

Article

MSFA: Multi-Strategy Fusion Algorithm for Data Cleaning and Its Application in Offshore Marine Environmental Monitoring

Kun Chen ^{1,2}, Ruikang Chang ², Li Ma ², Chao Ji ^{1,3} and Quan Liu ^{1,2,*}¹ State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan 430063, China; kunchen@whut.edu.cn (K.C.)² School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China³ National Engineering Research Center of Port Hydraulic Construction Technology, Tianjin Research Institute for Water Transport Engineering, Ministry of Transport, Tianjin 300456, China

* Correspondence: 11771@whut.edu.cn

Abstract

Marine monitoring records collected from buoys and nearshore sensors are often affected by missing values, abrupt spikes, and short-term fluctuations. These errors are difficult to remove with a single detection or interpolation rule, especially when local anomalies and global outliers occur in the same sequence. This study develops a Multi-Strategy Fusion Architecture (MSFA) for cleaning marine environmental time-series data. In MSFA, DBSCAN is not applied directly to the raw observations; instead, the time index and measurement value are first normalized into a common feature space, where local density anomalies can be detected more consistently. IQR screening is then used to identify global extreme values. After abnormal positions are marked, the repair result is estimated from two complementary sources: linear interpolation, which follows local temporal change, and a moving average based only on neighboring valid observations, which reduces random noise. Their contributions are adjusted according to local reliability rather than fixed manually. Because initial repair may still leave small residual errors, we further use a Combined Residual Metric (CRM) with a median/MAD-based threshold to recheck the repaired sequence and update the abnormal-position set when necessary. Experiments on the 2020 Dongying offshore buoy dataset and a self-collected nearshore dataset show that MSFA achieves AUROC/AUPRC/NRMSE values of 0.896/0.855/0.066 and 0.986/0.915/0.0653, respectively. Compared with DBSCAN+LOF, DBSCAN+Transformer, and IQR+Sigmoid, MSFA improves AUROC and AUPRC by about 12–25% on average and reduces NRMSE by more than 40%. These results indicate that the proposed method can improve the usability of noisy and incomplete marine monitoring data while keeping the cleaning process interpretable.

Keywords: data cleaning; Multi-Strategy Fusion; anomaly detection; combined residual metric

Academic Editor: Michele Melchiori

Received: 9 May 2026

Revised: 10 July 2026

Accepted: 14 July 2026

Published: 17 July 2026

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

1. Introduction

The rapid progress of the Internet of Things (IoT), artificial intelligence (AI), and big data analytics has brought intelligent transformation to many marine-related systems, including marine environmental monitoring, waterway safety management, and offshore navigation aids. In these application scenarios, sensor networks continuously collect time-

series information related to marine conditions, hydrology, and meteorology, providing essential data support for intelligent navigation and safety-oriented decision-making. At the same time, the growing volume and speed of data acquisition have also exposed a series of data quality problems. Marine environmental data are often nonlinear, spatio-temporally coupled, and easily affected by sensor failures, communication instability, and external environmental interference. As a result, missing values, redundant records, and abnormal observations frequently appear in practical datasets. Data cleaning has therefore become an indispensable preprocessing step for improving the reliability and usability of large-scale environmental data.

Traditional statistical methods usually identify anomalies or recover missing values by relying on distributional assumptions and robust statistical rules. For example, Xiang et al. [1] used a quartile-based strategy to remove abnormal records from wind turbine monitoring data before condition monitoring and anomaly detection. Zhang et al. [2] further combined quartile screening with RANSAC regression for cleaning abnormal wind speed–power data, showing that robust statistical boundaries can help improve the usability of wind turbine SCADA records. In energy monitoring scenarios, Martin Nascimento et al. [3] compared several commonly used outlier detection rules, including the 3σ rule, MAD-based thresholds, and boxplot-based screening. These approaches are easy to implement and highly interpretable. However, their limitations become apparent when different types of data quality problems occur together. In marine sensor sequences, local density anomalies, global extreme values, random disturbances, and consecutive missing segments may appear at the same time, making it difficult for a single fixed statistical threshold to capture the local temporal context of the data.

In addition to statistical rules, neighborhood- and density-based methods have also been widely used for anomaly detection. Morrison et al. [4] examined anomaly detection strategies for wind turbine SCADA data cleaning and power-curve preprocessing. Yue et al. [5] combined LOF and DBSCAN to detect abnormal data in gas-flow-meter measurements, indicating that local density deviation and density clustering can provide complementary information in noisy industrial data. Hou et al. [6] also applied DBSCAN clustering to large-scale wind turbine cluster modeling, further showing the usefulness of density-based grouping in complex energy-related time-series data. Nevertheless, these methods are still sensitive to neighborhood parameters and distance definitions. For non-stationary marine monitoring data, dense abnormal segments may be mistakenly grouped into normal clusters, whereas sparse but physically valid observations may be treated as noise. Recent benchmark studies have also shown that the performance of time-series anomaly detection methods can vary greatly across datasets and anomaly types [7].

Machine learning methods have recently been introduced into data cleaning and anomaly detection tasks. Isolation Forest, for instance, has been used to clean wind speed–power data by combining turbine operating characteristics with isolation-based anomaly scoring [8]. Broader surveys and benchmark studies have also shown that support-vector-based models, isolation-based methods, and deep neural networks can perform well in time-series anomaly detection under suitable data conditions [7,9,10]. However, their robustness is not guaranteed when the training data themselves contain dense, structured, or strongly correlated outliers. Once these abnormal samples affect the training distribution, the learned decision boundary, reconstruction behavior, or isolation structure may shift away from the true normal pattern. This problem is particularly relevant for marine monitoring data, where natural fluctuations and sensor-induced errors often overlap in the same sequence.

Beyond anomaly detection, data cleaning has also been investigated from the perspectives of data repair, error correction, and missing-value imputation. In rule-driven repair, Mecca et al. proposed BUNNI, which learns repair actions for rule-based data

cleaning and uses machine learning to identify dirty values and generate repair candidates [11]. Jäger and Biessmann connected data imputation with automated data cleaning through Conformal Data Cleaning, where imputation techniques and conformal prediction are combined to detect and correct erroneous values in a model-agnostic manner [12]. For multivariate time-series imputation, Du et al. proposed SAITS, a self-attention-based model designed to capture both temporal dependencies and cross-feature correlations [13]. Nie et al. later developed ImputeFormer, a low-rankness-induced Transformer for generalizable spatiotemporal imputation [14]. Wang et al. introduced an observed-value-consistent diffusion model, which explicitly considers the relationship between observed and missing values during multivariate time-series imputation [15].

Although these methods provide useful strategies for data repair, error correction, and time-series imputation, marine multi-source monitoring data present a more coupled cleaning problem. Missing values, local density anomalies, global extreme values, and random noise may appear simultaneously under strong hydrometeorological fluctuations and sensor instability. General repair or imputation methods usually focus on reconstructing missing or corrupted values after their positions have been determined, whereas marine monitoring data require anomaly identification, local temporal repair, and residual consistency checking to be considered together. Therefore, this study positions MSFA as a task-oriented marine time-series cleaning framework that integrates normalized anomaly detection, context-aware hybrid repair, and CRM-based residual correction within a unified detection–repair–correction workflow.

Marine environmental data are characteristically nonlinear, noisy, and frequently incomplete because of complex hydrometeorological dynamics, sensor instability, transmission interruption, and equipment malfunction. This study develops the Multi-Strategy Fusion Architecture (MSFA), a task-oriented data cleaning framework for marine multi-source time-series monitoring data.

The core idea of MSFA is to integrate anomaly detection, data repair, and residual correction into a coherent cleaning process. First, a normalized time-value feature space is constructed for DBSCAN, and IQR-based global screening is incorporated to identify both local density anomalies and global outliers. This design also avoids the scale mismatch between temporal indices and physical measurements. Second, MSFA adopts a context-aware hybrid repair strategy, where linear interpolation and valid-window moving average are adaptively combined according to the reliability estimated from neighboring valid observations, rather than from unavailable values at missing or abnormal positions. Third, a Combined Residual Metric (CRM) is introduced to measure the inconsistency between the fused repair result and its component estimates. Based on this metric, MSFA uses a robust iterative correction mechanism based on the median and Median Absolute Deviation (MAD) to refine residual anomalies that remain after the initial repair.

These modules are connected through shared invalid-position masks, valid observations, component estimates, and residual feedback, forming a closed detection–repair–correction workflow.

The remainder of this paper is organized as follows. Section 2 introduces the proposed MSFA methodology and the two datasets used in this study, including the 2020 Dongying offshore buoy dataset and the self-collected nearshore dataset.

Section 3 reports the comparative evaluation and ablation experiments. Section 4 concludes the paper.

2. Materials and Methods

This section introduces the experimental datasets and the implementation details of MSFA.

2.1. Dataset Description

We used two marine monitoring datasets in the experiments. The first one is the 2020 Dongying offshore buoy observation dataset, supported by the National Key R&D Program of China under Grant No. 2016YFC1402000. It contains 3697 records collected from a single offshore buoy. The variables include wind speed, air temperature, humidity, atmospheric pressure, wave height, wave period, and wave direction. The data were sampled once per hour, and each time point contains 24 monitoring indicators, giving a multidimensional description of the local marine environment. The second dataset comes from a nearshore monitoring scenario. It contains 7698 records and 34 monitoring indicators, including wind speed, wave direction, and water depth.

Artificial contamination was added to each monitoring indicator separately. During this process, the random seed was fixed to keep the experimental setting reproducible. For each indicator, 25% of the observations were randomly selected as contaminated positions. These positions were then assigned to three error types: extreme values, missing values, and noise interference. The proportions of the three error types for each monitoring indicator are given in Table 1.

Table 1. Error types and their proportions for all detection metrics in the two datasets.

Element	Missing Items	Extreme Values	Noise	Dataset1	Dataset2
1. Average wind speed (m/s)	8.98%	12.77%	3.95%	√	-
2. Wind direction of the average wind speed (°)	8.26%	8.54%	8.54%	√	√
3. Maximum wind speed (m/s)	9.37%	13.53%	4.02%	√	√
4. Wind direction of the maximum wind speed (°)	9.37%	7.77%	9.78%	√	√
5. Extreme wind speed (m/s)	8.47%	11.96%	3.98%	√	√
6. Wind direction of the extreme wind speed (°)	7.74%	13.44%	4.22%	√	√
7. Average air temperature (°C)	8.2%	13.01%	3.84%	√	-
8. Maximum air temperature (°C)	8.11%	12.31%	3.98%	√	-
9. Minimum air temperature (°C)	8.68%	12.79%	4.08%	√	-
10. Average humidity (%)	8.09%	13.04%	4.52%	√	-
11. Maximum humidity (%)	7.79%	11.96%	4.46%	√	-
12. Minimum humidity (%)	7.44%	13.01%	3.79%	√	-
13. Average barometric pressure (hPa)	9.28%	4.46%	11.66%	√	-
14. Maximum barometric pressure (hPa)	8.06%	5.17%	11.06%	√	-
15. Minimum barometric pressure (hPa)	8.06%	5.03%	12.58%	√	-
16. 1/10 or 1/3 wave height (m)	8.39%	12.71%	3.81%	√	√
17. 1/10 or 1/3 wave period (s)	8.74%	12.88%	4.14%	√	√
18. Average wave height (m)	8.28%	13.44%	3.22%	√	√
19. Average wave period (s)	9.01%	12.77%	4.27%	√	√
20. Maximum wave height (m)	8.41%	13.01%	3.81%	√	√
21. Maximum wave period (s)	7.71%	12.9%	3.98%	√	√
22. Significant wave height (m)	8.44%	13.15%	3.62%	√	√
23. Significant wave period (s)	8.66%	12.28%	4.33%	√	√
24. Main wave direction (°)	8.41%	13.01%	3.81%	√	-
25. Instant wind speed (m/s)	8.74%	12.98%	3.96%	-	√
26. Wind direction of the instant wind speed (°)	8.74%	7.15%	9.78%	-	√
27. Surface flow velocity (m/s)	9.96%	12.45%	6.22%	-	√
28. Surface flow direction (°)	9.13%	10.79%	5.81%	-	√
29. Average flow velocity in 0.2/0.4/0.6/0.8H (m/s)	8.51%	13.69%	3.63%	-	√
30. Average flow direction in 0.2/0.4/0.6/0.8H (°)	7.99%	12.76%	2.38%	-	√
31. Bottom flow velocity (m/s)	7.88%	13.28%	2.49%	-	√
32. Bottom flow direction (°)	10.37%	9.96%	3.73%	-	√
33. Water depth (m)	9.37%	10.69%	6.87%	-	√

34. Visibility (m)	9.37%	11.59%	6.52%	-	√
--------------------	-------	--------	-------	---	---

Note: Percentages indicate the proportions of synthetic error types among the selected contaminated positions.

For extreme-value errors, the mean and standard deviation were calculated from the valid observations of the current indicator. The selected observations were then replaced by values generated as $\mu + 10\sigma z$, where μ and σ denote the mean and standard deviation of the valid observations, and z follows a standard normal distribution. This type of contamination is used to represent severe sensor spikes or physically unreasonable measurements. Missing-value errors were produced by masking the selected observations as null values. Noise interference was generated by multiplying the selected observations by $1 + 0.25z$, where z also follows a standard normal distribution. In this case, random fluctuations are introduced while the original scale of the physical variable is largely maintained.

The contaminated positions were used as ground-truth anomaly labels for calculating AUROC and AUPRC. The original uncontaminated observations were kept as reference values for calculating NRMSE. All baseline methods and MSFA were tested on the same contaminated sequences and the same anomaly labels, so the reported differences mainly reflect the cleaning performance of the methods themselves. Systematic deviations were not considered as a separate error category in this study. Instead, they were represented by severe extreme-value shifts and scale-preserving multiplicative noise, which can reflect abrupt sensor offsets and persistent measurement disturbances commonly observed in marine monitoring scenarios.

2.2. The Proposed Data Cleaning Method

The proposed MSFA framework consists of three coupled modules: a DBSCAN+IQR anomaly detection module, an adaptive hybrid estimation module, and a CRM-based iterative correction module. The first module identifies missing values, local density anomalies, and global outliers, and then constructs the repair set and valid observation set. The second module reconstructs invalid positions by adaptively fusing linear interpolation and valid-window moving average according to local reliability estimated from neighboring valid observations. The third module evaluates the residual inconsistency of the repaired sequence through CRM and feeds newly detected residual anomalies back into the repair pipeline.

The coupling among these modules is realized through shared invalid-position masks, valid observation sets, component repair estimates, and residual feedback. Therefore, MSFA is formulated as a closed detection–repair–correction workflow for marine multi-source time-series cleaning. The specific workflow is shown in Figure 1.

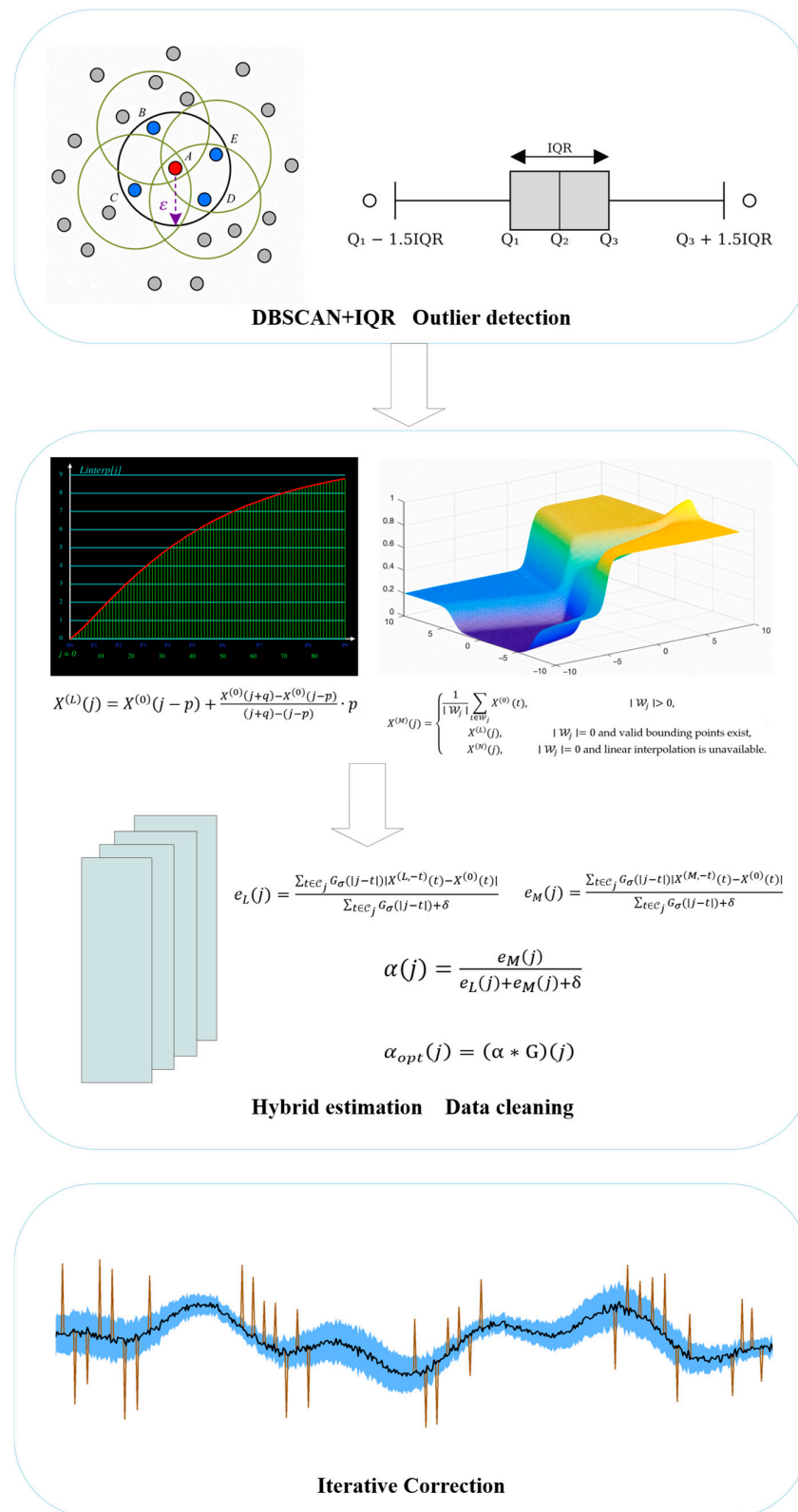


Figure 1. Overall workflow of the proposed MSFA framework. The framework first detects missing values, density-based anomalies, and global outliers using normalized DBSCAN and IQR. The detected invalid positions are then reconstructed by an adaptive hybrid estimation module that combines linear interpolation and valid-window moving average. Finally, a CRM-based iterative correction module further identifies residual anomalies and refines the repaired sequence.

2.2.1. DBSCAN and IQR Algorithms

Algorithm 1 presents the DBSCAN+IQR algorithm process.

Algorithm 1. DBSCAN+IQR Outlier Detection

Require: Contaminated observation sequence $X^{(0)} = \{X^{(0)}(j)\}_{j=1}^N$, small constant δ

Ensure: Masked observation sequence $X_{\text{mask}}^{(0)}$, repair set Ω , valid observation set $\mathcal{V}$

1: Initialize the original missing set Ω_{miss}

2: Compute $X_{\min}$ and $X_{\max}$ from the non-missing values of the current indicator.

3: **For** each non-missing observation $X^{(0)}(j)$, normalize the time coordinate and value coordinate:

$$\tilde{t}_j = \frac{j-1}{N-1},$$

$$\tilde{x}_j = \frac{X^{(0)}(j) - X_{\min}}{X_{\max} - X_{\min} + \delta}.$$

4: Construct the normalized DBSCAN input sample:

$$s_j = (\tilde{t}_j, \tilde{x}_j).$$

5: Form the normalized sample set:

$$S_{\text{norm}} = \{s_j \mid j \notin \Omega_{\text{miss}}\}.$$

6: Set the DBSCAN neighborhood radius:

$$\varepsilon_{\text{DB}} = 0.8 \cdot \text{std}(\tilde{x}).$$

7: Set the minimum number of points:

$$\text{Minpts} = \max(3, \lfloor 0.01N \rfloor).$$

8: Apply DBSCAN to S_{norm} using Euclidean distance in the normalized feature space.

9: Map the DBSCAN noise labels back to their original temporal positions and obtain the DBSCAN noise set Ω_{DB} .

10: Compute the first quartile Q_1 , the third quartile Q_3 , and the interquartile range:

$$\text{IQR} = Q_3 - Q_1.$$

11: Identify the IQR-based global outlier set:

$$\Omega_{\text{IQR}} = \{j \mid X^{(0)}(j) < Q_1 - 1.5\text{IQR} \text{ or } X^{(0)}(j) > Q_3 + 1.5\text{IQR}\}.$$

12: Construct the final repair set:

$$\Omega = \Omega_{\text{miss}} \cup \Omega_{\text{DB}} \cup \Omega_{\text{IQR}}.$$

13: Set all positions in Ω as invalid observations:

$$X_{\text{mask}}^{(0)}(j) = \text{NaN}, j \in \Omega.$$

14: Construct the valid observation set:

$$\mathcal{V} = \{1, 2, \dots, N\} \setminus \Omega.$$

15: **return** $X_{\text{mask}}^{(0)}$, Ω , and $\mathcal{V}$.

Algorithm 1 presents the anomaly detection procedure used in MSFA for handling the complex noise patterns in marine environmental data. The procedure combines density-based detection with robust statistical screening, so that different types of abnormal observations can be identified more reliably. Specifically, DBSCAN is first used to detect local density irregularities, and the Interquartile Range (IQR) method is then applied to screen global extreme deviations. Observations detected by either method are included in the invalid-position set and treated as missing values in the subsequent repair stage, which helps prevent abnormal measurements from affecting the estimation process. A similar need for combining statistical screening [16], clustering [17], and isolation-based detection [18] has also been reported in recent wind turbine and energy-monitoring data cleaning studies [19–22].

In the local density-based detection stage, DBSCAN is performed in a normalized two-dimensional time-value feature space, instead of being directly applied to the raw pair $(j, X^{(0)}(j))$. This normalization is needed because the time index and the corresponding physical measurement may differ greatly in numerical scale. If the Euclidean distance

is directly calculated on the unnormalized pair $(j, X^{(0)}(j))$, the distance measurement may be dominated by the time index, especially for long time-series sequences.

For each non-missing observation $X^{(0)}(j)$, the time index and the measurement value are first normalized as follows:

$$\tilde{t}_j = \frac{j-1}{N-1}, \quad (1)$$

$$\tilde{x}_j = \frac{X^{(0)}(j) - X_{\min}}{X_{\max} - X_{\min} + \delta'}, \quad (2)$$

Here, j denotes the temporal index of the current observation, N is the length of the time-series sequence, and $X^{(0)}(j)$ represents the observed value at time index j . $X_{\min}$ and $X_{\max}$ denote the minimum and maximum non-missing values of the current indicator, respectively. The term δ is a small positive constant used to avoid division by zero when $X_{\max} = X_{\min}$. After normalization, both $\tilde{t}_j$ and $\tilde{x}_j$ are scaled into the interval $[0, 1]$.

The normalized DBSCAN input sample is then constructed as:

$$s_j = (\tilde{t}_j, \tilde{x}_j), \quad (3)$$

where s_j denotes the normalized two-dimensional sample corresponding to the original observation $X^{(0)}(j)$. The normalized sample set is denoted as $S_{\text{norm}} = \{s_j\}$, which contains all non-missing observations used for DBSCAN detection.

DBSCAN is performed on S_{norm} using the standard Euclidean distance:

$$D(s_i, s_j) = \sqrt{(\tilde{t}_i - \tilde{t}_j)^2 + (\tilde{x}_i - \tilde{x}_j)^2}, \quad (4)$$

In this equation, $D(s_i, s_j)$ represents the distance between two normalized samples s_i and s_j . Since both dimensions are normalized to the same range, the distance calculation is no longer dominated by the raw time index.

The neighborhood radius is defined in the normalized feature space as:

$$\varepsilon_{\text{DB}} = 0.8 \cdot \text{std}(\tilde{x}), \quad (5)$$

where ε_{DB} denotes the DBSCAN neighborhood radius, and $\text{std}(\tilde{x})$ is the standard deviation of the normalized measurement values. The minimum number of points is set as:

$$\text{MinPts} = \max(3, \lfloor 0.01N \rfloor), \quad (6)$$

where MinPts denotes the minimum number of neighboring samples required to form a dense region, and $\lfloor 0.01N \rfloor$ represents the integer part of 1% of the sequence length. After normalization, the time index and the measurement value are brought to a comparable scale in the DBSCAN distance calculation. This reduces the risk that one dimension dominates the distance measurement simply because of its larger numerical range.

The coefficients in Equations (5) and (6) were determined through preliminary parameter trials and sensitivity analysis. During these trials, the neighborhood-radius coefficient of DBSCAN and the minimum-neighbor coefficient were adjusted around the adopted settings to examine the trade-off between anomaly detection sensitivity and reconstruction stability. A one-factor sensitivity analysis was then carried out based on the default DBSCAN configuration. The radius coefficient was set to 0.6, 0.8, and 1.0 in turn, with the MinPts coefficient fixed at 0.01. Next, the MinPts coefficient was set to 0.005, 0.01, and 0.02, while the radius coefficient remained fixed at 0.8. The averaged AUROC, AUPRC, and NRMSE results over all indicators from the two datasets are summarized in Table 2.

Table 2. Sensitivity analysis of DBSCAN parameter settings

Parameter Setting	AUROC/%	AUPRC/%	NRMSE
ε_{DB} coefficient = 0.6; MinPts = 0.01N	92.2	87.9	0.062
ε_{DB} coefficient = 0.8; MinPts = 0.01N	92.7	88.6	0.058
ε_{DB} coefficient = 1.0; MinPts = 0.01N	91.3	86.7	0.072
ε_{DB} coefficient = 0.8; MinPts = 0.005N	92.4	88.1	0.069
ε_{DB} coefficient = 0.8; MinPts = 0.02N	91.5	87.2	0.065

Note: Bold values indicate the best-performing results among the compared parameter settings or methods. Higher AUROC and AUPRC values and lower NRMSE values indicate better performance.

The sensitivity results suggest that the adopted parameter setting provides a relatively balanced performance. When the neighborhood-radius coefficient was set to 0.6, DBSCAN responded more strongly to small local variations. This improved sensitivity came at a cost: several normal observations were also flagged as abnormal, and the NRMSE increased accordingly. With the coefficient increased to 1.0, the detector became less sensitive to sparse local anomalies. Some of these anomalies were grouped with nearby normal samples, resulting in lower AUROC and AUPRC values.

MinPts showed a similar sensitivity–stability trade-off. A smaller MinPts value helped capture isolated abnormal points, but the subsequent repair became less stable. A larger MinPts value made the clustering criterion stricter, which caused short abnormal segments to be missed more easily. The default setting gave the most balanced results among the tested combinations, achieving the highest average AUROC and AUPRC and the lowest average NRMSE. Therefore, the parameters in Equations (5) and (6) were used in the following comparative and ablation experiments.

The operational logic is defined as follows:

Each non-missing observation is first represented as a normalized two-dimensional sample $s_j = (\tilde{t}_j, \tilde{x}_j)$. Since both coordinates are scaled to [0, 1], the Euclidean distance used by DBSCAN is not dominated by the raw time index.

1. For each normalized sample, the algorithm evaluates the number of samples within its ε_{DB} -neighborhood. If the number of neighboring samples is greater than or equal to *MinPts*, the sample is designated as a core point.
2. Samples that are directly or indirectly density-reachable from core points are aggregated into normal data clusters. A non-core point located within the ε_{DB} -neighborhood of a core point is classified as a border point, whereas a point that is not density-reachable from any core point is marked as a DBSCAN noise point and included in Ω_{DB} .

Density-based clustering is useful for identifying local spatial noise, but marine sensor data may also contain systematic deviations caused by equipment drift or measurement bias. For this reason, MSFA further introduces a global statistical screening step based on the IQR method. Since IQR uses order statistics, such as medians and quartiles, rather than the mean and variance, it is less sensitive to severe contamination and non-normal data distributions.

The radius circles shown in Figure 2 are schematic illustrations of the DBSCAN neighborhood in the normalized feature space. Their scale follows the neighborhood radius defined in Algorithm 1, and should not be interpreted according to the raw time index or the original measurement units. In this way, density reachability is evaluated only after the time and value dimensions have been normalized to comparable numerical ranges.

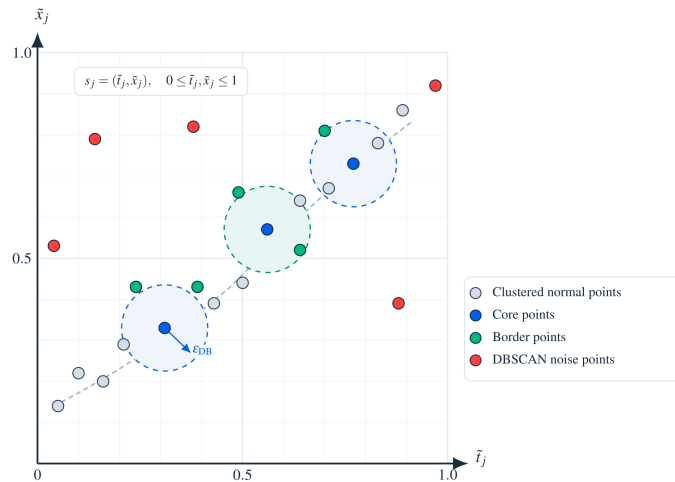


Figure 2. Illustration of DBSCAN anomaly detection in the normalized time-value feature space. The horizontal axis represents the normalized time index, and the vertical axis represents the normalized measurement value. Each observation is represented as a normalized sample in this two-dimensional space. The circles indicate the DBSCAN neighborhood radius in the normalized feature space. Points with sufficient neighboring samples are identified as core points, non-core points within the neighborhood of a core point are classified as border points, and points that are not density-reachable from any core point are identified as DBSCAN noise points.

Formally, let the dataset be arranged as an ascending sequence. The mechanism isolates global anomalies based on its ability to calculate the first Q_1 , second Q_2 , and third Q_3 quartiles:

$$Q_1 = \begin{cases} \frac{m+1}{4}, (m+1)/4 = 4k \\ \frac{3}{4}x_{k+1} + \frac{1}{4}x_{k+2}, (m+1)/4 \neq 4k' \end{cases} \quad (7)$$

In the above formula, Q_1 is the first quartile, and m is the number of data points in the sequence.

$$Q_2 = \begin{cases} \frac{x_{\frac{m+1}{2}}, m = 2k + 1, k = (0, 1, 2, 3, \dots) \\ \frac{1}{2}(x_{\frac{m}{2}} + x_m), m = 2k, k = (1, 2, 3, \dots) \end{cases} \quad (8)$$

In the above formula, Q_2 is the second quartile, and m is the number of data points in the sequence.

$$Q_3 = \begin{cases} \frac{3(m+1)}{4}, \frac{3(m+1)}{4} = 4k \\ \frac{1}{4}x_{3k+1} + \frac{3}{4}x_{3k+2}, \frac{3(m+1)}{4} \neq 4k' \end{cases} \quad (9)$$

In the above formula, Q_3 is the third quartile, and m is the number of data points in the sequence.

After calculating according to the above formula, the interquartile range IQR can be determined:

$$IQR = Q_3 - Q_1, \quad (10)$$

By combining the above formulas, the range of outliers for the data points can be obtained:

$$x_m \begin{cases} \text{outlier, } x_m < Q_1 - 1.5IQR \\ \text{outlier, } x_m > Q_3 + 1.5IQR' \end{cases} \quad (11)$$

Based on these rigorous statistical thresholds, any observation falling strictly outside the bounds defined by I_{QR} is identified as a global outlier and systematically removed alongside the density-based noise.

After the DBSCAN-IQR detection stage, all missing values and the points identified as density-based or statistical outliers are uniformly marked as invalid observations. The corresponding repair set is defined as:

$$\Omega = \Omega_{miss} \cup \Omega_{DB} \cup \Omega_{IQR}, \quad (12)$$

where Ω_{miss} , Ω_{DB} , and Ω_{IQR} denote the sets of original missing positions, DBSCAN-detected noise points, and IQR-detected global outliers, respectively. The valid observation set is then defined as:

$$\mathcal{V} = \{1, 2, \dots, N\} \setminus \Omega, \quad (13)$$

Here, N represents the length of the time-series sequence. The positions in Ω are regarded as target positions to be reconstructed, whereas the observations in $\mathcal{V}$ are regarded as reliable neighboring observations and are used for local reliability estimation in the subsequent hybrid repair stage.

2.2.2. Hybrid Estimation Method

Once the multi-scale detection phase isolates and masks the anomalies, the MSFA framework deploys an adaptive hybrid estimation layer (Algorithm 2) to reconstruct the missing data points. Traditional imputation techniques often force a compromise: linear interpolation tracks local trajectories effectively but is highly vulnerable to high-frequency noise, whereas moving averages suppress random fluctuations but fail to predict sharp local trends across significant data gaps. To resolve this dichotomy, MSFA orchestrates a dual-estimation baseline. Recent time-series imputation studies also suggest that reconstruction performance is closely related to the missingness pattern and temporal dependency structure [23–25].

Algorithm 2 presents the process of the hybrid estimation algorithm.

Algorithm 2. Hybrid Estimation with Gaussian-Smoothed Adaptive Weights

Require: Masked observation sequence $X^{(0)}$, repair set Ω , valid observation set $\mathcal{V}$, sliding window length k , context half-width h , Gaussian bandwidth σ , small constant δ

Ensure: Hybrid repaired signal $X^{(F,0)}$

1: Initialization.

2: **for** $j = 1$ to N **do**

3: **if** $j \in \Omega$ **then**

4: Find the nearest valid points $j - p$ and $j + q$, where $j - p \in \mathcal{V}$ and $j + q \in \mathcal{V}$, if they exist.

5: **if** valid bounding points exist **then**

6: Compute the linear interpolation estimate:

$$X^{(L)}(j) = X^{(0)}(j - p) + \frac{X^{(0)}(j + q) - X^{(0)}(j - p)}{(j + q) - (j - p)} \cdot p.$$

7: **else**

8: Set $X^{(L)}(j) = X^{(N)}(j)$ using the nearest valid observation.

9: **end if**

10: Construct the valid moving window $\mathcal{W}_j$.

11: **if** $|\mathcal{W}_j| > 0$ **then**

12: Compute the moving average estimate:

$$X^{(M)}(j) = \frac{1}{|\mathcal{W}_j|} \sum_{t \in \mathcal{W}_j} X^{(0)}(t).$$

```

13:   else if valid bounding points exist then
14:       Set  $X^{(M)}(j) = X^{(L)}(j)$ .
15:   else
16:       Set  $X^{(M)}(j) = X^{(N)}(j)$ .
17:   end if
18:   Construct the valid context window  $\mathcal{C}_j$ .
19:   Estimate  $e_L(j)$  and  $e_M(j)$  from neighboring valid context observations.
20:   Compute the initial adaptive fusion weight:
       
$$\alpha(j) = \frac{e_M(j)}{e_L(j) + e_M(j) + \delta}.$$

21:   else
22:       Retain the valid observation:
       
$$X^{(F,0)}(j) = X^{(0)}(j).$$

23:   end if
24: end for
25: Apply Gaussian smoothing to the initial weight sequence  $\alpha(j)$  to obtain the opti-
    mized weight sequence  $\alpha_{\text{opt}}(j)$ .
26: for each  $j \in \Omega$  do
27:   Compute the hybrid repair result:
       
$$X^{(F,0)}(j) = \alpha_{\text{opt}}(j)X^{(L)}(j) + [1 - \alpha_{\text{opt}}(j)]X^{(M)}(j).$$

28: end for
29: return  $X^{(F,0)}$ 

```

For any target position $j \in \Omega$, the linear interpolation component estimates the missing or abnormal value using the nearest valid observations before and after j . Let $j - p$ and $j + q$ denote the nearest preceding and succeeding valid positions, respectively. The linear interpolation estimate is defined as:

$$X^{(L)}(j) = X^{(0)}(j - p) + \frac{X^{(0)}(j + q) - X^{(0)}(j - p)}{(j + q) - (j - p)} \cdot p, \quad (14)$$

where $X^{(L)}(j)$ denotes the linear interpolation estimate at position j , $X^{(0)}$ denotes the observed sequence after anomaly masking, and p and q represent the temporal distances from j to the nearest preceding and succeeding valid observations, respectively. Specifically, $j - p \in \mathcal{V}$ and $j + q \in \mathcal{V}$.

Simultaneously, the framework computes a localized moving average to establish a noise-resistant statistical baseline. Different from the conventional moving average that directly averages all samples within a fixed window, the proposed implementation only averages valid observations that have not been marked as missing or anomalous. For a target position $j \in \Omega$, the valid moving window is defined as:

$$\mathcal{W}_j = \left\{ t \in \mathcal{V} \mid |t - j| \leq \frac{k-1}{2} \right\}, \quad (15)$$

where k is the sliding window length and is set as an odd integer. The moving average estimate is calculated according to the following boundary-aware rule:

$$X^{(M)}(j) = \begin{cases} \frac{1}{|\mathcal{W}_j|} \sum_{t \in \mathcal{W}_j} X^{(0)}(t), & |\mathcal{W}_j| > 0, \\ X^{(L)}(j), & |\mathcal{W}_j| = 0 \text{ and valid bounding points exist,} \\ X^{(N)}(j), & |\mathcal{W}_j| = 0 \text{ and linear interpolation is unavailable.} \end{cases} \quad (16)$$

Here, $X^{(M)}(j)$ denotes the moving average repair result at position j , $X^{(0)}(t)$ represents a valid observation within the local moving window, and $|\mathcal{W}_j|$ denotes the number of valid observations included in the window. The set $\mathcal{V}$ contains only valid obser-

variations after DBSCAN-IQR anomaly masking. Therefore, missing values and anomaly-flagged values are excluded from the moving average calculation.

When $|\mathcal{W}_j| > 0$, the moving average is computed only from valid observations, and the denominator is dynamically adjusted to $|\mathcal{W}_j|$, rather than the fixed window length k . When $|\mathcal{W}_j| = 0$, which may occur in long consecutive missing or corrupted segments, the moving average component is regarded as unavailable. Here, valid bounding points refer to the nearest valid observations before and after the target position j . In other words, linear interpolation is used only when there is at least one valid observation on both sides of j . If one side has no valid observation, the linear interpolation component is regarded as unavailable, and the nearest-valid-value estimate $X^{(N)}(j)$ is used as a conservative fallback.

The nearest-valid-value estimate $X^{(N)}(j)$ denotes the value of the valid observation closest to the target position j . This boundary-aware formulation avoids NaN values in the moving average component and keeps the hybrid estimation layer computable even when consecutive missing or corrupted segments occur. The fallback rule is only used to preserve computational continuity. It does not imply that very long missing blocks can always be accurately reconstructed.

Linear interpolation is useful for retaining local temporal trends, while the moving average method is better at reducing random fluctuations. Based on this complementarity, MSFA combines the two candidate repair results through an adaptive hybrid estimation strategy. The fusion weight is determined by the local reliability of the two components. Importantly, the local errors used for adaptive weighting are not computed at the target position $j \in \Omega$, where the true value is missing or unreliable. Instead, they are estimated from neighboring valid observations around j , and then transferred to the target position using Gaussian distance weighting.

For each target position $j \in \Omega$, the valid context window is defined as:

$$\mathcal{C}_j = \{t \in \mathcal{V} \mid 0 < |t - j| \leq h\}, \quad (17)$$

where h denotes the half-width of the valid context window. The condition $0 < |t - j|$ ensures that the target position j itself is not included in the context window.

The Gaussian distance weight is defined as:

$$G_\sigma(d) = \exp\left(-\frac{d^2}{2\sigma^2}\right), \quad (18)$$

where $d = |j - t|$ denotes the temporal distance between the target position j and a valid context point t , and σ controls the decay rate of the Gaussian distance weight.

Based on the valid context window, the local reliability errors of the linear interpolation and moving average components are estimated as:

$$e_L(j) = \frac{\sum_{t \in \mathcal{C}_j} G_\sigma(|j-t|) |X^{(L,-t)}(t) - X^{(0)}(t)|}{\sum_{t \in \mathcal{C}_j} G_\sigma(|j-t|) + \delta}, \quad (19)$$

$$e_M(j) = \frac{\sum_{t \in \mathcal{C}_j} G_\sigma(|j-t|) |X^{(M,-t)}(t) - X^{(0)}(t)|}{\sum_{t \in \mathcal{C}_j} G_\sigma(|j-t|) + \delta}, \quad (20)$$

Here, $X^{(L,-t)}(t)$ and $X^{(M,-t)}(t)$ denote the linear interpolation and moving average estimates of the valid context point t , respectively, obtained by temporarily excluding t from the local estimation process. The value $X^{(0)}(t)$ is the observed value of a valid neighboring point, rather than the unavailable true value at the target position j . The constant δ is a small positive value used to avoid division by zero.

The initial fusion weight assigned to the linear interpolation component is then calculated as:

$$\alpha(j) = \frac{e_M(j)}{e_L(j) + e_M(j) + \delta'} \quad (21)$$

This inverse-error weighting strategy assigns a larger weight to the component with a smaller estimated local error. When the linear interpolation component is more reliable, $e_L(j)$ becomes smaller and $\alpha(j)$ increases. Conversely, when the moving average component is more reliable, $e_M(j)$ becomes smaller and $\alpha(j)$ decreases, making the final repair rely more on the moving average estimate.

To avoid drastic fluctuations in the fused weights over time, a Gaussian convolution smoothing process is applied to the initial weight sequence $\alpha(j)$ to maintain continuity and stability of the weights over time:

$$\alpha_{opt}(j) = (\alpha * G)(j), \quad (22)$$

The Gaussian kernel function is defined as follows:

$$G(k) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(-\frac{k^2}{2\sigma_g^2}\right), \quad (23)$$

Based on the smoothed fusion weights, the final hybrid repair result is constructed:

$$X_{mix}(j) = \alpha_{opt}(j)X^{(L)}(j) + (1 - \alpha_{opt}(j))X^{(M)}(j), \quad (24)$$

This fusion operation only applies to the locations marked after anomaly detection; the remaining locations retain their original observations. The $X_{mix}(j)$ sequence calculated after the first run is $X^{(F,0)}(j)$ in the iterative correction.

2.2.3. Iterative Correction

The median, as a typical robust estimator, has a core advantage in its extremely high breakdown point. The median is less sensitive to extreme values than the arithmetic mean. In each local window, the samples are sorted, and the middle value is used as the median. If abnormal samples make up less than half of the window, they have only limited influence on this estimate. For this reason, the median is suitable for describing the local center of the residual sequence when impulse noise is present. This design is also consistent with recent robust filtering and residual-based anomaly correction studies for time-series data [26–31].

Algorithm 3 corrects residual errors that may remain after hybrid estimation. MSFA does not detect these errors from the raw observations. Instead, it uses the consistency of the repaired sequence as the basis for correction. This is necessary for marine environmental data, where short-term changes in wind, waves, or hydrological conditions may be part of the real signal. Directly checking the raw sequence could mistakenly treat these natural variations as repair errors.

The Combined Residual Metric (CRM), written as $\Delta^r(j)$, is computed from the differences between the fused repair result and the two candidate estimates: linear interpolation and valid-window moving average. Linear interpolation follows local temporal changes, while the moving average gives a smoother reference under random noise. If $\Delta^r(j)$ differs clearly from its local median, the fused result is regarded as inconsistent with its component estimates, and the corresponding position is marked as a possible residual repair anomaly. In this way, CRM provides a repair-oriented residual criterion rather than a raw-value anomaly criterion.

The detailed iterative correction procedure is presented in Algorithm 3.

Algorithm 3. Iterative correction via Combined Residual Metric (CRM)

Require: Initial fused sequence $X^{(F,0)}(j)$, linear component $X^{(L)}(j)$, moving-average component $X^{(M)}(j)$, window length L , threshold K , maximum iteration R_{max}

Ensure: Final refined sequence $X^{(F,R)}(j)$

- 1: $r \leftarrow 0$
- 2: **while** $r < R_{max}$ **do**
- 3: Compute residuals:

$$\delta_L^{(r)}(j) = X^{(F,r)}(j) - X^{(L)}(j)$$

$$\delta_M^{(r)}(j) = X^{(F,r)}(j) - X^{(M)}(j)$$
- 4: Combined residual metric:

$$\Delta^{(r)}(j) = \frac{|\delta_L^{(r)}(j)| + |\delta_M^{(r)}(j)|}{2}$$
- 5: Construct sliding window:

$$W_j = \{i | j - \frac{L-1}{2} \leq i \leq j + \frac{L-1}{2}\}$$
- 6: Compute local median:

$$M(j) = \text{median}\{\Delta^{(r)}(i)\}_{i \in W_j}$$
- 7: Compute MAD:

$$MAD(j) = \text{median}\{|\Delta^{(r)}(i) - M(j)|\}_{i \in W_j}$$
- 8: Detect anomaly if

$$|\Delta^{(r)}(i) - M(j)| > K \cdot 1.4826 \cdot MAD(j)$$
- 9: Re-execute Algorithm 2 to obtain $X^{(F,r+1)}(j)$
- 10: $r \leftarrow r + 1$
- 11: **end while**
- 12: **return** $X^{(F,r)}(j)$

In the r -th iteration, the residual deviations between the current fused repair sequence and the two component estimates are first calculated in Step 3 of Algorithm 3. Specifically, $\delta_L^{(r)}(j)$ denotes the deviation between $X^{(F,r)}(j)$ and the linear interpolation component $X^{(L)}(j)$, while $\delta_M^{(r)}(j)$ denotes the deviation between $X^{(F,r)}(j)$ and the moving average component $X^{(M)}(j)$. Based on these two residual terms, the combined residual metric used in the CRM sequence is defined as:

$$\Delta^{(r)}(j) = \frac{|\delta_L^{(r)}(j)| + |\delta_M^{(r)}(j)|}{2}, \quad (25)$$

Here, $\Delta^{(r)}(j)$ denotes the combined residual metric at position j during the r -th iteration. It is defined as the mean of the absolute deviations between the current fused repair result $X^{(F,r)}(j)$ and the two component estimates, namely, the linear interpolation estimate $X^{(L)}(j)$ and the moving-average estimate $X^{(M)}(j)$. A larger $\Delta^{(r)}(j)$ indicates greater inconsistency among the repair estimates and, consequently, a higher likelihood of a residual repair error.

To robustly identify such residual anomalies, the algorithm constructs a local sliding window around each position j . The window provides the local distribution of $\Delta^{(r)}(i)$ used for median/MAD-based thresholding:

$$W_j = \{i | j - \frac{L-1}{2} \leq i \leq j + \frac{L-1}{2}\}, \quad (26)$$

Here, L is the preset odd-numbered window length, so that the target position can be centered within the local residual window. The value of L controls the locality of residual evaluation. A smaller L is more sensitive to short-term residual changes, whereas a larger L provides more stable local statistics but may smooth out short residual fluctuations. Within this window, the local median $M(j)$ and the median absolute deviation $MAD(j)$ are calculated as follows. Here, MAD is computed using absolute deviations from the local median:

$$M(j) = \text{median}\{\Delta^{(r)}(i)\}_{i \in W_j}, \quad (27)$$

$$MAD(j) = \text{median}\{|\Delta^{(r)}(i) - M(j)|\}_{i \in W_j'} \quad (28)$$

The median $M(j)$ represents the local central level of the combined residual metric, while $MAD(j)$ provides a robust estimate of the local residual scale. Compared with the mean and standard deviation, the median and MAD are less sensitive to extreme residual values, making them suitable for detecting residual repair anomalies in noisy marine time-series data.

Based on these robust statistics, a position is identified as a residual anomaly in the current iteration when the following criterion is satisfied:

$$|\Delta^{(r)}(i) - M(j)| > K \cdot 1.4826 \cdot MAD(j), \quad (29)$$

In the above formula, K is the threshold gain coefficient used to control the sensitivity of residual anomaly detection and was set to 3 in the experiments. A larger K reduces false positives but may miss subtle residual errors, whereas a smaller K increases sensitivity but may over-correct normal fluctuations. The constant 1.4826 is the consistency factor that maps MAD to the standard-deviation scale under a normal distribution. The residual window length L is chosen as an odd number, approximately equal to 5% of the sequence length, and is not allowed to be smaller than 5.

For the positions that satisfy Equation (29), the corresponding observations are added to the invalid-position set. Algorithm 2 is then performed again to obtain the next-round fused repair sequence $X^{(F,r+1)}(j)$. To avoid unlimited correction, the maximum number of iterations $R_{\max}$ is fixed at 5. The process terminates once no additional residual anomalies are found, or when the maximum iteration number is reached. With this rule, the method can further correct residual repair errors while avoiding excessive repeated correction and possible over-smoothing.

3. Experimental Results and Discussion

This section evaluates MSFA against three baseline cleaning methods and analyzes the contribution of its repair-related modules through ablation experiments.

3.1. Overview of Comparison Algorithms

DBSCAN+LOF: This baseline combines density-based clustering and local density deviation analysis for anomaly detection. For each monitoring indicator, DBSCAN is first applied to identify sparse density-based noise points. The neighborhood radius is adaptively set to 0.3 times the interquartile range of the current indicator, and the minimum number of neighboring samples is set to $\max(5, \lfloor 0.01N \rfloor)$, where N is the sequence length. LOF is then used to further evaluate local density inconsistency, with the number of neighbors set to $\min(20, \max(5, \lfloor 0.05N \rfloor))$. Points whose LOF scores exceed the mean score plus two standard deviations are marked as local anomalies. The final anomaly set is obtained by combining the DBSCAN and LOF detection results. Detected anomalies are then repaired by linear interpolation, followed by Savitzky–Golay local smoothing and light moving-median smoothing to reduce residual spikes.

DBSCAN+Transformer (DBTf): This baseline is implemented as a staged density-detection and lightweight Transformer-style reconstruction pipeline. First, each indicator is standardized and processed by DBSCAN to remove isolated density anomalies, where the DBSCAN radius and minimum-neighbor settings are fixed as 1.5 and 5, respectively, in the standardized one-dimensional feature space. Missing or masked values are initially filled using the column mean to form a complete multivariate input matrix. A lightweight attention-based Transformer-style autoencoder with a hidden dimension of 16 is then used to reconstruct the multivariate sequence. Reconstruction errors are calculated for

each indicator, and the 95th percentile of the reconstruction error is used as the anomaly threshold. The final anomaly set is obtained by merging the DBSCAN-detected anomalies and reconstruction-error anomalies. The detected abnormal positions are repaired using boundary-aware linear interpolation, Savitzky–Golay smoothing with an adaptive odd window length $\max(5, \text{round}(0.05N))$, and moving-median smoothing with a window length $\max(3, \text{round}(0.01N))$. For brevity, DBSCAN+Transformer is abbreviated as DBTf in the following tables and figures.

IQR+Sigmoid: This baseline combines robust statistical screening and nonlinear compression-based repair. For each monitoring indicator, the first quartile Q1, the third quartile Q3, and the interquartile range are calculated. Observations outside the interval $[Q1 - 1.5IQR, Q3 + 1.5IQR]$, as well as missing values, are marked as abnormal. The detected abnormal values are then compressed toward the valid-data mean using a hyperbolic tangent Sigmoid-type mapping, $\mu + 3\sigma \tanh\left(\frac{x-\mu}{3\sigma}\right)$, where μ and σ denote the mean and standard deviation of non-anomalous observations. After the nonlinear correction, any remaining missing values are filled by linear interpolation. Savitzky–Golay smoothing and light moving-median smoothing are then used to improve local continuity and reduce isolated residual fluctuations.

All baseline methods were tested under the same conditions as MSFA. They used the same contaminated datasets, anomaly labels, and evaluation metrics. The synthetic contamination settings were also kept unchanged across methods, including the fixed random seed, the 25% contamination ratio, and the generation rules for extreme values, missing values, and noise interference. AUROC and AUPRC were calculated from the same ground-truth anomaly labels, while NRMSE was measured against the original uncontaminated observations. In this way, the comparison reflects the cleaning performance of each method, instead of differences caused by data generation or evaluation settings.

For the baseline methods, the detected abnormal positions are processed with the same post-detection repair procedure, which includes linear interpolation, Savitzky–Golay local smoothing, and light moving-median smoothing. Using a shared repair backbone makes the comparison more consistent, so the performance differences mainly reflect the anomaly detection and reconstruction strategies of each baseline. The SG filter smooths the sequence by fitting a low-order polynomial within a moving window, while still retaining peak and trend information. Its mathematical form is given as follows:

$$\hat{x}(t) = \sum_{k=-m}^m c_k x(t+k), \quad (30)$$

Here, m denotes the window radius. Since the sliding window is symmetric around the center point, the corresponding window length is $2m + 1$. A larger value of m usually produces a stronger smoothing effect, but it may also weaken rapid changes in the signal. The coefficients c_i are obtained from the local polynomial fitting process.

Compared with the traditional moving average, the Savitzky–Golay filter introduces less response lag while smoothing the sequence. This makes it more suitable for time-series signals with natural fluctuations, such as wind speed and flow velocity. In the comparative experiment, only the detected outliers were replaced:

$$x^*(t) = \begin{cases} \hat{x}(t), & t \in \Omega_a \\ x(t), & t \notin \Omega_a \end{cases} \quad (31)$$

Here, Ω_a denotes the invalid-position set identified by the anomaly detection stage. After interpolation and local repair, a light median smoothing step is further applied to suppress remaining isolated spikes and residual noise.

3.2. Comparative Experiment

We consolidated 34 types of self-collected nearshore marine data into nine key indicators (wind speed/direction, wave height/direction, wave period, visibility, water depth, and flow velocity/direction). Table 3 compares the data cleaning performance—measured by AUROC and AUPRC—of three models against the MSFA algorithm using this dataset.

Table 3. Comparison of Data Cleaning Results for the self-collected nearshore dataset

Method	DBSCAN+LOF		DBTf		IQR+Sigmoid		MSFA	
Indicator (%)	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Wind speed	93.5	89.4	93.4	88.7	93.0	88.2	99.2	95.9
Wind direction	95.6	89.5	96.1	89.5	95.9	89.1	98.6	92.9
Wave height	92.7	86.0	92.6	85.8	91.7	84.4	98.0	90.0
Wave direction	91.8	82.3	92.8	82.9	86.8	68.3	95.4	74.6
Wave period	96.4	92.0	95.7	91.5	95.6	90.6	99.6	96.7
Visibility	98.2	89.2	99.1	95.0	98.3	89.2	99.8	97.6
Water depth	99.5	98.4	99.2	98.1	99.1	98.0	99.9	99.3
Flow velocity	84.8	73.3	84.9	73.1	83.5	71.9	97.7	87.1
Flow direction	90.7	80.3	90.9	80.1	89.5	73.9	98.9	89.0

Note: Bold values indicate the best-performing results among the compared parameter settings or methods. Higher AUROC and AUPRC values indicate better performance.

We integrated 24 types of data from the 2020 Dongying offshore buoy observation dataset into 8 key indicators: wind speed, wind direction, temperature, humidity, atmospheric pressure, wave height, wave period, and wave direction. As shown in Table 4, it compares the data cleaning performance of three data cleaning models and the MSFA algorithm on the 2020 Dongying offshore buoy observation dataset.

Table 4. Comparison of Data Cleaning Results for the 2020 Dongying Offshore Buoy Observation Dataset

Method	DBSCAN+LOF		DBTf		IQR+Sigmoid		MSFA	
Indicator (%)	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Wind speed	74.5	63.4	74.2	63.0	73.9	62.7	84.4	78.9
Wind direction	78.0	63.1	77.9	62.7	78.2	63.3	85.6	81.4
Air temperature	93.2	88.2	92.6	85.8	93.0	88.0	95.0	93.9
Humidity	91.1	83.3	90.8	82.9	90.6	82.7	94.5	92.1
Atmospheric pressure	99.4	98.8	99.3	98.8	99.3	98.7	99.9	99.7
Wave height	79.1	64.8	78.9	65.7	77.7	63.1	82.4	71.6
Wave period	82.6	71.9	81.9	71.2	79.5	70.6	88.8	84.0
Wave direction	78.0	62.7	77.8	62.7	77.8	62.8	86.2	82.3

Note: Bold values indicate the best-performing results among the compared parameter settings or methods. Higher AUROC and AUPRC values indicate better performance.

The experimental results show that our data cleaning method outperforms the selected baseline methods. Specifically, DBSCAN+LOF mainly relies on density clustering and the local outlier factor to identify isolated points and local anomalies. This method has limited sensitivity to continuous anomalies, systematic deviations, and small-amplitude noise. Its correction strategy is also relatively simple, as it relies on a single cleaning mechanism and may therefore leave residual bias in the repaired sequence.

DBTf introduces sequence modeling and detects anomalies through reconstruction errors. However, its Transformer module is implemented as a lightweight approximate autoencoder, and the recovery of abnormal values still depends mainly on linear interpo-

lation. DBTf therefore shows limitations when applied to multi-source time-series data with nonlinear anomalies and strong noise coupling.

The IQR+Sigmoid method detects anomalies mainly through global statistical criteria. After detection, abnormal values are compressed by Sigmoid mapping and then repaired using linear interpolation. This procedure is relatively simple, but it is less responsive to local anomalies and isolated abnormal points. As a result, global deviations and local abnormal patterns are not always captured at the same time.

These results suggest that the joint use of local density detection, global statistical screening, hybrid repair, and CRM-based residual correction improves both anomaly identification and repaired-sequence quality.

AUROC and AUPRC mainly reflect the ability of a cleaning method to identify abnormal or erroneous values. However, they do not directly measure how close the repaired values are to the original uncontaminated observations. For this reason, NRMSE is further introduced to evaluate the repair accuracy of each cleaning method. The calculation of NRMSE is given as follows:

$$NRMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{x}_i - x_i)^2}}{\max(x) - \min(x)} \quad (32)$$

Here, x_i represents the real data, $\hat{x}_i$ represents the predicted data, N is the number of samples, and $\max(x)$ and $\min(x)$ represent the maximum and minimum values of the variable in the real data, respectively. According to this formula, a smaller NRMSE value indicates better cleaning performance and higher repair accuracy.

Figure 3 compares the NRMSE values of different cleaning methods on the two datasets. MSFA generally obtains lower NRMSE values than the baseline methods, indicating more accurate reconstruction on the tested datasets.

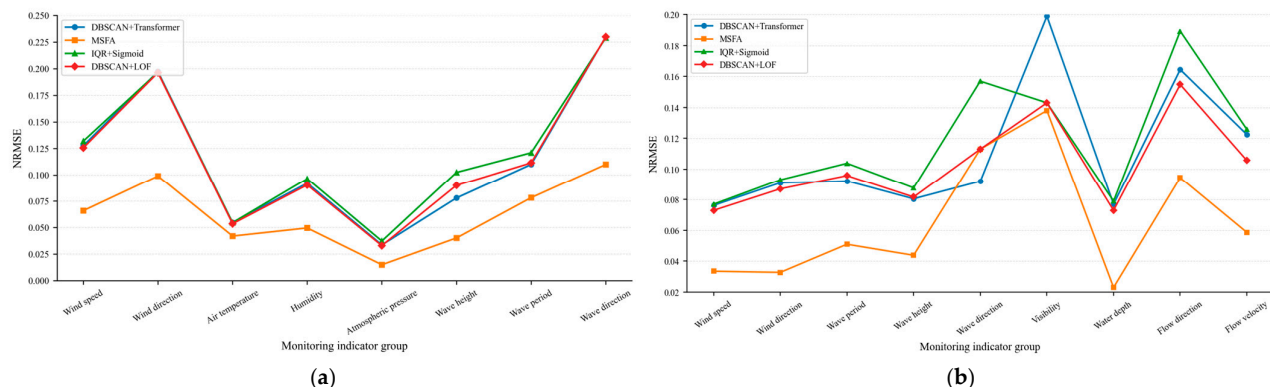


Figure 3. NRMSE comparison of different cleaning methods: (a) results on the 2020 Dongying offshore buoy observation dataset; (b) results on the self-collected nearshore dataset. The NRMSE values are averaged within each monitoring indicator category for clearer visualization, and lower values indicate higher reconstruction accuracy.

3.3. Ablation Experiment

Ablation experiments were performed on the DBSCAN+IQR+hybrid repair framework to examine the effect of each component on the cleaning results. The complete algorithm contains three parts:

1. DBSCAN+IQR anomaly detection: DBSCAN captures local density-based anomalies, and IQR screens global deviations and extreme outliers.
2. Linear interpolation and moving average repair: After abnormal positions are detected, repair is carried out using linear interpolation and moving average. Linear

interpolation preserves temporal continuity, while the moving average reduces random fluctuations.

3. Hybrid estimation and iterative correction: The two repair results are adaptively fused. CRM-based Hampel-type thresholding is then used to locate residual anomalies, which are returned to the hybrid repair process for further correction.

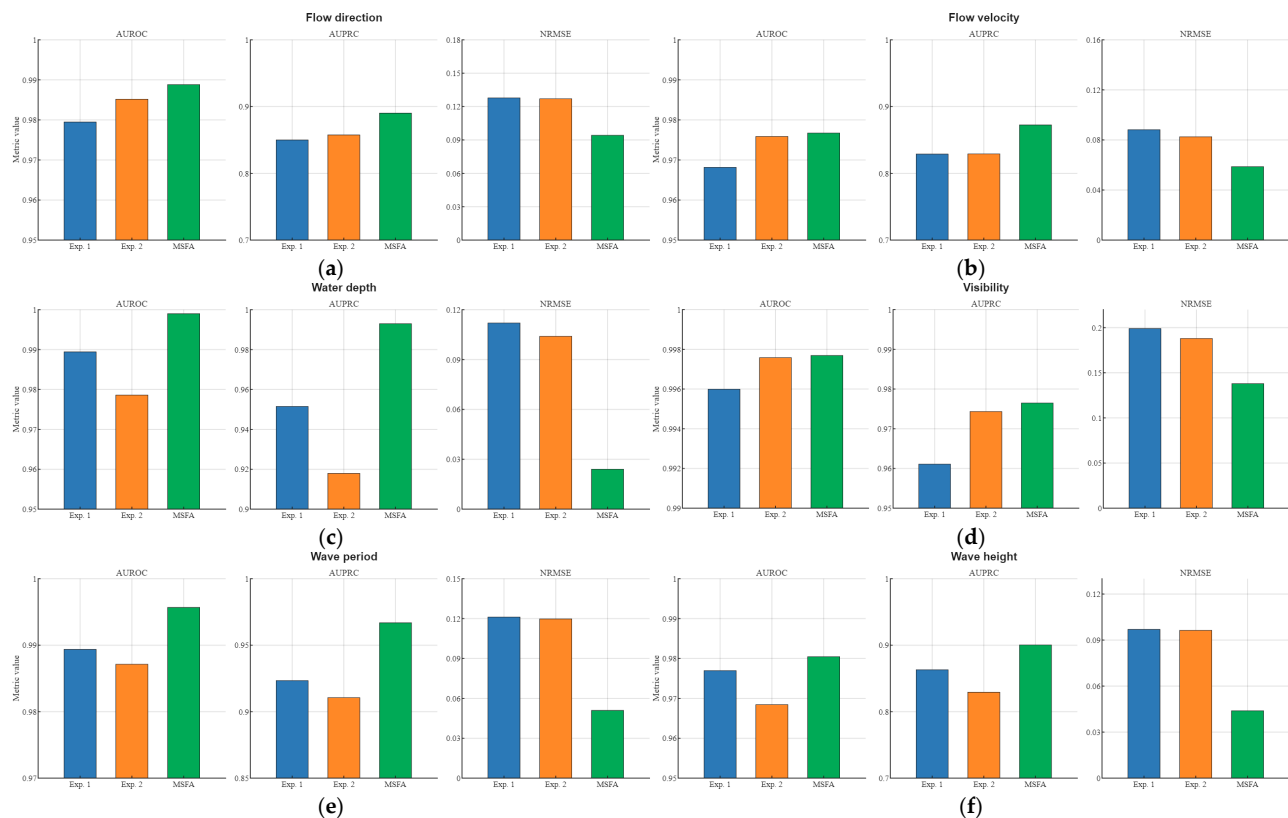
The DBSCAN+IQR detection stage was kept in all ablation settings, since it defines the invalid-position mask and the valid observation set for repair. Removing it would change both the anomaly-labeling rule and the repair target, making it impossible to isolate the effect of an individual repair component. Therefore, the ablation study focused on hybrid estimation and iterative correction. The experimental settings are as follows:

Experiment 1: The hybrid estimation module is removed. DBSCAN+IQR is retained for anomaly detection, and data cleaning depends on iterative correction.

Experiment 2: The iterative correction module is removed. DBSCAN+IQR anomaly detection and hybrid estimation repair are retained.

Experiment 3: The complete MSFA process is used, including DBSCAN+IQR anomaly detection, hybrid estimation repair, and iterative correction.

Figure 4 shows the ablation results on the self-collected nearshore dataset. Compared with the ablation variants, the complete MSFA framework obtains higher AUROC and AUPRC values, suggesting that it is more effective in anomaly detection and invalid-position identification. For NRMSE, MSFA records lower values in most cases, which points to more accurate restoration after fusion-based repair.



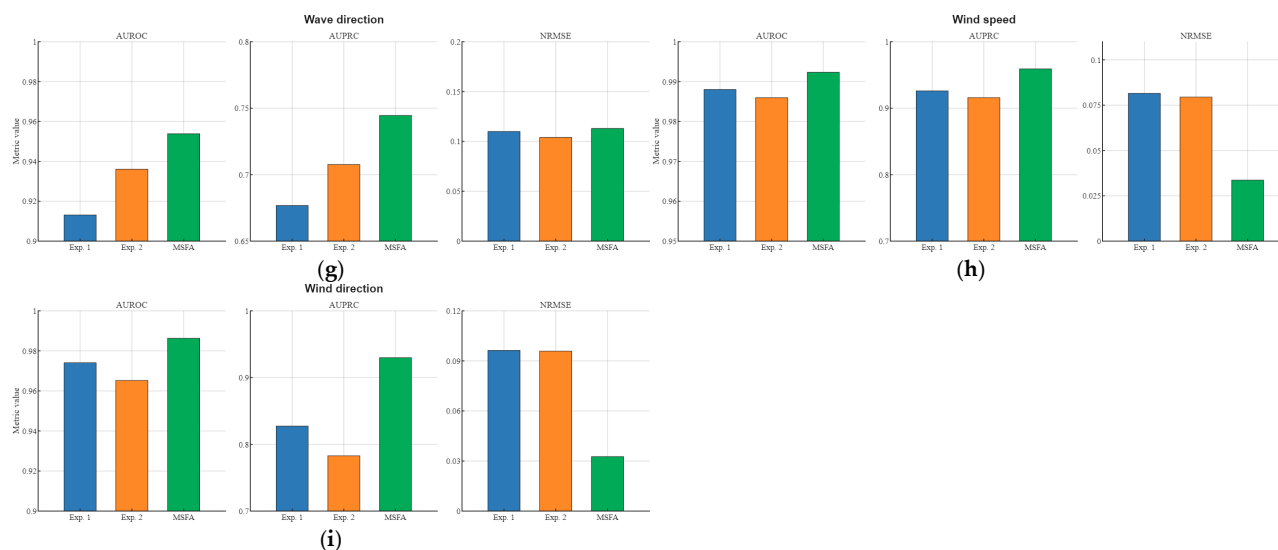


Figure 4. Ablation comparison on the self-collected nearshore dataset for (a) flow direction; (b) flow velocity; (c) water depth; (d) visibility; (e) wave period; (f) wave height; (g) wave direction; (h) wind speed; and (i) wind direction. Each subfigure compares Experiment 1, Experiment 2, and the complete MSFA framework in terms of AUROC, AUPRC, and NRMSE. Higher AUROC and AUPRC values and lower NRMSE values indicate better cleaning performance.

For water depth, visibility, wave period, wave height, and wind speed, MSFA gives better results on all three metrics. The difference is most evident in AUPRC, a metric that is sensitive to imbalanced anomaly distributions. For flow velocity, flow direction, and wave direction, the margins between MSFA and the ablation variants are smaller in some metrics. Still, the complete framework shows a steadier performance pattern.

The averaged ablation results are listed in Table 5.

Table 5. Comparison of Ablation Experiment results on the self-collected nearshore dataset.

Experiment	Avg AUROC/%	Avg AUPRC/%	Avg NRMSE
Experiment 1	97.5	86.8	0.115
Experiment 2	97.6	85.8	0.111
MSFA	98.6	91.5	0.0653

In Experiment 1, the results averaged over all involved data types were 97.5% for AUROC, 86.8% for AUPRC, and 0.1148 for NRMSE. Compared with the complete algorithm, AUROC decreased by 1.1 percentage points, AUPRC decreased by 4.7 percentage points, and NRMSE increased from 0.0653 to 0.1148, corresponding to a relative increase of 75.7%. These changes show that the hybrid estimation module plays an important role in combining different repair strategies. After this module is removed, iterative correction can still eliminate part of the residual anomalies, but the repair process depends more heavily on a single estimation mechanism. This makes the repaired sequence more likely to contain residual bias, especially when the data are strongly affected by noise.

In Experiment 2, the algorithm focused on verifying the role of the iterative correction module. We retained anomaly detection and hybrid estimation repair, and during data processing, the repaired data could be directly output without further removal of residual anomalies. After averaging the results over all involved indicators, AUROC was 97.6%, AUPRC was 85.8%, and NRMSE was 0.1108. Compared with the complete algorithm, AUROC decreased by 1.0 percentage point, AUPRC decreased by 5.6 percentage points, and NRMSE increased from 0.0653 to 0.1108, corresponding to a relative increase of 69.6%. The

results show that the iterative correction module is indispensable in handling residual anomalies. Especially after hybrid estimation, minor residual biases can be further detected by the CRM-based Hampel-type thresholding rule and then reprocessed through the repair pipeline, thereby enhancing robustness.

Figure 5 presents the ablation comparison on the 2020 Dongying offshore buoy dataset.

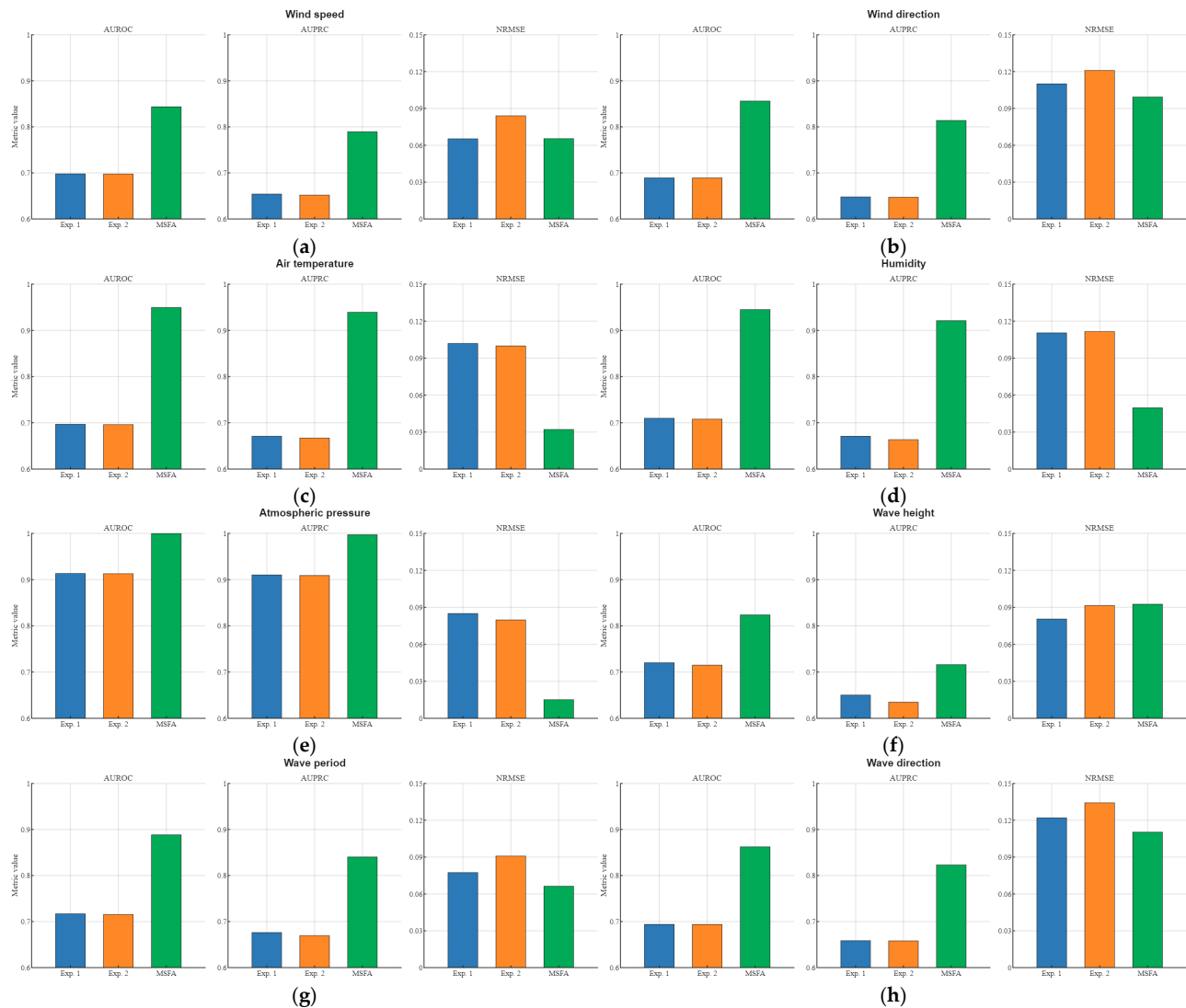


Figure 5. Ablation comparison on the 2020 Dongying offshore buoy dataset for (a) wind speed; (b) wind direction; (c) air temperature; (d) humidity; (e) atmospheric pressure; (f) wave height; (g) wave period; and (h) wave direction. Each subfigure compares Experiment 1, Experiment 2, and the complete MSFA framework in terms of AUROC, AUPRC, and NRMSE. Higher AUROC and AUPRC values and lower NRMSE values indicate better cleaning performance.

In Experiment 1, the hybrid estimation module was removed while the DBSCAN+IQR anomaly detection module and the iterative correction module were retained. This setting was used to examine whether relying on a single repair mechanism weakens the reconstruction stability on the 2020 Dongying offshore buoy dataset. As shown in Figure 5 and Table 6, removing the hybrid estimation module leads to clear performance degradation, especially in AUROC and AUPRC.

The averaged ablation results on the 2020 Dongying offshore buoy dataset are listed in Table 6.

Table 6. Comparison of Ablation Experiment Results on the 2020 Dongying buoy offshore dataset.

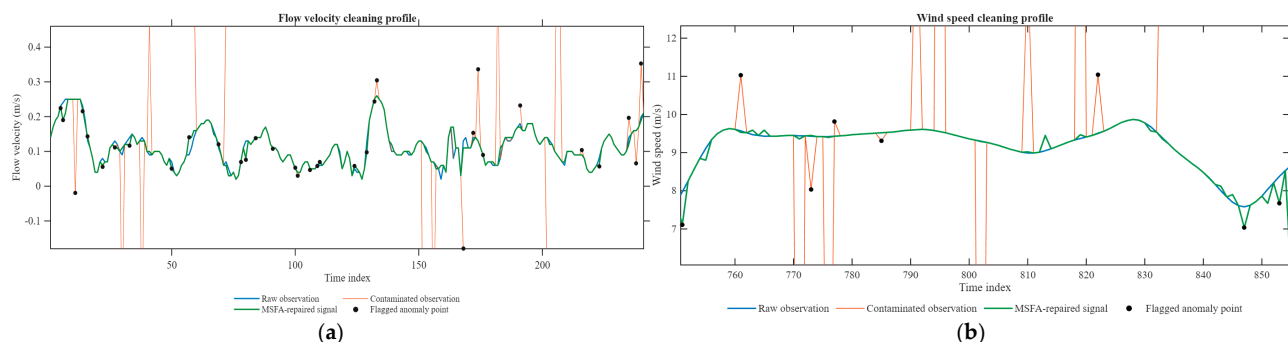
Experiment	Avg AUROC/%	Avg AUPRC/%	Avg NRMSE
Experiment 1	73.0	69.2	0.094
Experiment 2	72.8	68.8	0.102
MSFA	89.6	85.5	0.066

In Experiment 1, the results were averaged over all involved indicators. The AUROC, AUPRC, and NRMSE were 73.0%, 69.2%, and 0.094, respectively. Compared with the complete MSFA framework, AUROC decreased by 16.6%, AUPRC decreased by 16.3%, and NRMSE increased from 0.066 to 0.094, corresponding to a relative increase of 42.4%. These results indicate that the hybrid estimation module contributes substantially to both anomaly identification and reconstruction accuracy on the 2020 Dongying offshore buoy dataset.

In Experiment 2, the iterative correction module was removed, while the anomaly detection and hybrid estimation repair steps were retained. The repaired sequence was output after the first repair stage, without further correction of residual anomalies. Averaged over all indicators, AUROC, AUPRC, and NRMSE were 72.8%, 68.8%, and 0.102, respectively. Compared with the complete MSFA framework, AUROC decreased by 16.8 percentage points, AUPRC decreased by 16.7%, and NRMSE increased from 0.066 to 0.102, corresponding to a relative increase of 54.5%. This result shows that the CRM-based iterative correction module is important for reducing residual repair errors after the initial hybrid estimation stage.

In the ablation experiments, the complete MSFA framework achieved averaged AUROC, AUPRC, and NRMSE values of 98.6%, 91.5%, and 0.0653 on the self-collected near-shore dataset, and 89.6%, 85.5%, and 0.066 on the 2020 Dongying offshore buoy dataset. Compared with the two ablation variants, the complete framework shows more stable performance on both datasets. These results indicate that the hybrid estimation module and the CRM-based iterative correction module are both necessary for maintaining anomaly-identification capability and reducing reconstruction error.

Figure 6 presents representative local zoom-in examples for eight selected marine environmental indicators.



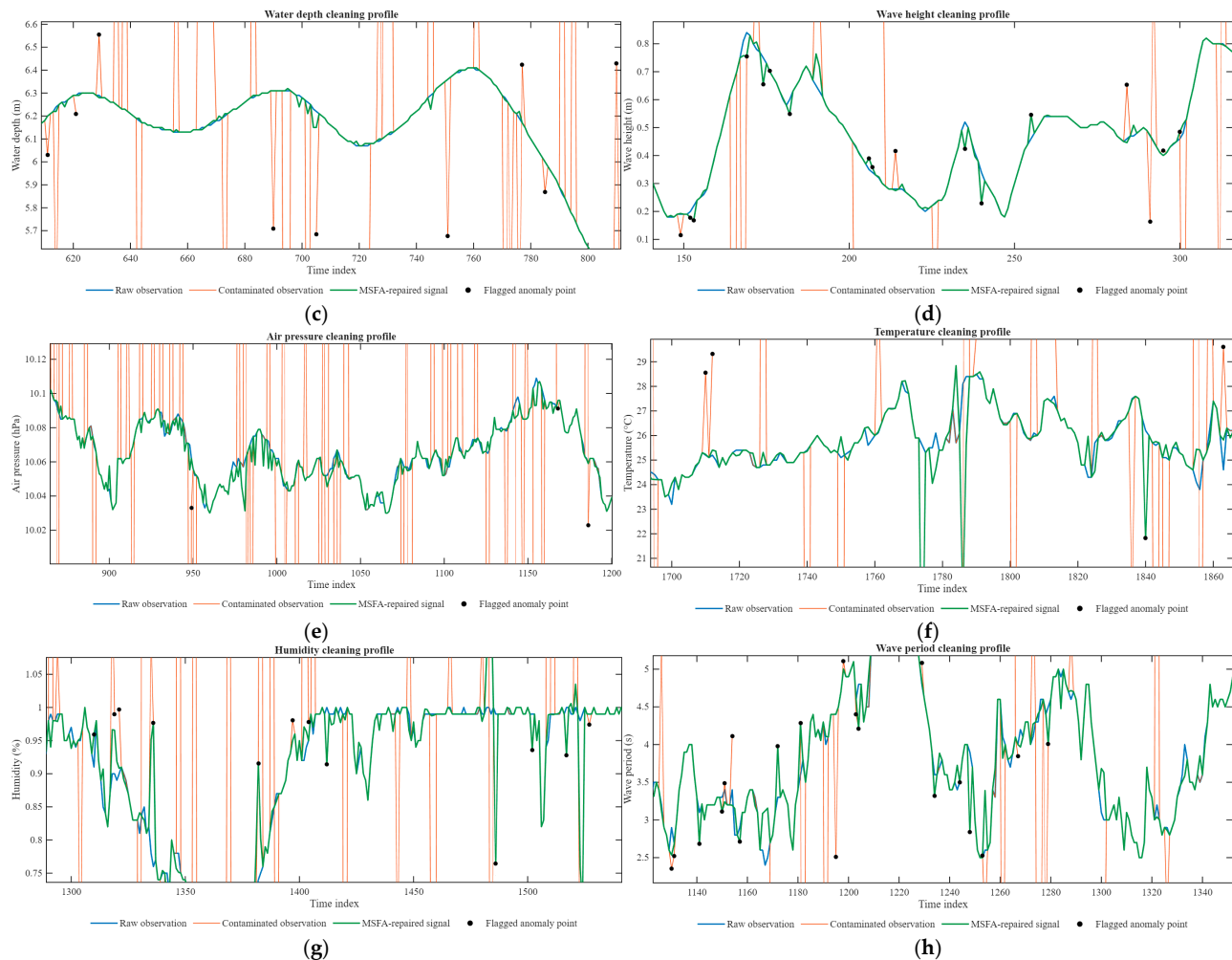


Figure 6. Representative zoom-in views of the MSFA cleaning results for selected marine environmental indicators: (a) flow velocity; (b) wind speed; (c) water depth; (d) wave height; (e) atmospheric pressure; (f) air temperature; (g) humidity; and (h) wave period. Each subfigure shows the raw observations, contaminated observations, MSFA-repaired signal, and flagged anomaly points, illustrating the local tracking and anomaly-repair performance of the proposed method.

4. Conclusions

The experimental results show that MSFA achieves competitive cleaning performance on both the 2020 Dongying offshore buoy dataset and the self-collected nearshore dataset. Compared with DBSCAN+LOF, DBTf, and IQR+Sigmoid, MSFA obtains higher AUROC and AUPRC values and lower NRMSE values in most evaluation settings, indicating improved anomaly identification and reconstruction accuracy for noisy marine monitoring data.

The results further indicate that the effectiveness of MSFA comes from the coordinated interaction among normalized anomaly detection, context-aware hybrid repair, and CRM-based residual correction. The ablation results show that removing either the hybrid estimation module or the iterative correction module leads to weaker cleaning performance, particularly in reconstruction accuracy.

The repaired data can provide more reliable support for sea-state assessment, buoy monitoring quality control, hydrological trend analysis, and downstream forecasting or early-warning tasks. In future work, the framework will be further tested on marine monitoring datasets from different sea areas and sensor platforms. Adaptive parameter selec-

tion under different sampling frequencies and contamination patterns will also be investigated. In addition, repeated random-seed contamination experiments will be conducted to report mean and standard deviation under different contamination ratios, missing-block lengths, and noise amplitudes.

Author Contributions: Conceptualization, K.C. and Q.L.; methodology, K.C., R.C. and C.J.; software, R.C.; validation, L.M.; formal analysis, R.C.; investigation, C.J.; resources, C.J.; data curation, L.M.; writing—original draft preparation, K.C., R.C. and L.M.; writing—review and editing, Q.L.; visualization, R.C.; supervision, Q.L.; project administration, Q.L.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2023YFB2603801), and the Open Research Project of State Key Laboratory of Maritime Technology and Safety (26-3-2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: One dataset presented in this study are openly available from <https://mds.nmdis.org.cn> (accessed on 1 May 2026). The other one dataset presented in this article are not readily available because the dataset is a self-collected nearshore dataset from a National Key Research and Development Program of China. Since the project is still ongoing, the data is not allowed to disclose at the moment.

Acknowledgments: The authors thank the National Science and Technology Infrastructure—China National Marine Data Center (<https://mds.nmdis.org.cn>, accessed on 1 May 2026) for providing data support. During the preparation of this manuscript, the authors used ChatGPT (GPT-5.5 Thinking, OpenAI) for the purpose of refining the language. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Xiang, L.; Yang, X.; Hu, A.; Su, H.; Wang, P. Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks. *Appl. Energy* **2022**, *305*, 117925. <https://doi.org/10.1016/j.apenergy.2021.117925>.
2. Zhang, F.; Zhang, X.; Xu, Z.; Dong, K.; Li, Z.; Liu, Y. Cleaning of abnormal wind speed power data based on quartile RANSAC regression. *Energies* **2024**, *17*, 5697. <https://doi.org/10.3390/en17225697>.
3. Martin Nascimento, G.F.; Wurtz, F.; Kuo-Peng, P.; Delinchant, B.; Jhoe Batistela, N. Outlier detection in buildings' power consumption data using forecast error. *Energies* **2021**, *14*, 8325. <https://doi.org/10.3390/en14248325>.
4. Morrison, R.; Liu, X.; Lin, Z. Anomaly detection in wind turbine SCADA data for power curve cleaning. *Renew. Energy* **2022**, *184*, 473–486. <https://doi.org/10.1016/j.renene.2021.11.118>.
5. Yue, X.; Wang, C.; Wang, Y.; Liu, H.; Zhang, Y. Gas flow meter anomaly data detection based on fused LOF-DBSCAN algorithm. In *Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition, Shanghai, China, 17–19 November 2022*; ACM: New York, NY, USA, 2022; pp. 503–508. <https://doi.org/10.1145/3581807.3581881>.
6. Hou, G.; Wang, J.; Fan, Y. Wind power forecasting method of large-scale wind turbine clusters based on DBSCAN clustering and an enhanced hunter-prey optimization algorithm. *Energy Convers. Manag.* **2024**, *307*, 118341. <https://doi.org/10.1016/j.enconman.2024.118341>.
7. Schmidl, S.; Wenig, P.; Papenbrock, T. Anomaly detection in time series: A comprehensive evaluation. *Proc. VLDB Endow.* **2022**, *15*, 1779–1797. <https://doi.org/10.14778/3538598.3538602>.
8. Lai, G.; Liao, L.; Zhang, L.; Li, T. Wind speed power data cleaning method for wind turbines based on fan characteristics and isolated forests. *J. Phys. Conf. Ser.* **2023**, *2427*, 012001. <https://doi.org/10.1088/1742-6596/2427/1/012001>.
9. Zamanzadeh Darban, Z.; Webb, G.I.; Pan, S.; Aggarwal, C.C.; Salehi, M. Deep learning for time series anomaly detection: A survey. *ACM Comput. Surv.* **2024**, *57*, 15. <https://doi.org/10.1145/3691338>.

10. Blázquez-García, A.; Conde, A.; Mori, U.; Lozano, J.A. A review on outlier/anomaly detection in time series data. *ACM Comput. Surv.* **2021**, *54*, 56. <https://doi.org/10.1145/3444690>.
11. Mecca, G.; Papotti, P.; Santoro, D.; Veltri, E. BUNNI: Learning repair actions in rule-driven data cleaning. *ACM J. Data Inf. Qual.* **2024**, *16*, 12. <https://doi.org/10.1145/3665930>.
12. Jäger, S.; Biessmann, F. From Data Imputation to Data Cleaning—Automated Cleaning of Tabular Data Improves Downstream Predictive Performance. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 2–4 May 2024; Proceedings of Machine Learning Research; PMLR: 2024; Volume 238, pp. 3394–3402. Available online: <https://proceedings.mlr.press/v238/jager24a.html> (accessed on 13 July 2026).
13. Du, W.; Côté, D.; Liu, Y. SAITS: Self-attention-based imputation for time series. *Expert Syst. Appl.* **2023**, *219*, 119619. <https://doi.org/10.1016/j.eswa.2023.119619>.
14. Nie, T.; Qin, G.; Ma, W.; Mei, Y.; Sun, J. ImputeFormer: Low-rankness-induced transformers for generalizable spatiotemporal imputation. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; ACM: New York, NY, USA, 2024; pp. 2260–2271. <https://doi.org/10.1145/3637528.3671751>.
15. Wang, X.; Zhang, H.; Wang, P.; Zhang, Y.; Wang, B.; Zhou, Z.; Wang, Y. An observed-value consistent diffusion model for imputing missing values in multivariate time series. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; ACM: New York, NY, USA, 2023; pp. 2409–2418. <https://doi.org/10.1145/3580305.3599257>.
16. Wang, W.; Yang, S.; Yang, Y. An improved data-efficiency algorithm based on combining isolation forest and mean shift for anomaly data filtering in wind power curve. *Energies* **2022**, *15*, 4918. <https://doi.org/10.3390/en15134918>.
17. Long, H.; Xu, S.; Gu, W. An abnormal wind turbine data cleaning algorithm based on color space conversion and image feature detection. *Appl. Energy* **2022**, *311*, 118594. <https://doi.org/10.1016/j.apenergy.2022.118594>.
18. Tuli, S.; Casale, G.; Jennings, N.R. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proc. VLDB Endow.* **2022**, *15*, 1201–1214. <https://doi.org/10.14778/3514061.3514067>.
19. Xu, J.; Wu, H.; Wang, J.; Long, M. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022. Available online: https://openreview.net/forum?id=LzQQ89U1qm_ (accessed on 13 July 2026).
20. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=ju_Uqw384Oq (accessed on 13 July 2026).
21. Wang, S.; Zhang, Z.; Wang, P.; Tian, Y. Short-term wind power prediction based on multidimensional data cleaning and feature reconfiguration. *Appl. Energy* **2021**, *292*, 116851. <https://doi.org/10.1016/j.apenergy.2021.116851>.
22. Liang, G.; Su, Y.; Chen, F.; Long, H.; Song, Z.; Gan, Y. Abnormal data cleaning for wind turbines by image segmentation based on active shape model and class uncertainty. *Renew. Energy* **2023**, *216*, 118965. <https://doi.org/10.1016/j.renene.2023.118965>.
23. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in time series: A survey. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 6778–6786. <https://doi.org/10.24963/ijcai.2023/759>.
24. Zhou, J.; Li, J.; Zheng, G.; Wang, X.; Zhou, C. MTSCI: A conditional diffusion model for multivariate time series consistent imputation. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; ACM: New York, NY, USA, 2024. <https://doi.org/10.1145/3627673.3679532>.
25. Oh, E.; Kim, T.; Ji, Y.; Khyalia, S. STING: Self-Attention-Based Time-Series Imputation Networks Using GAN. *arXiv* **2022**, arXiv:2209.10801. <https://doi.org/10.48550/arXiv.2209.10801>.
26. Toribio, M.R.-H.; Menéndez, A.G.; Toribio, S.R.-H.; García, I.Á. A novel approach to speed up Hampel filter for outlier detection. *Sensors* **2025**, *25*, 3319. <https://doi.org/10.3390/s25113319>.
27. Lei, L.; Wu, B.; Fang, X.; Chen, L.; Wu, H.; Liu, W. A dynamic anomaly detection method of building energy consumption based on data mining technology. *Energy* **2023**, *263*, 125575. <https://doi.org/10.1016/j.energy.2022.125575>.
28. Pehlivan, H.; Karabiber, F.; İnkaya, T. A novel outlier detection method based on Bayesian change point analysis for time series data. *EURASIP J. Adv. Signal Process.* **2024**, *2024*, 44. <https://doi.org/10.1186/s13634-023-01097-w>.
29. Boniol, P.; Paparrizos, J.; Palpanas, T. New Trends in Time-Series Anomaly Detection. In Proceedings of the 26th International Conference on Extending Database Technology, Ioannina, Greece, 28–31 March 2023; pp. 847–850. <https://doi.org/10.48786/edbt.2023.80>.

30. Mejri, N.; Lopez-Fuentes, L.; Roy, K.; Chernakov, P.; Ghorbel, E.; Aouada, D. Unsupervised Anomaly Detection in Time-Series: An Extensive Evaluation and Analysis of State-of-the-Art Methods. *arXiv* **2022**, arXiv:2212.03637. <https://doi.org/10.1016/j.eswa.2024.124922>.
31. Wang, F.; Jiang, M.; Qian, Y.; Zhang, Y.; Huang, Z.; Shi, Y. A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions. *Sensors* **2025**, *25*, 190. <https://doi.org/10.3390/s25010190>.

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.