

Ordinal temporal learning for operational quality control of significant wave-height observations from coastal buoy networks

Marcos Batista Figueredo

mbfigueredo@uneb.br

Universidade do Estado da Bahia

Jana Ribeiro Santana

Universidade do Estado da Bahia

Maria Dolores Ribeiro Orge

Universidade do Estado da Bahia

Leila Maria Mendes Santos

Universidade do Estado da Bahia

Lizandra Santos Moraes

Universidade do Estado da Bahia

Leticia Nayara Barreto dos Santos

Universidade do Estado da Bahia

Everton Vitor Almeida Monville

Universidade do Estado da Bahia

João Antonio Assis Santana Batista

Universidade do Estado da Bahia

Elaine Pereira Garrido

Universidade do Estado da Bahia

Marta Giovanetti

Università Campus Bio-Medico di Roma

Vagner Souza Fonseca

Universidade do Estado da Bahia

Research Article

Keywords: quality control, significant wave height, coastal buoy, sensor fault detection, ordinal classification, ocean observational systems

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11259902/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Ordinal temporal learning for operational quality control of significant wave-height observations from coastal buoy networks

Marcos Batista Figueredo^{1*}, Jana Ribeiro de Santana¹,
Maria Dolores Ribeiro Orge¹, Leila Maria Mendes Santos¹,
Lizandra Santos Morais¹, Leticia Nayara Barreto dos Santos¹,
Everton Vitor Almeida Monville¹, João Antonio Assis de
Santana Batista¹, Elaine Pereira Garrido¹, Marta Giovanetti⁴,
Vagner de Souza Fonseca^{1,3,5}

^{1*}Departamento de Ciencias Exatas e da Terra, Campus II,
Universidade do Estado da Bahia (UNEB), Alagoinhas, Bahia, Brazil.

³Centre for Epidemic Response and Innovation (CERI), School for Data
Science and Computational Thinking, Stellenbosch University,
Stellenbosch, South Africa.

⁴Department of Science and Bio-Technology, Università Campus
Bio-Medico di Roma, Rome, Italy.

⁵Laboratório de Arbovírus e Vírus Hemorrágicos, Instituto Oswaldo
Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil.

*Corresponding author(s). E-mail(s): mbfigueredo@uneb.br;

Abstract

Reliable sensor data are essential for coastal monitoring, navigation, port operations and infrastructure management. We present an online temporal ordinal classifier for real-time quality control of significant wave-height observations from coastal buoy networks. The method combines a univariate temporal predictor with one-sided residual and consistency features, accumulated temporal evidence and an ordinal decision head. Each observation is assigned to GOOD, SUSPECT or BAD from the current value and its preceding temporal context. We trained the model on pooled records from six public Brazilian Coastal Monitoring System buoys and evaluated it using chronological hold-out partitions, leave-one-buoy-out transfer, controlled sensor-failure scenarios and a sequential

streaming simulator. The evaluation included seven failure families, comparator criteria, and a predictor-configuration ablation. On the pooled controlled-failure test set, the classifier achieved a binary F1 score of 0.853, a BAD-class F1 of 0.614, precision of 0.890, recall of 0.819 and BAD-class AUPRC of 0.648. Leave-one-buoy-out transfer yielded a mean binary F1 of 0.856 across the six held-out stations. In streaming simulation, abrupt and sentinel failures were detected promptly, whereas drift and stuck-sensor episodes required accumulated evidence. The results support an auditable quality-control workflow for operational ocean-observation systems and identify the conditions under which human review remains necessary.

Keywords: quality control, significant wave height, coastal buoy, sensor fault detection, ordinal classification, ocean observational systems

1 Introduction

Coastal buoy networks provide significant wave height observations in near real time for navigation, port operations, coastal infrastructure management and environmental monitoring. Their records also support the statistical analyses used to characterise extremes, estimate design conditions and assess long-term changes in the marine environment (Bushnell, 2016; Bushnell, Waldmann, Seitz, Henslop, & Lara-Lopez, 2019). The reliability of these observations is therefore an engineering concern. An incorrect value can trigger an inappropriate operational decision, while an undetected failure retained in a historical archive can affect models, design statistics and subsequent management decisions (Bushnell et al., 2019; Tan, Zhang, Wu, Dong, & Cheng, 2022).

Measurements acquired by moored instruments are exposed to drift, saturation, freezing, communication errors and other failure mechanisms that produce anomalous or physically inconsistent observations (Abeyirigunawardena, Jeffries, Morley, Bui, & Hoeberechts, 2016; Diamant, Shachar, Makovsky, Ferreira, & Cruz, 2020). Automated fault detection and quality assurance are consequently part of the monitoring infrastructure itself, particularly when the volume and velocity of the data make continuous manual inspection impractical. Meeting this requirement in an operational setting increasingly means embedding learning-based, real-time automation directly in the data pipeline, rather than relying solely on offline statistical review.

Oceanographic quality assurance has traditionally relied on deterministic tests based on physical limits, climatology, rate of change, temporal continuity, spikes and local dispersion (Haines, Crout, Bosch, Symonds, & Thomas, 2011; Tan et al., 2022). Operational recommendations such as the Quality Assurance/Quality Control of Real-Time Oceanographic Data (QARTOD) procedures provide practical rules for flagging observations and directing them to later expert review (Haines et al., 2011; Kerfoot, Baltes, Bushnell, Campbell, & Knee, 2016). These procedures are valuable because they are transparent and easy to audit, but their performance depends on thresholds and reference distributions that may not transfer between regions (Bushnell et al., 2019; Tan et al., 2022).

Quality-control applications across coastal and ocean observing programmes indicate that deterministic rules often need to be interpreted within local hydrodynamic and climatic regimes. Fixed limits can therefore confuse genuine environmental variability with sensor faults, whereas broad physical envelopes may leave subtle or persistent failures undetected. The engineering task is consequently to distinguish environmental variability from measurement failure while maintaining a traceable decision process (Bushnell et al., 2019; Iwaniak, Zalewska, & Sapiega, 2025; Modi, Bhaskar, Pg, & Balakrishnan Nair, 2026; Tan et al., 2022).

Artificial intelligence has expanded the set of tools available for this problem (Castelão, 2021; Tan et al., 2022). Bayesian procedures and multiple classifier systems have been used to assess marine sensor reliability online, demonstrating that quality assessment can be formulated as an inference and decision problem rather than as a single threshold test (Rahman, Smith, & Timms, 2013; Smith, Timms, de Souza, & D’Este, 2012). Machine learning has also been applied to oceanographic data quality control, including probabilistic classification of hydrographic observations, deep learning support for ocean data management, and recurrent autoencoder models for buoy temperature anomalies (Castelão, 2021; Gao, Song, Hu, Song, & Cheng, 2023; Mieruch, Demirel, Simoncelli, Schlitzer, & Seitz, 2021). Recent contributions combine hybrid artificial intelligence with automated quality control of autonomous observing systems and use interpretable detector selection to support operational marine data pipelines (Aiwale, Stahl, & Sun, 2026; Nguyen, Skålvik, Sylligardos, Palpanas, & Alvsvåg, 2025).

These studies establish the practical value of learned representations (Aiwale et al., 2026; Castelão, 2021; Mieruch et al., 2021; Nguyen et al., 2025), but they also expose recurring requirements that are common to intelligent fault detection and diagnostics across engineering domains, not specific to ocean sensing. Rare-fault imbalance complicates training and evaluation (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; He & Garcia, 2009). Changing environmental and instrumental conditions create concept drift (Nguyen, Høidal, & Pelliccione, 2024), and a model developed for one station may not retain its performance at another station (Bauer, Su, Skovgaard Madsen, & Larsen, 2026; F. Wang, Ye, Chen, & Qiu, 2026).

For wave observations specifically, artificial intelligence studies often focus on forecasting significant wave height with recurrent, convolutional and hybrid decomposition architectures, often reporting improvements in global prediction error (Bekiryazıcı, Amarouche, Ozcan, & Akpınar, 2025; Huang, Wu, Hu, & Cui, 2026; Ji, Han, Jiang, & Liu, 2023; Ren, de Vos, Ravishanker, & O’Donnell, 2026; Zhao, Li, Zhang, & Teng, 2023). Other studies address bias correction or integrated wave prediction systems (da Silva, Barreto, Costa, & Camargo, 2025; K. Wang, Liu, Xing, & Lv, 2024; K. Xie & Zhang, 2025). Prediction accuracy alone, however, does not provide an operational quality flag (J. Xie, Jiang, Song, & Yang, 2023). A low average error can coexist with poor sensitivity to isolated spikes, missing values, stuck sensors and slowly evolving drift.

In quality-control pipelines, temporal prediction is most useful when its residuals are interpreted together with persistence, physical consistency and explicit decision rules, rather than being treated as an isolated forecast score. This combination has

been explored in wave-observation and buoy-monitoring applications, but the evidence remains heterogeneous in terms of temporal direction, station transfer and operational validation (Modi et al., 2026; Tan et al., 2022; J. Xie et al., 2023). These studies leave open the problem of a one-sided temporal and transferable decision system for heterogeneous buoy networks.

The unresolved engineering problem is therefore not wave-height prediction alone. It is the design of an intelligent monitoring system that converts streaming observations and prediction residuals into reliable, auditable and actionable fault assessments. Recent reviews of ocean-observation quality control emphasize the need to combine physical consistency, statistical evidence, expert knowledge and automated procedures across historical and real-time data workflows (Bushnell et al., 2019; Tan et al., 2022). Online probabilistic and classifier-based systems have demonstrated the feasibility of automated marine sensor assessment (Rahman et al., 2013; Smith et al., 2012). Stream-oriented approaches address changing data distributions and concept drift (Nguyen et al., 2024).

A system intended for heterogeneous buoy networks should therefore use only information available when each decision is made, represent different levels of fault severity, transfer across stations with distinct environmental regimes, and be evaluated under continuous operation. The need for transferability is supported by machine-learning quality-control studies that explicitly examine cross-station performance and domain shift (Bauer et al., 2026; F. Wang et al., 2026). Wave-specific quality-control studies further show the importance of connecting prediction residuals with explicit anomaly decisions (Modi et al., 2026; J. Xie et al., 2023). Within the literature reviewed for this study, these requirements are not addressed jointly, particularly the combination of one-sided temporal decision-making, ordinal severity assessment, cross-station transfer and feedback-aware streaming evaluation (Bauer et al., 2026; Nguyen et al., 2024; Tan et al., 2022; J. Xie et al., 2023).

Table 1 summarizes the positioning of the proposed system relative to representative approaches.

The objective of this study is to develop and evaluate an online temporal ordinal classifier for real-time quality control of significant wave-height observations in coastal buoy networks. The proposed system treats quality assessment as an engineering decision problem with three ordered outcomes, GOOD, SUSPECT and BAD, rather than as a single binary rejection. Its methodological contribution is the integration of a univariate temporal predictor, residual and consistency indicators, slow-deviation accumulators and explicit severity thresholds using information available at the decision time. Its engineering contribution is a transferable and auditable quality-assurance tool for significant wave-height measurements collected under heterogeneous coastal conditions.

The study makes three contributions. First, it introduces an online temporal architecture that combines instantaneous and accumulated evidence for graded quality assessment. Second, it evaluates alternative predictor configurations and controlled failure scenarios alongside statistical comparator criteria. Third, it establishes an engineering assessment of cross-station transferability and continuous monitoring under

Table 1 Comparison of representative approaches to oceanographic data quality control and marine sensor fault detection. A check mark indicates that the characteristic was explicitly addressed, whereas a cross indicates that it was not addressed or not reported. SE, streaming evaluation; CD, one-sided temporal decision; OS, ordinal severity; CT, cross-station transfer.

Study	Application and data	Main approach	SE	CD	OS	CT
Tan et al. (2022)	Ocean observations, review	Review of quality-control methods and future directions	–	×	×	×
Smith et al. (2012)	Marine sensors	Bayesian online quality assessment	Online	×	×	×
Rahman et al. (2013)	Marine sensor networks	Multiple-classifier quality assessment	–	×	×	×
Castelão (2021)	Oceanographic observations	Machine-learning quality control	–	×	×	×
Mieruch et al. (2021)	Ocean data management	Deep learning support for quality control	–	×	×	×
Aiwale et al. (2026)	Argo float data	Hybrid artificial intelligence quality control	–	×	×	×
Nguyen et al. (2025)	Marine data streams	Interpretable anomaly-detector selection	–	×	×	×
J. Xie et al. (2023)	Wave observations, four stations	Multivariate Long Short-Term Memory prediction and peak detection	×	×	×	×
Iwaniak et al. (2025)	Baltic Sea temperature and wave data	Regionally adapted deterministic tests	×	×	×	×
Nguyen et al. (2024)	Marine sensor streams	Concept-drift-adaptive anomaly detection	Yes	×	×	×
Bauer et al. (2026); F. Wang et al. (2026)	Sea-level and wave observations	Machine-learning quality control and domain-shift mitigation	–	×	×	✓
This study	Six coastal buoy records	Online temporal ordinal learning with temporal evidence and auditable decision rules	Yes	✓	✓	✓

sequential error feedback. Together, these contributions connect an applied artificial intelligence method to the operational requirement for reliable and interpretable quality control.

The empirical evaluation is designed to test both the artificial intelligence method and its operational usefulness. We train one model on pooled observations from six public buoys in the Brazilian Coastal Monitoring System and assess its performance on held-out data and under leave-one-buoy-out transfer. Documented synthetic corruption families establish a controlled event benchmark, while rule-derived labels on uncorrupted observations are retained as weak operational labels and reported separately.

We also compare threshold calibration strategies and conventional detectors, not as competing contributions but as engineering baselines for false-alarm control. Finally, a sequential streaming experiment measures onset latency, episode coverage and the effect of error feedback for abrupt, persistent and slowly evolving failure modes.

Accordingly, the study investigates three questions. First, does calibrating decision thresholds on the local target series improve false-alarm control relative to fixed constants inherited from another region? Second, can a single one-sided temporal ordinal model trained on pooled multi-station data transfer to unseen coastal buoys without station-specific adjustment? Third, how do online streaming and error feedback affect detection latency and episode coverage across distinct sensor-failure mechanisms? By answering these questions jointly, the study positions quality control as an applied artificial intelligence problem in engineering monitoring, while retaining the transparency and expert-review requirements of operational oceanographic data qualification.

2 Materials and Methods

This section formalises the computational architecture, statistical protocols, experimental benchmarks and validation metrics developed to evaluate real-time quality control of coastal wave observations. The workflow integrates multi-station observational records, formal learning objectives and auditable decision rules designed to address the research questions formulated above.

2.1 Study design

We evaluated automatic quality control of significant wave height (H_s) under three constraints that operational deployment imposes and that previous work has addressed separately (Bauer et al., 2026; Nguyen et al., 2024; Tan et al., 2022), namely thresholds calibrated on the series under analysis rather than inherited from another region, one model applicable to buoys with different wave regimes and sensor payloads, and validation under a continuous stream rather than on a static partition.

Two branches produced persisted quality-control labels and a third produced benchmark outputs, as shown in Fig. 1. The reference branch applied three statistical protocols to eleven observed variables. The proposed branch applied a one-sided temporal ordinal classifier to H_s alone. The benchmark branch supplied the comparison criteria listed in Table 1 and did not contribute labels to the persisted archive. Table 2 states the training data, evaluation data, ground truth and replication of every experiment. In this context, the proposed method denotes the one-sided temporal ordinal classifier; the calibrated criteria are benchmark methods and are not claimed as contributions.

2.2 Data and preprocessing

The BA-1 buoy (12.9895°S, 38.5418°W; Fig. 2) is moored in 30 m of water at the entrance of Baía de Todos os Santos, Brazil, within the SiMCosta coastal network. Semidiurnal tidal currents, river discharge and exchange with the open shelf combine there, so genuine and rapid increases in wave energy can resemble instrument failure, allowing the site to discriminate between methods that separate physical events from sensor error and methods that do not. Sea-state parameters came from a downward-looking acoustic Doppler current profiler (Nortek Aquadopp, 600 kHz), salinity, dissolved oxygen and near-surface temperature from a CTD, and turbidity,

Table 2 Experimental design. Each row states what the experiment trained on, what it was evaluated on, against which ground truth, and how many runs it comprised.

Experiment	Training data	Evaluation data	Ground truth	Runs
E1, comparator benchmark	BA-1, first 70% of a 1-yr window	BA-1, final 30%	Offline injection, 1 in 130	3×5
E2, proposed model	Six buoys pooled, first 60% of each; middle 20% for validation	Six buoys, final 20% of each	Strong labels, seven families	3×5
E3, transfer evaluation	Five buoys pooled	Held-out buoy	Strong labels, seven families	6 folds (LOBO)
E4, streaming	E2 frozen, not retrained	BA-1 real series, 8592 h	Episode onset times	10 seeds
E5, reconciliation	E2 frozen, not retrained	BA-1 real series, 8520 h	None, agreement only	1
E6, predictor ablation	BA-1, recent 8760 h	Controlled point events	3×5	

E1 uses the multivariate three-step benchmark network; E2 and E3 use the one-sided temporal ordinal classifier.

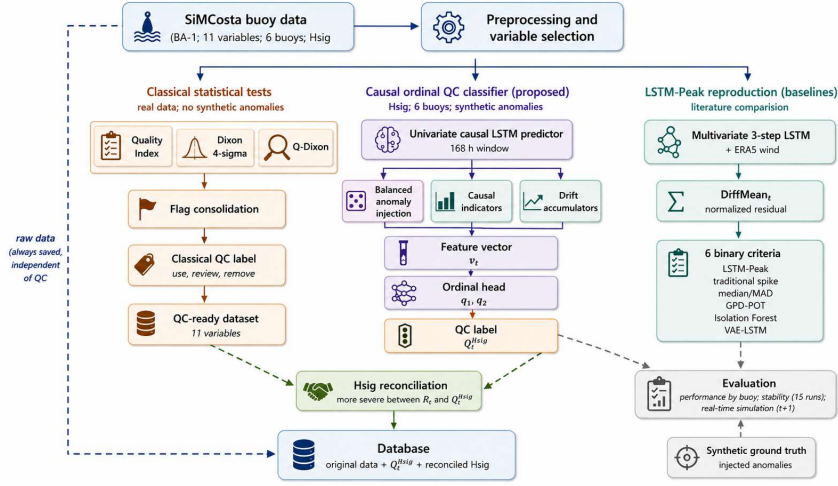


Fig. 1 Analysis pipeline. Three blocks share the same raw data and preprocessing. The reference and proposed branches converge on the persisted database; the comparator branch feeds evaluation only.

chlorophyll fluorescence and coloured dissolved organic matter from optical sensors. Because the platform carries no anemometer, ERA5 10 m wind speed and gust, retrieved through the Open-Meteo historical API at the nearest grid node, entered the benchmark branch as auxiliary covariates and are not in situ measurements.

The record contained 115,478 samples between 2 August 2019 and 25 July 2026, sampled irregularly at intervals of the order of 30 min. All series were resampled to hourly medians, gaps of up to 3 h were linearly interpolated for the historical benchmark, and longer gaps were retained as missing with an explicit mask. The interpolation mask was preserved as a model feature. This reconstruction step belongs to the offline preparation of the public archive; a live implementation should replace

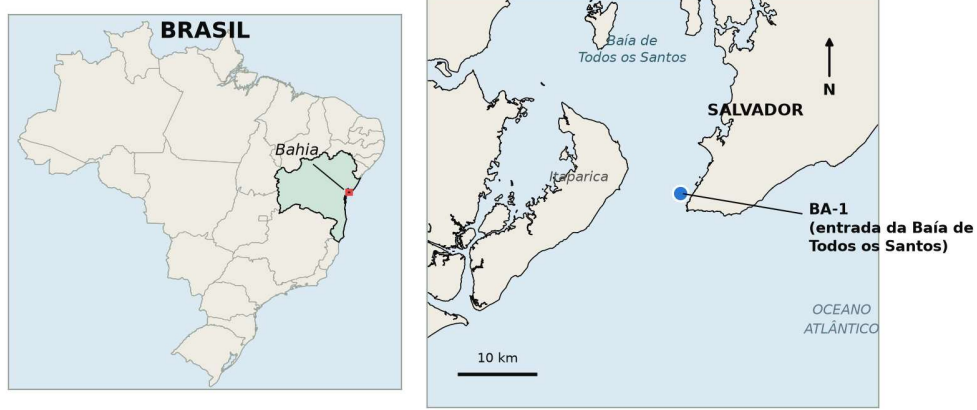


Fig. 2 Location of the BA-1 buoy at the entrance of Baía de Todos os Santos, Bahia, Brazil.

it with a one-sided temporal missing-data policy when a future endpoint is unavailable. Hourly coverage was 82.53% of 61,175 available hours for sea-state variables and 94.47 to 97.13% for water-column and optical variables (Table 3). Coverage of H_s fell to 65.45% in 2025 and to zero in 2026 while salinity, dissolved oxygen and temperature remained above 98%, indicating an ADCP-specific interruption rather than a platform failure; the last valid H_s observation is 2 September 2025, and all experiments terminate at that date (Fig. 3). Five further SimCosta buoys (ES-1, PR-1, RJ-1, RJ-2, RJ-4) were processed identically and pooled with BA-1 for E2 and E3. After preprocessing, five records contributed 8760 hourly observations, whereas ES-1 contributed 3015 because of its shorter valid record. Input variables were scaled by a min-max transform fitted on the training partition alone, ensuring that no statistic of the evaluation period informed model training.

Table 3 Descriptive statistics of the BA-1 variables over the complete raw series. H_s , the target variable, in bold.

Variable	Unit	Cov. (%)	Mean	SD	Min / Max
Significant wave height (H_s)	m	82.53	0.472	0.277	0.09 / 2.41
Peak wave period	s	82.53	8.080	2.988	1.70 / 28.60
Maximum wave height	m	82.53	0.784	0.454	0.14 / 4.71
Spectral wave height (HM0)	m	82.53	0.506	0.290	0.09 / 2.51
Salinity	psu	97.13	35.967	1.059	0.41 / 37.60
Dissolved oxygen	mL/L	96.91	4.139	0.373	0.04 / 5.63
Water temperature (CTD)	°C	96.95	27.181	1.138	23.31 / 31.49
Water temperature (ADCP)	°C	94.87	27.287	1.246	23.66 / 32.46
Turbidity	NTU	94.47	2.825	16.890	0.00 / 346.80
Chlorophyll fluorescence	µg/L	94.47	0.591	0.507	0.03 / 58.18
CDOM	ppb	94.47	0.940	1.872	-0.20 / 106.47

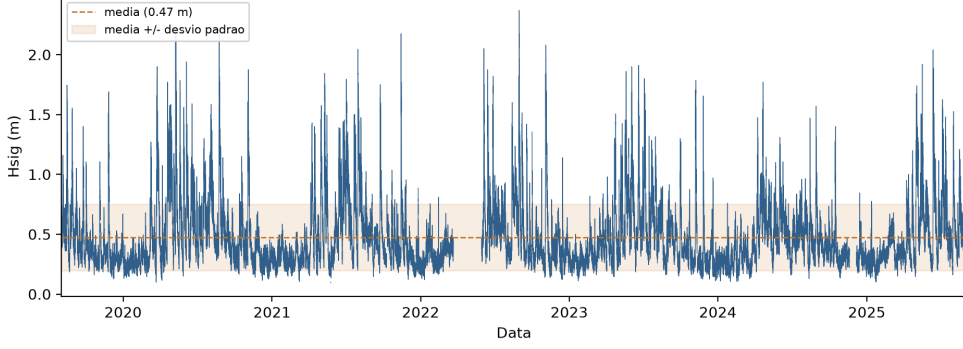


Fig. 3 Complete hourly H_s series, with reference mean and standard deviation.

2.3 Reference quality-control protocol

Three protocols were applied to the eleven variables over the complete real series, without synthetic corruption and using the same rule definitions as the operational archive. A continuous quality index (QI) combined seven tests, including classical spike, second difference, median/MAD, QARTOD, hourly rate of change, physical range and local dispersion. Five were normalised by their own thresholds, combined by their maximum and mapped to $QI = 100(1 - \text{score}/\text{score}_{\max})$, with $\text{score}_{\max}$ the 99.9th percentile per variable. Two outlier tests ran independently, consisting of the Dixon 4-sigma test on the global standardised score with cut points at 3 and 4, and the Q-Dixon test scoring gap/range in a centred 11-point window with cut points at 0.412 and 0.517.

Two properties of this construction bound the interpretation of its output. Four of the seven tests are deterministic functions of the spike statistic $|x_i - \frac{1}{2}(x_{i-1} + x_{i+1})|$, so the count of active tests is not a count of independent evidence and an isolated spike can activate three at once. The spike and second-difference thresholds are percentiles of the variable itself, which fixes the flag rate near 0.5% per test irrespective of the true data quality. The physical limits were fixed a priori from sensor specification and were not fitted; for most variables the observed range lies well inside them (Table 3), so those tests act as guards against gross failure rather than as discriminators here.

With miss_t and range_t binary indicators of missing value and range violation and n_t the number of active QI tests, the consolidated label followed a priority cascade (Fig. 4a) over

$$C_t^{\text{rem}} = \text{miss}_t \vee \text{range}_t \vee (QI_t < 50) \vee (n_t \geq 3) \vee (\text{Dixon}_t = \text{BAD}) \vee (\text{QDixon}_t = \text{BAD}), \quad (1)$$

$$C_t^{\text{rev}} = (n_t \geq 1) \vee (QI_t < 75) \vee (\text{Dixon}_t \neq \text{GOOD}) \vee (\text{QDixon}_t \neq \text{GOOD}). \quad (2)$$

The rule is deliberately asymmetric, since retaining a corrupted value costs more in an operational pipeline than inspecting a valid one. The spike statistic and the Q-Dixon window are centred, so this branch requires the subsequent observation and carries

a latency of one sample; it is a near-real-time product and not a strictly one-sided temporal one, a distinction that influences comparisons with predictive detectors.

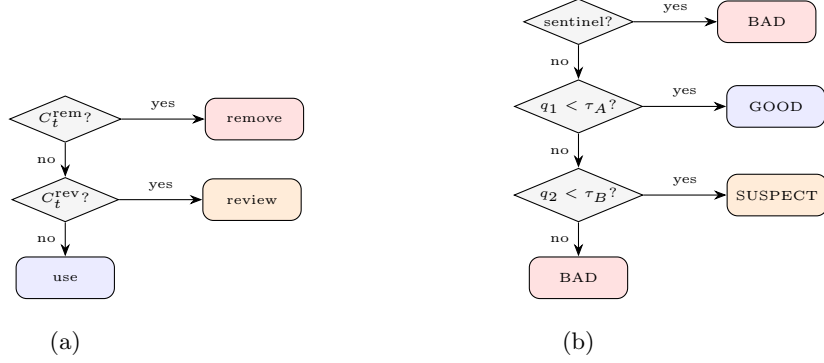


Fig. 4 Decision rules. (a) Reference label R_t , a priority cascade over Eqs. (1) and (2). (b) Proposed label Q_t from the ordinal probabilities q_1, q_2 and thresholds τ_A, τ_B , with missing and sentinel values resolved deterministically before the network.

2.4 Online temporal ordinal classifier

Predictive model. Let $\mathbf{X}_t = [x_{t-L}, \dots, x_{t-1}]$, which by construction excludes x_t and all later values. Three LSTM layers of decreasing width (128, 64, 32 units) followed by a linear layer gave

$$\mathbf{h}_t = \text{LSTM}_\theta(\mathbf{X}_t), \quad \hat{x}_t = g_\phi(\mathbf{h}_t), \quad (3)$$

trained with the Huber loss parameterized by threshold $\delta = 1.0$ in standardised space,

$$\mathcal{L}_{\text{Huber}}(x_t, \hat{x}_t) = \begin{cases} \frac{1}{2}(x_t - \hat{x}_t)^2, & \text{if } |x_t - \hat{x}_t| \leq \delta, \\ \delta|x_t - \hat{x}_t| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases} \quad (4)$$

We set $L = 168$ h, supplying a seven-day context that spans multiple tidal cycles and typical sea-state persistence before each prediction, and restricted the input to the H_s series itself. The predictor was trained with Adam for at most 25 epochs using batches of 64 observations and early stopping on validation loss with patience six. The plausible physical range was calibrated on the training partition as $[0, 3 \max(x^{\text{train}})]$ rather than fixed. The univariate input and the calibrated range are jointly what allow one trained model to run on any SiMCosta buoy reporting H_s , with no additional variable and no manual retuning.

Features. From the prediction residual $e_t = x_t - \hat{x}_t$ we computed six instantaneous indicators derived strictly from information available up to t , namely the one-sided second difference, deviation from the trailing median and MAD, rate of change, a binary out-of-range indicator, and robust and tail calibrations of the residual fitted

on training data alone. A second block accumulated slow deviation, to which the instantaneous residual is insensitive by construction, comprising three exponentially weighted moving averages of e_t ,

$$\text{EWMA}_t^{(\lambda)} = \lambda e_t + (1 - \lambda) \text{EWMA}_{t-1}^{(\lambda)}, \quad \lambda \in \{0.1, 0.3, 0.6\}, \quad (5)$$

a two-sided cumulative sum (CUSUM) with reference allowance $k = 0.5\sigma$,

$$S_t^+ = \max(0, S_{t-1}^+ + e_t - k\sigma), \quad S_t^- = \max(0, S_{t-1}^- - e_t - k\sigma), \quad (6)$$

with a reset cap at 20σ to bound growth under persistent bias, local linear slopes fitted over trailing windows of 6, 12, 24 and 48 h, and a sign-persistence counter. The feature vector $\mathbf{v}_t$ concatenated e_t , the six instantaneous indicators, the three EWMA values, the two CUSUM states, the four local slopes, the sign-persistence counter and the missing-value mask.

Ordinal head. The three classes are ordered, GOOD < SUSPECT < BAD, and we encoded that order in the output layer instead of discarding it with a softmax over unrelated categories. Two sigmoid units estimated

$$q_{1,t} = \Pr(Q_t \succeq \text{SUSPECT} \mid \mathbf{v}_t), \quad q_{2,t} = \Pr(Q_t = \text{BAD} \mid Q_t \succeq \text{SUSPECT}, \mathbf{v}_t), \quad (7)$$

from which the class probabilities follow. The head consisted of Dense(64, ReLU, dropout 0.2), Dense(32, ReLU, dropout 0.1) and two sigmoid outputs, trained by minimising the weighted ordinal loss function

$$\mathcal{L}_{\text{ordinal}} = -\frac{1}{N} \sum_{t=1}^N w_t [c_1 \ell_{\text{BCE}}(y_{t,1}, q_{1,t}) + \mathbb{I}(y_t \succeq \text{SUSPECT}) c_2 \ell_{\text{BCE}}(y_{t,2}, q_{2,t})], \quad (8)$$

where $y_{t,1} = \mathbb{I}(y_t \succeq \text{SUSPECT})$, $y_{t,2} = \mathbb{I}(y_t = \text{BAD})$, $\ell_{\text{BCE}}(y, q) = y \ln q + (1-y) \ln(1-q)$, w_t is the label confidence weight (1.0 for strong labels and 0.5 for weak labels), and c_1, c_2 are inverse class frequency balancing factors estimated from the training labels. The classifier was trained with Adam for at most 80 epochs using batches of 64 observations and early stopping on validation loss with patience ten. Labels followed the cascade of Fig. 4b with $\tau_A = 0.5$ and τ_B selected on validation from the grid 0.05, 0.10, ..., 0.95, 0.96, ..., 0.98 by maximising BAD-class F1; missing values and sentinel codes bypassed the network and were assigned BAD deterministically. The online inference path requires a fixed 168-h input context, 21 classifier features and constant-size feature updates at each hourly step. Repository-level behavioural verification reported a zero violation rate for the two one-sided temporal configurations. No hardware-level latency benchmark is included in the present study; the evidence therefore supports algorithmic online operation rather than a device-independent millisecond guarantee.

Training labels. Training targets combined two sources, both computed outside $\mathbf{v}_t$ to avoid circularity. Corrupted points received a strong label with confidence weight 1.0 determined by family and magnitude (Table 4), whereas uncorrupted points received a

weak label with confidence weight 0.5 assigned as GOOD or SUSPECT from a z-score against the training reference mean and standard deviation. Weak labels serve as a training device rather than ground truth, because they encode a threshold applied to data that is uncorrupted by construction, and a model reproducing them has reproduced a rule rather than detected a physical failure. All primary detection metrics were therefore computed on strongly labelled points against uncorrupted points, with the weakly labelled fraction of each partition reported separately as sensitivity analysis.

Table 4 Corruption model. t_0 is episode onset, σ the training standard deviation of H_s and u the stratified magnitude draw.

Family	Corruption at $t \geq t_0$	Magnitude	Duration
A, additive spike	$x_t \pm u\sigma$	$u \in \{0.5, 1, 2, 3, 4, 6\}$; SUSPECT if $u < 3$	1 h
B, multiplicative	$x_t f$	$f \in \{5, 10, 0.2, 0.1\}$	1 h
C, level shift	$x_t \pm u\sigma$	$u \in [1, 2.5] \cup [3, 4]$	$d \in \{12, 24, 48, 72\}$ h
D, drift	$x_t + \beta(t - t_0)$	cumulative 1–2.5 or 3–4 σ	$d \in \{12, 24, 48, 72\}$ h
E, stuck sensor	x_{t_0-1}	n/a	$d \in \{3, 6, 12, 24\}$ h
F, burst noise	$x_t + \varepsilon_t, \varepsilon_t \sim \mathcal{N}(0, u^2\sigma^2)$	$u \in [0.5, 2.5]$	$d \in \{3, 6, 12, 24\}$ h
G, missing/sentinel	NaN, sentinel or negative value	n/a	1 h

Pooled training and transfer validation. Two evaluation regimes were executed. The pooled regime (E2) split each buoy chronologically 60/20/20 within its own series, with the first segment used for training, the middle segment for validation and the final segment for testing. No timestamp was shared across these partitions, and the corresponding segments were pooled across all six buoys into unified training, validation and test sets. The predictor used a maximum of 25 epochs and patience six, while the classifier used a maximum of 80 epochs and patience ten. To test generalisation to unrepresented stations (RQ_2), the leave-one-buoy-out protocol (E3) iteratively trained the model on five pooled buoys and evaluated detection on the sixth held-out station across six independent retraining folds, ensuring that no observation from the target buoy informed parameter estimation. In E3, the predictor was trained with a maximum of five epochs and patience two because each fold used a reduced retraining budget; the ordinal classifier retained the training procedure described above.

2.5 Synthetic corruption model

To obtain reproducible controlled test cases, quantitative evaluation used a corruption model, and the conclusions of any such study are bounded by how faithfully that model represents instrument failure. We specified seven families spanning the failure modes documented for moored sensors (Table 4), each with its own budget of positions so that long-duration families did not exhaust the available space before short ones; level-shift and drift families drew stratified magnitudes so that BAD and SUSPECT examples arose in comparable proportion.

Two offline regimes were evaluated. The first corrupted about 1 in 130 test points with an additive perturbation of random sign and magnitude between 3 and 6 standard deviations, clipped at zero, matching the rate reported for real oceanographic records (J. Xie et al., 2023). Perturbations of that amplitude on an hourly series are detectable by essentially any local dispersion statistic, so a benchmark restricted to them measures false-alarm behaviour rather than detection capability; we therefore added a second regime at 1 to 2 standard deviations together with the drift and stuck-sensor families, and report both.

The streaming regime (E4) differs in kind. Corruption was generated as a sequential state process over one pass of the real hourly BA-1 series, treated as correct by definition. At each hour an episode began with probability 0.008, drawing family, severity and duration; while active, the real value was replaced by the family effect, and the corrupted value fed back into subsequent input windows exactly as a faulty sensor would, so the predictor progressively adapted to the corrupted level over long episodes. Because the real series is assumed correct, any genuine anomaly within it is scored as a false positive and precision on this experiment is biased downward accordingly.

2.6 Comparator methods

The first benchmark followed the configuration summarised in Table 1 and implemented the published multivariate wave-quality-control setup for comparison. Its detector uses constants calibrated for four Chinese stations, making it a relevant test of transferability to the present coastal network. A multivariate three-step LSTM (Fig. 5) took H_s and five auxiliary variables selected by the largest absolute Pearson correlation with H_s over the training period; HM0, $H_{\max}$, H_{avg} and H_{10} were excluded as order statistics of the same wave record ($r > 0.98$), which would render the prediction nearly tautological. The window ended at $t - 2$ so that no part of the target was visible, and the three predictions were averaged. Training used Adam and the mean squared error, up to 12 epochs with early stopping (patience 5), on the most recent 8760 valid hourly records (10 August 2024 to 2 September 2025), split 70/30.

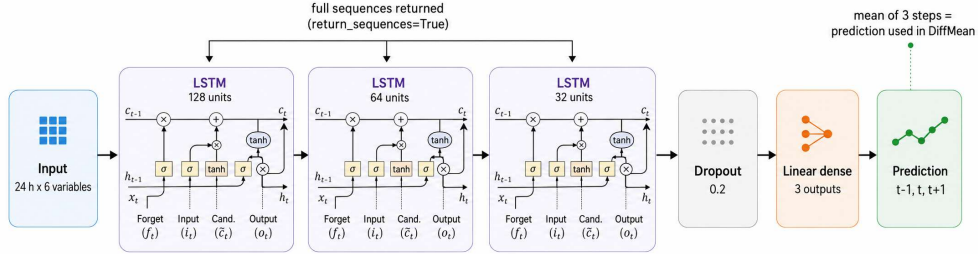


Fig. 5 Architecture of the multivariate three-step LSTM used in the benchmark implementation.

The detection score was the local mean-difference ratio

$$D_t = \frac{|x_t - \hat{x}_t|}{\frac{1}{3}(x_{t-1} + x_t + x_{t+1})}. \quad (9)$$

As defined by its authors, D_t uses the subsequent observation. The comparators below, like the reference branch, therefore operate with a one-step lookahead that the proposed classifier does not have, and every comparison here is conservative with respect to the proposed method.

Three comparators acted on D_t , consisting of the LSTM-Peak detector flagging local peaks that meet simultaneous minimum height, prominence and inter-peak distance with constants fixed by the original study ($H=1$, $T=1$, $d=50$), a robust median/MAD criterion converting D_t to a tail probability through a Student- t calibration with 4 degrees of freedom fitted on the clean training period, and a generalised Pareto criterion modelling the upper tail of the clean D_t above its 90th percentile. Three further comparators acted directly on the observed series or the variable vector, namely a robust spike test at six robust deviations, an Isolation Forest (200 trees, contamination 1/130) fitted on the clean preprocessed vector (Liu, Ting, & Zhou, 2008), and a VAE-LSTM reconstructing 24 h windows with an 8-dimensional latent space and $\beta = 10^{-3}$ (Lin et al., 2020). Since D_t is non-negative by construction, the calibrated criteria used the one-sided tail probability $p_t = 1 - F(z_t)$ and flagged $p_t < \alpha$, with $\alpha = 0.01$ as the operating point and a sensitivity sweep from 0.001 to 0.1; the Isolation Forest has no significance level and flagged the fraction set by its contamination parameter.

The Isolation Forest and the VAE-LSTM ingest the multivariate vector whereas the perturbation affects the target alone, a configuration that dilutes the anomalous signal across covariates of unequal relevance. We therefore also evaluated univariate variants of both, so that unfavourable input configuration can be distinguished from genuine failure to transfer.

2.7 Predictor ablation

A predictor-configuration ablation was performed on the BA-1 benchmark to quantify the effect of auxiliary variables, prediction horizon and lookback length. The baseline was the multivariate three-step predictor used in the comparator protocol. Each alternative changed one factor at a time: a univariate predictor without auxiliary variables, a single-step predictor, and multivariate predictors with 12-h or 30-h lookback windows. Each configuration was trained with three model seeds and evaluated over five anomaly-injection seeds. The ablation was used to inform model design; it was not used to replace the pooled six-buoy evaluation of the final temporal ordinal classifier.

2.8 Operational decisions and label reconciliation

The two persisted labels use an increasing operational severity scale. The reference protocol returns $R_t \in \{\text{use, review, remove}\}$, while the classifier returns $Q_t \in \{\text{GOOD, SUSPECT, BAD}\}$. These states are interpreted respectively as acceptable for routine use, requiring inspection, and requiring removal from downstream use.

For the H_s archive, the correspondence is use $\leftrightarrow$ GOOD, review $\leftrightarrow$ SUSPECT and remove $\leftrightarrow$ BAD. A third reconciled field was added without replacing either original, preserving the provenance of both decisions. When the branches disagreed, the more severe state prevailed. We chose the maximum over an average or a two-way majority because the branches use different evidence, one comparing an observation with its neighbourhood and historical distribution and the other with a one-sided temporal prediction conditioned on recent context. The rule was applied to the complete real BA-1 series without synthetic corruption.

2.9 Outcomes and statistical analysis

The primary outcome was binary-F1 on the pooled six-buoy test partition, restricted to strongly labelled points. Binary-F1 is the F1 obtained after collapsing SUSPECT and BAD into one positive class with GOOD as negative; it is the only quantity on which the three-class classifier and the binary comparators are directly comparable, and it was computed for both on the same partition, ground truth and seeds. Secondary outcomes were per-class F1, macro-F1, balanced accuracy, Matthews correlation coefficient, one-versus-rest AUROC and AUPRC, the per-buoy decomposition, leave-one-buoy-out transfer metrics, onset detection at $t+1$ by family, episode coverage and the false-alarm rate outside injected episodes. Predictive error was reported as MAE, RMSE and MAPE, each with its test window and prediction horizon, since the benchmark and the proposed model use different windows and horizons and their errors are not interchangeable.

Uncertainty was estimated from the replication scheme rather than from the pooled runs. The 15 runs of E1 and E2 form a crossed design of 3 model seeds by 5 injection seeds, not 15 independent replicates, and resampling them as independent understates uncertainty and conflates two variance sources. We apportioned variance descriptively into model-seed, injection-seed and a combined residual/interaction component by a two-way sum-of-squares allocation, not a fitted random-effects model, since the crossed 3×5 design has one observation per cell and cannot separate interaction from residual error; we report the percentage of variance attributed to each component. The reported 95% confidence intervals for aggregate stability metrics use 5,000 simple bootstrap resamples of the 15 crossed run-level values, while intervals for streaming onset rates use 5,000 bootstrap resamples of the ten per-seed rates. Differences smaller than the observed run-to-run variation were not interpreted as practically meaningful. No pairwise inferential test was used because the available replications are crossed seed combinations rather than independent observational units. The threshold τ_B was selected on validation and never on test data; α was fixed at 0.01, with other levels reported as sensitivity analysis and not as alternative primary results.

The complete training and inference pipeline, together with the streaming simulator and reproducible scripts, are made available to support independent verification.

Taken together, the design separates model development, benchmark comparison, cross-station transfer and operational streaming assessment, while preserving the provenance of each quality-control decision. This structure enables the results to be interpreted both as an artificial-intelligence evaluation and as an engineering test of a deployable monitoring workflow.

Generative artificial-intelligence tools were used during preparation of the study for language editing, code inspection and organisation of the computational workflow; they were not used to generate scientific content, analyse data or draw conclusions, and all methodological decisions, analyses and interpretations were independently reviewed and validated by the authors.

3 Results and Discussion

This section presents the empirical validation across the six experimental configurations, addressing predictive residual accuracy, local baseline calibration (RQ_1), generalisation and cross-station transferability (RQ_2), and streaming detection under continuous error feedback (RQ_3), followed by a discussion of these findings.

3.1 Predictive error

The comparator network evaluated on BA-1 reached low dispersion across retraining seeds relative to the magnitude of the error (Table 5), indicating stable predictive residuals across the evaluated initialisations. The univariate single-step architecture trained on pooled multi-buoy records produced lower aggregate prediction error on its own test window, although the two rows use different stations, horizons and test periods and are not directly comparable.

Table 5 Prediction error of the two networks. Test window and horizon differ between rows and the values are not interchangeable.

Model	Test window	MAE (m)	RMSE (m)	MAPE (%)
E1, multivariate, 3 steps	BA-1, 30%	0.075 ± 0.002	0.112 ± 0.004	11.40 ± 0.21
E2, univariate, 1 step	Six buoys, 20%	0.059	0.081	7.03

3.2 Reference protocol on the real series

The three reference protocols applied to the eleven BA-1 variables over the complete uncorrupted series yielded lower median quality indices for the three sea-state variables than for salinity and water temperature (Table 6). The fraction classified BAD remained below one per cent for H_s . Flags raised by the median/MAD test for H_s were distributed across the full observed range rather than concentrating at the upper tail, indicating responsiveness to local step discontinuities relative to the temporal neighbourhood.

Consolidated over the 8520 h where both branches are defined, the reference label assigned 72.3% of hours to use, 22.4% to review and 5.3% to remove.

Table 6 Reference quality-control indicators per variable on BA-1, over the complete real series.

Variable	Median QI	Dixon BAD (%)	Q-Dixon BAD (%)
Significant wave height (H_s)	84.6	0.62	0.07
Spectral wave height (HM0)	84.8	0.58	0.06
Maximum wave height	85.1	0.65	0.34
Peak wave period	92.4	0.01	0.91
Salinity	100.0	0.06	0.13
Dissolved oxygen	94.6	1.06	0.14
Water temperature (CTD)	90.9	0.00	0.06
Water temperature (ADCP)	90.6	0.01	0.03
Turbidity	100.0	0.89	2.19
Chlorophyll fluorescence	98.0	0.21	0.42
CDOM	100.0	0.38	0.66

3.3 Comparator criteria and threshold calibration

Comparing the six detection criteria on BA-1 under offline injection across 15 runs revealed substantial performance divergence (Fig. 6). Three criteria decide on a threshold that does not depend on α , LSTM-Peak with the published fixed constants, the traditional robust spike test, and Isolation Forest, and are reported in Table 7; their F1 scores are also drawn as constant reference lines in Fig. 6 for direct visual comparison against the other three. The remaining three criteria, the robust median/MAD test with local Student- t calibration, the GPD-POT tail test, and VAE-LSTM, have their decision threshold set by α and are therefore plotted across the full range $\alpha \in [0.001, 0.1]$ instead of tabulated at a single point. At the calibrated operating point ($\alpha = 0.01$, dashed vertical line), the robust median/MAD criterion reached the highest F1 of all six criteria, by a wide margin over the LSTM-Peak detector. This ranking remained unchanged across the evaluated model seeds and across the displayed α sweep, providing the comparative result for RQ_1 .

Table 7 Comparator criteria whose decision threshold does not depend on the significance level α , on BA-1 under the offline injection regime (mean $\pm$ standard deviation over 3 models $\times$ 5 injection replicates). The other three criteria are calibrated by α and reported across its full range in Fig. 6.

Criterion	Precision	Recall	F1
LSTM-Peak	1.000 ± 0.000	0.603 ± 0.069	0.750 ± 0.054
Traditional robust spike	0.286 ± 0.008	1.000 ± 0.000	0.445 ± 0.010
Isolation Forest	0.016 ± 0.006	0.067 ± 0.024	0.026 ± 0.009

Three of the six criteria achieved perfect recall across all runs for the evaluated large perturbations (3 to 6 standard deviations). The performance separation shown in Fig. 6 and Table 7 therefore arose mainly through differences in false-positive control rather than in sensitivity to those severe anomalies. The calibrated criteria produced different false-alarm and detection trade-offs as α varied.

Two comparator methods exhibited poor performance under the evaluated configuration. The VAE-LSTM generated over 600 false flags per run because a single point

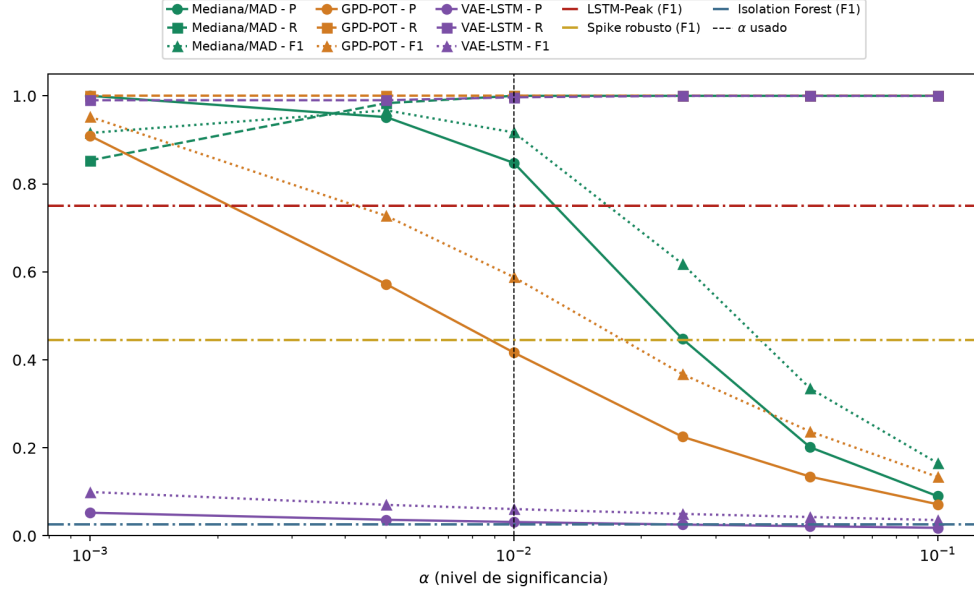


Fig. 6 Precision (P, solid), recall (R, dashed) and F1 (dotted) of the three α -calibrated comparator criteria (median/MAD, GPD-POT, VAE-LSTM) as a function of the significance level α . The three dash-dot horizontal lines give the F1 of the criteria whose threshold does not depend on α (LSTM-Peak, traditional robust spike, Isolation Forest); exact precision and recall for these three are given in the following table. The vertical dashed line marks the $\alpha = 0.01$ operating point used throughout.

anomaly propagated across up to 24 consecutive reconstruction windows. The Isolation Forest produced very low F1 when the target signal was diluted across unperturbed covariates. These behaviours were also visible in the representative observed/predicted trajectories and comparative flag series.

An independent physical event occurred on 12 June 2025 within the test period, when an abrupt storm elevated wave heights to 2.15 m. The calibrated median/MAD criterion flagged the onset hours, while the Dixon 4-sigma protocol flagged the subsequent elevated phase as a plausible physical extreme rather than sensor failure. The two outputs therefore provided different assessments of the same energetic sea-state episode.

3.4 Predictor ablation

The predictor ablation showed that the univariate 24-h configuration achieved the highest binary detection F1 among the tested BA-1 variants, while the single-step configuration reduced prediction error (Table 8). The multivariate three-step baseline was competitive but did not improve quality-control F1 over the univariate alternative. The 12-h window retained high F1 with a small recall reduction, whereas the 30-h window had the largest variability and the lowest mean F1. These results support the final multi-buoy design choice of using only H_s : auxiliary variables were not required for the benchmarked quality-control task and would reduce the set of stations to which a common model could be transferred.

Table 8 Predictor-configuration ablation on the BA-1 controlled point-event benchmark. Values are mean $\pm$ standard deviation over three model seeds and five anomaly-injection seeds.

Configuration	Precision	Recall	Binary-F1	RMSE (m)
Multivariate, three-step, 24 h (baseline)	0.847 ± 0.047	1.000 ± 0.000	0.917 ± 0.027	0.112 ± 0.004
Univariate, three-step, 24 h	0.924 ± 0.037	1.000 ± 0.000	0.960 ± 0.020	0.103 ± 0.001
Single-step, multivariate, 24 h	0.857 ± 0.061	1.000 ± 0.000	0.922 ± 0.036	0.080 ± 0.006
Multivariate, three-step, 12 h	0.921 ± 0.046	0.980 ± 0.025	0.949 ± 0.029	0.106 ± 0.007
Multivariate, three-step, 30 h	0.838 ± 0.084	0.983 ± 0.024	0.903 ± 0.052	0.117 ± 0.015

3.5 Multi-buoy generalisation and leave-one-buoy-out transfer

The general ordinal model trained on the pooled partitions of the six buoys was evaluated both on the pooled multi-buoy test set (E2) and under a rigorous leave-one-buoy-out transfer protocol (E3), testing cross-station transferability without local recalibration (RQ_2).

Table 9 General multi-buoy model performance under pooled evaluation (E2) and leave-one-buoy-out transfer validation across unseen stations (E3). Precision and recall refer to the binary GOOD versus non-GOOD decision.

Evaluation	Station	n	M-F1	F1 BAD	B-F1	Prec.	Recall	AUPRC BAD
Pooled (E2)	All (strong labels)	6456	0.742	0.614	0.853	0.890	0.819	0.648
	BA-1	1555	0.697	0.557	0.689	0.763	0.629	0.595
	ES-1	251	0.751	0.429	0.920	0.912	0.929	0.560
	PR-1	584	0.540	0.315	0.755	0.992	0.610	0.443
	RJ-1	1293	0.880	0.800	0.913	0.926	0.901	0.841
	RJ-2	1537	0.703	0.738	0.945	0.955	0.934	0.939
	RJ-4	1236	0.711	0.438	0.836	0.779	0.902	0.544
LOBO transfer (E3)	Mean unseen	6409	0.784	0.735	0.856	0.925	0.810	0.801
	BA-1 (unseen)	1542	0.898	0.928	0.940	0.931	0.949	0.935
	ES-1 (unseen)	256	0.681	0.560	0.779	0.984	0.645	0.625
	PR-1 (unseen)	635	0.754	0.667	0.794	0.994	0.661	0.658
	RJ-1 (unseen)	1254	0.803	0.652	0.904	0.909	0.899	0.751
	RJ-2 (unseen)	1540	0.763	0.829	0.830	0.880	0.786	0.968
	RJ-4 (unseen)	1182	0.805	0.775	0.887	0.855	0.922	0.867

Table 10 Confusion matrix for the pooled controlled-failure test set (E2), with rows representing the event class and columns the predicted class.

Actual class	GOOD	SUSPECT	BAD	Total
GOOD	5446	89	4	5539
SUSPECT	113	348	69	530
BAD	53	130	204	387
Total	5612	567	277	6456

On the pooled controlled-failure test partition, the model discriminated the three ordered states with F1 scores of 0.977 for GOOD, 0.634 for SUSPECT and 0.614 for BAD (Table 9). The evaluation contained 6456 controlled-event points and 1899 weak operational points; the latter were retained as a sensitivity subset and were not used to define the primary estimate. The confusion matrix shows that most errors were between adjacent severity levels. Only four GOOD observations were assigned BAD, whereas 53 BAD observations were assigned GOOD, which is consistent with the lower BAD-class F1 of 0.614 and identifies the main safety-relevant error mode.

The leave-one-buoy-out transfer evaluation (E3) directly answered RQ_2 . With one station held out at a time, binary-F1 ranged from 0.779 on ES-1 to 0.940 on BA-1, and the unweighted mean across the six unseen stations was 0.856. Three stations transferred as well as or better than when pooled, whereas ES-1 and RJ-2 declined by more than ten points relative to their pooled estimates. Transfer quality therefore depends on the station being withheld rather than changing uniformly across the network.

One contrast is worth stating before interpreting either evaluation. On the pooled test partition the model assigned 77.4% of points to GOOD, whereas on the real BA-1 series without injected anomalies it assigned 97.9%. The two figures describe different data, but the gap is large enough that the aggregate numbers above should be read as primary controlled-event detection performance, not as unconditional performance on naturally occurring faults.

3.6 Stability across retraining

Fifteen complete runs combining three model seeds with five injection seeds described the stability of the architecture (Table 11). Precision and recall remained balanced, with standard deviations of 0.038 and 0.026, respectively. For binary-F1, the descriptive allocation attributed 27.3% of the observed variation to model seed, 16.0% to injection seed and 56.7% to the inseparable interaction/error remainder. For F1 BAD, 93.2% was associated with injection seed. These percentages are descriptive rather than inferential, because the crossed 3×5 design has a single observation per cell and cannot separate an interaction term from residual error.

Table 11 Sensitivity of the general model over 15 crossed runs (3 model seeds $\times$ 5 injection seeds), computed on the complete synthetic test labels. The variance allocation is descriptive.

Metric	Mean $\pm$ SD
Macro-F1	0.760 ± 0.043
F1 GOOD	0.980 ± 0.004
F1 SUSPECT	0.665 ± 0.092
F1 BAD	0.634 ± 0.079
Binary-F1	0.884 ± 0.030
Binary precision	0.915 ± 0.038
Binary recall	0.856 ± 0.026
AUPRC BAD	0.651 ± 0.148
AUROC BAD	0.928 ± 0.036

The BAD class exhibited the highest relative variance due to its low baseline prevalence, where individual event realisations produce measurable changes in sample statistics. The binary-F1 of 0.853 and BAD-class F1 of 0.614 reported for the pooled E2 result refer to the fixed final model evaluated on the primary controlled-event test partition. The binary-F1 of 0.884 ± 0.030 and BAD-class F1 of 0.634 ± 0.079 in Table 11 are means across the 15 crossed retraining and injection-seed runs; they describe run-to-run stability and are not alternative estimates of the same single-model result.

3.7 Streaming evaluation and onset latency

The streaming simulator processed 8592 h of the real BA-1 series under continuous error propagation. Aggregating ten independent simulation seeds and 596 distinct failure episodes, the overall onset detection rate at $t+1$ sat well below the near-perfect rates reached by the most abrupt families, reflecting the drag exerted by drift and stuck-sensor episodes on the pooled figure (Table 12, Fig. 7).

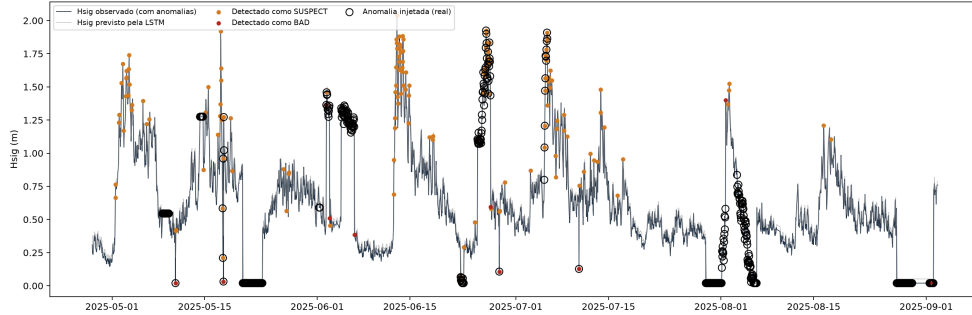


Fig. 7 Excerpt of the streaming evaluation on BA-1 showing observed and predicted series, points flagged SUSPECT or BAD, and injected anomalies.

Onset detection separated sharply by failure mechanism (RQ_3). Abrupt step changes, sentinel codes, multiplicative scaling and large level shifts were detected promptly at the initial hour of occurrence. Conversely, drift and stuck-sensor modes exhibited near-zero onset detection, because the initial perturbation lies within the

Table 12 Onset detection at $t+1$ by failure family, aggregated over 10 simulation seeds (596 distinct episodes).

Family	Episodes	Rate at $t+1$	95% CI across seeds
A, additive spike	76	0.579	[0.447, 0.708]
B, multiplicative	99	0.747	[0.678, 0.831]
C, level shift	80	0.838	[0.704, 0.941]
D, drift	88	0.011	[0.000, 0.027]
E, stuck sensor	84	0.048	[0.000, 0.095]
F, burst noise	98	0.510	[0.348, 0.609]
G, missing/sentinel	71	1.000	[1.000, 1.000]
Overall	596	0.529 ± 0.082	mean over 10 seeds

natural ambient variance of the preceding sea state, rendering the fault statistically indistinguishable from valid observations at t_0 .

In the reference streaming run, 159 of 7841 clean hours were flagged outside injected episodes, corresponding to a false-alarm rate of 2.03%, while the mean temporal coverage across the 45 complete injected episodes reached 47.85%. The 8592-h and 596-episode totals in Table 12 aggregate the ten simulation seeds, whereas the 7841 clean hours and 45 episodes belong to the reference run used for the false-alarm and coverage summaries. These denominators therefore describe different aggregations of the streaming experiment. The temporal accumulators enabled gradual detection of persistent drift and stuck states as evidence accrued over subsequent observation steps.

3.8 Reconciliation on the real series

Applying the reconciliation cascade to 8520 h of uncorrupted BA-1 observations demonstrated 72.7% exact agreement between the reference and predictive branches (Table 13). This is a separate analytical window from the 8592-h streaming simulation and was used to compare the two persisted label branches on the real, uncorrupted series. The reference branch escalated severity in 26.9% of hours, whereas the one-sided temporal classifier escalated the reference assessment in 0.4% of hours. The reconciled archive yielded 71.89% GOOD, 22.81% SUSPECT and 5.31% BAD observations, aligning closely with the reference statistical distribution while preserving auditable provenance.

Table 13 Agreement between the reference label R_t and the classifier label Q_t over 8520 h of the real BA-1 series, with no synthetic corruption.

$R_t \setminus Q_t$	GOOD	SUSPECT	BAD
use	6125	33	1
review	1838	72	0
remove	374	77	0

The reference branch establishes the auditable archive classification, while the strictly one-sided temporal classifier provides instantaneous triage at the leading edge of the real-time stream. The Dixon 4-sigma comparison in the storm example is an auxiliary physical-extreme illustration and is not one of the six formal controlled-failure criteria in the primary benchmark.

Across the experiments, the proposed system maintained its main engineering properties under pooled testing, station transfer and sequential error feedback. The results also identify the operational boundary of the method: abrupt and sentinel failures can be assessed at onset, whereas slowly evolving faults require accumulated evidence and should be evaluated through detection latency and episode coverage.

3.9 Discussion

The findings of this study provide evidence addressing the three core research questions and the positioning established in Table 1. The results jointly substantiate four engineering properties. First, the classifier makes a one-sided temporal decision using the current observation and preceding context, and the repository-level behavioural verification reported zero violations for the tested one-sided configurations. Second, its ordinal output preserves three operational severity states, with most classification errors occurring between adjacent classes rather than between GOOD and BAD. Third, the univariate predictor configuration performed at least as well as the multivariate alternatives in the dedicated BA-1 ablation, supporting a common input design across stations. Fourth, the same model transfers across unseen buoys and remains usable under sequential error feedback. These properties are supported by the temporal construction, the confusion matrix in Table 10, the ablation in Table 8, and the transfer and streaming results reported above.

Regarding local calibration (RQ_1), the benchmark comparison shows that fixed detection constants inherited across distinct coastal regions can produce an unfavourable precision–recall trade-off (Table 7). The LSTM-Peak detector maintained high precision but sacrificed recall, whereas calibrating detection thresholds directly on local baseline data using robust Student- t dispersion produced the strongest F1 among the evaluated criteria at the selected operating point.

Regarding cross-station transferability (RQ_2), the multi-buoy leave-one-buoy-out evaluation (E3) indicates that a single temporal classifier can generalise to unseen stations within the Brazilian coastal network without local retuning, on average. Learned representations of sea-state persistence, one-sided second differences and temporal accumulators transferred with a mean binary-F1 of 0.856, but station-level variation remained substantial, from 0.779 on ES-1 to 0.940 on BA-1. This range should be read as network-level transferability rather than evidence of universal geographical generalisation; the study does not test whether bathymetry, energy spectra, sensor configuration or regional climate explains the variation.

Regarding online streaming dynamics (RQ_3), the streaming evaluation highlights fundamental operational constraints imposed by failure phenomenology. Abrupt step discontinuities, multiplicative anomalies and sentinel codes are detected at the initial onset sample ($t+1$). Conversely, slow sensor drift and stuck-sensor faults exhibit near-zero immediate onset detection, because the initial erroneous increment lies within the ambient variance of the preceding sea state. Under real-time error feedback, these slow-evolving modes require the accumulation of temporal evidence via EWMA and CUSUM statistics. Consequently, detection latency under bounded false-alarm rates represents the correct operational benchmark for persistent sensor degradation rather than instantaneous onset rate.

Reconciling the statistical reference protocol with the temporal classifier on uncorrupted observational records shows the two branches agreeing on most hours (Table 13), with the reference branch escalating severity far more often than the classifier does. The two branches serve complementary operational roles: the centred reference protocol supports archival quality control, whereas the one-sided temporal classifier enables online triage at the leading edge of an observational data stream.

Several methodological constraints qualify these conclusions. Quantitative evaluation relies on controlled corruption models spanning seven documented marine failure families, so the reported scores are conditional on how faithfully those scenarios represent operational sensor behaviour. Controlled event labels provide a reproducible benchmark, while weak labels capture conservative statistical boundaries. The storm example is a qualitative demonstration that a physical extreme and a sensor fault can lead to different quality assessments; it is not independent validation of a naturally occurring sensor failure. The historical preprocessing also includes limited short-gap interpolation for the offline archive; a live implementation should replace this step with a one-sided missing-data policy. Finally, evaluating streaming detection over historical series assumes the baseline record is uncorrupted, so any genuine anomaly within it is scored as a false alarm, biasing the reported false-alarm rate upward rather than understating it. The absence of a hardware timing study further limits claims about deployment speed.

In operational deployment, the ordinal three-class formulation transforms rigid binary rejection into a tiered triage queue: GOOD observations can proceed to routine processing, SUSPECT observations can be prioritised for review, and BAD observations can be withheld from downstream wave statistics until resolved. This layer can complement deterministic or QARTOD-style checks while preserving the raw observation, the quality flag and the decision provenance. The temporal classifier therefore provides an auditable, transferable tool for coastal monitoring infrastructure. The present evidence supports hourly online decision logic and sequential error-feedback evaluation; it does not establish hardware-independent latency, quantify downstream engineering benefit, or replace operational acceptance procedures.

4 Conclusions

The objective of this study was to develop and evaluate a transferable one-sided temporal ordinal learning system for real-time fault detection in significant wave height observations from heterogeneous coastal buoy networks. The system was designed to combine a univariate temporal predictor, residual and consistency indicators, accumulated evidence and explicit severity decisions using only information available at the decision time.

This objective was achieved at the experimental validation level. On the pooled controlled-event test partition, the classifier obtained a binary-F1 of 0.853, with F1 values of 0.977 for GOOD, 0.634 for SUSPECT and 0.614 for BAD. These results show that the system can preserve graded quality-control decisions instead of reducing every nonconforming observation to a single binary rejection. The comparison of six detection criteria also showed that local threshold calibration produced the strongest

F1 trade-off at the selected operating point, although the available replications support a comparative result rather than a claim of statistical superiority.

Cross-station transferability was supported within the tested six-station Brazilian network under leave-one-buoy-out validation, where the same one-sided temporal model achieved a mean binary-F1 of 0.856 without station-specific retraining. Performance ranged from 0.779 to 0.940 across held-out buoys, supporting transferability while also showing that station-dependent variation remains relevant. In the hourly streaming experiment, onset detection varied by failure mechanism, from 0.011 for drift and 0.048 for stuck-sensor episodes to 1.000 for missing or sentinel values. The false-alarm rate outside injected episodes was 2.03%, and mean episode coverage was 47.85%, indicating that accumulated temporal evidence is necessary for slowly evolving faults.

Reconciling the one-sided temporal ordinal classifier with the statistical reference protocol yielded 72.7% exact agreement over 8,520 uncorrupted BA-1 observations while preserving both original labels and their provenance. The results therefore support the proposed system as an auditable triage layer for real-time engineering monitoring and full-scale ocean-observation workflows, not as a replacement for operational acceptance procedures or sensor-maintenance diagnostics. The principal limitation is reliance on controlled corruption models and the assumption that the historical baseline is largely correct; future work should evaluate the classifier with independently adjudicated in situ events, quantify its effect on downstream wave statistics and test a one-sided missing-data policy. The complete codebase, trained models and streaming simulator are available for independent verification.

Declarations. Ethics approval and consent to participate Not applicable. This study used publicly available oceanographic sensor records and did not involve human participants, human data, human tissues, or animals.

Consent for publication Not applicable. This manuscript contains no identifiable data, images, or videos of individual persons.

Availability of data and materials The datasets analysed during the current study are publicly available from SiMCosta (Sistema de Monitoramento da Costa Brasileira) at <https://www.simcosta.furg.br> and from the ERA5 reanalysis through the Open-Meteo historical API. The six buoy records analysed in this study are identified by their original filenames in the accompanying computational repository. The retrieval and preprocessing scripts, trained model, complete pipeline, and streaming simulator are available at <https://github.com/marcosFigueredo/simcosta-wave-qc> and archived at <https://doi.org/10.5281/zenodo.22135590>.

Competing interests The authors declare that they have no competing interests.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. V.F. was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, grant numbers 444903/2024-0 and 314941/2025-8.

Authors' contributions M.B.F.: conceptualization, methodology, software, formal analysis, writing – original draft. J.R.S.: conceptualization, methodology, software, formal analysis, writing – original draft. M.D.R.O.: conceptualization, methodology,

software, formal analysis, writing – review and editing. V.S.F.: investigation, conceptualization, methodology, software, formal analysis, writing – review and editing. L.M.M.S.: investigation, data curation, writing – review and editing. L.S.M.: investigation, data curation, software. L.N.B.S.: investigation, data curation, validation. E.V.A.M.: investigation, validation, software. J.A.A.S.B.: investigation, formal analysis, visualization. E.P.G.: investigation, supervision, writing – review and editing. M.G.: conceptualization, formal analysis, writing – review and editing, supervision. All authors read and approved the final manuscript.

Acknowledgements The authors gratefully acknowledge CAPES, for financial support, and the Universidade do Estado da Bahia (UNEB) for institutional support.

Authors’ information Not applicable.

References

- Abeyisirigunawardena, D.S., Jeffries, M.A., Morley, M.G., Bui, A.O.V., Hoeberechts, M. (2016, feb 8). Data quality control and quality assurance practices for Ocean Networks Canada observatories. *OCEANS 2015 - MTS/IEEE Washington*, , <https://doi.org/10.23919/oceans.2015.7404600>
- Aiwale, S., Stahl, F.T., Sun, L. (2026). *Had-QC: A Hybrid AI Approach for Automated Quality Control of Argo Float Data* (Vol. 16302 LNAI).
- Bauer, F., Su, J., Skovgaard Madsen, K.S., Larsen, M.A.D. (2026, may 1). Seaqc-X: Transferability of a machine learning-based sea level quality control framework. *Applied Ocean Research*, 170, , <https://doi.org/10.1016/j.apor.2026.105065>
- Bekiryazıcı, S., Amarouche, K., Ozcan, N., Akpınar, A. (2025). An innovative deep learning-based approach for significant wave height forecasting. *Ocean Engineering*, , <https://doi.org/10.1016/j.oceaneng.2025.120623>
- Bushnell, M.H. (2016, feb 8). Quality Assurance/Quality Control of Real-Time Oceanographic Data. *OCEANS 2015 - MTS/IEEE Washington*, , <https://doi.org/10.23919/oceans.2015.7404613>
- Bushnell, M.H., Waldmann, C., Seitz, S., Henslop, E., Lara-Lopez, A.L. (2019, nov 19). Quality Assurance of Oceanographic Observations: Standards and Guidance Adopted by an International Partnership. *Frontiers in Marine Science*, 6, , <https://doi.org/10.3389/fmars.2019.00706>
- Castelão, G.P. (2021, oct 1). A machine learning approach to quality control oceanographic data. *Computers and Geosciences*, 155, , <https://doi.org/10.1016/j.cageo.2021.104803>

- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002, jan 1). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357, <https://doi.org/10.1613/jair.953>
- da Silva, M.B.L., Barreto, F.T.C., Costa, M.C.D.O., Camargo, R. (2025). Bias correction of significant wave height with LSTM neural networks. *Ocean Engineering*, , <https://doi.org/10.1016/j.oceaneng.2024.120015>
- Diamant, R., Shachar, I., Makovsky, Y., Ferreira, B.M., Cruz, N.A. (2020, nov 1). Cross-sensor quality assurance for marine observatories. *Remote Sensing*, 12(21), 1–16, <https://doi.org/10.3390/rs12213470>
- Gao, S., Song, M., Hu, W., Song, H., Cheng, K. (2023, jan 1). Research on Anomaly Detection of Offshore Buoy Water Temperature Data Based on LSTM-DAE Model. *2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence, RICAI 2023*, 858–863, <https://doi.org/10.1109/RICAI60863.2023.10489464>
- Haines, S.M., Crout, R.L., Bosch, J., Symonds, D.R., Thomas, J. (2011, may 27). A summary of quality control tests for waves and in situ currents and their effectiveness. *2011 IEEE/OES/CWTM 10th Working Conference on Current, Waves and Turbulence Measurement, CWTM 2011*, 100–106, <https://doi.org/10.1109/CWTM.2011.5759534>
- He, H., & Garcia, E.A. (2009, sep 1). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284, <https://doi.org/10.1109/TKDE.2008.239>
- Huang, Y., Wu, R., Hu, Z., Cui, Y. (2026). A synergistic VMD-LSTM-LightGBM model for significant wave height prediction in the Northern Taiwan Strait. *Ocean Engineering*, , <https://doi.org/10.1016/j.oceaneng.2026.126749>
- Iwaniak, M., Zalewska, T., Sapięga, P. (2025). Quality control of time-series seawater temperature and wave data adapted to the regional conditions of the baltic sea. *Oceanological and Hydrobiological Studies*, 54(1), 59–78, <https://doi.org/10.26881/oahs-2025.1.06>

- Ji, Q., Han, L., Jiang, L., Liu, Y. (2023). Short-term prediction of the significant wave height and average wave period based on the variational mode decomposition–temporal convolutional network–long short-term memory (VMD–TCN–LSTM) algorithm. *Ocean Science*, , <https://doi.org/10.5194/os-19-1561-2023>
- Kerfoot, J., Baltes, R.E., Bushnell, M.H., Campbell, L., Knee, K.R. (2016, nov 28). The U.S. National glider network: Application of QARTOD recommended quality control methods to glider CTD data sets. *OCEANS 2016 MTS/IEEE Monterey, OCE 2016*, , <https://doi.org/10.1109/OCEANS.2016.7761356>
- Lin, S., Clark, R., Birke, R., Schönborn, S., Trigoni, N., Roberts, S. (2020). Anomaly detection for time series using VAE-LSTM hybrid model. *Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4322–4326).
- Liu, F.T., Ting, K.M., Zhou, Z.-H. (2008). Isolation forest. *Proceedings of the 2008 eighth ieee international conference on data mining (icdm)* (pp. 413–422).
- Mieruch, S., Demirel, S., Simoncelli, S., Schlitzer, R., Seitz, S. (2021, apr 28). Salaciaml: A Deep Learning Approach for Supporting Ocean Data Quality Control. *Frontiers in Marine Science*, 8, , <https://doi.org/10.3389/fmars.2021.611742>
- Modi, A., Bhaskar, T.V.S.U., Pg, R., Balakrishnan Nair, T.M.B. (2026). Quality control of wave rider buoy data and its applications to cyclones and swell surges in the Indian Ocean. *Ocean Dynamics*, , <https://doi.org/10.1007/s10236-026-01772-0>
- Nguyen, N.T., Heldal, R., Pelliccione, P. (2024, dec 1). Concept-drift-adaptive anomaly detector for marine sensor data streams. *Internet of Things (The Netherlands)*, 28, , <https://doi.org/10.1016/j.iot.2024.101414>
- Nguyen, N.T., Skålvik, A.M., Sylligardos, E., Palpanas, T., Alvsvåg, S.J. (2025, jan 1). Interpretable Multivariate Anomaly Detector Selection for Automatic Marine Data Quality Control. *Proceedings of the IEEE International Conference on Big Data, BigData(2025)*, 1494–1499, <https://doi.org/10.1109/BigData66926.2025.11401894>
- Rahman, A., Smith, D.V., Timms, G.P. (2013, aug 9). Multiple classifier system for automated quality assessment of marine sensor data. *Proceedings of the 2013 IEEE 8th International Conference on Intelligent Sensors, Sensor Networks and Information Processing: Sensing the Future, ISSNIP 2013*, 1, 362–367, <https://doi.org/10.1109/ISSNIP.2013.6629262>

- Ren, Z., de Vos, M., Ravishanker, N., O'Donnell, J. (2026). Noise-augmented and probabilistic deep learning for significant wave height forecasting with attention-based LSTM models. *Applied Ocean Research*, , <https://doi.org/10.1016/j.apor.2026.105016>
- Smith, D.V., Timms, G.P., de Souza, P.A., D'Este, C. (2012, jul 1). A Bayesian framework for the automated online assessment of sensor data quality. *Sensors (Switzerland)*, 12(7), 9476–9501, <https://doi.org/10.3390/s120709476>
- Tan, Z., Zhang, B., Wu, X., Dong, M., Cheng, L. (2022, feb 1). Quality control for ocean observations: From present to future. *Science China Earth Sciences*, 65(2), 215–233, <https://doi.org/10.1007/s11430-021-9846-7>
- Wang, F., Ye, Q., Chen, B., Qiu, L. (2026). Mitigating domain shift for cross-station wave height prediction in South China Sea island and reef waters: Proposal and systematic evaluation of three novel methods. *Journal of Ocean Engineering and Science*, , <https://doi.org/10.1016/j.joes.2026.05.013>
- Wang, K., Liu, Y., Xing, Q., Lv, M. (2024). An integrated system to significant wave height prediction: Combining feature engineering, multi-criteria decision making, and hybrid kernel density estimation. *Expert Systems with Applications*, , <https://doi.org/10.1016/j.eswa.2023.122351>
- Xie, J., Jiang, H., Song, W., Yang, J. (2023). A novel quality control method of time-series ocean wave observation data combining deep-learning prediction and statistical analysis. *Journal of Sea Research*, 195, 102439, <https://doi.org/10.1016/j.seares.2023.102439>
- Xie, K., & Zhang, T. (2025). Forecasting significant wave height intervals along China's coast based on hybrid modal decomposition and CNN-BiLSTM. *Journal of Marine Science and Engineering*, , <https://doi.org/10.3390/jmse13061163>
- Zhao, L., Li, Z., Zhang, J., Teng, B. (2023). An integrated complete ensemble empirical mode decomposition with adaptive noise to optimize LSTM for significant wave height forecasting. *Journal of Marine Science and Engineering*, , <https://doi.org/10.3390/jmse11020435>