule comparator (C1) and the target-representation error-propagation analysis (C5); the multi-model benchmark (C3) is included for rigor and transparency rather than as the headline novelty.

5.3. Calibration Requires a Validation-Residual Variance Term

MC-Dropout and small deep ensembles capture mainly epistemic uncertainty; on this signal their raw intervals are far too narrow and badly under-cover (epistemic-only PICP@90 is 0.447 for MC-Dropout and 0.238 for the ensemble). Adding a homoscedastic aleatoric variance estimated from validation residuals [20] restores calibration and makes the three methods comparable. Much of the reported coverage therefore comes from the shared residual term rather than from method-specific epistemic spread, so the small differences among the UQ rows should not be over-interpreted.

5.4. Coverage Is Conditional, So Conformal Recalibration Is a Deployment Requirement

A marginal PICP near 0.90 conceals where the interval actually holds. Sliced by condition, coverage decays with horizon, drops at the most energetic sea state, and falls to 0.83 in the large-inclination tail the alarm cares about (Table 8). The nominal interval is thus optimistic in exactly the regime of interest, which is the relevant caveat for an exceedance application: an alarm that quotes a 90% interval but covers only ~ 0.83 of tail outcomes will systematically under-state risk where it matters most. We therefore treat conformal recalibration not as an optional refinement but as a required deployment-time layer. Split conformal prediction and CQR pull empirical coverage back onto the target, and CQR also tightens the band (Table 9), so a deployed system must wrap the predictor in a conformal layer calibrated on held-out data before its intervals can be trusted at the operating thresholds. We did not fold this into the headline forecasting numbers because the temporal calibration split shrinks the evaluation set; a dedicated calibration record would remove that cost without changing the conclusion that recalibration is necessary.

5.5. Operational Alert Burden and How to Manage It

Scored at the event level with a leakage-free operating point (Table 11), recall is near-complete (≥ 0.99) at the $\text{FPR} \leq 0.10$ budget, so the operational question is not whether events are caught but how many false alerts accompany them. Here we distinguish *case-level* false alarms (individual decision cells firing wrongly, controlled by the FPR budget) from *event-level* false-alert episodes (contiguous spurious-alarm runs counted per hour) since an operator experiences the latter. At equal recall the probabilistic alarm cuts false-alert episodes per hour by roughly 16–33% and raises precision at every threshold. The trend baseline's higher event-level precision at 15° and 20° reflects this episode accounting rather than better discrimination: firing choppy, it fragments each sustained exceedance into several short true-alert episodes—inflating the precision numerator even while it misses roughly a third of the events (recall 0.65–0.82)—whereas the probabilistic alarm covers each event with a single sustained alarm. The burden is nonetheless heavy in absolute terms (even the probabilistic alarm raises ~ 113 – 154 false-alert episodes per hour at this fixed 10% budget). Two complementary levers reduce it. Tightening the false-alarm budget to $\text{FPR} = 0.05$ cuts the rate by about 30% at 15° (from ~ 113 to ~ 80 episodes per hour) while keeping recall above 0.97. More effective is temporal evidence accumulation: requiring k

of the last n exceedance predictions before arming the alarm (Section 4.3.2) filters isolated spurious crossings, and a $k = 3$ -of- $n = 10$ rule at 15° reduces the burden roughly seven-fold (from ~ 113 to ~ 16 episodes per hour) while retaining ~ 0.94 event recall and raising event-level precision from 0.55 to 0.72. This accumulation rule is cheap to implement onboard—a short sliding counter on the alarm stream—and, combined with a conservative FPR budget, substantially reduces the residual nuisance-alert load without sacrificing the detection of sustained exceedance events, pending evaluation in onboard field trials. A fielded alarm would set both levers to an operator-acceptable rate rather than adopt the comparison value used here.

5.6. Target Representation

The agreement of the direct and GZ paths is expected: the simulator-specific $GZ = GM \sin(\theta_{TIA})$ is a deterministic monotone transform of the target (a constant- GM proxy, not a full hydrostatic righting-arm curve), so predicting GZ is predicting θ_{TIA} in different units. The reconstruction penalty is genuine and is explained by Equation (4): the Euclidean norm is convex, so reconstructing from noisy angle predictions carries a positive expected curvature bias for the estimated error covariance, largest near zero inclination where the Jacobian is ill-conditioned. The analytic prediction matches Monte Carlo in the large-inclination regime; near zero inclination the first-order expansion breaks down (the r^{-3} terms diverge), so we restrict the empirical comparison to the regime where it is valid.

5.7. Limitations

The study is based on 6-DOF simulation, not measured IMU data, and the wave forcing uses a simulator-specific scale, so absolute angle magnitudes should be read as relative rather than field-calibrated. The 1 Hz sampling rate and 5 s forecast horizon are properties of this simulator study; real IMU systems usually operate at higher rates, so measured-data transfer should revisit both decimation and horizon selection. The grid spans five vessels and three sea states, now sampled across six independent wave-phase realizations, so wave-phase variability has been substantially addressed within the sampled spectrum and six realizations—the leave-one-realization-out folds confirm phase robustness ($\Delta R^2 = +0.001$); but the five-vessel, three-sea-state condition envelope is still modest, and the leave-one-out folds characterize interpolation and extrapolation within this grid rather than transfer to arbitrary hulls or spectra. The joint $H_s \times$ vessel holdout (Table 13) sharpens this: the model interpolates jointly (the center cell keeps near-ID skill) but every joint fold involving the 10 t vessel collapses regardless of sea state, identifying the smallest vessel as the dominant failure mode. That fold is also confounded with the simulator's $GM_{\min} = 0.30$ m clamp, which affects only that vessel. A plausible contributing mechanism is that the floor raises this hull's metacentric height by about 16% over its unclamped value ($0.258 \rightarrow 0.300$ m) and so shortens its natural roll period by roughly 7% (the period scales as $GM^{-1/2}$); a model trained on the consistent GM -period relationship of the 20–50 t hulls then carries a small frequency mismatch that accumulates into a growing phase error over the 5 s horizon, consistent with the steep R^2 drop at $t + 5$. We tested this mechanism directly rather than leaving it as conjecture: re-simulating the 10 t vessel with the floor removed (true $GM = 0.258$ m) and repeating the fold improves the overall R^2 by 0.527 (from -0.911 to -0.384 ; Section 4.4). A substantial part of the smallest-vessel failure is therefore a hydrostatic-discontinuity artifact, although the fold remains negative after unclamping, so a residual genuine small-hull extrapolation difficulty persists. The bootstrap intervals (Table 15) temper the headline skill but less than in the single-realization data: resampling the 90 series gives a tight interval on the pooled (micro) overall R^2 (0.677 , 95% CI $\approx \pm 0.011$), still wider than the seed spread but

no longer an order of magnitude larger, while the condition-balanced (macro) average (0.617) sits below the pooled value (0.677). The benchmark numbers therefore still describe this particular five-vessel, three-sea-state condition grid—now sampled across six wave phases—rather than a population of arbitrary vessels and spectra. At the operator-facing event level both quantities are aggregated symmetrically: each detected exceedance event contributes a single lead-time value and each contiguous spurious-alarm run a single false-alert episode (Table 11); the per-cell lead times of Table 10 are retained only as a finer diagnostic. Comparisons at the 25° threshold should additionally allow for the different achieved false-alarm rates between the two alarms. We compare against one non-learned trend-extrapolation alarm; other operational rules, such as a current-GZ threshold or vessel-specific capsizes criteria, could be added, though the present margin already indicates the learned alarm is not merely re-expressing a simple threshold- or rate-based operational rule. Finally, the early-warning thresholds (15°/20°/25°) are percentile-based choices for this dataset, deliberately treated as operational large-inclination exceedance levels rather than as vessel-specific capsizes criteria; mapping them onto a hull's dynamic-stability limits is left to future work. Taken together, these points define an explicit deployment gap: the present results establish the decision layer as a simulation benchmark with a stated operating envelope (modified Pierson–Moskowitz spectrum, $T_0 = 8$ s, 140° heading, zero forward speed, restricted loading), but a fielded warning system additionally requires validation on measured seakeeping data, a wider vessel and spectral population, the threshold-to-stability mapping above, and the conformal recalibration layer made mandatory by the tail-coverage shortfall. Closing that gap, not adding model capacity, is the priority for moving from this benchmark toward operational use.

6. Conclusions

This study developed a reproducible simulation benchmark for uncertainty-aware short-horizon warning of large-inclination exceedance—the crossing of fixed operational heel thresholds, not a vessel-specific capsizes criterion—in small fishing vessels. Across a leakage-safe, seed-averaged comparison of seven primary forecasters and two window-regression baselines, the LSTM provided the strongest in-distribution reference forecaster ($R^2 = 0.677$ pooled over the 1–5 s horizon, with $R^2 = 0.427$ at $t + 5$), while quantile regression gave the sharpest prediction intervals, marginally calibrated in distribution; conditional and conformal analyses then showed that coverage falls in the large-inclination tail and must be restored by a conformal recalibration layer at deployment. Converting the forecast into a probabilistic exceedance alarm and evaluating it at the event level lowered the false-alert burden relative to point alarms at comparable recall, with a receiver-operating advantage that is threshold-specific—clear at 15°, marginal at 20°, and reversed at the rarest 25°—so the alarm is preferred for its lower event-level false-alert burden and higher precision rather than for uniformly longer lead time or better discrimination. The absolute burden nonetheless remains operationally high—of order 110–160 false-alert episodes per hour at the fixed 10% budget—so the probabilistic alarm reduces, but does not by itself resolve, the nuisance load, which still requires the explicit suppression mechanisms below. Leave-one-realization-out folds showed the forecaster to be robust to wave phase, whereas the smallest-vessel (10 t) fold failed; a controlled re-simulation traced a substantial part of that failure to the simulator's hydrostatic GM -floor discontinuity, with a residual genuine small-hull difficulty remaining. A simple k -of- n evidence-accumulation rule cut the residual nuisance-alert load roughly seven-fold at ~ 0.94 recall, showing that the high absolute alert rate can be substantially reduced, pending evaluation in onboard field trials. A closed-form, Monte-Carlo-validated error-propagation analysis additionally explains the cost of reconstructing the target from its roll and pitch components. Together with the

high absolute alert rate and the smallest-vessel failure, these results show that the present system should be read as a simulation-tested methodology rather than a deployment-ready warning system: operational decision support remains limited until validation on measured data and until the inclination-exceedance thresholds are mapped to vessel-specific stability criteria.

Future Work

Future work should prioritize validation on measured IMU/seakeeping data, expansion to a broader vessel and sea-state population (and to varied headings, forward speeds, and loading conditions beyond the fixed operating envelope used here), heteroscedastic and physics-informed uncertainty models, additional rule-based alarm comparators, and coupling the calibrated alarm to vessel-specific dynamic-stability criteria for at-sea deployment. The two onboard levers identified here for managing alarm load—a conservative false-alarm budget and k -of- n evidence accumulation—should be tuned against operator-acceptable nuisance rates on real data.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jmse14131195/s1>.

Author Contributions: Conceptualization, H.J. and T.I.; methodology, H.J. and T.I.; software, H.J.; validation, H.J. and T.I.; formal analysis, H.J.; investigation, H.J.; resources, B.S., K.P. and K.K.; data curation, H.J.; writing—original draft, H.J.; writing—review and editing, B.S., K.P., K.K. and T.I.; visualization, H.J.; supervision, T.I.; project administration, T.I.; funding acquisition, B.S., K.P. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries (RS-2022-KS221571), Republic of Korea.

Data Availability Statement: The simulation data, processing pipeline, model-training and evaluation code, and the scripts that regenerated every table and figure in this article are publicly available at GitHub (https://github.com/taehoim/JMSE_Data, accessed on 24 June 2026). The repository contains the 90 6-DOF motion series (five tonnages $\times$ three sea states $\times$ six wave-phase realizations); the windowing and leakage-safe splitting code; the benchmark, uncertainty-quantification, early-warning, generalization, and representation experiments; and the conformal-recalibration and hysteresis analyses. The results can be reproduced end-to-end by running the documented pipeline from the raw series; the released CSVs underlying each table and figure are included for direct verification.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TIA	Total inclination angle ($\theta_{TIA} = \sqrt{\phi^2 + \theta_p^2}$; ϕ roll, θ_p pitch)
DOF	Degree of freedom
GM/GZ	Metacentric height/righting arm
LSTM/GRU	Long short-term memory/gated recurrent unit
RNN/MLP	Recurrent neural network/multilayer perceptron
TCN	Temporal convolutional network
AR/GBM	Autoregressive model/gradient-boosting machine
IMU/RAO	Inertial measurement unit/response amplitude operator
UQ	Uncertainty quantification
MC	Monte Carlo (dropout)
PICP	Prediction-interval coverage probability
MPIW	Mean prediction-interval width

CRPS	Continuous ranked probability score
CQR	Conformalized quantile regression
ECE	Expected calibration error
ROC/PR	Receiver operating characteristic/precision–recall
AUC	Area under the curve
FPR/TPR	False-positive/true-positive rate
ID/OOD	In-distribution/out-of-distribution
LOHO/LOVO	Leave-one- H_s -out/leave-one-vessel-out
LORO	Leave-one-realization-out
MSS	Marine Systems Simulator (MATLAB)
RMSE/MAE	Root-mean-square error/mean absolute error

References

- Wang, J.; Pillay, A.; Kwon, Y.S.; Wall, A.D.; Loughran, C.G. An analysis of fishing vessel accidents. *Accid. Anal. Prev.* **2005**, *37*, 1019–1024. <https://doi.org/10.1016/j.aap.2005.05.005>.
- Davis, B.; Colbourne, B.; Molyneux, D. Analysis of fishing vessel capsizing causes and links to operator stability training. *Saf. Sci.* **2019**, *118*, 355–363. <https://doi.org/10.1016/j.ssci.2019.05.017>.
- Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1050–1059.
- Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Beach, California, USA, 4–9 December 2017; Volume 30, pp. 6402–6413. <https://arxiv.org/abs/1612.01474>.
- Koenker, R.; Bassett, G., Jr. Regression quantiles. *Econometrica* **1978**, *46*, 33–50. <https://doi.org/10.2307/1913643>.
- Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. <https://doi.org/10.3115/v1/D14-1179>.
- Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, <https://doi.org/10.48550/arXiv.1803.01271>.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008, <https://arxiv.org/abs/1706.03762>.
- International Maritime Organization. *Resolution MSC.267(85): Adoption of the International Code on Intact Stability, 2008 (2008 IS Code)*; International Maritime Organization: London, UK, 2008. Available online : [https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/MSCResolutions/MSC.267\(85\).pdf](https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/MSCResolutions/MSC.267(85).pdf) (accessed on 4 December 2008).
- Fossen, T.I. *Handbook of Marine Craft Hydrodynamics and Motion Control*; John Wiley & Sons: Chichester, UK, 2011. <https://doi.org/10.1002/9781119994138>.
- Sun, Q.; Tan, J.; Zhou, Y. A GRNN Neural Network-Based Surrogate Model for Ship Dynamic Stability Calculation. *J. Mar. Sci. Eng.* **2025**, *13*, 1777. <https://doi.org/10.3390/jmse13091777>.
- Yang, K.; Lin, J.; Ding, J.; Zheng, B.; Qin, L. An Integrated Safety Monitoring and Pre-Warning System for Fishing Vessels. *J. Mar. Sci. Eng.* **2025**, *13*, 1049. <https://doi.org/10.3390/jmse13061049>.
- Chen, Z.; Liu, X.; Ji, X.; Gui, H. Real-Time Prediction of Multi-Degree-of-Freedom Ship Motion and Resting Periods Using LSTM Networks. *J. Mar. Sci. Eng.* **2024**, *12*, 1591. <https://doi.org/10.3390/jmse12091591>.
- Li, S.; Wang, T.; Li, G.; Skulstad, R.; Zhang, H. Short-term ship roll motion prediction using the encoder–decoder Bi-LSTM with teacher forcing. *Ocean Eng.* **2024**, *295*, 116917. <https://doi.org/10.1016/j.oceaneng.2024.116917>.
- Lee, H.; Ahn, Y. Comparative Study of RNN-Based Deep Learning Models for Practical 6-DOF Ship Motion Prediction. *J. Mar. Sci. Eng.* **2025**, *13*, 1792. <https://doi.org/10.3390/jmse13091792>.
- Hou, X.; Xia, S. Short-Term Prediction of Ship Roll Motion in Waves Based on Convolutional Neural Network. *J. Mar. Sci. Eng.* **2024**, *12*, 102. <https://doi.org/10.3390/jmse12010102>.
- Fu, H.; Gu, Z.; Wang, Y. Ship Pitch Prediction Based on Bi-ConvLSTM-CA Model. *J. Mar. Sci. Eng.* **2022**, *10*, 840. <https://doi.org/10.3390/jmse10070840>.
- Yan, H.; Chu, Z.; Tang, J. A short-term ship motion prediction method based on quaternions and Transformer-LSTM model. *Ocean Eng.* **2025**, *342*, 122874. <https://doi.org/10.1016/j.oceaneng.2025.122874>.

20. Kendall, A.; Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, California, USA, 4–9 December 2017, Volume 30, pp. 5580–5590. <https://arxiv.org/abs/1703.04977>.
21. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. <https://doi.org/10.1198/016214506000001437>.
22. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017, Volume 70, pp. 1321–1330.
23. Wei, Y.; Chen, Z.; Zhao, C.; Chen, X.; Yang, R.; He, J.; Zhang, C.; Wu, S. Deterministic and probabilistic ship pitch prediction using a multi-predictor integration model based on hybrid data preprocessing, reinforcement learning and improved QRNN. *Adv. Eng. Inform.* **2022**, *54*, 101806. <https://doi.org/10.1016/j.aei.2022.101806>.
24. Kim, G.y.; Lim, C.; Oh, S.j.; Nam, I.h.; Lee, Y.m.; Shin, S.c. Deep Learning-Based Prediction of Ship Roll Motion with Monte Carlo Dropout. *J. Mar. Sci. Eng.* **2025**, *13*, 2378. <https://doi.org/10.3390/jmse13122378>.
25. Li, J.; Han, Y.; Li, K.; Gong, Q.; Teng, Y.; Chu, S. WT-A-LSTM-KAN: A novel hybrid deep learning framework for high-precision short-term prediction of moored ship roll and pitch motions. *Ocean Eng.* **2026**, *346*, 123968. <https://doi.org/10.1016/j.oceaneng.2025.123968>.
26. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>.
27. Romano, Y.; Patterson, E.; Candès, E. Conformalized quantile regression. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 3538–3548.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.