electronics

MDPI

Article

Receiver-Side Reconstruction of Vessel Roll and Pitch Under Simulated Packet Loss and Delay: Accuracy–Freshness Trade-Offs

Siniša Zorica [1,*], Ivan Zorica [2], Marinko Lipovac [1] and Toni Jončić [1]

[1] Department of Professional Studies, University of Split, 21000 Split, Croatia
[2] Coast Guard of the Republic of Croatia, 21000 Split, Croatia
* Correspondence: szorica@oss.unist.hr

Abstract

Vessel roll and pitch measurements transmitted over communication links may be delayed or lost. We compared six receiver-side reconstruction methods: hold, linear interpolation, linear extrapolation, adaptive selection, constant-velocity Kalman filtering, and Rauch–Tung–Striebel smoothing. Recorded motion data from MITO 31 and Seafar were processed under three synthetic communication scenarios: independent 10% packet loss with gamma delays (S1), bursty 10% loss with time-varying delays (S2), and bursty 20% loss (S3). Reconstruction used only packets received before reporting deadlines, with permitted lags of 0–1 s. Parameters were fixed on training data; 50 paired channel realisations were evaluated per segment. Increasing lag from 0.2 to 0.5 s reduced interpolation error by approximately 40–77%. Under S1 at 0.5 s, smoothing reduced MITO 31 roll and pitch errors by approximately 40% and 52% relative to interpolation. Under S3 at 0.5 s, smoothing increased MITO 31 roll error by 28.5%, while adaptive selection improved interpolation by 3.3% for roll and 2.8% for pitch. Doubling bursty loss increased interpolation error by approximately 33–47% at $L = 0.5$ s across the evaluated datasets and axes. The Kalman process-noise setting selected on MITO 31 was not optimal for Seafar. No method consistently dominated: smoothing benefited from timely post-target observations but deteriorated across prolonged bursts. Interpolation with hold fallback remained robust under bursty loss. Method selection should reflect vessel motion and measured link characteristics.

Keywords: vessel attitude; IMU; packet loss; communication delay; receiver-side reconstruction; Kalman smoothing

Academic Editor: Paulo Pereira Monteiro

Received: 22 October 2026
Revised: 25 November 2026
Accepted: 27 November 2026
Published: 29 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Roll and pitch, measured by an onboard inertial measurement unit (IMU), are used for seakeeping assessment, motion monitoring, data logging and as inputs to downstream processing ashore or in other onboard systems [1]. When these time-stamped samples are transported over a non-ideal link, for example, a shipboard wireless segment or a ship-to-shore channel, some packets are lost, and the remainder arrive with variable delay and possibly out of acquisition order [2]. A receiving application must then decide what value to report for a given instant using only the packets that it has actually received.

Two quantities govern this decision. The first is the information available at the receiver, which depends on the loss pattern and the delay distribution of the link. The second is how long the application is willing to wait before it reports a value. If the estimate for target time τ is reported at $t = \tau + L$, a larger permitted lag L allows packets acquired after τ to arrive and be used, at the cost of a less fresh output. L is therefore an application-level design parameter, not a measured network latency. The engineering question is which reconstruction rule to deploy for a given link and lag, and how much accuracy is gained by waiting longer.

Estimation under intermittent and delayed observations is well established for networked control and tracking [3–5], and Kalman filtering and Rauch–Tung–Striebel (RTS) smoothing are the standard model-based tools [6,7]. Recent maritime studies have examined missing ship-motion data, but mainly in terms of their effect on subsequent motion forecasting [8,9]. There is less evidence on a simpler question that arises at every receiver: for recorded vessel attitude, how do simple rules compare with model-based estimators when they must use only the packets delivered by a reporting deadline, and how does this comparison change with the loss pattern?

The aim of this study was to quantify how the permitted reporting lag, the packet-loss structure and the reconstruction method jointly determine roll and pitch reconstruction error. We hypothesised that (H1) a longer permitted lag reduces error for methods that can exploit post-target observations available before the reporting deadline, (H2) model-based extrapolating estimators benefit from independent losses but are penalised by bursty losses, and (H3) a Kalman parameter tuned on one vessel does not transfer unchanged to another. The specific objectives were (i) to define a reproducible, deadline-constrained benchmark with explicit synthetic channel models; (ii) to evaluate six reconstruction methods on two independently recorded vessel datasets, with strict separation of training and evaluation data; (iii) to explain the observed rankings by decomposing errors according to the information available at the deadline; and (iv) to document the reproducible workflow and the availability of code and derived data. The main result is that no method dominated: smoothing was best when post-target packets were usually available, whereas interpolation combined with holding the last value was more robust under bursty loss.

Related Work

Networked estimation research established that packet arrival is itself a random process that must be included in estimator design. Sinopoli et al. [3] showed that the Kalman filter error covariance can diverge when the arrival probability falls below a critical value, and Schenato et al. [5] and Hespanha et al. [4] analysed estimation and control over lossy, delayed networks. Zhang et al. [10] surveyed network-induced constraints, including delay, loss and packet disorder. Bar-Shalom [11] derived exact updates for out-of-sequence measurements. Bursty loss is commonly represented by a two-state Markov (Gilbert) channel [12], which we adopt for S2 and S3. Clock synchronisation, for example, in accordance with IEEE 1588 [13], is a separate problem; here, acquisition timestamps are assumed to be available.

In the maritime domain, state estimation research focuses mainly on navigation and sensor fusion. Gorska and Theotokatos [14] fused GNSS and IMU measurements with an extended Kalman filter for an unmanned surface vehicle, and Alamoush and Ölçer [15] reviewed navigation architectures for maritime autonomous surface ships, in which reliable connectivity is a recognised requirement [2]. Gong et al. [8] evaluated simple filling strategies for missing data before LSTM ship-motion prediction, and Ni et al. [9] proposed an online feature-reconstruction module to maintain forecasting accuracy when roll features are unavailable. Those studies evaluate forecasting accuracy. The present study is

complementary: it evaluates the reconstruction error of already acquired roll and pitch values under an explicit packet-arrival deadline, using established estimators rather than proposing a new one.

2. Materials and Methods

Figure 1 shows the receiving-side setting. Recorded attitude sequences were passed through synthetic packet-delivery models, reconstructed at a receiver under a reporting deadline, and compared with the recorded values. All communication impairments were simulated; no operational maritime link was measured.

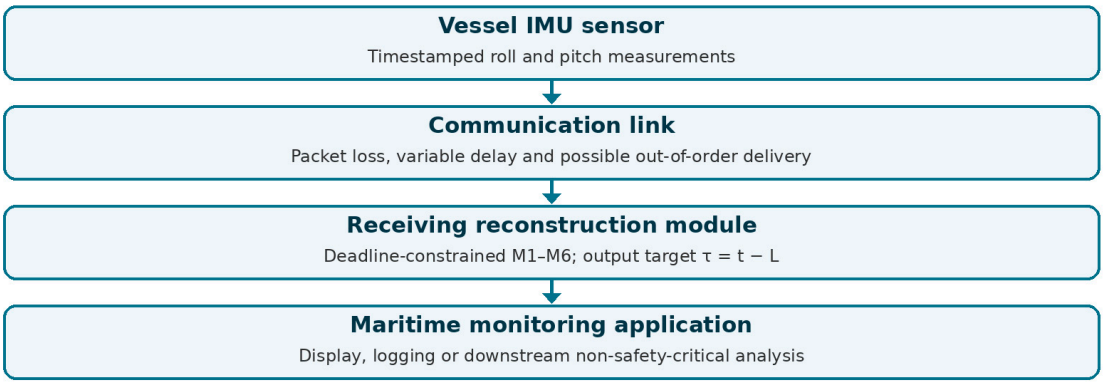


Figure 1. Receiver-side reconstruction setting. The reconstruction module reports, at time t, an estimate for the earlier target time $\tau = t - L$ using only packets that have arrived by t. Loss and delay are simulated in this study.

2.1. Recorded Attitude Data

The MITO 31 records come from the openly deposited full-scale sea-trial dataset of a 9.35 m rigid inflatable boat (maximum breadth 3.35 m; loaded displacement 3200 kg) [16]. Roll and pitch were taken from the IMU PRDID sentences in the processed exports (median sampling interval 0.02 s). The first 30 s of the free-drift record (zero speed; hindcast significant wave height about 0.36 m) and of the straight high-speed run were used only for training. The two turning manoeuvres (43.2 s and 28.0 s) were held out for evaluation. The sensor make, model and calibration are not documented in the dataset, and in the source processing the attitude messages inherit GNSS-derived UTC timestamps. The processed timestamps are therefore treated as the acquisition-time reference.

The Seafar GPS/IMU record [17] contains 143,318 samples at 10 Hz over 3.98 h. Fourteen non-increasing timestamps were removed (each sample whose timestamp did not exceed that of the previously retained sample). Four 60 s evaluation windows started at 10, 35, 60 and 85% of the record duration. For a sensitivity analysis over more physical segments, 19 windows starting at 5, 10, …, 95% were also used; this set contains the four primary windows. Roll and pitch are reported with a resolution of 0.01°. The vessel's dimensions are not documented, so no vessel-size inference is made.

MITO 31 was evaluated at nominal 10 Hz and 20 Hz, and Seafar at 10 Hz. For each nominal grid time, the preprocessing code selects the nearest recorded sample within a tolerance of 0.051 s, removes duplicate selected indices, and retains the selected samples with their original recorded timestamps rather than replacing them with nominal grid times. Consequently, the resulting sequence need not have exactly uniform spacing, and a source observation selected by two nearby grid points is retained only once. The 0.051 s value is the configured maximum admissible mismatch, not a separately measured maximum mismatch. Reference values were never interpolated, and no noise was added. The recorded IMU values serve as the evaluation reference; they are not independent ground-

truth attitude. Table 1 summarises the evaluation segments; motion amplitudes differ by about an order of magnitude between datasets.

Table 1. Recorded segments used for training and evaluation; MITO 31 motion amplitudes are about ten times those of Seafar. SD = standard deviation of the recorded series.

Segment	Duration (s)	Median Δt (s)	Roll SD (°)	Pitch SD (°)	Roll Range (°)	Pitch Range (°)
MITO 31 free drift (training, first 30 s)	30.0	0.02	1.921	1.175	8.19	6.65
MITO 31 straight run (training, first 30 s)	30.0	0.02	1.753	0.874	7.17	4.75
MITO 31 turn 1 (evaluation)	43.2	0.02	1.838	0.946	13.47	5.12
MITO 31 turn 2 (evaluation)	28.0	0.02	2.830	0.992	12.76	5.68
Seafar window 1 (10%)	59.9	0.10	0.370	0.081	1.38	0.31
Seafar window 2 (35%)	59.9	0.10	0.109	0.034	0.52	0.14
Seafar window 3 (60%)	59.9	0.10	0.174	0.112	0.76	0.38
Seafar window 4 (85%)	59.9	0.10	0.113	0.031	0.51	0.14

2.2. Synthetic Packet-Delivery Models

Each sample i acquired at time t_i is transmitted as one packet that is either dropped or delivered with delay d_i; its arrival time is

$$a_i = \begin{cases} t_i + d_i, & \text{packet } i \text{ delivered} \\ +\infty, & \text{packet } i \text{ dropped} \end{cases} \quad (1)$$

Delivered packets may therefore arrive out of acquisition order. Three scenarios were defined (Table 2). In S1, losses are independent with probability $\pi = 0.10$ and delays are gamma-distributed (shape 1, scale 0.2 s). In S2 and S3, losses follow a two-state Markov chain [12], with

$$P(\text{loss}_i \mid \text{loss}_{i-1}) = p_{LL}, \qquad P\left(\text{loss}_i \mid \text{delivery}_{i-1}\right) = p_{DL} = \frac{\pi\,(1 - p_{LL})}{1 - \pi} \quad (2)$$

where π is the stationary loss probability, and the expected loss-burst length is $1/(1 - p_{LL}) = 5$ samples. The delay in S2 and S3 adds a slowly varying deterministic component to the gamma term,

$$d_i = g_i + 0.14\left[1 + \sin\left(\frac{2\pi\,(t_i - t_0)}{37\ \text{s}} + 0.5\right)\right]\ \text{s}, \qquad g_i \sim \Gamma(k = 1,\ \theta = 0.2\ \text{s}) \quad (3)$$

which ranges from 0 to 0.28 s and mimics a periodically congested link. S3 differs from S2 only in the stationary loss probability (0.20 instead of 0.10), so the S2–S3 contrast isolates loss severity. The parameter values are controlled assumptions chosen to span independent and bursty regimes; they were not estimated from measured maritime traffic. In the Markov model, the first loss state is sampled from the stationary loss probability π; the transition probability from delivery to loss is $\pi(1 - p_{LL})/(1 - \pi)$. In Equation (3), t_0 is the acquisition time of the first sample of the segment, so the phase of the periodic delay is reset for every segment (initial phase 0.5 rad).

Table 2. Synthetic channel scenarios; S2 and S3 share the delay model and burst persistence and differ only in loss probability. Empirical loss rates, burst lengths and mean delays were verified on 10^6 simulated packets (Section 2.7).

Scenario	Loss Model	Stationary Loss π	p_{LL}	p_{DL}	Mean Burst (Samples)	Delay Model	Mean Delay (s)
S1 iid_gamma	independent	0.10	–	–	1.11	gamma(1, 0.2 s)	0.20
S2 bursty_variable_delay	Markov	0.10	0.80	0.0222	5	gamma + periodic term	0.34
S3 severe_bursty_loss	Markov	0.20	0.80	0.0500	5	gamma + periodic term	0.34

To pair scenario contrasts, the same pseudorandom stream (uniform variates for loss, and gamma variates for delay) was used for S1, S2 and S3 of a given segment and replicates (common random numbers). Fifty replicates were simulated per evaluation segment, scenario and sampling rate. Replicates perturb the same recorded trajectories; they are not independent voyages.

2.3. Reporting Deadline and Information Set

A target is the acquisition time $\tau_k = t_k$ of a transmitted sample, and its estimate is reported at $t = \tau_k + L$. The eligible information set is

$$\tau = t - L, \qquad \mathcal{R}(t) = \{\, i : a_i \le t \,\} \tag{4}$$

No method may use a packet that arrives after t, but a packet acquired after τ may be used if it has arrived. Permitted lags $L \in \{0, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0\}$ s were evaluated; 0.2 s and 0.5 s were pre-specified as the primary lags. Targets within 1 s of either segment end were excluded. Methods were compared on the common set of targets for which all six estimates existed (at least one eligible sample acquired at or before τ); this excluded 0% of MITO 31 targets and 0.01% of Seafar targets. Figure 2 illustrates the rule.

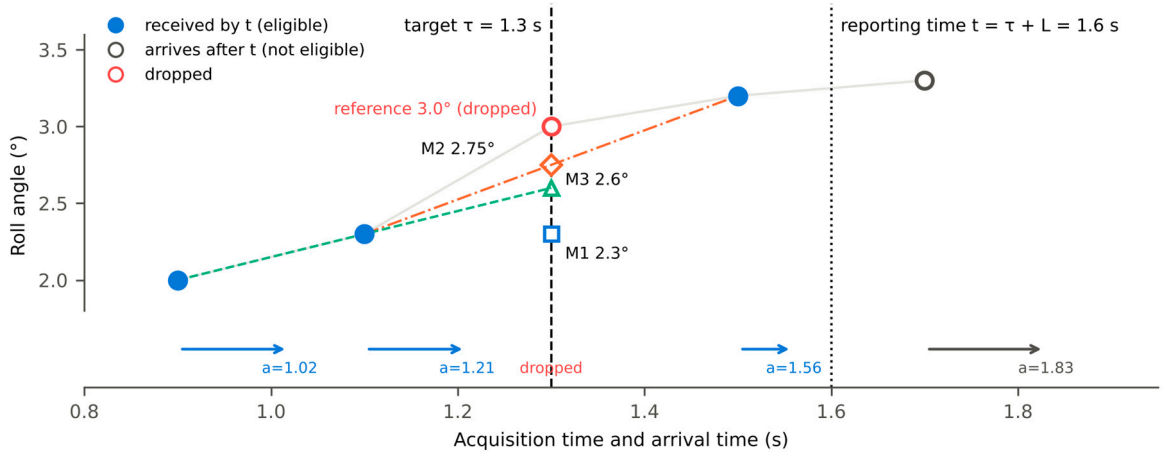


Figure 2. Only packets received by the reporting time can be used, which determines what each method returns (illustrative values). The target sample (3.0°) was dropped; the 1.5 s packet arrived at 1.56 s and is eligible, the 1.7 s packet arrives after t = 1.6 s. M1 returns 2.3° (error −0.7°), M3 extrapolates to 2.6° and M2 interpolates to 2.75°. Arrows show acquisition-to-arrival delay. Filled blue circles mark eligible packets, the open grey circle marks a packet arriving after t and the open red circle marks the dropped target sample; the square, diamond and triangle mark the M1, M2 and M3 estimates, and a denotes the arrival time (s).

2.4. Reconstruction Methods

Let l be the latest eligible sample with $t_l \le \tau$, p the eligible sample before it, and r the earliest eligible sample with $t_r > \tau$. If the target sample itself is eligible, M1–M4 return it exactly.

M1, last-observation hold, returns u_l. M2, linear interpolation, and M3, linear extrapolation, return

$$\hat{u}_{\mathrm{M2}}(\tau) = u_l + \frac{\tau - t_l}{t_r - t_l}(u_r - u_l), \qquad \hat{u}_{\mathrm{M3}}(\tau) = u_l + \frac{u_l - u_p}{t_l - t_p}(\tau - t_l) \tag{5}$$

where M2 falls back to M1 if r does not exist and M3 falls back to M1 if p does not exist. M4, adaptive selector, chooses among M1–M3 using a fixed lookup policy indexed by whether r exists and by the age $\tau - t_l$ of the latest sample (bins ≤ 0.1 s, 0.1–0.3 s, > 0.3 s), giving six categories. In particular, the M3 slope is $(u_l - u_p)/(t_l - t_p)$, using two distinct earlier eligible observations; it is multiplied by $(\tau - t_l)$ and added to u_l. The index p denotes the immediately preceding eligible observation in acquisition time, which need not be the preceding raw-data index.

M5, Kalman filter, and M6, RTS smoother, use, for each axis, the constant-angular-rate model

$$\mathbf{x} = \begin{bmatrix}\phi \\ \dot{\phi}\end{bmatrix}, \quad \mathbf{F}(\Delta t) = \begin{bmatrix}1 & \Delta t \\ 0 & 1\end{bmatrix}, \quad \mathbf{Q}(\Delta t) = q\begin{bmatrix}\Delta t^3/3 & \Delta t^2/2 \\ \Delta t^2/2 & \Delta t\end{bmatrix}, \quad z = [1 \quad 0]\mathbf{x} + v,\ v \sim \mathcal{N}(0, \sigma_v^2) \tag{6}$$

with assumed observation-noise standard deviation σ_v = 0.15° for roll and 0.10° for pitch (filter settings, not instrument specifications). For each target, the eligible samples with recorded indices from min (k − 12, l) to k (M5), or to k + 12 (M6), were processed in acquisition order, with a prediction node inserted at τ. Starting the window at min(k − 12, l) ensures that the latest eligible pre-target sample is included even after an outage longer than 12 samples. The filter was initialised at the first sample in the window with zero rate and rate variance 100 (°/s)². M5 reports a deadline-constrained, retrospectively computed filtered estimate at τ: it may incorporate observations acquired no later than τ that arrived after τ but by $\tau + L$, rather than representing a strictly real-time filter evaluated at τ. M6 reports the RTS-smoothed estimate at τ and additionally uses eligible observations acquired after τ, provided they arrived by $\tau + L$. Thus, M5 and M6 differ in their use of the eligible information set, not solely in computational complexity. Because the window is recomputed for every target, out-of-order arrivals are handled exactly.

The methods are compared as receiver-side reporting strategies under the same packet-arrival deadline, rather than as estimators with identical information usage. M5 is a forward filter that does not incorporate eligible post-target observations into its estimate at the target time, whereas M6 can use such observations through backward smoothing. Any M6 advantage therefore reflects both the smoothing operation and its ability to exploit the additional observations permitted by the reporting lag; it should not be interpreted as an isolated algorithmic superiority over M5.

2.5. Training, Parameter Selection and the M4 Policy

The process-noise intensity q was selected from the grid $\{10^0, \ldots, 10^6\}$ deg² s⁻³ by minimising the mean M5 and M6 RMSE on the two MITO 31 training segments only, under S1 and S2, at both primary lags and both sampling rates (eight replicates from a separate seed domain). The selected value, q = 1000 deg² s⁻³, was frozen and used for both state-space methods, both MITO 31 sampling rates and Seafar. Seafar data and S3 were never used for selection.

The observation-noise standard deviations and the selected process-noise intensity are modelling and tuning assumptions, not independently measured IMU sensor-noise specifications. The same selected settings were retained for Seafar to test transfer across

recordings; this does not establish the best attainable Kalman performance after vessel-specific retuning.

The M4 policy was learned on synthetic signals that were independent of both vessels: 120 s sums of three sinusoids (amplitudes 0.3–4°, periods 2–12 s) plus white noise (SD 0.03°), at 10 and 20 Hz, under S1 and S2, with 40 training replicates per lag. For each permitted lag and category, the candidate (M1, M2 or M3) with the lowest mean squared error was selected; unpopulated categories fall back to M3 without and M2 with a post-target sample. The learned policy selects M2 whenever a post-target sample is available and M3 otherwise, except that, for $L \geq 0.5$ s, it selects M1 when the latest sample is older than 0.3 s.

2.6. Error Metrics and Statistical Analysis

For each segment, scenario, replicate, lag, method and axis $u \in \{\text{roll, pitch}\}$, the run-level error is

$$\mathrm{RMSE}_u = \sqrt{\frac{1}{N_{\mathrm{valid}}} \sum_{k=1}^{N_{\mathrm{valid}}} \left(\hat{u}(\tau_k) - u(\tau_k)\right)^2}, \qquad \mathrm{NRMSE}_u = \frac{\mathrm{RMSE}_u}{\sigma_u} \tag{7}$$

where σ_u is the standard deviation of the 10 Hz recorded segment, and the sum runs over the targets of the common evaluation set of the run (Section 2.3). Reported means average run-level RMSE values. NRMSE is used only in the lag-sweep figure in Section 3.3 to place both datasets on a common scale. Method comparisons were paired (same segment, replicate, scenario and lag). Because MITO 31 provides only two held-out manoeuvres, an exploratory within-recording comparison was performed on 19 Seafar windows: replicate-averaged RMSE per window was compared with the Friedman test across the six methods [18] and with two-sided Wilcoxon signed-rank tests for M4 and M6 against M2 [19], with Holm correction across the 24 Wilcoxon tests [20]. These windows are sampled from one recording, not independent voyages or vessels. Consequently, the p values are descriptive, conditional on this recording and the chosen windows; temporal dependence among windows may affect nominal test calibration. They do not support population-level inference across voyages or vessels.

For the factorial and oscillatory-model uncertainty summaries, differences were formed within the same recorded segment and channel-randomisation replicate. For each replicate, factorial contrasts were averaged over the fixed evaluation segments (two MITO 31 manoeuvres or four Seafar windows), while the oscillatory-model contrasts used two segments per dataset; the reported 95% intervals were calculated as the replicate-mean contrast plus or minus the Student-t 0.975 quantile, with 49 degrees of freedom times its standard error across 50 replicate indices. Thus, each interval is $\bar{d} \pm t_{0.975,49} \frac{s_d}{\sqrt{50}}$, where $\bar{d}$ and s_d are the mean and standard deviation of the 50 replicate-level differences. Factorial contrasts comprise the equally weighted bursty-minus-independent loss-process effect, the gamma-plus-periodic-minus-gamma-only delay-model effect, and their difference-in-differences interaction. With $R(X, Y)$ denoting the mean RMSE under loss process X (I = independent, B = bursty) and delay model Y (G = gamma only, P = gamma plus periodic), the loss effect is $[R(\mathrm{B, G}) - R(\mathrm{I, G}) + R(\mathrm{B, P}) - R(\mathrm{I, P})]/2$, the delay effect is $[R(\mathrm{I, P}) - R(\mathrm{I, G}) + R(\mathrm{B, P}) - R(\mathrm{B, G})]/2$, and the interaction is $[R(\mathrm{B, P}) - R(\mathrm{B, G})] - [R(\mathrm{I, P}) - R(\mathrm{I, G})]$. All 144 factorial effect estimates use 50 paired randomisation replicates per dataset, axis, method and reporting lag. These intervals quantify conditional synthetic-channel Monte Carlo uncertainty and do not provide cluster-level uncertainty across vessels. The exploratory Seafar window-level Wilcoxon tests use a Holm adjustment over 24 comparisons; the 95% intervals for the paired, factorial and oscillatory-model contrasts are pointwise, not familywise-simultaneous, intervals.

The same pairing procedure was also applied to the principal M4 − M2 and M6 − M2 comparisons at L = 0.2 and 0.5 s (48 conditional intervals; Appendix A, Table A1). At L = 0.5 s for MITO 31 roll, M6 − M2 was −0.0310° (95% CI −0.0343 to −0.0277°) under S1, compared with +0.1827° (95% CI +0.1197 to +0.2458°) under S3. Negative values favour M6. All 48 contrasts and their intervals are listed in Table A1; the paired sample counts are given in the supplementary file main_paired_method_conditional_CI.csv.

For the primary 10 Hz benchmark, the two MITO 31 manoeuvres contained 411 and 259 target times, and the four Seafar windows contained 580, 580, 579 and 580 target times, respectively, after the defined boundary exclusions. These are within-recording target counts, not independent experimental units. At each dataset–scenario–lag–axis–method condition, 50 simulated channel realisations per segment generated 100 run-level RMSE values for MITO 31 (two segments) and 200 for Seafar (four windows). For the paired conditional confidence intervals, the difference between methods was averaged over the fixed segments for each matched channel-realisation index, yielding 50 paired replicate-level differences per contrast (49 degrees of freedom); the number of eligible estimates within each run can be smaller than the target count.

2.7. Implementation and Verification

The benchmark was implemented in Python (NumPy, pandas, SciPy and Numba). The complete pipeline, comprising data preparation, policy training, q selection, evaluation and analysis, was timed in the reported implementation (the benchmark stage took approximately 24 s in the original environment). The measured runtimes (approximately 0.8 µs per target and axis for all six methods in the original benchmark environment, averaged over 200 repeated reconstructions of one segment at L = 0.5 s, and approximately 1.5 µs for one 25-sample RTS window) are implementation- and hardware-dependent benchmarks, not guaranteed real-time performance on maritime equipment. Twenty-five automated checks passed. They covered the Figure 2 example; invariance of every estimate when values of packets arriving after the deadline were replaced by random numbers; exactness of M1–M4 on a lossless channel; exactness of M2 and M3 for linear motion; agreement of M5 and M6 with an independent dense Kalman/RTS implementation in 200 random cases (maximum relative difference below 10^{-13}); empirical loss rates, burst lengths and mean delays of all three channels; decimation without interpolation; and consistency of M4 with its policy.

The complete computational workflow was re-executed by the authors on a separate computer (Windows 11; CPython 3.12.7, NumPy 2.5.5, pandas 3.1.1, SciPy 1.18.1, Numba 0.68.1) from a fresh extraction of the supplement, using the derived MITO 31 and Seafar segments it contained; the archived outputs had been produced on Linux with Python 3.13, NumPy 2.5.3, pandas 3.0.5 and Numba 0.68.0. All 18 workflow steps completed: the 25 automated checks (all passed), the baseline M1–M6 benchmark and analysis, the twelve additional analysis scripts, the 50-realisation oscillatory-model workflow and the confidence-interval tables. All 53 regenerated CSV files were compared with the archived files after normalising line endings. Twenty-six, including the 100,800-row main run table and most run-level tables of the additional analyses, had identical content; the remaining 27 files (aggregated tables, the M4 policy and q-selection tables, the channel-sensitivity channel audit, and the run-level outputs of the oscillatory-model, independent-reference and target-retention analyses) differed only by floating-point rounding (maximum relative difference 5×10^{-10}), which does not affect any reported value. The file-by-file comparison is given in REPRODUCIBILITY_AUDIT/RERUN_WINDOWS_2026-11-19/FILE_COMPARISONS_ALL.csv. These checks did not re-download or preprocess the external raw recordings or validate the synthetic communication assumptions.

Engineering interpretation of runtime and window size**.** At a nominal 10 Hz, reporting both roll and pitch requires approximately 20 target-axis estimates per second (40 at 20 Hz). Multiplying these rates by the measured approximately 0.8 µs per target and axis for the combined six-method benchmark gives approximately 16 µs and 32 µs of benchmark computation time per second, respectively, excluding data acquisition, communication, scheduling and other application overhead. These figures are illustrative extrapolations from the reported implementation, not measured end-to-end receiver latency or a guarantee for embedded hardware. The RTS evaluation uses a fixed window of 12 recorded sample indices before and after the target, rather than enlarging the window automatically as L increases. At 10 Hz, this corresponds nominally to up to 1.2 s on either side, and at 20 Hz to up to 0.6 s; the actual observations included remain constrained by packet arrival at the reporting deadline. Thus, increasing L changes which post-target observations are eligible within the fixed window but does not itself increase the configured number of RTS nodes. Increasing the window length would increase the per-target computational work and requires a separate timing assessment.

3. Results

3.1. Illustrative Reconstruction

Figure 3 shows one S2 realisation for MITO 31 roll at L = 0.5 s. Over the 12–26 s interval, the RMSE was 0.169° for M1, 0.069° for M2, 0.059° for M4 and 0.058° for M6. M1 errors occurred mainly when consecutive packets were lost or delayed beyond the deadline.

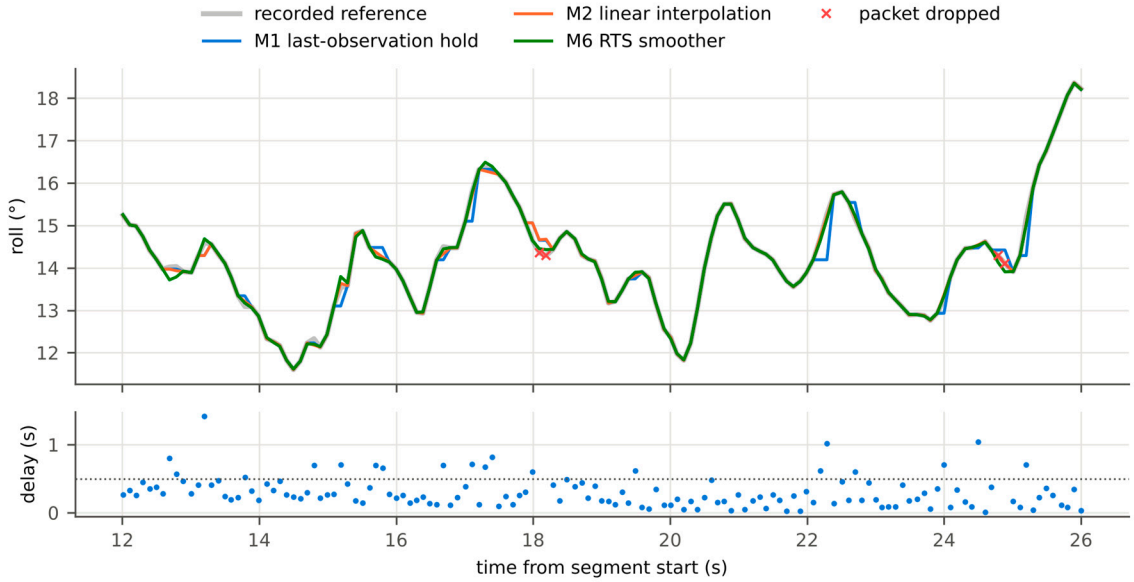


Figure 3. Hold (M1) produces step errors during loss bursts, whereas interpolation (M2) and smoothing (M6) follow the recorded roll more closely. MITO 31 turn 1, scenario S2, L = 0.5 s, 10 Hz, first replicate. Top: recorded reference (grey), reconstructions and dropped packets (×). Bottom: delay of delivered packets; the dotted line marks L.

3.2. Accuracy at the Primary Permitted Lags

Tables 3 and 4 give mean RMSE at the two primary lags. For MITO 31 (Table 3), the smoother (M6) had the lowest error in every S1 condition, whereas in S2 and S3 the lowest error was obtained by M2 at L = 0.2 s and by M4 at L = 0.5 s. Under bursty loss at 0.2 s, M6 roll error was 1.345° against 1.130° for M2 (S3). Extrapolation (M3) and the forward Kalman filter (M5) were never best in MITO 31.

Table 3. MITO 31 mean RMSE (°) at 10 Hz: smoothing (M6) is best under independent loss; interpolation (M2) or the selector (M4) under bursty loss. Lowest value per row in bold; 100 runs per cell (2 manoeuvres × 50 replicates).

Axis	Scenario	*L* (s)	M1	M2	M3	M4	M5	M6
Roll	S1	0.2	0.428	0.345	0.326	0.263	0.321	**0.256**
Roll	S1	0.5	0.225	0.078	0.150	0.075	0.146	**0.047**
Roll	S2	0.2	0.984	**0.977**	1.204	1.194	1.032	1.024
Roll	S2	0.5	0.559	0.450	0.790	**0.424**	0.652	0.517
Roll	S3	0.2	1.139	**1.130**	1.665	1.647	1.358	1.345
Roll	S3	0.5	0.776	0.642	1.261	**0.621**	1.014	0.825
Pitch	S1	0.2	0.338	0.276	0.297	0.240	0.281	**0.223**
Pitch	S1	0.5	0.178	0.065	0.132	0.063	0.124	**0.031**
Pitch	S2	0.2	0.713	**0.707**	1.034	1.027	0.904	0.899
Pitch	S2	0.5	0.381	0.300	0.614	**0.283**	0.516	0.416
Pitch	S3	0.2	0.793	**0.785**	1.362	1.344	1.145	1.134
Pitch	S3	0.5	0.517	0.425	0.981	**0.413**	0.809	0.663

For Seafar (Table 4), M2 had the lowest pitch error in all conditions. For roll, M4 was best under S1, and M6 under S2 and S3, but under bursty loss M2, M4 and M6 differed by at most 16% at 0.2 s and 5% at 0.5 s. Seafar pitch errors (0.0012–0.014°) are of the same order as the quantisation error of the 0.01° recorded resolution ($0.01/\sqrt{12} \approx 0.003°$), so Seafar pitch rankings should not be over-interpreted.

Table 4. Seafar mean RMSE (°) at 10 Hz: interpolation (M2) is best for pitch, and differences among M2, M4 and M6 for roll are small. Lowest value per row in bold; 200 runs per cell (4 windows × 50 replicates).

Axis	Scenario	*L* (s)	M1	M2	M3	M4	M5	M6
Roll	S1	0.2	0.0091	0.0075	0.0062	**0.0052**	0.0067	0.0056
Roll	S1	0.5	0.0048	**0.0017**	0.0034	**0.0017**	0.0036	0.0020
Roll	S2	0.2	0.0227	0.0224	0.0209	0.0207	0.0195	**0.0193**
Roll	S2	0.5	0.0148	0.0113	0.0149	0.0109	0.0136	**0.0108**
Roll	S3	0.2	0.0282	0.0278	0.0286	0.0282	0.0254	**0.0250**
Roll	S3	0.5	0.0211	0.0166	0.0224	0.0162	0.0197	**0.0158**
Pitch	S1	0.2	0.0030	**0.0027**	0.0047	0.0039	0.0054	0.0045
Pitch	S1	0.5	0.0017	**0.0012**	0.0026	0.0013	0.0029	0.0015
Pitch	S2	0.2	0.0060	**0.0059**	0.0113	0.0111	0.0124	0.0123
Pitch	S2	0.5	0.0037	**0.0031**	0.0069	0.0036	0.0073	0.0057
Pitch	S3	0.2	0.0069	**0.0068**	0.0139	0.0137	0.0144	0.0143
Pitch	S3	0.5	0.0049	**0.0041**	0.0099	0.0047	0.0101	0.0082

3.3. Accuracy–Freshness Trade-Off

Figure 4 shows normalised error as a function of the permitted lag. Error decreased monotonically with *L* for M2, M4 and M6 in every condition, consistent with H1 for methods that exploit eligible post-target observations. Between 0.2 s and 0.5 s, M2 RMSE fell by 40–77%; the gain was largest under S1 and smallest under S3, where long loss bursts leave targets without a post-target sample even at 0.5 s. Increasing *L* from 0.75 s to 1.0 s reduced error by 1–12% for M1, M3 and M5, which gain only from late-arriving pre-target packets, and by 5–26% for M2, M4 and M6, which also use post-target samples. At $L = 0$, when no post-target sample is available, extrapolating methods (M3–M6) were better than hold for roll in MITO 31 under S1 and in Seafar under all scenarios, but worse than hold for MITO 31 roll under bursty loss and for pitch in both datasets.

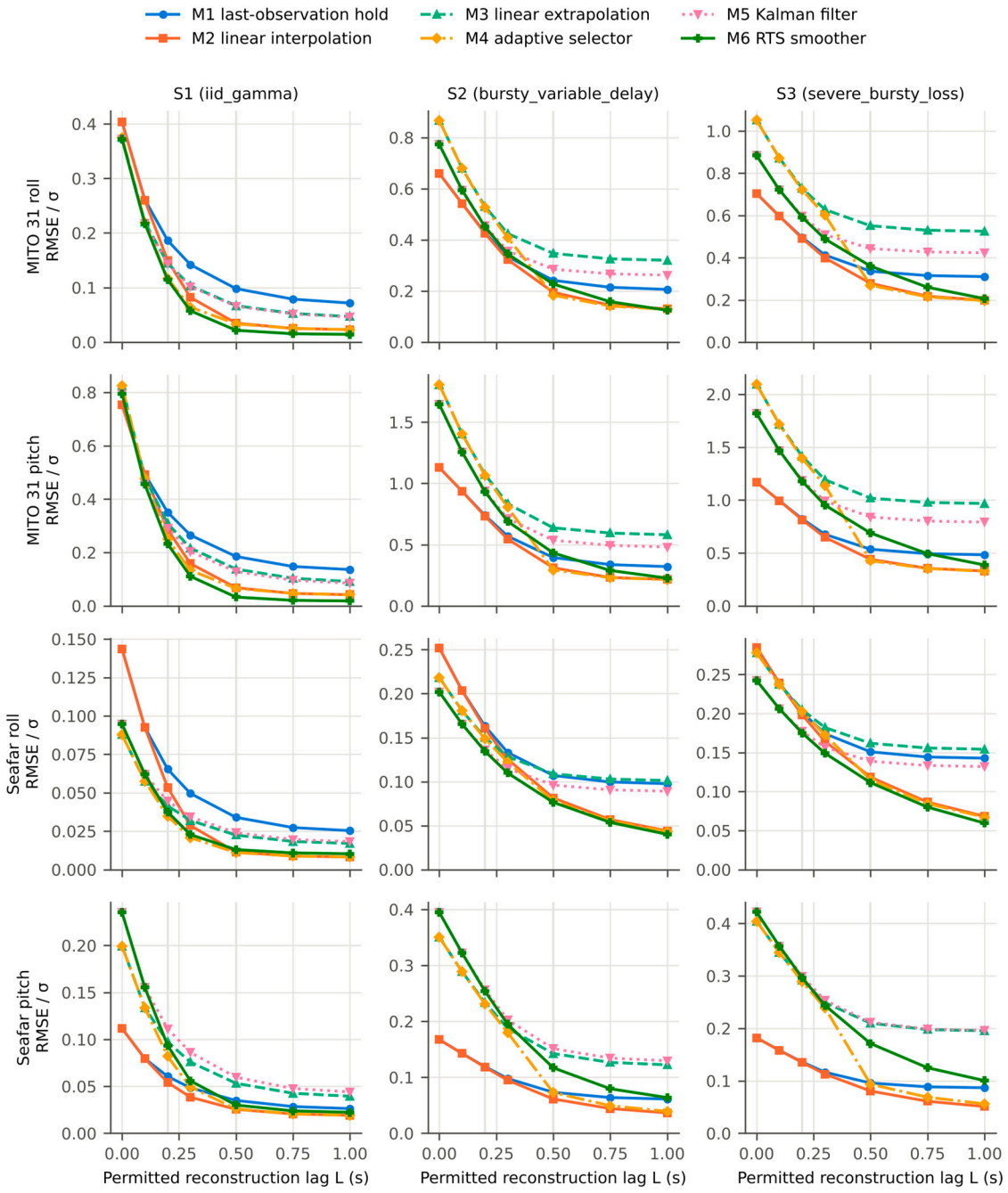


Figure 4. Waiting longer reduces error, and the method ranking depends on the loss scenario. Mean NRMSE (RMSE divided by the segment standard deviation) versus permitted lag L at 10 Hz; columns are channel scenarios, and rows are dataset and axis. Vertical grey lines mark the primary lags 0.2 s and 0.5 s.

3.4. Why Rankings Change: Information Available at the Deadline

Table 5 splits the roll errors at $L = 0.5$ s according to whether an eligible post-target sample existed. When one existed, M6 had the lowest error for MITO 31 in every scenario, and M2 and M6 were nearly equal for Seafar. When none existed, every method had to extrapolate or hold. Under independent loss, these targets were rare (1.8% for MITO 31, S1). Under bursty loss, they were common (32.8%, S3) and their errors were large. For MITO 31 in S3, these targets had an RMSE of 1.159° with hold, 2.127° with extrapolation and 1.584° with the smoother. These targets therefore dominated the overall RMSE and explain why M6 lost its advantage under S2 and S3, consistent with H2. For MITO 31, M4

was best overall under bursty loss at 0.5 s because its policy combines interpolation with hold for stale samples.

Table 5. Bursty loss leaves many targets without a post-target sample, and errors can increase when no eligible post-target sample is available. Roll RMSE (°) at L = 0.5 s and 10 Hz, split by availability of an eligible post-target sample. Lowest value per row in bold; M5 omitted (never lowest).

Dataset	Scenario	Post-Target Sample	Share of Targets (%)	M1	M2	M3	M4	M6
MITO 31	S1	yes	98.2	0.225	0.076	0.155	0.076	**0.049**
MITO 31	S1	no	1.8	0.208	0.208	**0.117**	**0.117**	0.118
MITO 31	S2	yes	73.4	0.399	0.132	0.653	0.132	**0.101**
MITO 31	S2	no	26.6	0.953	0.953	1.671	**0.915**	1.209
MITO 31	S3	yes	67.2	0.548	0.181	1.008	0.181	**0.148**
MITO 31	S3	no	32.8	1.159	1.159	2.127	**1.130**	1.584
Seafar	S1	yes	98.2	0.0049	**0.0016**	0.0034	**0.0016**	0.0019
Seafar	S1	no	1.8	0.0047	0.0047	**0.0036**	0.0037	0.0038
Seafar	S2	yes	73.1	0.0121	**0.0023**	0.0128	**0.0023**	0.0024
Seafar	S2	no	26.9	0.0248	0.0248	0.0288	0.0242	**0.0238**
Seafar	S3	yes	66.6	0.0168	**0.0029**	0.0189	**0.0029**	**0.0029**
Seafar	S3	no	33.4	0.0312	0.0312	0.0359	0.0306	**0.0296**

The M4 decision audit agreed with this explanation. For MITO 31 under S1, M4 selected M2 in 36% of decisions at 0.2 s and in 98% at 0.5 s. Under S3 at 0.5 s, it selected M2 in 67%, M3 in 25% and M1 in 8% of decisions. The full decision counts and category-wise errors for all lags are provided with the code.

3.5. Paired Comparisons and Segment-Level Consistency

Figure 5 shows the distribution of paired relative differences of M4 and M6 with respect to M2. At L = 0.5 s under S1, M6 was better than M2 in 98/100 MITO 31 roll runs and 99/100 pitch runs. Under S3, the M6 differences were widely dispersed in both directions, showing that its performance depended strongly on the realised loss pattern. M4 was better than M2 in 83/100 MITO 31 roll runs in S3 at 0.5 s, by a small margin.

Table 6 reports the segment-level analysis on the 19 Seafar windows. For pitch, M2 ranked first in all 19 windows in every condition (Friedman $p < 0.001$), and M4 and M6 had higher errors than M2, with Holm-adjusted p values below 0.001 in the exploratory window-level comparisons. For roll, M4 had the best mean rank under S1 (best in 14 of 19 windows at 0.2 s). Under S2 and S3, M6 or M4 had the best mean rank, but the corresponding Holm-adjusted p values did not fall below 0.05. Because the 19 windows originate from a single recording and may be temporally dependent, the nominal p values are exploratory, their calibration may be affected by dependence, and they do not support inference across independent voyages or vessels. The broader within-recording analysis showed broadly similar patterns to the four primary windows, without establishing representativeness beyond this recording.

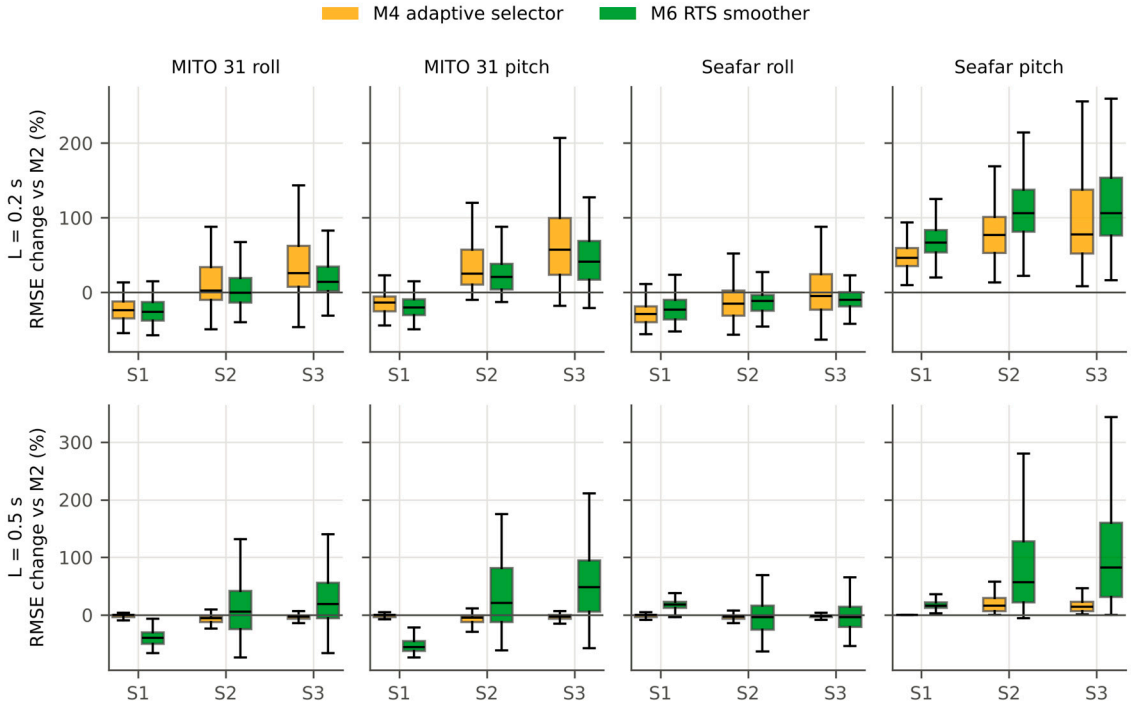


Figure 5. Relative to interpolation (M2), smoothing (M6) shows condition-dependent differences, particularly under bursty loss, whereas the selector (M4) stays close to M2 at L = 0.5 s. Boxes show the interquartile range and median of paired run-level RMSE differences (%), and whiskers the empirical 2.5th–97.5th percentile range of individual run differences (not a confidence interval); MITO 31 n = 100 and Seafar n = 200 per box. Values below zero favour the alternative method.

Table 6. Across 19 Seafar windows, interpolation ranks first for pitch in every condition, whereas roll rankings depend on the scenario. Mean ranks (1 = best) of replicate-averaged RMSE; median relative difference to M2 with exploratory Holm-adjusted Wilcoxon signed-rank p values; 20 replicates per window. Windows are drawn from one recording and do not represent independent voyages.

Axis	Scenario	L (s)	Mean Rank M1/M2/M3/M4/M5/M6	Friedman χ^2 (p)	M4 vs. M2: Median % (p_{Holm})	M6 vs. M2: Median % (p_{Holm})
Roll	S1	0.2	5.0/3.6/3.7/1.5/4.7/2.6	46.9 (<0.001)	−24.9 (0.059)	−18.0 (0.259)
Roll	S1	0.5	5.5/1.7/4.3/1.4/5.3/2.9	84.7 (<0.001)	−1.1 (0.510)	+16.9 (0.014)
Roll	S2	0.2	4.4/3.4/4.4/3.4/3.2/2.2	19.7 (0.001)	−8.4 (1.000)	−13.9 (0.510)
Roll	S2	0.5	4.8/2.6/5.0/1.9/4.5/2.2	54.6 (<0.001)	−4.2 (0.160)	−7.4 (1.000)
Roll	S3	0.2	4.1/2.9/5.0/3.9/3.1/2.0	30.0 (<0.001)	+4.1 (1.000)	−8.9 (0.844)
Roll	S3	0.5	4.7/2.6/5.2/2.2/4.3/2.0	53.2 (<0.001)	−2.5 (0.184)	−5.2 (1.000)
Pitch	S1	0.2	2.0/1.0/4.9/3.0/6.0/4.1	94.5 (<0.001)	+47.6 (<0.001)	+68.2 (<0.001)
Pitch	S1	0.5	4.0/1.0/5.0/2.0/6.0/3.0	95.0 (<0.001)	+1.3 (<0.001)	+17.6 (<0.001)
Pitch	S2	0.2	2.0/1.0/4.1/3.0/6.0/4.9	94.5 (<0.001)	+99.5 (<0.001)	+121.7 (<0.001)
Pitch	S2	0.5	2.4/1.0/5.2/2.6/5.8/4.0	90.7 (<0.001)	+19.8 (<0.001)	+91.1 (<0.001)
Pitch	S3	0.2	2.0/1.0/4.4/3.3/5.7/4.6	83.3 (<0.001)	+103.1 (<0.001)	+130.0 (<0.001)
Pitch	S3	0.5	2.4/1.0/5.2/2.6/5.8/4.0	90.7 (<0.001)	+16.4 (<0.001)	+96.5 (<0.001)

3.6. Effect of Loss Severity

Because S2 and S3 used common random numbers and differ only in loss probability, their contrast is paired. Increasing the bursty loss probability from 0.10 to 0.20 increased mean M2 RMSE by 33–47% at L = 0.5 s and by 11–24% at 0.2 s across datasets and axes. S3

was worse than S2 in at least 99% of paired M2 runs. The increase was larger for M6 (45–60% at 0.5 s), consistent with its sensitivity to long gaps.

3.7. Sampling Rate

At 20 Hz the same loss probability and delay distribution produce shorter gaps in time between eligible samples. Table 7 compares M2, M4 and M6 for MITO 31 at $L = 0.5$ s. All errors were lower at 20 Hz. The advantage of M6 under S1 persisted, and under S2 and S3 at 20 Hz M6 and M4 were close. The 20 Hz results were analysed separately and were not averaged with the 10 Hz results. The controlled rate-matched comparison reported in Section 3.9 shows that this original comparison also changed the physical duration of lost-packet runs and the estimator window; the apparent rate benefit should not be interpreted as a pure sampling-frequency effect.

Table 7. A higher sampling rate lowers reconstruction error for the three methods shown (M2, M4 and M6). MITO 31 mean RMSE (°) at $L = 0.5$ s; lowest value per sampling rate in bold.

Axis	Scenario	M2 10 Hz	M4 10 Hz	M6 10 Hz	M2 20 Hz	M4 20 Hz	M6 20 Hz
Roll	S1	0.078	0.075	**0.047**	0.023	0.023	**0.018**
Roll	S2	0.450	**0.424**	0.517	0.221	0.201	**0.195**
Roll	S3	0.642	**0.621**	0.825	0.340	0.316	**0.314**
Pitch	S1	0.065	0.063	**0.031**	0.019	0.019	**0.011**
Pitch	S2	0.300	**0.283**	0.416	0.174	**0.160**	0.170
Pitch	S3	0.425	**0.413**	0.663	0.258	**0.239**	0.245

3.8. Transfer of the Kalman Tuning

On the MITO 31 training data, the criterion was flat for $q \geq 10^3$ (Figure 6, left). On the held-out MITO 31 manoeuvres, $q = 1000$ was also close to the best grid value. For Seafar, which was not used for selection, the diagnostic held-out curve had its minimum at $q = 10$, where M6 RMSE was 20% lower than at the selected value (Figure 6, right). This is consistent with H3: the process-noise scale reflects the motion bandwidth and amplitude of the training vessel and did not transfer to the smaller-amplitude Seafar motion. The diagnostic curve was not used to retune the reported Seafar results.

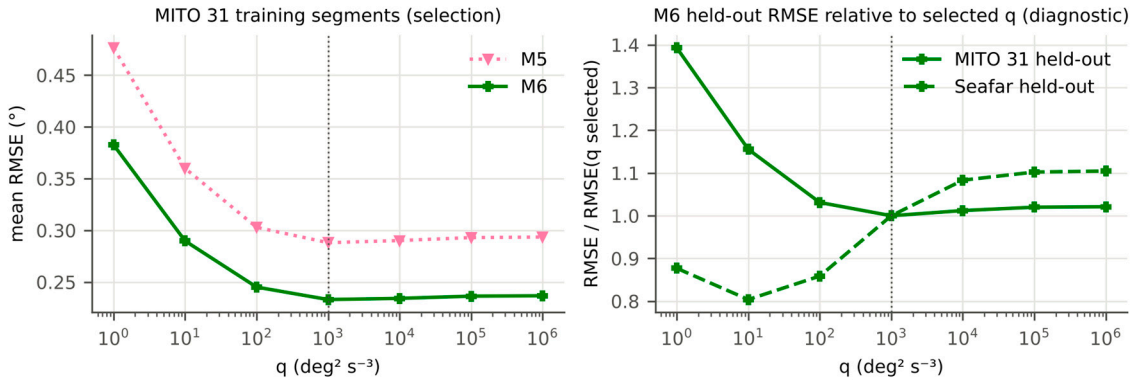


Figure 6. The process-noise scale selected on MITO 31 is too large for Seafar. Left: training criterion on MITO 31 training segments (mean RMSE of M5 and M6); the dotted line marks the selected q. Right: diagnostic held-out M6 RMSE relative to its value at the selected q (10 Hz, both primary lags, all scenarios); these values were not used for selection.

3.9. Additional Controlled Experiments

To separate channel factors that changed simultaneously in the original S1–S2 contrast, we conducted a 2 × 2 analysis, crossing independent versus Markov packet loss with

gamma-only versus gamma-plus-periodic delay at the same nominal 10% loss rate. The results show that both the loss process and the delay model affect reconstruction error. Table 8 collects the principal values of these additional experiments, and Table 9 lists the factorial contrasts for all methods. A second comparison matched the expected mean delay at 0.34 s; the resulting differences cannot be attributed to periodicity alone because the full delay distributions were not matched.

Table 8. Summary of the additional controlled experiments. RMSE values are expressed in degrees (°); the oscillatory-model comparisons report M5, M6 and oscillatory RTS in that order. iid = independent loss; burst = Markov loss; gamma = gamma-only delay; periodic = gamma-plus-periodic delay; OSC = oscillatory-state RTS.

Controlled Comparison	Observed RMSE
Factorial loss/delay, MITO 31 roll, M4, 10 Hz, L = 0.5 s	iid/gamma 0.0749°; iid/periodic 0.1493°; burst/gamma 0.3212°; burst/periodic 0.3950°
Matched expected outage duration, MITO 31 roll, M4, S2-type, L = 0.5 s	10 Hz 0.3602°; 20 Hz 0.3308°
OSC sensitivity, S2, 10 Hz, L = 0.5 s—MITO 31 roll (M5/M6/OSC)	0.655°/0.511°/0.515°
OSC sensitivity, S2, 10 Hz, L = 0.5 s—MITO 31 pitch (M5/M6/OSC)	0.534°/0.408°/0.402°
OSC sensitivity, S2, 10 Hz, L = 0.5 s—Seafar roll (M5/M6/OSC)	0.0098°/0.0079°/0.0091°
OSC sensitivity, S2, 10 Hz, L = 0.5 s—Seafar pitch (M5/M6/OSC)	0.0047°/0.0040°/0.0032°

Table 9. Main effects and interaction of the 2 × 2 loss–delay factorial at 10 Hz (°). Loss = Markov minus independent loss; Delay = gamma-plus-periodic minus gamma-only delay; Int. = interaction (difference-in-differences). Values are means of 50 paired channel realisations, averaged over the two MITO 31 manoeuvres or the four Seafar windows; † marks contrasts whose pointwise 95% conditional confidence interval includes zero. Interval limits for all 144 contrasts are given in the supplementary file factorial_effects_conditional_CI.csv.

Dataset	Method	Axis	L = 0.2 s			L = 0.5 s		
			Loss	Delay	Int.	Loss	Delay	Int.
MITO 31	M1	roll	+0.137	+0.389	−0.078	+0.227	+0.074	−0.036
MITO 31	M1	pitch	+0.088	+0.295	−0.063	+0.160	+0.057	−0.029
MITO 31	M2	roll	+0.144	+0.456	−0.093	+0.240	+0.106	−0.025
MITO 31	M2	pitch	+0.095	+0.343	−0.075	+0.175	+0.071	−0.039
MITO 31	M3	roll	+0.503	+0.398	−0.101	+0.606	+0.056	−0.031
MITO 31	M3	pitch	+0.406	+0.392	−0.128	+0.520	+0.055	−0.037
MITO 31	M4	roll	+0.486	+0.457	−0.086	+0.246	+0.074	−0.001 †
MITO 31	M4	pitch	+0.394	+0.445	−0.122	+0.182	+0.054	−0.029
MITO 31	M5	roll	+0.347	+0.381	−0.122	+0.462	+0.054	−0.032
MITO 31	M5	pitch	+0.297	+0.377	−0.144	+0.418	+0.052	−0.035
MITO 31	M6	roll	+0.344	+0.433	−0.124	+0.382	+0.095	+0.024
MITO 31	M6	pitch	+0.295	+0.427	−0.146	+0.347	+0.084	+0.007 †
Seafar	M1	roll	+0.0074	+0.0072	−0.0020	+0.0099	+0.0013	−0.0009
Seafar	M1	pitch	+0.0011	+0.0019	−0.0002	+0.0014	+0.0005	−0.0002
Seafar	M2	roll	+0.0075	+0.0084	−0.0023	+0.0085	+0.0023	+0.0002 †
Seafar	M2	pitch	+0.0011	+0.0022	−0.0002	+0.0012	+0.0006	+0.0001
Seafar	M3	roll	+0.0106	+0.0053	−0.0007	+0.0115	+0.0010	−0.0004
Seafar	M3	pitch	+0.0029	+0.0038	−0.0005	+0.0036	+0.0008	−0.0003

Seafar	M4	roll	+0.0103	+0.0062	−0.0005	+0.0086	+0.0019	+0.0007
Seafar	M4	pitch	+0.0028	+0.0044	−0.0005	+0.0013	+0.0010	−0.0002
Seafar	M5	roll	+0.0083	+0.0055	−0.0013	+0.0098	+0.0010	−0.0006
Seafar	M5	pitch	+0.0024	+0.0045	−0.0009	+0.0034	+0.0010	−0.0004
Seafar	M6	roll	+0.0082	+0.0065	−0.0013	+0.0076	+0.0019	+0.0008
Seafar	M6	pitch	+0.0024	+0.0053	−0.0010	+0.0029	+0.0013	+0.00003 †

For MITO 31 roll at a 0.5 s reporting lag, the 50-replicate factorial analysis gave M4 RMSE values of 0.0749°, 0.1493°, 0.3212° and 0.3950° for independent/gamma, independent/gamma-plus-periodic, Markov/gamma and Markov/gamma-plus-periodic conditions, respectively. These controlled comparisons supplement rather than replace the original scenario definitions. The factorial runs used a separate random-number stream, so the Markov/gamma-plus-periodic cell is not numerically identical to scenario S2 in Table 3 (0.424°).

Table 9 reports the loss-process main effect, the delay-model main effect and their interaction for all six methods, both datasets, both axes and both primary lags, with the pointwise conditional intervals defined in Section 2.6. At L = 0.5 s, the loss-process effect was the largest contrast in all 24 method–dataset–axis combinations. At L = 0.2 s, the delay-model effect exceeded the loss-process effect in 16 of 24 combinations, namely for MITO 31, both axes with M1, M2, M5 and M6 and pitch with M4 (for M2 roll, +0.456° versus +0.144°) and, for Seafar, pitch with all six methods and roll with M2. With the shorter permitted lag, the additional periodic delay component moves more packets beyond the reporting deadline, so the delay distribution matters as much as the loss process. All 24 interactions at L = 0.2 s were negative, and their intervals excluded zero, indicating that the combined impairment was smaller than the sum of the two main effects; at L = 0.5 s, 20 of 24 intervals excluded zero, and the sign depended on the method. For example, for MITO 31 roll with M4 at L = 0.5 s, the loss-process effect was +0.2461° (95% CI 0.2076 to 0.2845°), the delay-model effect +0.0741° (0.0672 to 0.0809°) and the interaction −0.0006° (−0.0147 to 0.0135°), one of the four interactions whose interval included zero.

The lower and upper interval limits of all 144 contrasts are provided in the supplementary file factorial_effects_conditional_CI.csv; they are pointwise, not simultaneous, intervals.

A sampling-rate sensitivity analysis (MITO 31 roll, S2-type bursty loss and delay, L = 0.5 s) approximately matched the expected physical loss-run duration (0.5 s) and the nominal estimator window duration across 10 and 20 Hz. M4 RMSE decreased from 0.3602° to 0.3308° under the matched configuration, compared with 0.4214° to 0.2057° under the original configuration (recomputed with the sensitivity script's own random-number stream; Table 7 reports 0.424° and 0.201°). Thus, part of the originally reported 20 Hz advantage reflects changes in the physical durations of outages and estimator windows. The additional test matched expected, not realised, outage durations.

The evaluation reference is the recorded IMU sequence, not an independent measure of true vessel attitude. On target-available samples, M1–M4 can reproduce that sequence exactly, so zero error against the recorded sequence is not evidence of zero physical estimation error. We therefore separated target-available and target-unavailable performance and added a synthetic test with a known noiseless reference and independent observation noise. The synthetic test is a diagnostic and does not establish real-vessel ground-truth accuracy.

As a control for the asymmetry in exact-target handling, variants M5-R and M6-R return the recorded target observation when its packet has arrived by the reporting deadline, while retaining the original M5 or M6 estimate otherwise. This was evaluated on the same paired channel realisations and recorded segments as the original methods. Across

the 48 dataset–scenario–lag–axis–method groups in the supplementary comparison, the modified variants did not increase group-mean RMSE. This is a conditional robustness check against the recorded IMU reference, not proof of a lower physical attitude-estimation error.

Across the 24 primary dataset–scenario–lag–axis combinations, substituting M5-R and M6-R did not change the lowest-RMSE method. The largest mean-run M6 reduction was 3.060% for MITO 31 roll and 4.935% for Seafar roll (both S1, $L = 0.5$ s). Detailed results for all 48 method-specific groups and the 24 ranking checks are provided in the supplementary CSV files.

Kalman/RTS sensitivity tests varied window length, initial rate variance and assumed process/observation noise. The RTS result depended strongly on window length, so the original window is not universally optimal. In a within-dataset training/test split, the selected process-noise intensity was $q = 1000$ for MITO 31 and $q = 10$ for Seafar; the Seafar calibration is within one recording and does not demonstrate cross-vessel generalisation. The initial angle equals the first eligible measurement; initial angular rate is zero; initial angle variance equals the assumed observation variance; initial rate variance is 100 $(°/s)^2$; and initial angle–rate covariance is zero. Out-of-order handling is exact only within the finite maintained window.

Tail-error analysis for MITO 31 roll under S2 at $L = 0.5$ s, restricted to target-unavailable samples, gave a mean run-level 95th percentile absolute error of 1.850° for M4, compared with 1.958° for M2 and 2.403° for M6. The maximum observed error was 9.12° for M4 and 18.49° for M6. Channel sensitivity analysis confirmed that longer loss bursts increased error even at fixed nominal loss probability. Channel Monte Carlo replicates quantify uncertainty conditional on the recorded motion segments, not variability among independent vessels or voyages.

The computational re-execution of the benchmark and of the additional analyses is described in Section 2.7. The retained QA.txt and QA_STATUS.json files document checks with different scopes; REPRODUCIBILITY_AUDIT/QA_RECORDS_GUIDE.md explains the scope of each record. These checks confirm that the archived outputs can be regenerated from the supplied code and derived segments; they do not validate the physical attitude reference or the synthetic channel assumptions.

Additional Oscillatory-Model Control and Computational Cost

As an additional model-sensitivity check, we compared an oscillatory-state RTS smoother with the original constant-angular-rate Kalman filter (M5) and RTS smoother (M6). Model parameters were selected on separate training segments and then held fixed during evaluation. We used 50 paired channel realisations per evaluation segment (two MITO 31 manoeuvres and Seafar windows 3 and 4), under scenario S2 at 10 Hz and a reporting lag of $L = 0.5$ s. All methods were evaluated on the same target samples, for which all three returned valid estimates; reported values are means of run-level RMSE. Because M5 and M6 were re-tuned in this control ($q = 100$ deg^2 s^{-3} for MITO 31 and $q = 10$ deg^2 s^{-3} for Seafar, selected on the MITO 31 free-drift and straight-run segments and on Seafar windows 1 and 2), a separate random-number stream was used, and the comparison was restricted to a common target set and segment subset, the M5 and M6 values differ from those in Tables 3 and 4. For MITO 31 roll, mean RMSE was 0.655° (M5), 0.511° (M6), and 0.515° (oscillatory RTS); for pitch it was 0.534°, 0.408°, and 0.402°, respectively. For Seafar roll, the corresponding values were 0.0098°, 0.0079°, and 0.0091°; for pitch, 0.0047°, 0.0040°, and 0.0032°. Oscillatory RTS had lower error for pitch and higher error for roll than M6 in both datasets. Each dataset includes only two evaluation segments, and the training selection used two channel realisations per training segment. Thus, this

exploratory check does not establish model superiority, statistical significance across independent voyages, or cross-vessel generalisation.

Paired run-level comparisons of oscillatory RTS against the original M6 yielded a mean RMSE difference (oscillatory RTS minus M6) of +0.0039° for MITO 31 roll (conditional 95% CI −0.0406° to +0.0485°) and −0.0055° for MITO 31 pitch (−0.0431° to +0.0321°). The corresponding Seafar differences were +0.0012° for roll (+0.0007° to +0.0017°) and −0.0009° for pitch (−0.0011° to −0.0007°). The exploratory comparison therefore cannot demonstrate general model superiority.

A separate implementation-specific timing check covered four segments under S2 at 10 Hz, L = 0.5 s, and a 12-sample window, using 25 measurements after two warm-up repetitions. The median per-target runtime ranged from 10.83 to 11.77 μs for the oscillatory RTS implementation; the separately timed combined call for all six original methods ranged from 0.47 to 0.50 μs per target (roll axis only). The two timings involve substantially different computational workloads and implementations. They are reported only to document measurement conditions, not to compare intrinsic computational efficiency, provide an isolated M6 runtime or establish suitability for real-time deployment. This check also used a different software environment and timing protocol (median of 25 repetitions per segment) from the original benchmark estimate in Section 2.7 (approximately 0.8 μs per target and axis, mean of 200 repetitions), so the two baseline figures are not directly comparable.

4. Discussion

4.1. Interpretation

The results support a simple explanation of when each method helps. Waiting (larger L) helps because it increases the fraction of targets with a post-target sample, which turns an extrapolation problem into an interpolation or smoothing problem. When such a sample is available, the RTS smoother can use more eligible observations than two-point interpolation, and it gave the lowest MITO 31 roll errors in the stratified comparisons of Table 5; this advantage was not universal across both datasets and axes [7]. When it is not available, model-based estimators must predict across the gap. Under independent loss, some prediction-based methods outperformed hold in the evaluated roll conditions, but this did not hold for every axis and recording. Under bursty loss, longer unavailable intervals were associated with larger constant-rate prediction errors in the evaluated recordings, whereas hold showed smaller errors in those particular comparisons. This qualitative dependence on packet availability is consistent with the broad context of intermittent-observation estimation discussed by Sinopoli et al. [3], in which estimator quality depends on the arrival process and not only on the model. Estimator complexity alone therefore does not guarantee lower error; the benefit of a dynamic model depends on the loss structure of the link.

The cross-dataset results should be read within each dataset. The two datasets differ in vessel type, sea conditions, motion amplitude, sampling and sensor resolution, so their absolute RMSE values are not comparable. The method ordering for Seafar pitch, in particular, is measured at the quantisation limit of the recording. The consistent findings across both datasets are qualitative: the benefit of a longer lag, the sensitivity of method rankings to independent and bursty loss, and the poor transfer of a fixed process-noise scale.

4.2. Practical Guidance

The results suggest a design procedure rather than a universal method. First, characterise the link: estimate the loss probability, the burst-length distribution and the delay

distribution relative to the intended reporting lag, not only the mean loss rate, because the tested loss and delay structures affected rankings in these recordings. Second, set the permitted lag from the application's freshness requirement, using curves such as in Figure 4; in these data, 0.5 s instead of 0.2 s reduced interpolation error by 40–77%. Third, for short bursts, consider a smoother tuned on representative motion from the target vessel; for longer bursts, interpolation with hold, such as the evaluated M4 policy, is a robust candidate in these recordings. Because Table 5 shows that the best method differs between targets with and without a post-target sample, switching rules by target-information category could be investigated in future work. Such a hybrid rule was not one of the prespecified methods, and its error and computational cost have not been evaluated here. Finally, validate the chosen rule on the actual link before use. None of the evaluated uses is safety-critical, and no benefit for closed-loop control is claimed.

The observed gains and method rankings are conditional on the two evaluated recordings, the chosen synthetic channel distributions and the reporting deadlines. The factorial contrasts, retained-target Kalman/RTS controls, and exploratory oscillator comparison help explain sensitivity but do not establish a generally superior estimator for different vessels or operating conditions.

4.3. Limitations

The channel models are synthetic and were not fitted to measured maritime traffic; real links can show heavier-tailed delays, correlated delay and loss, and packet aggregation. Each packet carried a single sample, whereas practical systems may batch samples. MITO 31 provided only two held-out manoeuvres, and Seafar windows come from one recording, so between-vessel generalisation cannot be established; segment-level tests on Seafar describe variation within one recording, and their nominal p values may be affected by dependence among windows. Replicates describe Monte Carlo variation conditional on the recordings. The references are IMU outputs with undocumented calibration, and MITO 31 attitude timestamps are derived from GNSS time in the source processing. The state-space model is a deliberately simple constant-rate model with fixed observation noise; the added oscillatory-state control is exploratory and limited to selected segments, one sampling frequency, one reporting lag and one channel scenario; broader validation of oscillatory models and adaptive noise estimation remains future work. Finally, only roll and pitch were considered. Additional controlled tests address several confounds, but neither independent vessel-attitude ground truth nor a measured maritime packet-trace model was available. Exact-target-retaining M5-R/M6-R controls were evaluated; they did not change the best-method ranking in the 24 primary combinations. These controls assess agreement with recorded IMU values and do not establish physical attitude-estimation accuracy. Conditional confidence intervals quantify channel Monte Carlo variability for the evaluated fixed segments, not uncertainty across independent vessels or voyages.

5. Conclusions

The comparison demonstrates that the reporting deadline and the structure of packet loss must be considered together when selecting a receiver-side reconstruction strategy. Longer permitted lag improved interpolation accuracy, but its practical value depends on the freshness required by the application. Smoothing was advantageous for MITO 31 under independent loss, whereas interpolation with a hold fallback was more robust in the evaluated burst-loss conditions. The different behaviour on Seafar also shows why a process-noise setting calibrated on one recording should not be assumed transferable to another vessel.

For maritime monitoring, these findings favour link-aware method selection and vessel-specific parameter calibration rather than a universal estimator. Before deployment,

the approach should be tested using measured onboard or ship-to-shore traffic, additional vessels and sea states, and motion models that account for oscillatory dynamics. The results concern offline evaluation under simulated communication impairments and do not establish suitability for safety-critical or closed-loop applications.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics15235606/s1, Reproducibility_Supplement.zip: reproducibility archive containing the Python implementations and configurations for the six reconstruction methods (M1–M6), random seeds, derived MITO 31 and Seafar motion segments, preprocessing instructions, archived numerical outputs, additional computational analyses, and the REPRODUCIBILITY_AUDIT directory with verification logs and pointwise conditional confidence-interval tables. The retained QA records document their respective verification scopes (REPRODUCIBILITY_AUDIT/QA_RECORDS_GUIDE.md), and REPRODUCIBILITY_AUDIT/RERUN_WINDOWS_2026-11-19/ documents the complete re-execution of the workflow. The archive includes the analyses described in Sections 2.6, 2.7 and 3.9, including M5-R/M6-R target-retention controls, the factorial loss–delay analysis and the oscillatory-RTS sensitivity control.

Author Contributions: Conceptualization, S.Z., I.Z., M.L. and T.J.; methodology, S.Z., I.Z., M.L. and T.J.; software, S.Z.; formal analysis, S.Z. and M.L.; writing—original draft preparation, T.J.; writing—review and editing, I.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original MITO 31 and Seafar recordings are available from the source repositories cited in References [16] and [17]. The derived segments used in the simulations, together with preprocessing instructions, experimental configurations, Python code, seeds and numerical outputs, are provided in Reproducibility_Supplement.zip. A complete re-execution of the workflow on a separate computer reproduced all 53 CSV files written by the workflow, each compared with its archived counterpart, with identical content or differences only due to floating-point rounding (relative difference $\leq 5 \times 10^{-10}$; Section 2.7; file-by-file comparison in REPRODUCIBILITY_AUDIT/RERUN_WINDOWS_2026-11-19/FILE_COMPARISONS_ALL.csv). The external raw recordings were not reprocessed in this check, and no measured maritime communication traces were evaluated. The scope of each verification record is described in REPRODUCIBILITY_AUDIT/QA_RECORDS_GUIDE.md.

Acknowledgments: The authors thank the creators of the openly available MITO 31 and Seafar datasets [16,17]. During the preparation of this manuscript, the authors used ChatGPT (GPT-5 and GPT-6, OpenAI) and Claude (Claude Opus 5.5, Anthropic) to assist with manuscript writing and software development. The authors reviewed and verified the AI-assisted content and take full responsibility for the final manuscript.

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

Appendix A

Table A1 reports the paired conditional confidence intervals for the principal method comparisons defined in Section 2.6.

Table A1. Paired differences in mean RMSE (°) of the selector (M4) and the smoother (M6) relative to interpolation (M2) at 10 Hz, with pointwise conditional 95% confidence intervals (50 paired channel realisations, averaged over the two MITO 31 manoeuvres or the four Seafar windows; Section 2.6). Each cell gives the mean difference and, below it, the interval; Seafar values are given to five decimals because of their small magnitude. Negative values favour M4 or M6; † marks intervals that

include zero. The intervals describe simulated-channel variability conditional on the recorded segments, not variability between vessels.

Dataset	Axis	Scenario	M4 − M2 (L = 0.2 s)	M6 − M2 (L = 0.2 s)	M4 − M2 (L = 0.5 s)	M6 − M2 (L = 0.5 s)
MITO 31	roll	S1	−0.082 (−0.093, −0.070)	−0.088 (−0.100, −0.076)	−0.003 (−0.005, −0.001)	−0.031 (−0.034, −0.028)
MITO 31	roll	S2	+0.216 (+0.094, +0.338)	+0.047 † (−0.006, +0.099)	−0.026 (−0.032, −0.020)	+0.067 (+0.016, +0.119)
MITO 31	roll	S3	+0.517 (+0.369, +0.666)	+0.215 (+0.148, +0.283)	−0.021 (−0.026, −0.016)	+0.183 (+0.120, +0.246)
MITO 31	pitch	S1	−0.036 (−0.044, −0.027)	−0.052 (−0.061, −0.044)	−0.002 (−0.004, −0.001)	−0.034 (−0.037, −0.031)
MITO 31	pitch	S2	+0.320 (+0.226, +0.414)	+0.192 (+0.147, +0.237)	−0.017 (−0.021, −0.012)	+0.116 (+0.066, +0.166)
MITO 31	pitch	S3	+0.559 (+0.448, +0.671)	+0.349 (+0.296, +0.403)	−0.012 (−0.016, −0.008)	+0.238 (+0.179, +0.298)
Seafar	roll	S1	−0.00231 (−0.00242, −0.00220)	−0.00189 (−0.00200, −0.00177)	−0.00004 (−0.00006, −0.00002)	+0.00027 (+0.00024, +0.00029)
Seafar	roll	S2	−0.00173 (−0.00288, −0.00058)	−0.00310 (−0.00356, −0.00263)	−0.00037 (−0.00042, −0.00031)	−0.00051 (−0.00101, -8.7×10^{-8})
Seafar	roll	S3	+0.00042 † (−0.00105, +0.00189)	−0.00276 (−0.00322, −0.00230)	−0.00039 (−0.00043, −0.00034)	−0.00074 (−0.00124, −0.00023)
Seafar	pitch	S1	+0.00128 (+0.00122, +0.00134)	+0.00185 (+0.00177, +0.00193)	+0.00002 (+0.00001, +0.00003)	+0.00022 (+0.00021, +0.00024)
Seafar	pitch	S2	+0.00520 (+0.00466, +0.00575)	+0.00636 (+0.00601, +0.00671)	+0.00052 (+0.00047, +0.00058)	+0.00257 (+0.00219, +0.00295)
Seafar	pitch	S3	+0.00691 (+0.00615, +0.00766)	+0.00744 (+0.00704, +0.00785)	+0.00058 (+0.00053, +0.00063)	+0.00409 (+0.00360, +0.00458)

References

1. Fossen, T.I. *Handbook of Marine Craft Hydrodynamics and Motion Control*, 2nd ed.; John Wiley & Sons: Chichester, UK, 2021. https://doi.org/10.1002/9781119575016.
2. Höyhtyä, M.; Huusko, J.; Kiviranta, M.; Solberg, K.; Rokka, J. Connectivity for Autonomous Ships: Architecture, Use Cases, and Research Challenges. In Proceedings of the 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 18–20 October 2017; pp. 345–350. https://doi.org/10.1109/ICTC.2017.8191000.
3. Sinopoli, B.; Schenato, L.; Franceschetti, M.; Poolla, K.; Jordan, M.I.; Sastry, S.S. Kalman filtering with intermittent observations. *IEEE Trans. Autom. Control* **2004**, *49*, 1453–1464. https://doi.org/10.1109/TAC.2004.834121.
4. Hespanha, J.P.; Naghshtabrizi, P.; Xu, Y. A survey of recent results in networked control systems. *Proc. IEEE* **2007**, *95*, 138–162. https://doi.org/10.1109/JPROC.2006.887288.
5. Schenato, L.; Sinopoli, B.; Franceschetti, M.; Poolla, K.; Sastry, S.S. Foundations of control and estimation over lossy networks. *Proc. IEEE* **2007**, *95*, 163–187. https://doi.org/10.1109/JPROC.2006.887306.
6. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45. https://doi.org/10.1115/1.3662552.
7. Rauch, H.E.; Tung, F.; Striebel, C.T. Maximum likelihood estimates of linear dynamic systems. *AIAA J.* **1965**, *3*, 1445–1450. https://doi.org/10.2514/3.3166.
8. Gong, J.; Ni, P.; Fu, Z.; Qin, Y. Study on the prediction performance of ship motion in waves by LSTM under missing data. *Brodogradnja* **2026**, *77*, 77311. https://doi.org/10.21278/brod77311.
9. Ni, P.; Gong, J.; Vladimir, N.; Fan, A.; Fu, Z.; Qin, Y. Real-time ship motion prediction under missing-feature conditions via MLP-assisted online feature reconstruction and LSTM. *Ocean Eng.* **2026**, *366*, 127500. https://doi.org/10.1016/j.oceaneng.2026.127500.
10. Zhang, L.; Gao, H.; Kaynak, O. Network-induced constraints in networked control systems—A survey. *IEEE Trans. Ind. Inform.* **2013**, *9*, 403–416. https://doi.org/10.1109/TII.2012.2219540.

11. Bar-Shalom, Y. Update with out-of-sequence measurements in tracking: Exact solution. *IEEE Trans. Aerosp. Electron. Syst.* **2002**, *38*, 769–777. https://doi.org/10.1109/TAES.2002.1039398.
12. Gilbert, E.N. Capacity of a burst-noise channel. *Bell Syst. Tech. J.* **1960**, *39*, 1253–1265. https://doi.org/10.1002/j.1538-7305.1960.tb03959.x.
13. *1588-2019*; IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems; IEEE: Piscataway, NJ, USA, 2020. https://doi.org/10.1109/IEEESTD.2020.9120376.
14. Gorska, K.; Theotokatos, G. Multi-sensor data fusion for state estimation of unmanned surface vehicles. *Ships Offshore Struct.* 2026, *in press*. https://doi.org/10.1080/17445302.2026.2727610.
15. Alamoush, A.S.; Ölçer, A.I. Maritime Autonomous Surface Ships: Architecture for Autonomous Navigation Systems. *J. Mar. Sci. Eng.* **2025**, *13*, 122. https://doi.org/10.3390/jmse13010122.
16. Niazmand Bilandi, R.; Roshan, F.; Mancini, S.; Vitiello, L. *Dataset for Sea-Trial Data Processing and Manoeuvring Analysis of a Rigid Inflatable Boat (RIB) MITO 31. Mendeley Data, Version 1*; Mendeley: London, UK, 2026. https://doi.org/10.17632/pz37h8vtjt.1.
17. Chernyavskiy, A. *Realistic GPS and IMU Data. Zenodo, Version 0.0*; Zenodo: Geneva, Switzerland, 2022. https://doi.org/10.5281/zenodo.6557994.
18. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. https://doi.org/10.1080/01621459.1937.10503522.
19. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. https://doi.org/10.2307/3001968.
20. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70. Available online: https://www.jstor.org/stable/4615733 (accessed on 29 November 2026).

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.