frontiers | Research Topics

Data-driven ocean environmental perception with its applications

Edited by
Jianchuan Yin, Yu Jiang, Phoebe Koundouri
and Hongde Qin

Published in
Frontiers in Marine Science

ISSN 1664-8714
ISBN 978-2-8325-6161-4
DOI 10.3389/978-2-8325-6161-4

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Data-driven ocean environmental perception with its applications

Topic editors
Jianchuan Yin — Guangdong Ocean University, China
Yu Jiang — Jilin University, China
Phoebe Koundouri — Athens University of Economics and Business, Greece
Hongde Qin — Harbin Engineering University, China

Citation
Yin, J., Jiang, Y., Koundouri, P., Qin, H., eds. (2025). *Data-driven ocean environmental perception with its applications*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6161-4

Table of contents

04 Robust sensor selection based on maximum correntropy criterion for ocean data reconstruction
Qiannan Zhang, Huafeng Wu, Li'nian Liang, Xiaojun Mei and Jiangfeng Xian

22 ISSA optimized spatiotemporal prediction model of dissolved oxygen for marine ranching integrating DAM and Bi-GRU
Wenjing Liu, Ji Wang, Zhenhua Li and Qingjie Lu

43 Bio-optical variability of particulate matter in the Southern Ocean
Juan Li, David Antoine and Yannick Huot

56 Optimizing data-driven arctic marine forecasting: a comparative analysis of MariNet, FourCastNet, and PhyDNet
Aleksei V. Buinyi, Dias A. Irishev, Edvard E. Nikulin, Aleksandr A. Evdokimov, Polina G. Ilyushina and Natalia A. Sukhikh

66 Modeling Atlantic herring distribution in the Northeast Atlantic for informed decision-making towards sustainable fisheries
Ward Standaert, Rutendo Musimwa, Martha Stevens, Jesus Alonso Guerra, Carlota Muñiz, Elisabeth Debusschere, Steven Pint and Gert Everaert

82 Research on adaptive dimming management methods for intelligent lighting systems in port traffic based on ocean weather perception
Haoyu Jiang, Xiaolong Zhao, Zeguo Zhang and Jiacheng Ji

101 Significant wave height prediction in monsoon regions based on the VMD-CNN-BiLSTM model
Wengeng Shen, Zongquan Ying, Yiming Zhao and Xuegang Wang

118 Risk performance analysis model of escort operation in Arctic waters via an integrated FRAM and Bayesian network
Zhuang Li, Xiaoming Zhu, Shiguan Liao, Kaixian Gao and Shenping Hu

133 Path planning for unmanned surface vehicles in anchorage areas based on the risk-aware path optimization algorithm
Hongbo Wang, Shuaiwei Mao, Xiaoguang Mou, Jinfeng Zhang and Ronghui Li

146 Improved deep learning method and high-resolution reanalysis model-based intelligent marine navigation
Zeguo Zhang, Liang Cao and Jianchuan Yin

TYPE Original Research
PUBLISHED 04 October 2024
DOI 10.3389/fmars.2024.1467519

Check for updates

OPEN ACCESS

EDITED BY
Jianchuan Yin,
Guangdong Ocean University, China

REVIEWED BY
Hongchu Yu,
Wuhan University of Technology, China
Hailong Feng,
China Maritime Service Center, China

*CORRESPONDENCE
Huafeng Wu
✉ hfwu@shmtu.edu.cn
Jiangfeng Xian
✉ jfxian@shmtu.edu.cn

RECEIVED 20 July 2024
ACCEPTED 10 September 2024
PUBLISHED 04 October 2024

CITATION
Zhang Q, Wu H, Liang L, Mei X and Xian J (2024) Robust sensor selection based on maximum correntropy criterion for ocean data reconstruction.
Front. Mar. Sci. 11:1467519.
doi: 10.3389/fmars.2024.1467519

Robust sensor selection based on maximum correntropy criterion for ocean data reconstruction

Qiannan Zhang[1], Huafeng Wu[1*], Li'nian Liang[1], Xiaojun Mei[1] and Jiangfeng Xian[2*]

[1]Merchant Marine College, Shanghai Maritime University, Shanghai, China, [2]Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai, China

Selecting an optimal subset of sensors that can accurately reconstruct the full state of the ocean can reduce the cost of the monitoring system and improve monitoring efficiency. Typically, in data-driven sensor selection processes, the use of Euclidean distance to evaluate reconstruction error is susceptible to non-Gaussian noise and outliers present in ocean data. This paper proposes a Robust Sensor Selection (RSS) evaluation model based on the Maximum Correntropy Criterion (MCC) through subspace learning, enabling the selection of robust sensor measurement subsets and comprehensive data reconstruction. To more accurately quantify the impact of varying noise magnitudes, noise weights were incorporated into the model's objective function. Additionally, the local geometric structure of data samples is utilized to further enhance reconstruction accuracy through the selected sensors. Subsequently, the MCC_RSS algorithm is proposed, which employs the Block Coordinate Update (BCU) method to achieve the optimal solution for the proposed model. Experiments conducted using ocean temperature and salinity datasets validate the proposed MCC_RSS algorithm. The results demonstrate that the sensor selection method proposed in this paper exhibits strong robustness, outperforming comparative methods under varying proportions of outliers and non-Gaussian noise.

KEYWORDS

sensor selection$_1$, Maximum Correntropy Criterion (MCC)$_2$, robust$_3$, data reconstruction$_4$, ocean$_5$, subspace learning$_6$

1 Introduction

In the field of oceanography, optimizing sensor selection is a critical area of research. Effective sensor selection can directly impact sensor deployment and enhance our understanding of the oceanic physical parameters. By tailoring sensor selection to meet specific requirements, various objectives can be achieved, including cost reduction (Emily et al., 2020; Saito et al., 2023), energy efficiency (Ghosh et al., 2021), conservation of communication resource (Yang et al., 2015), assistance in localization (Mei et al., 2024), improved field reconstructions (Santini and Colesanti, 2009; Zhang et al., 2018; Nguyen et al., 2021; Santos et al., 2023) and enhanced state predictions (Saucan and Win, 2020; Patan et al., 2022), among others.

The sensor selection problem involves selecting the optimal p positions from n candidate positions to achieve the desired outcomes, a task recognized as NP-hard (Chamon et al., 2021). This implies that an exhaustive search would need to traverse up to $n!/[p!(n-p)!]$ combinations, which is nearly impossible when the number of candidate positions is large in ocean monitoring. General solutions to the sensor selection problem include the following: convex optimization (Joshi and Boyd, 2009), statistical methods (Chepuri and Leus, 2015; Lin et al., 2019; Yamada et al., 2021), heuristic methods (Khokhlov et al., 2019; Zhao et al., 2021; Meray et al., 2023), information theory (Krause et al., 2008; Prakash and Bhushan, 2023), dimensionality reduction (Yildirim et al., 2009; Manohar et al., 2018; Jayaraman et al., 2019), machine learning-based clustering (Kalinić et al., 2022), among others.

Data-driven sensor selection provides an excellent optimization solution for selecting sensors from a large pool of candidate locations in ocean monitoring. By analyzing the intrinsic characteristics of known data, it identifies the most critical geographical locations for reconstructing the entire physical field, without requiring precise modeling or complex statistical analysis of the monitoring object or requirements. However, these methods typically evaluate the reconstruction effect based on the Euclidean distance between the original and reconstructed data, which is highly sensitive to non-Gaussian noise and outliers. This sensitivity is particularly problematic in ocean monitoring, where specific sudden events (such as tsunamis causing sensor failure, communication interruptions, or data loss) can significantly impact data quality. Consequently, noise in the data can severely affect the effectiveness of sensor deployment. Moreover, greedy algorithms such as Proper Orthogonal Decomposition (POD) and QR decomposition cannot guarantee globally optimal results.

Building on the work of Zhou et al. (2019) on Maximum Correntropy Criterion-based sparse subspace learning for feature selection, we propose a novel sparse sensor selection method. This method quantifies the similarity between the original data and the reconstructed data using correntropy, thereby effectively mitigating the impact of outliers on the feature selection process. Additionally, the subspace learning approach allows for the simultaneous updating of the feature selection matrix and the reconstruction matrix, enhancing the accuracy of the reconstruction.

This work employs subspace learning based on the Maximum Correntropy Criterion (MCC) for sensor selection. The main contributions of this study are as follows:

- The application of the MCC for evaluating reconstruction error supersedes the traditional Euclidean distance, thereby enhancing the stability of results in the presence of non-Gaussian noise and outliers. Additionally, noise weight is employed to measure the MCC, and the higher entropy of noise weight is utilized to achieve a noise distribution that more accurately represents the distribution of real system variables.
- In order to further improve reconstruction accuracy, a term that preserves the local geometric structure between samples was incorporated into the objective function to minimize the similarity between the selected measurements.
- The adoption of subspace learning allows for the simultaneous determination of both the sensor selection matrix and the mapping for data reconstruction from low-dimensional measurements to high-dimensional measurements corresponding to this selection matrix.
- Experiments conducted on ocean temperature and salinity datasets demonstrate that the proposed sparse sensor selection method exhibits robust performance.

Subsequently, we review the related work in Section 2. Section 3 introduces the sparse sensor deployment model based on MCC, with the solution algorithm detailed in Section 4. The proposed algorithm is validated using ocean temperature and salinity datasets in Section 5. Finally, Section 6 provides a summary and discussion.

2 Related works

The Euclidean distance is frequently utilized as a criterion for measuring the reconstruction error in sensor selection problems. Specifically, this involves using the Frobenius norm of the difference between the original data and the reconstructed data, as follows:

$$C = \arg\min_{C} \| X - \hat{X} \|_F \tag{1}$$

where $X \in \mathbb{R}^{n\times m}$ represents the original data, $\hat{X} \in \mathbb{R}^{n\times m}$ represents the reconstructed data, $C \in \mathbb{R}^{p\times n}$ represents sensor selection matrix, n represents the number of all candidate locations for sensor selection, m represents the number of samples and p represents the number of sensors to be selected. Typically, once the sensor selection matrix C is established, the sensor's measurement data can be acquired, which can be expressed as: $Y = CX$. By designing an appropriate mapping based on the measurement data Y, the reconstruction data $\hat{X}$ can be obtained.

There is extensive research on data reconstruction aimed at determining the mapping from measurement data to original data. Examples include fluid reconstruction based on sparse representation (Callaham et al., 2019; Xue et al., 2019) and autoencoder networks (Erichson et al., 2020; Sahba et al., 2022). In these studies, the subset of locations is typically selected in a random manner. Some research focuses on mapping the original fluid data to low-dimensional features using deep neural networks (Özbay and Laizet, 2022; Zhang et al., 2023). These features reside in a subspace of the high-dimensional space and are not directly related to the sensor positions. Other research employs sensor selection by designing sensor positions according to specific partition rules, such as Voronoi tessellation (Fukami et al., 2021) or predetermined positions in a divided grid (Model and Zibulevsky, 2006), among others.

Algorithms for sensor selection and dimension reduction, such as the POD (Jayaraman et al., 2019) and QR decomposition (Manohar et al., 2018; Zhang et al., 2023), primarily map high-dimensional matrices to low-dimensional subspaces to obtain low-dimensional location indices. However, POD relies on a base matrix derived from Singular Value Decomposition (SVD) for data reconstruction, with sensors typically selected at random. In contrast, QR decomposition generally employs a greedy approach to identify low-dimensional location indices with the highest energy (e.g., spectral norm) to determine the measurement subset that can best reconstruct the original data. While a greedy approach focuses on the benefit of each individual step in the solution process, it often neglects the impact on the overall solution.

There are also sensor selection methods for reconstruction that integrate both dimension reduction and data reconstruction, such as data-driven sparse sensing (Jayaraman and Mamun, 2020), clustering for sensor select and regressive reconstruction in (Dubois et al., 2022) and compress sensing (Carmi and Gurfil, 2013; Joneidi et al., 2020). According to the research by Peherstorfer et al (Peherstorfer et al., 2020), the presence of noise in the data exacerbates the impact of the noise on the results as the number of selected locations increases. Furthermore, since these methods utilize Euclidean distance for similarity measurement, they are particularly susceptible to non-Gaussian noise or outliers in real-world marine monitoring scenarios.

To minimize the impact of noise, (Zhou et al. (2019) proposed a sparse subspace learning method based on MCC, which simultaneously searches for the feature selection matrix and the mapping. However, this method is primarily used for feature selection in image and sound data. Generally, MCC, grounded in the concept of correntropy from information theory, is adept at capturing nonlinear relationships and complex structures within data. This endows MCC with a significant advantage in handling complex datasets, enabling it to more accurately reflect the true characteristics of the data. By maximizing correntropy, MCC can effectively mitigate the influence of outliers on the model. Additionally, MCC does not depend on the specific distribution form of noise, thereby exhibiting excellent performance when dealing with non-Gaussian noise. Conversely, Guo et al. (Guo and Lin (2018) minimize the impact of noise by identifying the noise indicator of the maximum entropy distribution during low-rank matrix decomposition. These studies suggest that MCC and entropy-based noise indicators can provide a feasible solution for the problem of robust sparse sensor selection.

3 Model of robust sensor selection based on MCC

This section introduces a model for robust sensor selection. Initially, an error measure based on the Maximum Correntropy Criterion (MCC) is proposed to enhance the robustness of sensor selection. Subsequently, an objective function for the robust sensor selection model is formulated utilizing this error measure. To further augment the robustness of the model, noise indicators are established, which impose additional constraints on the objective function through the noise matrix.

3.1 Reconstruction error based on MCC

In Information Theoretic Learning (ITL), correntropy has proven effective in mitigating the impact of non-Gaussian noise and outliers (Liu et al., 2007). The MCC has demonstrated its efficacy in robust compressive sensing reconstruction (He et al., 2019). Consequently, within this context, MCC is utilized as a standard to evaluate the similarity between the original data and the reconstructed data for robust sensor selection, as follows:

For any two random variables A and B, the correntropy is defined as:

$$V(A,B) = E[\kappa(A,B)] \tag{2}$$

where $E[\cdot]$ represents the expectation operator, $\kappa(\cdot,\cdot)$ represents kernel function which map the original variables to the Hilbert functional space.

Generally, $\kappa(\cdot,\cdot)$ is adopted as a Gaussian kernel function. For two given discrete variables a_i and b_i, then:

$$\kappa(a_i,b_i) = \kappa_\sigma(a_i - b_i) = \exp\left(-\frac{(a_i - b_i)^2}{2\sigma^2}\right) \tag{3}$$

where σ represents kernel bandwidth.

The similarity between variables a_i and b_i can be measured using the correntropy estimator as follows:

$$\tilde{V}_\sigma(A,B) = \frac{1}{m}\sum_{i=1}^{m} \kappa_\sigma(a_i - b_i) \tag{4}$$

where m represents sample number.

MCC aims to find the maximum correntropy of the difference between two variables, which is utilized to estimate probability distributions with maximum correntropy under given constraints.

According to the principles of linear subspace learning, once the data representation in a low-dimensional subspace is obtained via the feature selection matrix, the data can be reconstructed using a transformation matrix that maps the low-dimensional data back to the high-dimensional space. Consequently, the reconstruction of data from the low-dimensional measurements Y to high-dimensional estimated data $\hat{X}$ is defined through the transformation matrix $T \in \mathbb{R}^{n\times p}$, as follows:

$$\hat{X} = TY = TCX \tag{5}$$

According to Equations 1, 4, 5, the error measure of data reconstruction based on MCC is defined as follows:

$$J_{MCC} = \sum_{i=1}^{m} \exp\left(\frac{-\| s_i^T - TCs_i^T \|_2}{2\sigma^2}\right) \tag{6}$$

where, s_i represents the i-th sample of original data X, TCs_i^T represents the i-th sample of reconstructed data $\hat{X}$. $(\cdot)^T$ denotes the transpose of the matrix.

3.2 Model of robust sparse sensor selection

Building on the aforementioned content, the robust sensor selection model employing MCC is formulated to determine an optimal selection matrix C, such that the correntropy error specified in Equation 6 is maximized, as follows:

$$\begin{aligned}\hat{C} &= \arg\max_{C} \frac{1}{2}\sum_{i=1}^{m} \exp\left(\frac{-\| s_i^T - TCs_i^T \|_2}{2\sigma^2}\right)\\ s.t.\quad & C \in \{0,1\}^{p\times n}, C\mathbf{1}_{n\times 1} = \mathbf{1}_{p\times 1},\\ & \| C\mathbf{1}_{p\times 1} \|_0 = p.\end{aligned} \tag{7}$$

For ease of solution, as suggested in reference (Zhou et al., 2016), the binary variables of C in the constraint conditions are relaxed to a continuous form. Additionally, to further enhance reconstruction accuracy, the local geometric structure preservation term, as utilized in feature selection (Liu et al., 2014), is incorporated. Based on the representation form of the reconstructed data in Equation 5, this local geometric structure preservation term is transformed into: $Tr(CXLX^TC^T)$. Then:

$$\begin{aligned}\hat{C} &= \arg\max_{C} \frac{1}{2}\sum_{i=1}^{m} \exp\left(\frac{-\| s_i^T - TCs_i^T \|_2}{2\sigma^2}\right) - \frac{\mu}{2} Tr(CXLX^TC^T)\\ s.t.\ & C \in \mathbb{R}_+^{p\times n}\end{aligned} \tag{8}$$

where μ represents a predefined coefficient, $L \in \mathbb{R}^{m\times m}$ refers to the graph Laplacian matrix that captures the local geometric structure of all data samples. To better measure the relationship between samples, the Linear Preserve Projection (LPP) method is employed to obtain the L matrix, as described in (Liu et al., 2014). Additionally, C is a non-negative matrix.

Simultaneously, to constrain the sparsity of the solution, a sparse regularization term for the selection matrix C is incorporated:

$$\begin{aligned}\hat{C} &= \arg\max_{C} \frac{1}{2}\sum_{i=1}^{m} \exp\left(\frac{-\| s_i^T - TCs_i^T \|_2}{2\sigma^2}\right) - \frac{\mu}{2} Tr(CXLX^TC^T) - \alpha \| C \|_{2,1}\\ s.t.\ & C \in \mathbb{R}_+^{p\times n}\end{aligned} \tag{9}$$

Here, the $\ell_{2,1}$-norm of the selection matrix C is introduced to control its column sparsity and prevent the selection of too many redundant sensor positions. α represents the sparse coefficient of selection matrix C.

3.3 Model enhancement based on noise weight

Moreover, the noise weight matrix has been demonstrated to effectively enhance the robustness of outlier estimation during the process of low-rank matrix decomposition (Guo and Lin, 2018). The sensor selection problem can be conceptualized as a full state reconstruction leveraging the sparse characteristics of the low-rank matrix. Consequently, we estimate noise using both severe noise and smaller noise weight matrices, respectively, to further mitigate the impact of non-Gaussian noise and outliers on the sensor selection process, as well as the model and measurement noises. Under this condition, the smaller noise weight matrix is incorporated into the error evaluation based on MCC as follows:

$$J_{MCC} = \sum_{i=1}^{m} \exp\left(\frac{-\| W_i \odot (s_i^T - TCs_i^T) \|_2}{2\sigma^2}\right) \tag{10}$$

where W_i represents the i-th columns of the smaller noise weight matrix $W \in \mathbb{R}^{n\times m}$, $\odot$ represents Hadamard product operator.

Simultaneously, to mitigate the impact of severe noise (such as outliers) on the results, we have incorporated a regularization term $\| \bar{W} \|_1$ for the severe noise matrix $\bar{W} \in \mathbb{R}^{n\times m}$, ensuring its sparsity. Furthermore, according to the maximum entropy theory, a higher entropy of the noise distribution better represents the actual distribution of system variables. Consequently, we have included an entropy term for both severe and minor noise to align the results more closely with the true distribution. Therefore, Equation 9 is modified as follows:

$$\begin{aligned}C \leftarrow & \arg\max_{C} \frac{1}{2}\sum_{i=1}^{m} \exp\left(\frac{-\| \sqrt{W_i} \odot (s_i^T - TCs_i^T) \|_2}{2\sigma^2}\right) - \frac{\mu}{2} Tr(CXLX^TC^T)\\ & - \alpha \| C \|_{2,1}\\ & - \beta \| \bar{W} \|_1 - \gamma \sum_{i,j}(w_{ij}\log w_{ij} + \bar{w}_{ij}\log \bar{w}_{ij})\\ & \text{s.t.}\quad W + \bar{W} = \mathbf{1},\quad W\ and\quad \bar{W} \in [0,1]^{n\times m}\\ & C \in \mathbb{R}_+^{p\times n}\end{aligned} \tag{11}$$

where $w_{ij} \in W$ and $\bar{w}_{ij} \in \bar{W}$, β represents coefficient of regularization term $\| \bar{W} \|_1$ and γ represents coefficient of entropy of noise. Equation 11 presents the final model for our robust sensor selection.

4 Algorithm for robust sensor selection

To address the Gaussian kernel function in the model, the half-quadratic optimization technique was employed to simplify the objective function in Equation 11. Subsequently, due to the presence of non-convex components that render direct solution challenging, the Block Coordinate Update (BCU) iterative method (Xu and Yin, 2013), is utilized to resolve the problem in Equation 11.

4.1 Reformulation via half-quadratic optimization

For the correntropy utilizing the Gaussian kernel function, the maximum value calculation through sample accumulation can be interpreted as Welch's M-estimation. Consequently, it can be approximated using half-quadratic optimization techniques. Let:

$$x = \frac{\| \sqrt{W_i} \odot (\mathbf{s}_i^T - TC\mathbf{s}_i^T) \|_2}{2\sigma^2} \tag{12}$$

According to the half-quadratic optimization (He et al., 2014), we obtain:

$$\phi(x) = \sup_{q_i}\{q_i x - \varphi(q_i)\} \tag{13}$$

where q_i represents a scalar variable, $\phi(x) = \exp(-x)$ is denoted as the kernel function satisfies the condition of finding minimum correntropy. Consequently, we obtain:

$\varphi(q_i) = q_i - q_i \ln(-q_i)$, and:

$$\exp\left(\frac{-\| \sqrt{W_i} \odot (\mathbf{s}_i^T - TC\mathbf{s}_i^T) \|_2}{2\sigma^2}\right)$$

$$= \sup_{q_i}\left\{ q_i \frac{-\| \sqrt{W_i} \odot (\mathbf{s}_i^T - TC\mathbf{s}_i^T) \|_2^2}{2\sigma^2} - \varphi(q_i) \right\} \tag{14}$$

where $i = 1, 2, \cdots, m$. In order to streamline the description process, let:

$$F_1^{MCC}(C, T, W, \mathbf{q})$$

$$= \frac{1}{2}\sum_{i=1}^{m}\left(q_i \frac{-\| \sqrt{W_i} \odot (\mathbf{s}_i^T - TC\mathbf{s}_i^T) \|_2^2}{2\sigma^2} - \varphi(q_i) \right) \tag{15}$$

Then, let:

$$F(C, T, W, \mathbf{q}) = F_1^{MCC}(C, T, W, \mathbf{q}) + \frac{\mu}{2} Tr(CXLX^T C^T) \tag{16A}$$

$$E(W) = \beta \| \bar{W} \|_1 + \gamma \sum_{i,j} (w_{ij} \log w_{ij} + \bar{w}_{ij} \log \bar{w}_{ij}) \tag{16B}$$

Consequently, the objective function of Equation 11 can be reformulated as:

$$\begin{aligned} & C \leftarrow \arg\max_{C} F(C, T, W, \mathbf{q}) - \alpha \| C \|_{2,1} - E(W) \\ & \text{s.t. } W + \bar{W} = \mathbf{1}, \quad W \textit{ and } \bar{W} \in [0, 1]^{n \times m} \\ & C \in \mathbb{R}_+^{p \times n} \end{aligned} \tag{17}$$

4.2 Iterative method by BCU

According to the BCU method described in (Xu and Yin, 2013), the objective function of Equation 17 can be optimized by sequentially updating and iterating the variables C, T, W and $\boldsymbol{q}$. During the update of one variable, the remaining three variables are held constant. The iterative process continues until the termination condition is satisfied, which occurs when the objective function reaches its maximum value and no further significant updates can be made.

Let $\hat{G}^k = \nabla_C F(\hat{C}^k, T^k, W^k, \mathbf{q}^k)$ denote the block-partial gradient of function $F(\cdot)$ at $\hat{C}^k$ during the k-th iteration. Throughout the iteration process, the variables are updated as follows:

$$C^{\mathrm{k+1}} = \arg\max_{C \in \mathbb{R}_+^{P \times N}} \langle \hat{G}^k, C - \hat{C}^k \rangle - \frac{L_C^k}{2} \| C - \hat{C}^k \|_F^2 - \alpha \| C \|_{2,1} \tag{18A}$$

$$T^{k+1} = \arg\max_{T} F_1^{MCC}(C^{k+1}, T^k, W^k, \mathbf{q}^k) \tag{18B}$$

$$W^{k+1} = \arg\max_{W} F_1^{MCC}(C^{k+1}, T^{k+1}, W^k, \mathbf{q}^k) + E(W^k) \tag{18C}$$

$$\mathbf{q}^{k+1} = \arg\max_{\mathbf{q}} F_1^{MCC}(C^{k+1}, T^{k+1}, W^{k+1}, \mathbf{q}^k) \tag{18D}$$

In our algorithm, L_C^k is defined as follows:

$$L_C^k = \| T^k \|_2^2 \| X^k \|_2^2 \| W^k \|_2 + \mu \| XLX^T \|_2 \tag{19}$$

And $L_C^k > 0$ denotes the Lipschitz constant of $\hat{G}^k$, which can be determined according to Equation 41 in the Appendix.

In Equation 18A, $\hat{C}^k$ represents an extrapolated point for the update of C:

$$\hat{C}^k = C^k + \omega_C^k (C^k - C^{k-1}) \tag{20}$$

where $\omega_C^k \geq 0$ represents the extrapolation weight as defined in the BCU method (Xu, 2015), and it is typically set as follows:

$$\omega_C^k = \min(\hat{\omega}_C^k, \delta_\omega \sqrt{L_C^{k-1} / L_C^k}) \tag{21}$$

where $\delta_\omega < 1$ and $\hat{\omega}_C^k = (t^{k-1} - 1)/t^k$, with:

$$t^k = \left(1 + \sqrt{1 + 4(t^{k-1})^2}\right)/2 \tag{22}$$

and $t^0 = 1$.

In the aforementioned iterative update process, the treatment of C differs from that of the other three variables. Specifically, C is updated using a block proximal gradient method, whereas the remaining variables are updated directly through block maximization. The primary reason for this distinction is that C is a matrix composed of binary elements (0 and 1), making it challenging to solve directly. The detail solution process for each variable is as follows:

4.2.1 Solution for sensor selection matrix

In order to facilitate the determination of sensor selection matrix C, we first derive the equivalent form of Equation 18A as follows:

$$\max_{C \in \mathbb{R}_+^{p \times n}} \frac{1}{2} \| C - \left(\hat{C}^k - \frac{\hat{G}^k}{L_C^k} \right) \|_F^2 + \frac{\alpha \| C \|_{2,1}}{L_C^k} \tag{23}$$

Let $Z = \hat{C}^k - \hat{G}^k / L_C^k$ and $\lambda = \alpha / L_C^k$. For any given column $\mathbf{c} \in C, \mathbf{z} \in Z$, by decomposing the problem in Equation 23 into n independent subproblems, each subproblem can be solved corresponding to a column of matrices C and Z, respectively, as referenced in (Zhou et al., 2016; Zhou et al., 2019) as follows:

$$\arg\min_{\mathbf{c} \geq 0} \frac{1}{2} \| \mathbf{c} - \mathbf{z} \|_2^2 + \lambda \| \mathbf{c} \|_2 \tag{24}$$

Equation 24 can be resolved by applying Theorem 1 as presented in reference (Zhou et al., 2016), as follows:

Theorem 1 (Zhou et al., 2016). Given $\mathbf{z}$, let Ω represents the index set of the positive elements of $\mathbf{z}$. Then the solution $\mathbf{c}$ of Equation 24 is given as:

(A). For any $i \notin \Omega$, $\mathbf{c}_i^* = 0$;

(B). If $\| \mathbf{z}_\Omega \|_2 \leq \lambda$, then $\mathbf{c}_\Omega^* = 0$; otherwise, $\mathbf{c}_\Omega^* = (\| \mathbf{z}_\Omega \|_2 - \lambda)$ $\mathbf{z}_\Omega / \| \mathbf{z}_\Omega \|_2$.

Based on the aforementioned Theorem 1, after updating each column's variable **c** and subsequently combining all columns, the updated matrix C can be obtained.

4.2.2 Solution for transformation matrix

The solution for transformation matrix T can be obtained by directly maximizing Equation 18B in a block-wise manner, as follows:

$$T^{k+1} = \arg\max_{A} \frac{1}{2} \sum_{i=1}^{m} \left(q_i \frac{- \| \sqrt{W_i} \odot (\mathbf{s}_i^T - TC\mathbf{s}_i^T) \|_2^2}{2\sigma^2} - \varphi(q_i) \right) \tag{25}$$

Equation 25 is equivalent to:

$$T^{k+1} = \arg\max_{A} \frac{1}{2} \| \sqrt{W^k} \odot (X^k - TC^{k+1}X^k) \|_F^2 \tag{26}$$

By taking the first-order partial derivative of the right-hand of Equation 26 with respect to T, and setting the result to zero, we obtain the following expression:

$$W^k \odot (X^k - TC^{k+1}X^k)(C^{k+1}X^k)^T = 0 \tag{27}$$

The solution to Equation 27 can be derived as follows:

$$T^{k+1} = X^k(C^{k+1}X^k)^T(C^{k+1}X^k(C^{k+1}X^k)^T)^{\dagger} \tag{28}$$

where $(\cdot)^{\dagger}$ represents the pseudoinverse, X^k represents updated data matrix under impact of intermediate variable **q** which will be introduced later.

4.2.3 Solution for noise weight matrix

With respect to the noise weight matrix W subproblem, solving Equation 18C is equivalent to solving the following equation:

$$\begin{aligned} W^{k+1} &\leftarrow \arg\max_{W} F_1^{MCC}(C^{k+1}, T^{k+1}, W^k, \mathbf{q}^k) + E(W^k) \\ \text{s.t.} \quad & W + \bar{W} = \mathbf{1}, \quad W \; and \; \bar{W} \in [0,1]^{n\times m} \end{aligned} \tag{29}$$

In order to facilitate the solution, the Lagrange multiplier method is employed to relax the aforementioned equation, yielding the following result:

$$\begin{aligned} L(w_{ij}, \bar{w}_{ij}, \rho_i) = {} & \tfrac{1}{2} w_{ij}[X^k - T^{k+1}C^{k+1}X^k]_{ij}^2 + \beta \bar{w}_{ij} + \gamma(w_{ij}\log w_{ij} + \bar{w}_{ij}\log \bar{w}_{ij}) \\ & + \rho_i(w_{ij} + \bar{w}_{ij} - 1) \end{aligned} \tag{30}$$

where ρ_i denotes the Lagrange multiplier.

$$\begin{aligned} \tfrac{\partial L}{\partial w_{ij}} &= \tfrac{1}{2}[X^k - T^{k+1}C^{k+1}X^k]_{ij}^2 + \gamma \log w_{ij} + \gamma + \rho_i = 0, \\ \tfrac{\partial L}{\partial \bar{w}_{ij}} &= \beta + \gamma \log \bar{w}_{ij} + \gamma + \rho_i = 0, \\ \tfrac{\partial L}{\partial \rho_i} &= w_{ij} + \bar{w}_{ij} - 1 = 0 \end{aligned} \tag{31}$$

Further derivation of the solution to Equation 31 yields:

$$w_{ij}^{k+1} \leftarrow \frac{1}{\exp\left(([X - T^{k+1}C^{k+1}X^k]_{ij}^2/2 - \beta)/\gamma\right) + 1} \tag{32}$$

At the same time, $\bar{w}_{ij}$ can be updated as: $\bar{w}_{ij}^{k+1} = 1 - w_{ij}^{k+1}$.

4.2.4 Solution for q

By computing the partial derivative of Equation 13 with respect to q_i, we obtain:

$$q_i = -\exp(-x) \tag{33}$$

Substituting Equation 12 into Equation 33, we have:

$$\mathbf{q}^{k+1} = -\exp\left(\frac{- \| \sqrt{W_i} \odot (\mathbf{s}_i^T - TC\mathbf{s}_i^T) \|_2^2}{2\sigma^2} \right) \tag{34}$$

Simultaneously, update X^k to:

$$X^{k+1} = Diag\left(\sqrt{-\frac{\mathbf{q}^{k+1}}{2\sigma^2}} \right) X^k \tag{35}$$

The entire iterative method proposed by BCU for solving Equations 18A–D is referred to as the Maximum Correntropy Criterion-based Robust Sensor Selection (MCC_RSS) algorithm. To elucidate the iterative process of the MCC_RSS algorithm more clearly, we present it in the form of a flowchart, as depicted in Figure 1. Herein, the output J represents the locations of selected sensors. For the sake of clarity, the total objective function in Equations 18A-D is expressed as follows:

$$O(C, T, W, \mathbf{q}) = F(C, T, W, \mathbf{q}) - \alpha \| C \|_{2,1} - E(W) \tag{36}$$

4.3 Theoretical analysis

4.3.1 Convergence analysis

To facilitate the convergence analysis, we present **Theorem 2** and **Lemma 1** as follows:

Lemma 1: At k-th iteration with fixed C and T, the solutions of W in Equation 32 are global optimal.

Proof: The W obtained by Equation 32 is the global optimal because it is solved by Lagrange multiplier method and the Equation 29 is convex with the fixed C and T.

Theorem 2: The sequence of $\{O(C^k, T^k, W^k, \mathbf{q}^k)\}$, which is generated by the whole objective function in Equation 36 converges monotonically.

Proof: According to the BCU principle and Lemma 1, in the process of iterative optimization, we have:

$$\begin{aligned} & \{O(C^k, T^k, W^k, \mathbf{q}^k)\} \le \{O(C^{k+1}, T^k, W^k, \mathbf{q}^k)\} \le \{O(C^{k+1}, T^{k+1}, W^k, \mathbf{q}^k)\} \\ & \le \{O(C^{k+1}, T^{k+1}, W^{k+1}, \mathbf{q}^k)\} \le \{O(C^{k+1}, T^{k+1}, W^{k+1}, \mathbf{q}^{k+1})\} \end{aligned} \tag{37}$$

During each iteration, the energy of the objective function progressively increases through four sequential updates. Additionally, the objective function has an upper bound. Consequently, the MCC_RSS algorithm exhibits monotonic convergence.

4.3.2 Computational complexity

For the MCC_RSS algorithm, its computational complexity is determined by the number of samples m, the number of location features n in the original data matrix X, and the number of sensors

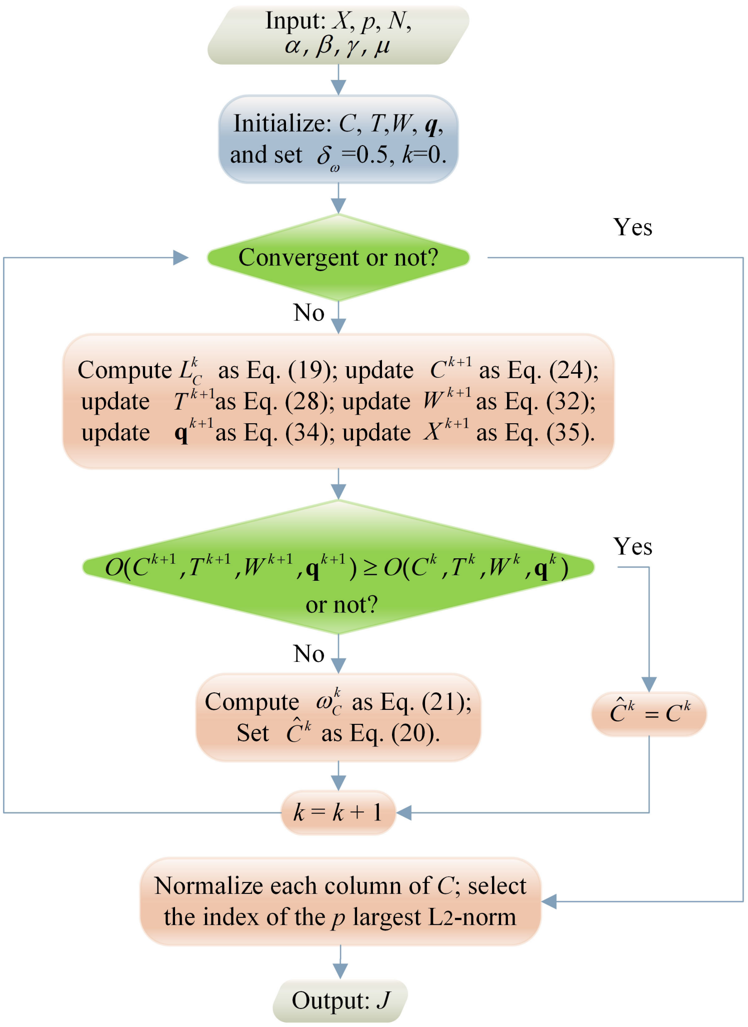

Input: X, p, N, α, β, γ, μ
Initialize: C, T,W, $\mathbf{q}$, and set δ_ω=0.5, k=0.
Convergent or not?
Yes
No
Compute L_C^k as Eq. (19); update C^{k+1} as Eq. (24); update T^{k+1}as Eq. (28); update W^{k+1} as Eq. (32); update $\mathbf{q}^{k+1}$as Eq. (34); update X^{k+1} as Eq. (35).
$O(C^{k+1},T^{k+1},W^{k+1},\mathbf{q}^{k+1})\geq O(C^k,T^k,W^k,\mathbf{q}^k)$ or not?
Yes
No
Compute ω_C^k as Eq. (21); Set $\hat{C}^k$ as Eq. (20).
$\hat{C}^k = C^k$
$k = k + 1$
Normalize each column of C; select the index of the p largest L2-norm
Output: J

FIGURE 1
Flowchart of MCC_RSS algorithm.

to be selected p. The complexity of each variable update process is as follows:

Update sensor selective matrix C: $np^2 + nm^2 + m^2 + nm + n^2 + n^3$

Update transformation matrix T: $pm + p^2 + p^3 + 2np$

Update noise weight matrix W: $n^2 + 2nm$

Update variable $\mathbf{q}$ and X: $2nm + nm^2$Disregarding the sparsity of the original data matrix X, and by omitting the lower-order terms, the resultant time complexity is given by: $O(n^3 + nm^2 + np^2 + p^3)$.

5 Experimental evaluation and results

The MCC_RSS algorithm we proposed is compared with the QR-based sensor selection outlined in (Manohar et al., 2018), POD, and two random selection method. In these methods, data reconstruction is carried out by SVD basis (RS) and sparse representation [SR (Callaham et al., 2019)] respectively. To better demonstrate the robustness of the MCC_RSS method, we also compared the proposed algorithm with the MSE_RSS method [where MSE refers to the use of the Frobenius norm to evaluate the difference between the original data and the reconstructed data as in (Zhang et al., 2024)].

5.1 Dataset and experimental description

5.1.1 Datasets description

5.1.1.1 Ocean temperature

The ocean temperature data utilized in this study is derived from the IAP Global Ocean Temperature Dataset of version IAPv4 (Cheng et al., 2024a) provided by Institute of Atmospheric Physics (IAP), Chinese Academy of Sciences. This dataset includes bias-corrected data from various observational systems within the World Ocean Database as well as data obtained through model simulations by research group of IAP (Cheng and Jiang, 2016; Cheng et al., 2017). Together, these ensemble data constitute the full-state global ocean temperature data. Due to the extensive matrix operations involved in the algorithm and the limitations of our computer memory, a subset of the dataset was selected. Specifically, ocean temperature data from the North Pacific region was used here, with a geographical range of 65°N latitude to 10° S latitude, and 78°W longitude to 99°E longitude. The spatial resolution accuracy is 1°×1°, encompassing a total of 10,188 geographical coordinates as the sensor selection locations. In this study, sea surface temperature at vertical levels of 0m is used to conduct the experiments. In addition, the temporal resolution is monthly, with a total of 996 samples spanning from 1940 to 2022. Of these, the first 800 samples are used as the training dataset, and the remaining samples are used as the test dataset.

5.1.1.2 Ocean salinity

The ocean salinity data utilized in this study is also derived from the IAP Global Ocean Salinity Dataset (Cheng et al., 2024b). This dataset also includes bias-corrected data from the World Ocean Database and the IAP research group, as well as model simulation data (Cheng and Jiang, 2016; Cheng et al., 2020). Similar to the temperature data, salinity data from the North Pacific region, sharing the same geographical range, were extracted. The geospatial resolution is 1°×1°. This ocean salinity dataset encompasses 41 vertical levels ranging from 0 to 2000 meters. For this experiment, the salinity data from the first vertical level were used. The temporal resolution of this dataset is monthly, spanning from January 1940 to December 2021, comprising a total of 984 samples. Of these, the first 800 samples are used as training data, while the remaining samples are used as test data.

5.1.2 Quality of reconstruction

The performance of the proposed method is evaluated by reconstruction errors, which are represented as follows:

$$R_{error} = \frac{\| Test - \hat{T}est \|_2}{\| Test \|_2} \tag{38}$$

Wherein $Test$ is input test data from the test set, $\hat{T}est$ is reconstructed by T from Equation 28 and the sensor's measurement data $Y_{test} = C_J \times Test$, as $\hat{T}est = T \times Y_{test}$. J is obtained from the sensor selection methods and C_J is the corresponding sensor selection matrix.

5.1.3 Experimental setting

The hardware and software environment used in the experiment is shown in Table 1.

The specific parameter settings for the MCC_RSS algorithm are as follows: α=1×10^{6}, β=1×10^{-5}, γ=1×10^{-4}, μ=1×10^{-4}, with the maximum number of iterations set to 400. During the execution of the MCC_RSS algorithm, the data is first normalized, followed by iterative updates of each subproblem solution based on BCU. The selection of these parameters is determined according to the algorithm's iterative process. Specifically, inappropriate parameters can lead to non-convergence of the objective function or premature termination of iterations. For instance, the value of α affects the solution process of Equation 23; an unsuitable α will prevent effective updates of matrix C. We determined the specific value of α by observing the algorithm's iterative process during experiments. Similarly, the values of β and γ influence the solution of the weight matrix W. Inappropriate values can cause the elements w_{ij} of Equation 32 to quickly converge to infinity or a constant, such as 1/2 (this conclusion can be easily derived by analyzing the relative relationship between β and γ in Equation 32). The value of μ is selected based on the overall distribution range of the objective function, ensuring it does not affect the convergence speed of the objective function value. Finally, among several alternative parameter combinations, the aforementioned parameters were selected as they exhibited the lowest error in the absence of noise.

TABLE 1 Experimental environment.

Hardware	Memory	16.0 GB
	CPU	AMD Ryzen 5 5600G @3.9GHz
Software	Programming Language	Matlab
	Operating System	Windows 11 Professional

To compare the robustness of different methods, we introduced varying proportions of outliers into the training data to simulate the loss conditions of actual oceanographic data. Considering the impact of non-Gaussian noise, we use the α-stable distribution to simulate heavy-tailed non-Gaussian noise, setting the signal-to-noise ratio parameter to 60. The alpha value (denoted as α_0 to avoid confusion with the model parameter α) is used to control the magnitude of the heavy tail, with α_0 set to1.

In the following experiments, Po=20% indicates that the proportion of outliers is 20%. Meanwhile, Sn=60 means that the signal-to-noise ratio of non-Gaussian noise is 60.

5.2 Reconstruction for ocean temperature

5.2.1 Compared with comparative methods

5.2.1.1 Reconstruction for different test snapshot

Figure 2 illustrates the comparison of reconstruction errors between the proposed method and the comparative methods for different snapshots in the test set. The number of selected sensors is set to 10. Due to the presence of random components in the comparative methods, each baseline method was executed 10 times, and the median error of the results was taken for comparison. Referring to Figure 2A, when there are outliers and noise in the training data, the reconstruction errors of the comparative methods increase rapidly. This indicates that the effectiveness of the QR and SR methods in the comparative methods is highly dependent on the quality of the training dataset. In contrast, the proposed MCC_RSS method can still minimize the impact of noise and maintain a low reconstruction error even in the presence of outliers and noise, achieving relatively stable reconstruction of test snapshots. Referring to Figure 2B, when the proportion of outliers in the training data increases and noise is still present, the proposed MCC_RSS method still exhibits the lowest reconstruction error compared to the comparative methods. Although the reconstruction error increases slightly

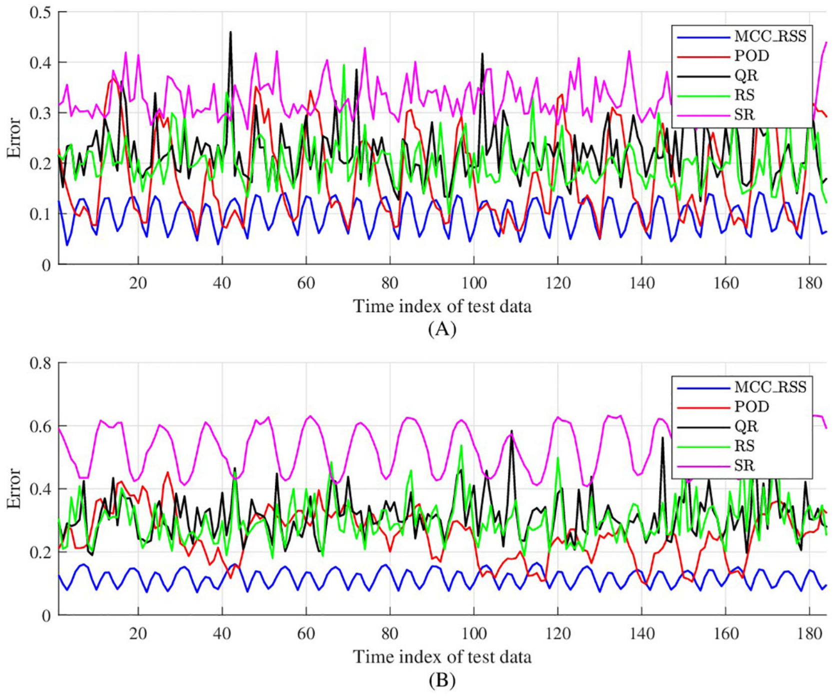


FIGURE 2
Reconstruction error for temperature comparation. **(A)** Po =20%, Sn=60; **(B)** Po =40%, Sn=60.

compared to the case with weaker noise, the overall difference is small. This fully demonstrates that the proposed MCC_RSS method is minimally affected by noise in the training dataset during data reconstruction, and its sparse sensor selection process has good robustness.

Figure 2 also illustrates that the reconstruction errors of different methods fluctuate over different time periods. Despite the varying degrees of noise contamination in the training data, the proposed MCC_RSS method effectively captures these temporal fluctuations with only 10 selected sensors, demonstrating superior stability.

5.2.1.2 Reconstruction for one test snapshot

To better reflect the sensitivity of different methods to outliers, a 10-fold cross-validation approach was employed. The results for each method, based on a single snapshot with p = 10, are compared and illustrated in Figure 3. Figure 3A demonstrates that the overall reconstruction error of the proposed method is consistently than that of other methods after multiple validations. Figure 3B indicates that even as the number of outliers increases, the reconstruction error of the proposed method remains lower than that of the other three methods, with only the POD method occasionally achieving lower reconstruction error. However, overall, the results of the proposed method are highly stable, with outcomes remaining concentrated even after multiple experiments. In contrast, the results of the comparative method exhibit a larger distribution range and lack stability across multiple validations. This stability is primarily due to the iterative optimization algorithm proposed in this paper, which focuses on gradually approaching the optimal solution until the algorithm termination condition is met. In the comparative method, the reconstructing based on the basis or orthogonal basis of SVD decomposition is significantly influenced by the data itself, leading to the instability of the solution.

Based on Figure 4, we present a randomly selected snapshot from the test set along with the corresponding reconstruction maps using different methods. In this scenario, the outlier ratio is set to 20%, and the signal-to-noise ratio is 60. The red dots in each reconstruction map indicate the sensor locations selected by the respective method. As shown in Figure 4B, the method proposed in this paper can effectively reconstruct the sea surface temperature distribution in the North Pacific region using only 10 selected sensors for this snapshot. Among the compared methods, only the POD method can relatively reconstruct the temperature distribution for this snapshot, but it still contains numerous noise points. Naturally, the reconstruction results vary for different snapshots, as indicated by the numerical comparison of reconstruction errors mentioned above. Although the POD method performs relatively well for this particular snapshot, the numerical results demonstrate that its reconstruction error is still higher than that of the proposed method when only 10 sensors are selected, and its stability is compromised by the randomly chosen sensor locations.

5.2.1.3 Reconstruction error by different number of sensors

Figure 5 presents a comparison of reconstruction errors for different methods when varying the numbers of selected sensors, under noise conditions of Po=20% and Sn=60%. To mitigate the

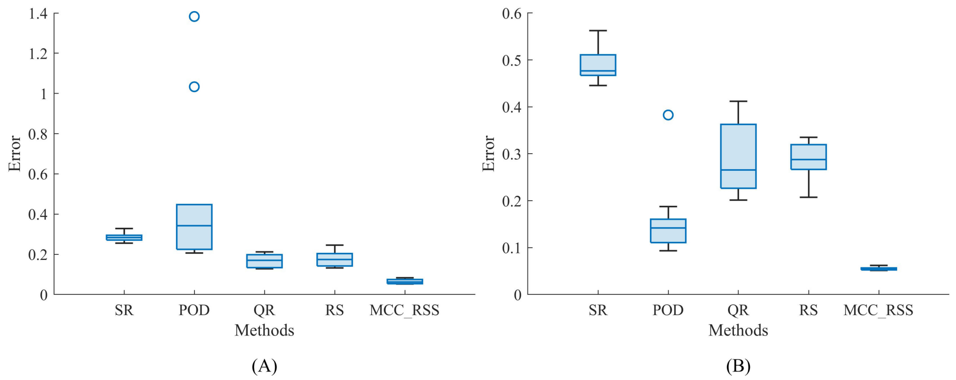


FIGURE 3
Reconstruction error of temperature for a snapshot. **(A)** Po =20%, Sn=60; **(B)** Po =40%, Sn=60.

influence of random factors, the comparative methods were subjected to 10-fold cross-validation. The error comparison results in Figure 5 indicate that when the training data contains noise, the proposed MCC_RSS method consistently achieves significantly lower reconstruction errors than other comparative methods, regardless of the number of sensors selected. Additionally, while the reconstruction errors of the comparative methods decrease as the number of sensors increases, the reconstruction error obtained by the proposed method shows almost no significant change. The primary reason for this is that, in the proposed method, after obtaining a *C* matrix through subspace learning, the column indices (i.e., sensor locations) are determined by selecting the

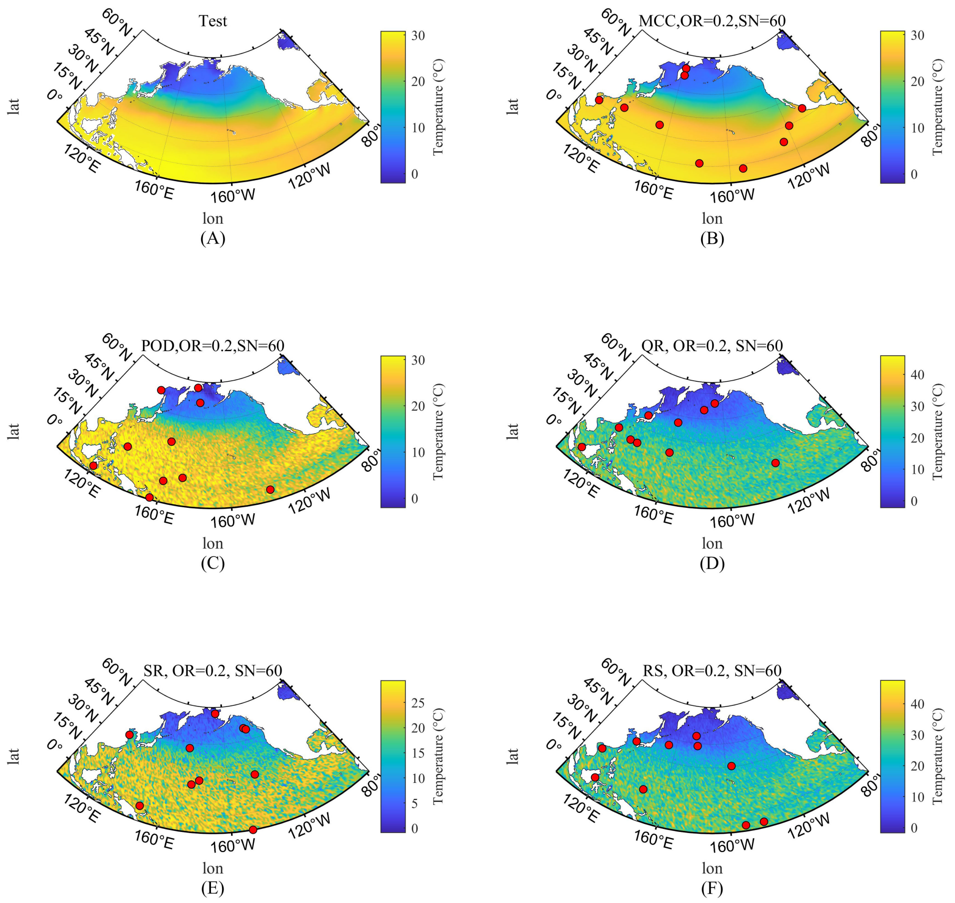


FIGURE 4
Reconstruction error of temperature for a snapshot. **(A)** Snapshot of test; **(B)** Reconstructed temperature by MCC_RSS; **(C)** Reconstructed temperature by POD; **(D)** Reconstructed temperature by QR; **(E)** Reconstructed temperature by SR; **(F)** Reconstructed temperature by RS.

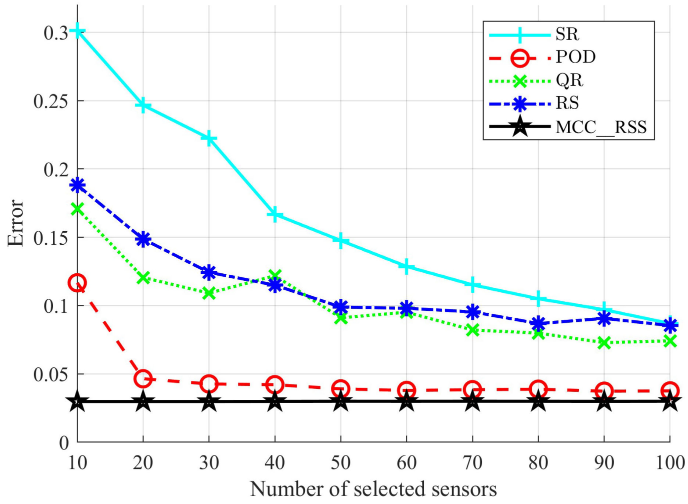


FIGURE 5
Reconstruction error of temperature by different number of sensors.

columns with the largest 2-norms for a given number of sensors. Therefore, once the training data is given, the low-dimensional subspace obtained through subspace learning is fixed, and selecting more sensors does not contribute additional useful information to the identified subspace. This results in the reconstruction error remaining nearly constant regardless of the number of sensors. Consequently, a very small number of sensors can still achieve good reconstruction performance. In contrast, the comparative methods increase the number of features used as the number of sensors increases, leading to a reduction in reconstruction error. Therefore, the proposed method is more suitable for scenarios requiring a limited number of sensors.

5.2.2 Compared with MSE_RSS methods

To better demonstrate the effectiveness of the MCC method in improving robustness, we compare the proposed MCC_RSS method with the MSE_RSS method, as shown in Figure 6. The primary difference between MSE_RSS and MCC_RSS lies in the measurement of the discrepancy between the original and reconstructed data, with MSE_RSS lacking the local geometric structure preservation

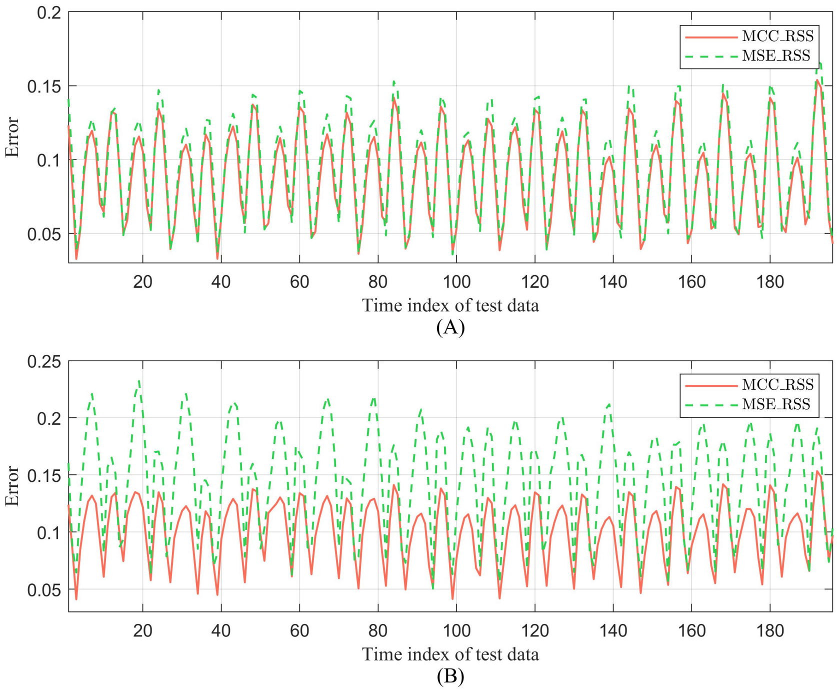


FIGURE 6
Comparison between MCC_RSS and MSE-RSS of ocean temperature. **(A)** No additional noise; **(B)** Po =20%, Sn=60.

term. The update formulas for Lipschitz constant of MSE_RSS are presented as: $L_C^k = \| A^k \|_2^2 \| X \|_2^2 \| W^k \|_2$, where X remains unchanged during the iteration process.

The reconstruction error results shown in Figure 6A indicate that even for subspace learning on training data without added noise, the sensor subset selected by the proposed MCC_RSS method achieves superior data reconstruction performance compared to the MSE_RSS method. This is primarily because, even without additional noise in the ocean temperature training data, the original data inherently contains model noise introduced during the ocean data assimilation process. The sensor selection method based on MCC proposed in this paper can minimize the impact of such noise as much as possible. Furthermore, Figure 6B presents the reconstruction results of these two methods when the training data contains 40% outliers and non-Gaussian noise. The results demonstrate that, with more severe noise, the difference in reconstruction performance between the sensor subset selected by the proposed MCC_RSS method and the MSE_RSS method further increases. This indicates that the proposed MCC_RSS method, by using MCC as the measure of the difference between the original and reconstructed data, is better able to mitigate the impact of noise on the results when the training data contains noise.

5.3 Reconstruction for ocean salinity

5.3.1 Compared with comparative methods

5.3.1.1 Reconstruction for different test snapshot

Figure 7 presents a comparison of the reconstruction errors between the proposed method and the comparative methods for ocean salinity data, with the number of sensors selected being 10. From Figures 7A, B, it can be observed that when the training data contains varying levels of noise, the reconstruction errors of the proposed MCC_RSS method are consistently lower than those of the comparative methods. Additionally, the reconstruction errors still reflect the periodicity of the ocean data to a certain extent. As the level of noise contamination in the training data increases, the reconstruction errors of all methods decrease. However, compared to the comparative methods, the decrease in reconstruction error for the proposed MCC_RSS method is less significant. This further demonstrates that, when selecting sensors for ocean salinity data, the proposed MCC_RSS method is less affected by the noise present in the data compared to the comparative methods.

5.3.1.2 Reconstruction for one test snapshot

Figure 8 presents a comparison of reconstruction error for a randomly selected sample (snapshot) using 10-fold cross-validation, with p=10. From Figures 8A, B, it can be observed that despite variations in outliers and noise distribution in the ocean salinity training data during multiple implementations of both the proposed method and the comparison method, the reconstruction error distribution of the proposed MCC_RSS method remains relatively concentrated, indicating better algorithm stability. In contrast, the reconstruction error distribution of the comparison method becomes more dispersed when the noise distribution in the training data changes. Additionally, the proposed method consistently achieves the lowest reconstruction error. This result further demonstrates that the MCC_RSS algorithm, based on MCC subspace learning, can iteratively learn a relatively stable low-

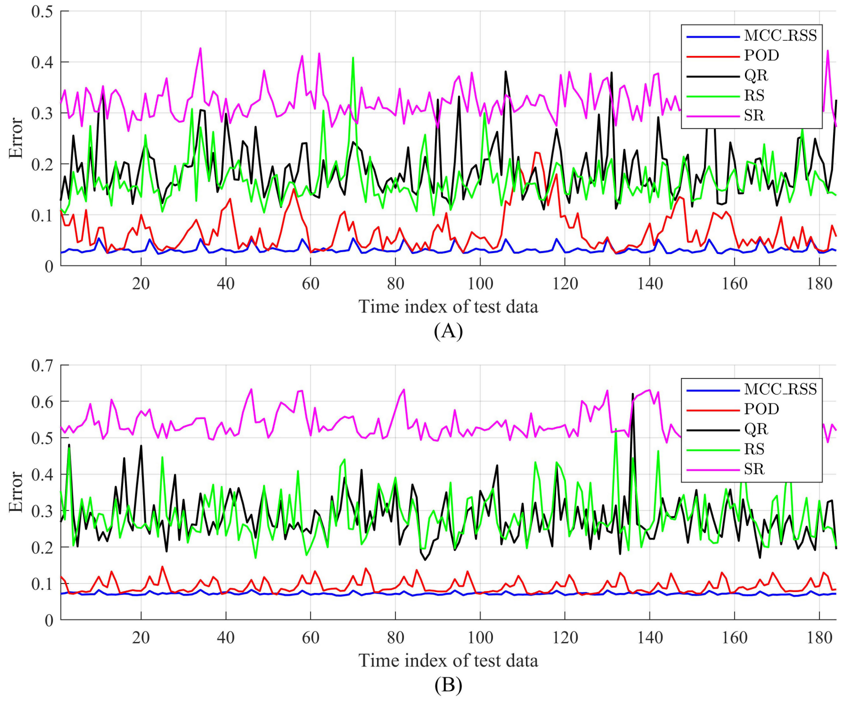


FIGURE 7
Reconstruction error for salinity comparation. **(A)** Po =20%, Sn=60; **(B)** Po =40%, Sn=60.

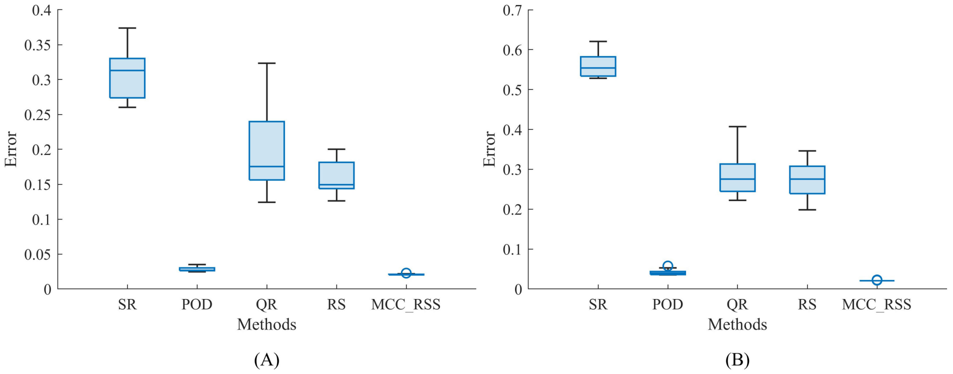


FIGURE 8
Reconstruction error of salinity for a snapshot. **(A)** Po =20%, Sn=60; **(B)** Po =40%, Sn=60.

dimensional subspace under different conditions, thereby ensuring that the selected subset of sensor measurements exhibits good robustness and achieves better data reconstruction.

Figure 9 presents a comparison of the reconstruction effects of different methods on the aforementioned randomly selected snapshot, with the noise in the training data set to Po=20% and Sn=60%. The red dots indicate the positions of the sensors selected by the different methods. As shown in Figure 9B, the proposed MCC_RSS method achieves effective reconstruction of ocean salinity data with only a subset of 10 sensors, successfully capturing the main characteristics of the salinity distribution in the North Pacific region when compared to the test snapshot. The POD method, while slightly inferior to the proposed method, also generally reflects the main patterns of salinity distribution in the

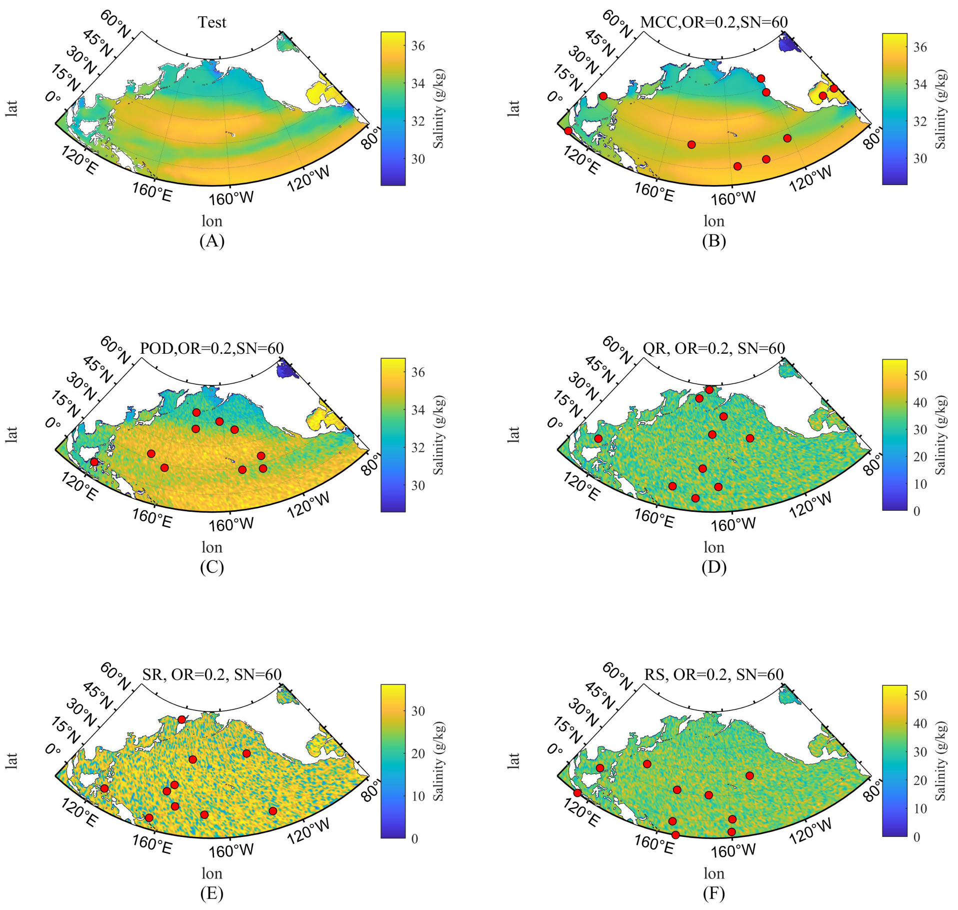


FIGURE 9
Reconstruction error of salinity for a snapshot. **(A)** Snapshot of test; **(B)** Reconstructed salinity by MCC_RSS; **(C)** Reconstructed salinity by POD; **(D)** Reconstructed salinity by QR; **(E)** Reconstructed salinity by SR; **(F)** Reconstructed salinity by RS.

North Pacific region. However, the other three comparative methods fail to capture the salinity distribution characteristics with only a subset of 10 sensors. This indicates that, even with a certain level of noise in the training data and a limited number of sensors, the sensor subset selected by the proposed MCC_RSS method can still achieve effective data reconstruction.

5.3.1.3 Reconstruction error by different number of sensors

Figure 10 presents a comparison of the reconstruction errors for different methods when selecting varying numbers of sensors. The noise in the training data is set to Po=40% and Sn=60. As shown in the figure, the proposed MCC_RSS method consistently achieves the lowest reconstruction error compared to the comparative methods, regardless of the number of sensors selected. Additionally, as the number of sensors increases, the reconstruction error remains relatively stable. As previously mentioned, once the proposed MCC_RSS method determines the matrix C corresponding to the low-dimensional subspace, the indices of the selected sensors, regardless of their number, are derived from the entries of matrix C with the largest 2-norms of the columns. This selection process does not significantly alter the obtained subspace, further demonstrating that the low-dimensional subspace derived from the proposed method is relatively stable. Consequently, it is more suitable for scenarios with fewer sensors compared to the comparative methods.

In contrast, for the comparative methods, particularly the QR and RS methods, the reconstruction error decreases rapidly as the number of selected sensors increases. However, they are still significantly affected by noise, and their reconstruction errors are not as favorable as those of the proposed method. The SR method, which relies more heavily on the library established from the training data, is the most affected by noise. Comparatively, the POD method performs closer to the proposed method in terms of ocean salinity reconstruction and can reasonably reconstruct salinity data with different numbers of sensors. Nevertheless, its error remains significantly higher than that of the proposed method.

Therefore, utilizing the sensors selected by the proposed MCC_RSS method for data reconstruction can achieve more desirable results, particularly when the number of sensors is limited.

5.3.2 Compared with MSE_RSS methods

Figure 11 shows the experimental results of the proposed MCC_RSS method and the corresponding MSE_RSS method on global ocean salinity data, using 10 sensors. As shown in Figure 11A, when no additional noise is introduced to the training data, there is no significant difference in the reconstruction errors between the two methods. Differences are observed only in specific time samples, such as in the trough region between sample indices 100 and 140, where the error of the MCC_RSS method is smaller than that of the corresponding MSE_RSS method. In Figure 11B, when the training data contains noise, it is evident that the overall fluctuation of the reconstruction error of the MCC_RSS method is significantly smaller than that of the MSE_RSS method. The average error of the MCC_RSS method is 0.0375, while the average error of the MSE_RSS method is 0.0391. This further demonstrates that the proposed method can more effectively mitigate the impact of noise.

6 Conclusion and discussion

Considering the distinct low-rank characteristics of ocean data, we explored how to optimally utilize subspace learning methods to derive a more reasonable low-dimensional subspace of high-dimensional ocean data. This approach facilitates the selection of low-dimensional measurements from sensors that better meet the requirements. Based on this premise, we develop a robust sensor selection method that establishes an evaluation function based on the Maximum Correntropy Criterion (MCC) and selects sensor

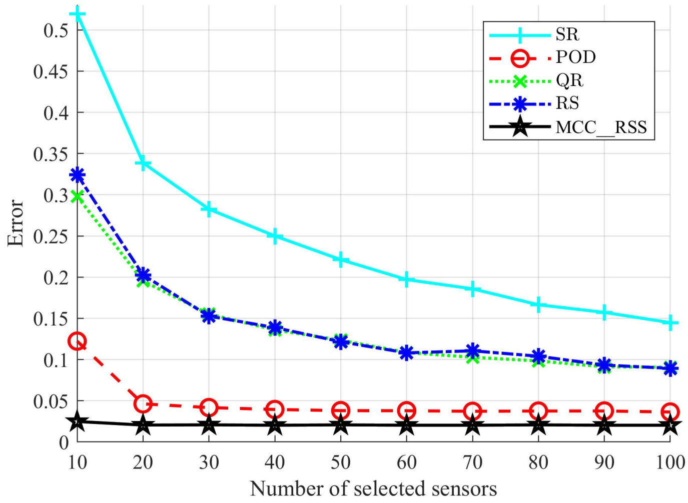


FIGURE 10
Reconstruction error of salinity by different number of sensors.

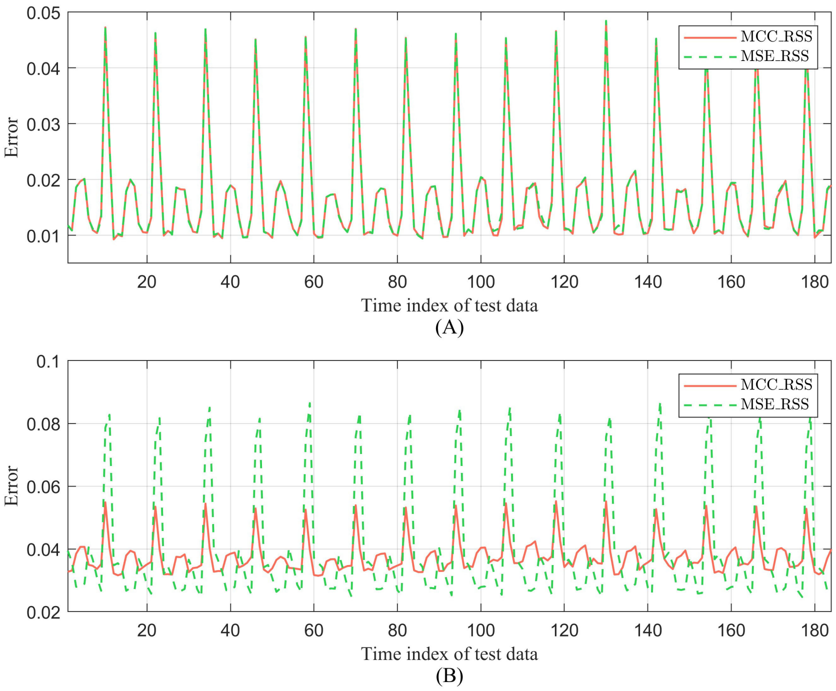


FIGURE 11
Comparison between MCC_RSS and MSE-RSS of ocean salinity. **(A)** No additional noise; **(B)** Po =20%, Sn=60.

subsets to reconstruct the full state ocean data through subspace learning. Compared to the Euclidean distance used in existing methods, MCC demonstrates superior robustness in evaluating the discrepancies between reconstructed data and original data, particularly in the presence of varying levels of noise in the original data. The model also incorporates noise weighting and optimizes noise distribution using entropy terms, effectively controlling sparse severe noise and mitigating the impact of non-Gaussian noise and outliers. The use of noise weighting in the proposed method allows for better identification of varying levels of noise during the subspace learning process. This reduces the impact on the learned subspace, resulting in more stable reconstruction outcomes for sensor selection under different noise conditions.

Furthermore, the integration of the local geometric structure of data samples further enhances the reconstruction accuracy achieved by the selected sensors. By minimizing the similarity of the selected sensor measurement subset through the graph Laplacian matrix between samples, the reconstruction capability of the selected sensors for the full state data is further improved. To better solve the model's evaluation function, the half-quadratic BCU method was employed, effectively addressing the challenge of solving the non-convex parts of the objective function. During the iterative solving process, the selection matrix, transformation matrix, and noise weighting matrix continuously evolve towards the optimal solution. This ultimately results in the learned low-dimensional subspace, along with the corresponding selection and transformation matrices, achieving superior data reconstruction outcomes. Additionally, the model effectively converges to the optimal solution with a low number of iterations.

Compared to the benchmark methods, our approach performs better and yields highly robust solutions under varying noise conditions. Specifically, the proposed method demonstrates that even with data containing different levels of noise, it can achieve effective data reconstruction using a smaller number of sensors. This makes it particularly suitable for ocean data reconstruction where the number of sensors is limited. This provides a valuable reference for future ocean environment monitoring systems on how to deploy fewer sensors more efficiently.

In our future work, we will explore how to improve the method proposed in this paper to reduce its computational complexity. For example, after preliminary screening of location features using statistical methods such as variance analysis and correlation coefficients, BCU iterative solving can be performed, or location features can be grouped and optimized separately before combining the results. For the parameter selection, we will also explore more scientific methods, such as grid search and Bayesian methods, to obtain parameter values that can achieve the optimal convergence results of the objective function. In addition, the method proposed in this paper does not make a significant contribution to the results when the number of sensors increases. Therefore, with the increase in the number of selected sensors, further exploration is needed to obtain a better low-dimensional subspace that can introduce more effective information. Potential improvements include incorporating oceanographic knowledge to screen location features, thereby identifying the most valuable candidate locations for monitoring. Alternatively, oceanographic models can be used to assess the value of each location feature, facilitating the optimization of a data-driven sensor selection model.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

QZ: Conceptualization, Formal Analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. HW: Funding acquisition, Project administration, Supervision, Writing – review & editing. LL: Investigation, Writing – review & editing. XM: Formal Analysis, Writing – review & editing. JX: Writing – review & editing, Supervision.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (Grant Nos. 52331012, 52071200, 52201401, 52201403, and 52102397), in part by the National Key Research and Development Program (Grant No. 2021YFC2801002), in part by the Shanghai Committee of Science and Technology, China (Grant No. 23010502000), in part by the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No. 23CGA61), in part by the Top-Notch Innovative Program for Postgraduates of Shanghai Maritime University under Grant 2022YBR012.

Acknowledgments

We acknowledge the use of ChatAI (version Chat GPT 4o) for translation purposes in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Callaham, J. L., Maeda, K., and Brunton, S. L. (2019). Robust flow reconstruction from limited measurements via sparse representation. *Phys. Rev. Fluids.* 4. doi: 10.1103/PhysRevFluids.4.103907

Carmi, A., and Gurfil, P. (2013). Sensor selection via compressed sensing. *Automatica (Oxf).* 49, 3304–3314. doi: 10.1016/j.automatica.2013.08.032

Chamon, L. F. O., Pappas, G. J., and Ribeiro, A. (2021). Approximate supermodularity of kalman filter sensor selection. *IEEE Trans. Automat. Contr.* 66, 49–63. doi: 10.1109/TAC.9

Cheng, L., Trenberth, K. E., Fasullo, J. T., Boyer, T., Abraham, J. P., Zhu, J., et al. (2024a). *Data from: Institute of Atmospheric Physics, Chinese Academy of Sciences*. Available online at: http://www.ocean.iap.ac.cn/ftp/cheng/IAPv4_IAP_Temperature_gridded_1month_netcdf/ (Accessed June 02, 2024).

Cheng, L., Trenberth, K. E., Gruber, N., Abraham, J. P., Fasullo, J. T., Li, G., et al. (2024b). *Data from: Institute of Atmospheric Physics, Chinese Academy of Sciences*. Available online at: http://www.ocean.iap.ac.cn/ftp/cheng/CZ16_v0_IAP_Salinity_gridded_1month_netcdf/ (Accessed June 2, 2024).

Cheng, L., and Jiang, Z. (2016). Benefits of CMIP5 multimodel ensemble in reconstructing historical ocean subsurface temperature variations. *J. Clim.* 29, 5393–5416. doi: 10.1175/JCLI-D-15-0730.1

Cheng, L., Trenberth, K. E., Fasullo, J. T., Boyer, T., Abraham, J. P., and Zhu, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. *Sci. Adv.* 3. doi: 10.1126/sciadv.1601545

Cheng, L., Trenberth, K. E., Gruber, N., Abraham, J. P., Fasullo, J. T., Li, G., et al. (2020). Improved estimates of changes in upper ocean salinity and the hydrological cycle. *J. Clim.* 33, 10357–10381. doi: 10.1175/JCLI-D-20-0366.1

Chepuri, S. P., and Leus, G. (2015). Sparsity-promoting sensor selection for non-linear measurement models. *IEEE Trans. Signal Process.* 63, 684–698. doi: 10.1109/TSP.2014.2379662

Dubois, P., Gomez, T., Planckaert, L., and Perret, L. (2022). Machine learning for fluid flow reconstruction from limited measurements. *J. Comput. Phys.* 448. doi: 10.1016/j.jcp.2021.110733

Emily, C., Steven, L. B., and Kutz, J. N. (2020). Multi-fidelity sensor selection-Greedy algorithms to place cheap and expensive sensors with cost constraints. *IEEE Sens. J.* 21, 600–611. doi: 10.1109/JSEN.2020.3013094

Erichson, N. B., Mathelin, L., Yao, Z., Brunton, S. L., Mahoney, M. W., and Kutz, J. N. (2020). Shallow neural networks for fluid flow reconstruction with limited sensors. *Pro. Roy Soc A.* 476. doi: 10.1098/rspa.2020.0097

Fukami, K., Maulik, R., Ramachandra, N., Fukagata, K., and Taira, K. (2021). Global field reconstruction from sparse sensors with Voronoi tessellation-assisted deep learning. *Nat. Mach. Intell.* 3, 945–951. doi: 10.1038/s42256-021-00402-2

Ghosh, S., De, S., Chatterjee, S., and Portmann, M. (2021). Learning-based adaptive sensor selection framework for multi-sensing WSN. *IEEE Sens. J.* 21, 13551–13563. doi: 10.1109/JSEN.2021.3069264

Guo, X., and Lin, Z. (2018). Low-rank matrix recovery via robust outlier estimation. *IEEE Trans. Image Process.* 27, 5316–5327. doi: 10.1109/TIP.2018.2855421

He, R., Hu, B., Yuan, X., and Wang, L. (2014). "Correntropy and linear representation," in *Robust recognition via information theoretic learning* (SpringerBriefs in Computer Science: Springer, Cham), 45–60.

He, Y., Wang, F., Wang, S., Cao, J., and Chen, B. (2019). Maximum correntropy adaptation approach for robust compressive sensing reconstruction. *Inform. Sci.* 480, 381–402. doi: 10.1016/j.ins.2018.12.039

Jayaraman, B., Al Mamun, S. M. A., and Lu, C. (2019). Interplay of sensor quantity, placement and system dimension in POD-based sparse reconstruction of fluid flows. *Fluids.* 4. doi: 10.3390/fluids4020109

Jayaraman, B., and Mamun, S. M. A. A. (2020). On data-driven sparse sensing and linear estimation of fluid flows. *Sensors.* 20. doi: 10.3390/s20133752

Joneidi, M., Zaeemzadeh, A., Shahrasbi, B., Qi, G.-J., and Rahnavard, N. (2020). E-optimal sensor selection for compressive sensing-based purposes. *IEEE Trans. Big Data.* 6, 51–65. doi: 10.1109/TBigData.6687317

Joshi, S., and Boyd, S. (2009). Sensor selection via convex optimization. *IEEE Trans. Signal Process.* 57, 451–462. doi: 10.1109/TSP.2008.2007095

Kalinić, H., Ćatipović, L., and Matić, F. (2022). Optimal sensor placement using learning models—A mediterranean case study. *Remote Sens.* 14. doi: 10.3390/rs14132989

Khokhlov, I., Pudage, A., and Reznik, L. (2019).Sensor selection optimization with genetic algorithms. In: *2019 IEEE SENSORS* (Montreal, QC, Canada) (Accessed 27-30 October 2019). 2019 IEEE SENSORS.

Krause, A., Singh, A., and Guestrin, C. (2008). Near-optimal sensor placements in gaussian processes theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.* 9, 235–284. doi: 10.5555/1390681.1390689

Lin, X., Chowdhury, A., Wang, X., and Terejanu, G. (2019). Approximate computational approaches for Bayesian sensor placement in high dimensions. *Inform Fusion.* 46, 193–205. doi: 10.1016/j.inffus.2018.06.006

Liu, W., Pokharel, P. P., and Principe, J. C. (2007). Correntropy: properties and applications in non-gaussian signal processing. *IEEE Trans. Signal Process.* 55, 5286–5298. doi: 10.1109/TSP.2007.896065

Liu, X., Wang, L., Zhang, J., Yin, J., and Liu, H. (2014). Global and local structure preservation for feature selection. *IEEE Trans. Neural Netw. Learn Syst.* 25, 1083–1095. doi: 10.1109/TNNLS.2013.2287275

Manohar, K., Brunton, B. W., Kutz, J. N., and Brunton, S. L. (2018). Data-driven sparse sensor placement for reconstruction: demonstrating the benefits of exploiting known patterns. *IEEE Control Syst.* 38, 63–86. doi: 10.1109/MCS.2018.2810460

Mei, X., Han, D., Saeed, N., Wu, H., Han, B., and Li, K.-C. (2024). Localization in underwater acoustic ioT networks: dealing with perturbed anchors and stratification. *IEEE Internet Things J.* 11, 17757–17769. doi: 10.1109/JIOT.2024.3360245

Meray, A., Boza, R., Siddiquee, M. R., Reyes, C., Amini, M. H., and Prabakar, N. (2023). Subset sensor selection optimization: A genetic algorithm approach with innovative set encoding methods. *IEEE Sens. J.* 23, 28462–28473. doi: 10.1109/JSEN.2023.3322596

Model, D., and Zibulevsky, M. (2006). Signal reconstruction in sensor arrays using sparse representations. *Signal Process.* 86, 624–638. doi: 10.1016/j.sigpro.2005.05.033

Nguyen, L., Thiyagarajan, K., Ulapane, N., and Kodagoda, S. (2021). "Multimodal sensor selection for multiple spatial field reconstruction," in *2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA).* (Chengdu, China: IEEE). 1181–1186. doi: 10.1109/ICIEA51954.2021.9516255

Özbay, A. G., and Laizet, S. (2022). Deep learning fluid flow reconstruction around arbitrary two-dimensional objects from sparse sensors using conformal mappings. *AIP Advances.* 12. doi: 10.1063/5.0087488

Patan, M., Klimkowicz, K., and Patan, K. (2022). "Optimal sensor selection for prediction-based iterative learning control of distributed parameter systems," in *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, (Singapore, Singapore: IEEE). 449–454. doi: 10.1109/ICARCV57592.2022.10004370

Peherstorfer, B., Drmač, Z., and Gugercin, S. (2020). Stability of discrete empirical interpolation and gappy proper orthogonal decomposition with randomized and deterministic sampling points. *SIAM J. Sci. Comput.* 42, A2837–A2864. doi: 10.1137/19M1307391

Prakash, O., and Bhushan, M. (2023). Kullback-Leibler divergence based sensor placement in linear processes for efficient data reconciliation. *Comput. Chem. Eng.* 173. doi: 10.1016/j.compchemeng.2023.108181

Sahba, S., Wilcox, C. C., Mcdaniel, A., Shaffer, B., Brunton, S. L., and Kutz, J. N. (2022). Wavefront sensor fusion via shallow decoder neural networks for aero-optical predictive control." in *SPIE Optical Engineering + Applications*. (San Diego, California, United States. Interferometry XXI) Vol 12223. doi: 10.1117/12.2631951 (accessed October 03, 2022).

Saito, Y., Nakai, K., Nagata, T., Yamada, K., Nonomura, T., Sakaki, K., et al. (2023). Sensor selection with cost function using nondominated-solution-based multiobjective greedy method. *IEEE Sens. J.* 23, 31006–31016. doi: 10.1109/JSEN.2023.3328005

Santini, S., and Colesanti, U. (2009). "Adaptive random sensor selection for field reconstruction in wireless sensor networks," in *Proceedings of the Sixth International Workshop on Data Management for Sensor Networks*, Lyon, France, August 2009. (New York, NY, USA: Association for Computing Machinery). doi: 10.1145/1594187.1594195

Santos, J. E., Fox, Z. R., Mohan, A., O'Malley, D., Viswanathan, H., and Lubbers, N. (2023). Development of the Senseiver for efficient field reconstruction from sparse observations. *Nat. Mach. Intell.* 5, 1317–1325. doi: 10.1038/s42256-023-00746-x

Saucan, A. A., and Win, M. Z. (2020). Information-seeking sensor selection for ocean-of-things. *IEEE Internet Things J.* 7, 10072–10088. doi: 10.1109/JIoT.6488907

Xu, Y. (2015). Alternating proximal gradient method for sparse nonnegative Tucker decomposition. *Math. Program. Comput.* 7, 39–70. doi: 10.1007/s12532-014-0074-y

Xu, Y., and Yin, W. A. (2013). Block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* 6, 1758–1789. doi: 10.1137/120887795

Xue, J., Zhao, Y., Liao, W., and Chan, J. (2019). Nonlocal tensor sparse representation and low-rank regularization for hyperspectral image compressive sensing reconstruction. *Remote Sens.* 11, 193. doi: 10.3390/rs11020193

Yamada, K., Saito, Y., Nankai, K., Nonomura, T., Asai, K., and Tsubakino, D. (2021). Fast greedy optimization of sensor selection in measurement with correlated noise. *Mech. Syst. Signal Process.* 158. doi: 10.1016/j.ymssp.2021.107619

Yang, C., Wu, J., Ren, X., Yang, W., Shi, H., and Shi, L. (2015). Deterministic sensor selection for centralized state estimation under limited communication resource. *IEEE Trans. Signal Process.* 63, 2336–2348. doi: 10.1109/TSP.2015.2412916

Yildirim, B., Chryssostomidis, C., and Karniadakis, G. E. (2009). Efficient sensor placement for ocean measurements using low-dimensional concepts. *Ocean Model.* 27, 160–173. doi: 10.1016/j.ocemod.2009.01.001

Zhang, J., Liu, J., and Huang, Z. (2023). Improved deep learning method for accurate flow field reconstruction from sparse data. *Ocean Eng.* 280, 114902. doi: 10.1016/j.oceaneng.2023.114902

Zhang, P., Nevat, I., Peters, G. W., Septier, F., and Osborne, M. A. (2018). Spatial field reconstruction and sensor selection in heterogeneous sensor networks with stochastic energy harvesting. *IEEE Trans. Signal Process.* 66, 2245–2257. doi: 10.1109/TSP.78

Zhang, Q., Wu, H., Liang, L., Mei, X., Xian, J., and Zhang, Y. A. (2024). Robust sparse sensor placement strategy based on indicators of noise for ocean monitoring. *J. Mar. Sci. Eng.* 12, 1220. doi: 10.3390/jmse12071220

Zhang, Q., Wu, H., Mei, X., Han, D., Marino, M. D., Li, K. C., et al. (2023). A sparse sensor placement strategy based on information entropy and data reconstruction for ocean monitoring. *IEEE Internet Things J.* 10, 19681–19694. doi: 10.1109/JIOT.2023.3281831

Zhao, X., Du, L., Peng, X., Deng, Z., and Zhang, W. (2021). Research on refined reconstruction method of airfoil pressure based on compressed sensing. *Theor. Appl. Mechanics Letters.* 11. doi: 10.1016/j.taml.2021.100223

Zhou, N., Xu, Y., Cheng, H., Fang, J., and Pedrycz, W. (2016). Global and local structure preserving sparse subspace learning: An iterative approach to unsupervised feature selection. *Pattern Recogn.* 53, 87–101. doi: 10.1016/j.patcog.2015.12.008

Zhou, N., Xu, Y., Cheng, H., Yuan, Z., and Chen, B. (2019). Maximum correntropy criterion-based sparse subspace learning for unsupervised feature selection. *IEEE Trans. Circ. Syst. Vid.* 29, 404–417. doi: 10.1109/TCSVT.76

Appendix A

The Lipschitz constant L_C^k could be obtained by computing the derivative of C in Equation 18A $\hat{G}^k = \nabla_C F(\hat{C}^k, T^k, W^k, \mathbf{q}^k)$. Through matrix calculation, it is easy to derive:

$$\begin{aligned}&\nabla_C F(C, T, W, \mathbf{q}^k)\\&\quad = T^T[W \odot (X^k - TCX^k)](X^k)^T - \mu CXLX^T\end{aligned} \tag{39}$$

where X^k is the updated data at i-th iteration by variable $\mathbf{q}$. Given two matrix variables $\hat{C}$ and $\tilde{C}$, then we have:

$$\begin{aligned}&\| \nabla_C F(\hat{C}, T, W) - \nabla_C F(\tilde{C}, T, W) \|_F\\&= \| T^T[W \odot (X^k - T\hat{C}X^k)](X^k)^T - \mu\hat{C}XLX^T - T^T\\&\ \ [W \odot (X^k - T\tilde{C}X^k)](X^k)^T + \mu\tilde{C}XLX^T \|_F\\&= \| T^T\{W \odot [T(\hat{C} - \tilde{C})X^k]\}(X^k)^T + \mu(\tilde{C} - \hat{C})XLX^T \|_F\\&\leq \| T^T\{W \odot [T(\hat{C} - \tilde{C})X^k]\}(X^k)^T \|_F + \mu \| (\tilde{C} - \hat{C})XLX^T \|_F\\&\leq \| T \|_2^2 \| X^k \|_2^2 \| W \|_2 \| \hat{C} - \tilde{C} \|_F + \mu \| XLX^T \|_2 \| \tilde{C} - \hat{C} \|_F\\&= \left(\| T \|_2^2 \| X^k \|_2^2 \| W \|_2 + \mu \| XLX^T \|_2\right) \| \hat{C} - \tilde{C} \|_F\end{aligned} \tag{40}$$

The inequality part in above equation is transformed according to the Cauchy-Schwarz inequality. By Equation 40, we have the Lipschitz constant L_C^k as:

$$L_C^k = \| T^k \|_2^2 \| X^k \|_2^2 \| W^k \|_2 + \mu \| XLX^T \|_2 \tag{41}$$

Appendix B

To facilitate reading, a nomenclature listing used in this study is provided here; please refer to Table A1.

TABLE A1 Abbreviations and Full Term.

Abbreviation	Full Term
MCC	Maximum Correntropy Criterion
RSS	Robust Sensor Selection
BCU	Block Coordinate Update
NP-hard	Non-deterministic Polynomial-time hard
POD	Proper Orthogonal Decomposition
SVD	Singular Value Decomposition
ITL	Information Theoretic Learning
LPP	Linear Preserve Projection
SR	Sparse Representation
RS	Random Selection
MSE	Mean Square Error

TYPE Original Research
PUBLISHED 17 October 2024
DOI 10.3389/fmars.2024.1473551

OPEN ACCESS

EDITED BY
Yu Jiang,
Jilin University, China

REVIEWED BY
Yahui Guo,
Central China Normal University, China
Hasbi Yasin,
Diponegoro University, Indonesia
Salim Heddam,
University of Skikda, Algeria

*CORRESPONDENCE
Ji Wang
13902576499@163.com

RECEIVED 31 July 2024
ACCEPTED 30 September 2024
PUBLISHED 17 October 2024

CITATION
Liu W, Wang J, Li Z and Lu Q (2024) ISSA optimized spatiotemporal prediction model of dissolved oxygen for marine ranching integrating DAM and Bi-GRU.
Front. Mar. Sci. 11:1473551.
doi: 10.3389/fmars.2024.1473551

ISSA optimized spatiotemporal prediction model of dissolved oxygen for marine ranching integrating DAM and Bi-GRU

Wenjing Liu[1,2], Ji Wang[1,2*], Zhenhua Li[1,2] and Qingjie Lu[1,2]

[1]School of Electronic and Information Engineering, Guangdong Ocean University, Zhanjiang, Guangdong, China, [2]Guangdong Province Smart Ocean Sensor Network and Equipment Engineering Technology Research Center, Guangdong Ocean University, Zhanjiang, Guangdong, China

In marine ranching aquaculture, dissolved oxygen (DO) is a crucial parameter that directly impacts the survival, growth, and profitability of cultured organisms. To effectively guide the early warning and regulation of DO in aquaculture waters, this study proposes a hybrid model for spatiotemporal DO prediction named PCA-ISSA-DAM-Bi-GRU. Firstly, principal component analysis (PCA) is applied to reduce the dimensionality of the input data and eliminate data redundancy. Secondly, an improved sparrow search algorithm (ISSA) based on multi strategy fusion is proposed to enhance the optimization ability and convergence speed of the standard SSA by optimizing the population initialization method, improving the location update strategies for discoverers and followers, and introducing a Cauchy-Gaussian mutation strategy. Thirdly, a feature and temporal dual attention mechanism (DAM) is incorporated to the baseline temporal prediction model Bi-GRU to construct a feature extraction network DAM-Bi-GRU. Fourthly, the ISSA is utilized to optimize the hyperparameters of DAM-Bi-GRU. Finally, the proposed model is trained, validated, and tested using water quality and meteorological parameter data collected from a self-built LoRa+5G-based marine ranching aquaculture monitoring system. The results show that: (1) Compared with the baseline model Bi-GRU, the addition of PCA, ISSA and DAM module can effectively improve the prediction performance of the model, and their fusion is effective; (2) ISSA demonstrates superior capability in optimizing model hyperparameters and convergence speed compared to traditional methods such as standard SSA, genetic algorithm (GA), and particle swarm optimization (PSO); (3) The proposed hybrid model achieves a root mean square error (RMSE) of 0.2136, a mean absolute percentage error (MAPE) of 0.0232, and a Nash efficient (NSE) of 0.9427 for DO prediction, outperforming other similar data-driven models such as IBAS-LSTM and IDA-GRU. The prediction performance of the model meets the practical needs of precise DO prediction in aquaculture.

KEYWORDS

marine ranching, dissolved oxygen prediction, improved sparrow search algorithm (ISSA), dual attention mechanism, Bi-GRU

1 Introduction

As one of the crucial indicators of water quality, dissolved oxygen directly determines the health status of the water environment in marine ranching, and then affects the overall aquaculture benefits. Its concentration is influenced by factors such as air temperature, atmospheric pressure, and water body conditions, exhibiting nonlinear, coupled, and time-varying characteristics (Cuenco et al., 1985; Lipizer et al., 2014). When the DO concentration in water is too high or insufficient, it can directly or indirectly alter other water quality indicators, affecting the health status of aquacultured species, leading to decreased resistance, slow growth, stagnation, or even death (Abdel-Tawwab et al., 2019; Neilan and Rose, 2014; Jiang et al., 2021). Therefore, through real-time monitoring and effective prediction of DO concentration in water aquaculture, precise regulation of the water quality environment can be achieved, reducing the aquaculture risks in marine farms and enhancing their economic benefits.

Currently, artificial intelligence technology is widely used for modeling complex nonlinear systems (Zhu et al., 2019; Choi et al., 2021; Than et al., 2021; Guo et al., 2022, 2023). Scholars have proposed various methods for water quality prediction in different environments and achieved certain results. Wu et al. (2018) used a BP neural network model optimized by particle swarm optimization (PSO) for dissolved oxygen prediction. Zhu et al. (2017) established a dissolved oxygen prediction model based on the least squares support vector regression (LSSVR) model and fruit fly optimization algorithm (FOA). Li et al. (2023) applied a prediction model combining PCA with particle swarm optimization-based LSSVM to dissolved oxygen prediction in the Yangtze River Basin in Shanghai. Kuang et al. (2020) proposed a hybrid DO prediction model KIG-ELM consisting of K-means, improved genetic algorithm (IGA), and extreme learning machine (ELM). Cao et al. (2021a) proposed a method based on k-means clustering, PSO, and an improved soft ensemble extreme learning machine (SELM). The BP, SVM, LSSVM, and ELM prediction methods mentioned above all belong to shallow machine learning models. They have fast training speeds and can achieve high accuracy, but their representation capabilities for complex functions are limited under limited samples and computing units. Their generalization ability for complex classification problems is also constrained to a certain extent.

Additionally, scholars have also proposed an adaptive network-based fuzzy inference system (ANFIS), which combines the characteristics of fuzzy logic and neural networks. By learning the fuzzy rules and weight parameters from data, ANFIS can predict unknown data. Sharad et al. (2018) introduced two data-driven adaptive neuro-fuzzy systems: fuzzy C-means and ANFIS based on subtractive clustering, which were used to predict sensitive parameters in monitoring stations that could lead to changes in existing water quality index values. Arora and Keshari (2021) employed ANFIS with grid partitioning (ANFIS-GP) and subtractive clustering (ANFIS-SC) to simulate and predict high-dimensional river characteristics. The results showed that both ANFIS models could fully and accurately predict DO. However, ANFIS lacks adaptability, precise control over complex systems, and may encounter high computational complexity when dealing with complex problems.

In recent years, the development of deep learning models has provided an effective solution for the prediction of dissolved oxygen in aquaculture. Deep learning can achieve complex function approximation by learning a deep nonlinear network structure and mine the implicit information in data. Compared with machine learning methods with shallow structures, it has stronger learning and generalization abilities and demonstrates a strong ability to learn the essential features of data sets from a small number of samples. Among them, the recurrent neural network (RNN) based on deep learning, as a powerful tool for modeling sequential data, has received widespread attention and application. By introducing a recurrent structure within the network, RNN can model the temporal dependencies in sequential data, thereby capturing temporal dependencies and contextual information. However, due to parameter sharing and multiple multiplications, RNN is prone to the problems of gradient vanishing or gradient explosion during backpropagation, making it difficult to train the model or causing it to fail to converge. Long short-term memory (LSTM) and gated recurrent unit neural network (GRU), as the most popular variants of RNN, can effectively address the issues of gradient vanishing and gradient explosion during RNN training, and have become the mainstream for time series prediction (Li et al., 2021; Liu P. et al., 2019). Compared to LSTM, GRU consists of an update gate and a reset gate with simpler structure and fewer number of hyperparameters. Liu Y. et al., (2019) conducted research on short-term and long-term DO predictions using attention-based RNN, indicating that the proposed model outperformed five attention-based RNN methods and five baseline methods. Zhang et al., 2020 introduced a DO prediction model, kPCA-RNN, which combines Kernel PCA and RNN demonstrating that the model's prediction performance surpassed current feedforward neural networks (FFNNs), support vector regression (SVR), and general regression neural networks (GRNN). Sun et al., 2021 proposed a DO prediction model that integrates an improved beetle antennae search algorithm (IBAS) with LSTM networks. Cao et al. (2021b) proposed a LSTM prediction model based on K-means clustering and improved particle swarm optimization (IPSO). Huan et al., 2022 systematically discussed and compared GRU water quality prediction methods based on the attention mechanism. The results showed that its performance in DO prediction surpassed that of LSTM based on the attention mechanism, as well as five traditional baseline algorithms: ANFISR, BF-AN, ELM, SVR, and ANN. However, only the feature attention mechanism was utilized in their study. Chen et al. (2022) established an attention-based LSTM model (AT-LSTM) to predict water quality in the Burnett River in Australia. The research findings indicated that the incorporation of the attention mechanism enhanced the prediction performance of the LSTM model. Only the temporal attention mechanism was used in their study. Tan et al. (2022) constructed a neural network model combining CNN and LSTM to predict DO demonstrating that this model achieved more accurate peak fitting predictions than traditional LSTM models. Yang and Liu (2022) utilized an improved whale optimization algorithm (IWOA) to optimize a

GRU, creating a water quality prediction model for sea cucumber aquaculture. Experimental results showed that this model surpassed prediction models such as Support Vector Regression (SVR), Random Forest (RF), CNN, RNN, and LSTM networks in terms of prediction accuracy and generalization performance. Jiange et al. (2023) proposed a prediction model combining improved grey relational analysis (IGRA) with LSTM optimized by the ISSA named IGRA-ISSA-LSTM. Results indicated that the proposed model achieved higher determination coefficients (R2) for predicting DO, pH, and KMnO4 compared to the IGRA-BP, IGRA-LSTM, and IGRA-SSA-LSTM models. Zhang et al. (2023) introduced an DO spatio-temporal prediction model based on an improved RGU with a dual attention mechanism (IDA-GRU) and an improved inverse distance weighting (IIDW) interpolation algorithm.

Existing research has shown that various models can be employed for DO prediction, with deep learning-based models outperforming shallow machine learning models and ANFIS. The critical aspects of building an efficient and accurate DO prediction model focus on preprocessing of input data, model selection and improvement and hyperparameter optimization (Wang et al., 2023). Based on these findings, this paper proposes an hybrid model, named PCA-ISSA-DAM-Bi-GRU, to predicting DO in marine aquaculture farms. Specifically, PCA is utilized for dimensionality reduction of the model input data, while the DAM integrating both temporal and feature attention, is fused with the bidirectional gated recurrent unit (Bi-GRU) neural network for feature extraction. Furthermore, an enhanced ISSA incorporating multiple strategies is employed to search and optimize the hyperparameters of the Bi-GRU, aiming to enhance the model's prediction precision. Finally, the accuracy and reliability of the model are validated using data collected from a self-built LoRa+5G-based marine aquaculture farm monitoring system.

2 Materials and methods

2.1 Marine ranching environment monitoring system based on LoRa+5G

This experiment has independently established a marine ranching environment monitoring system based on LoRa+5G, which integrates functions such as data collection, remote transmission, storage management, remote monitoring, and data analysis. The overall architecture is shown in Figure 1 and can be functionally divided into a perception layer, a network layer, and an application layer. The perception layer utilizes various sensors to collect water quality parameters and meteorological parameters. The network layer transmits the collected data to the application layer through the LoRa sensor network combined with 5G communication technology. The application layer stores and analyzes the collected data, providing a user interface as needed.

For this experiment, the monitoring system was deployed at an aquaculture farm in Xiayang Town, Xuwen County, Zhanjiang City, Guangdong Province, China, covering a sea area of 40m in length and 40m in width. To collect three-dimensional distribution data of the aquaculture area, nine water quality sensors were placed at corresponding locations above and below water depths of 0.8m and 1.6m. The monitor point distribution is shown in Figure 2.

The data collected by the water quality sensors include dissolved oxygen, water temperature, conductivity, pH value, ammonia nitrogen content, and turbidity. The meteorological monitoring station, located near the aquaculture farm, gathers data on atmospheric temperature, atmospheric relative humidity, atmospheric pressure, wind speed, wind direction, solar radiation, and rainfall. During the data collection process, factors such as the aquaculture environment, sensor malfunctions, and fluctuations in network signals can lead to the presence of abnormal values and a small number of missing values in the sample data. In this study, the mean smoothing method is adopted to eliminate abnormal data, and the linear interpolation method is used to fill in missing values. Additionally, a min-max normalization process is applied to each variable to ensure consistent scaling for analysis.

2.2 Construction of dissolved oxygen prediction model

2.2.1 Principal component analysis

On the basis of ensuring the integrity, validity, and accuracy of the input data, dimensionality reduction can be applied to eliminate redundancy in the input data, effectively reduce the complexity of the model structure, and enhance the model's learning performance and prediction accuracy. Principal Component Analysis (PCA) is a commonly used data analysis method that transforms data from a high-dimensional space to a low-dimensional space. It recombines numerous indicators with certain correlations into a new set of uncorrelated comprehensive indicators, thereby achieving the goals of removing redundant information and noise reduction. Assuming the input raw data is in the form of a matrix, the specific steps for PCA to extract the principal components are as follows:

1. Data Decentralization: subtract the mean of each feature from itself $X'_{ij} = X_{ij} - \bar{X}_i$;
2. Compute the Covariance Matrix: $X'_{ij}X'^{T}_{ij}$;
3. Calculate Eigenvalues and Eigenvector;
4. Select Principal Components: sort the eigenvalues from largest to smallest and select the top k eigenvalues;
5. Construct Projection Matrix: combine the eigenvectors corresponding to the selected eigenvalues to form the projection matrix;
6. Dimensionality Reduction: multiply the original matrix by the projection matrix to obtain a new set of samples that retains most of the representative feature information from the original samples.

2.2.2 Bi-directional gated recurrent unit neural network

The GRU network is a simplified variant of the LSTM network. It consists of an update gate and a reset gate, resulting in a simpler structure with fewer hyperparameters. GRU networks take

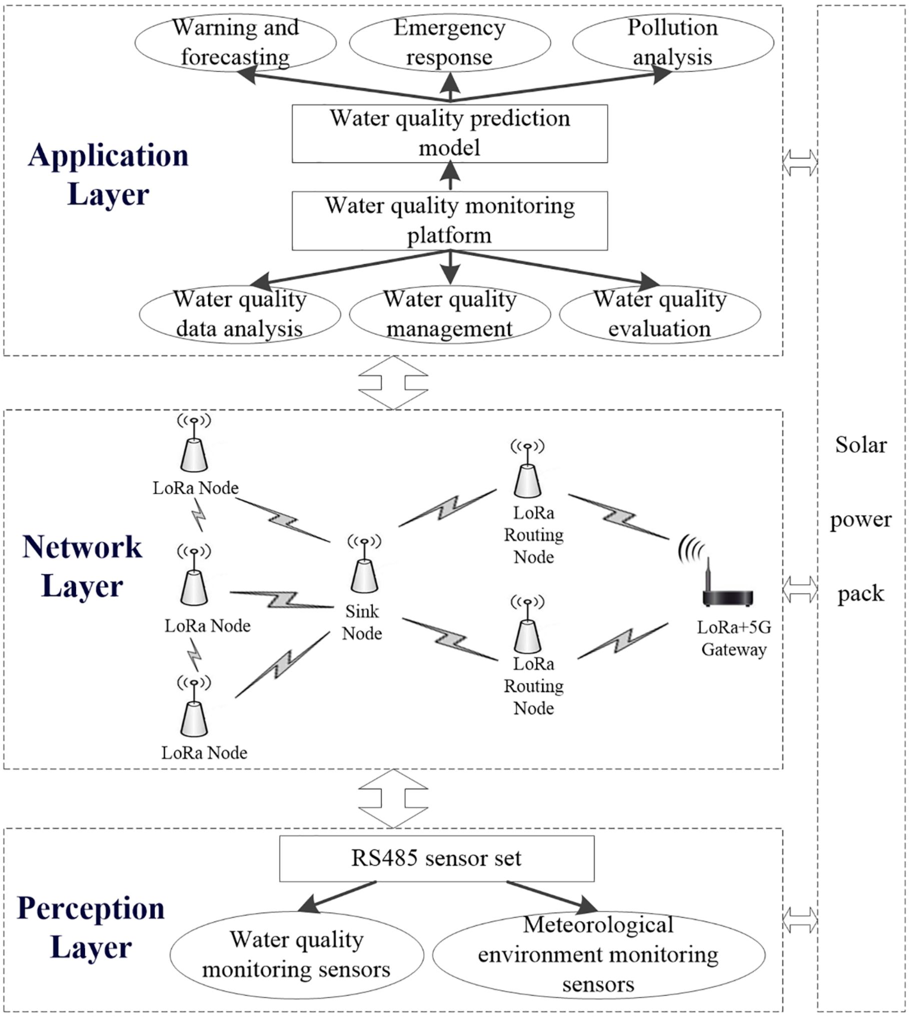


FIGURE 1
Overall structure of the aquaculture environmental monitoring system.

sequential data as input and utilize recurrent convolutional neural networks for feature extraction, making them well-suited for time series prediction. The specific structure of the GRU network cycle unit is illustrated in Figure 3. The input of the network unit includes the current input x_t and the hidden state h_{t-1} passed down from the previous time step. The output is both the output for the current time step and the hidden state h_t passed to the next time step. The specific calculation process is described by Equations 1–4:

$$r_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1} + b_r) \tag{1}$$

$$z_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1} + b_z) \tag{2}$$

$$h_t^{'} = \tanh(W_{hx}x_t + W_{hr}r_t h_{t-1} + b_h) \tag{3}$$

$$h_t = (1 - z_t)h_{t-1} + z_t h_t^{'} \tag{4}$$

where r_t, z_t, $h_t^{'}$ and h_t represent the output of the reset gate, the output of the update gate, the candidate state, and the hidden state, respectively. W_{rx} and W_{rh} are the weight matrices of the reset gate, W_{zx} and W_{zh} are the weight matrices of the update gate, and W_{hx} and W_{hr} are the weight matrices of the candidate output. b_r, b_z and b_h are the bias vectors for the reset gate, the update gate, and the candidate output, respectively. σ and tanh denote the sigmoid activation function and the hyperbolic tangent function, respectively.

Since GRU can only establish unidirectional associations in time series, the concentration of dissolved oxygen at a given moment should be related to both the preceding and following water quality and meteorological factors. The bidirectional GRU (Bi-GRU) can simultaneously mine the sequential correlation and reverse correlation of the time series, and comprehensively extract the timing features. Therefore, this study employs bi-directional GRU (Bi-GRU), which simultaneously explores the sequential and inverse sequential correlations in the time series, comprehensively extracting temporal features. The Bi-GRU network comprises two independently and symmetrically structured GRUs with identical inputs but opposite information transmission directions. The outputs from these two GRUs, which are independent and do not interact with each other, are concatenated to form the output for each time step, as shown in Figure 4.

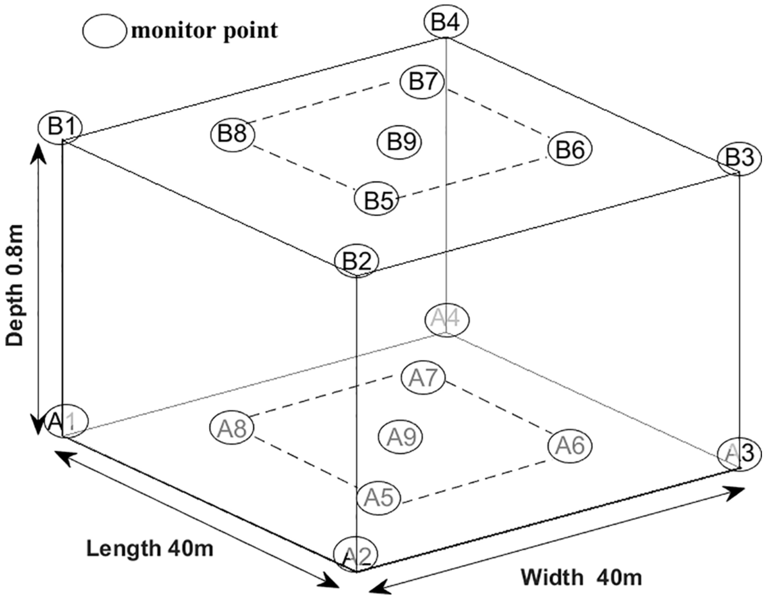


FIGURE 2
The distribution of monitor points.

2.2.3 Dual attention mechanism

The attention mechanism in deep learning is a biomimetic technique that mimics the selective attention behavior in human reading, listening and speaking. Integrating attention mechanisms into neural network can make it autonomously learn and pay more attention to the important information in model input, and enhances the model's feature extraction capabilities, robustness, and generalization ability by assigning different weights to the model's inputs. In the DO prediction, the importance of each environmental factor is different, and the influence weight of the same environmental factor on DO concentration at different time points is also different. Furthermore, environmental factors at different historical moments have different importance in influencing current DO concentrations. Therefore, in this study, a feature attention mechanism is introduced at the Bi-GRU encoder stage to adaptively assign weights to different environmental factors at each time step. This mechanism enables the model to focus on the most influential factors for DO prediction. Additionally, a temporal attention mechanism is introduced at the decoder stage of the fully connected layer to dynamically adjust the weights of different time steps' influence on the current DO concentration, so as to better capture the key information in the time series data. The combination of these two attention mechanisms allows for a more comprehensive and nuanced understanding of the complex relationships between environmental factors and DO concentrations over time.

The feature attention mechanism in the encoder utilizes multi-layer perceptron operations to quantify the feature attention weights, as illustrated in Figure 5. Its input comprises *n*

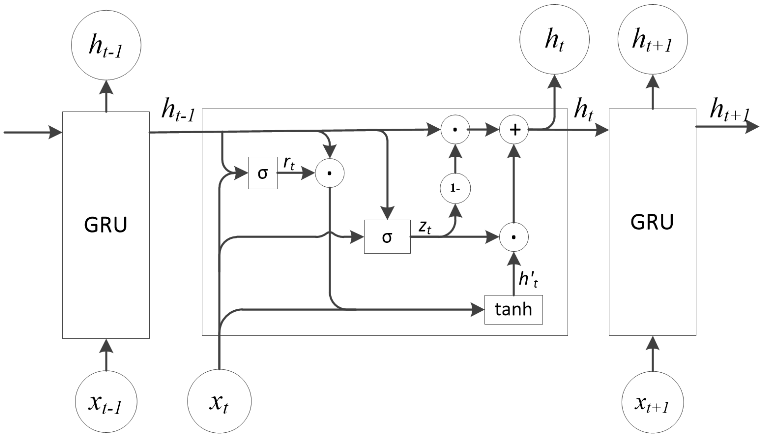


FIGURE 3
Basic structure of GRU.

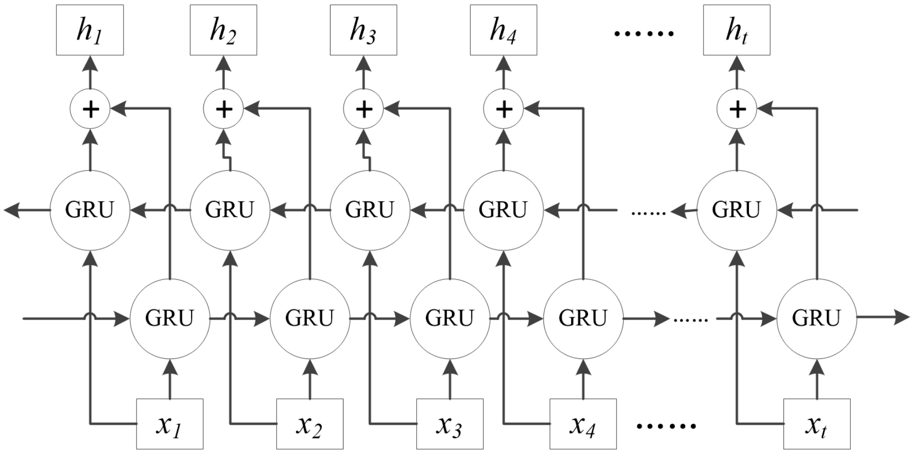


FIGURE 4
Bi-GRU network structure.

environmental feature vectors $\boldsymbol{x}_t = (x_t^1, x_t^2, \cdots, x_t^n)$ at time t and the hidden layer state $h_{t\text{-}1}$ output by the encoder at the previous time step. The output is the attention weight of each feature at this time step $\boldsymbol{\alpha}_t = (\alpha_t^1, \alpha_t^2, \cdots, \alpha_t^n)$, where α_t^k assesses the importance of the k-th feature. Subsequently, the updated $\tilde{\boldsymbol{x}}_t = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \cdots, \alpha_t^n x_t^n)$ is employed as the encoder input for time t. The specific calculation process is outlined in Equations 5 and 6:

$$r_t^k = \boldsymbol{V}_r^T \tanh(\boldsymbol{W}_r h_{t-1} + \boldsymbol{U}_r \boldsymbol{x}^k + \boldsymbol{b}_r) \tag{5}$$

$$\alpha_t^k = \mathrm{softmax}(r_t^k) = \frac{\exp(r_t^k)}{\sum_{i=1}^{n} \exp(r_t^i)} \tag{6}$$

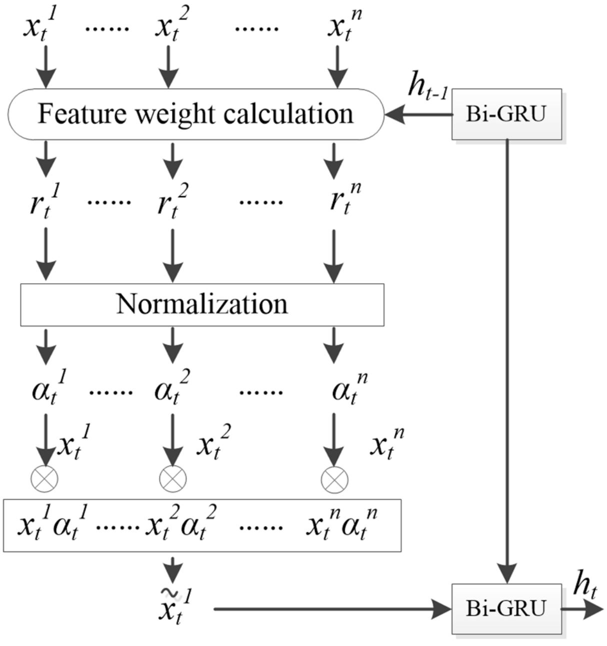

x_t^1 ...... x_t^2 ...... x_t^n
Feature weight calculation
h_{t-1}
Bi-GRU
r_t^1 ...... r_t^2 ...... r_t^n
Normalization
α_t^1 ...... α_t^2 ...... α_t^n
x_t^1
x_t^2
x_t^n
$x_t^1\alpha_t^1$ ......$x_t^2\alpha_t^2$ ...... $x_t^n\alpha_t^n$
$\tilde{x}_t^1$
Bi-GRU
h_t

FIGURE 5
Structural diagram of the feature attention mechanism.

where $\boldsymbol{V}_r^T$, $\boldsymbol{W}_r$ and $\boldsymbol{U}_r$ represents the network feature weights that need to be learned, and $\boldsymbol{b}_r$ is the bias parameters. The softmax function is applied for normalization, ensuring that the sum of all weights equals 1.

The temporal attention mechanism structure in the decoder is illustrated in Figure 6. Take the encoder's historical hidden state $H = (\boldsymbol{h}_1, \cdots, \boldsymbol{h}_t \cdots, \boldsymbol{h}_T)$ and the decoder's hidden layer state at the previous moment $d_{t\text{-}1}$ as the input of the temporal attention mechanism to obtain the temporal attention weight coefficient $\boldsymbol{\beta}_t = (\beta_t^1, \beta_t^2, \cdots, \beta_t^T)$ at the current moment. β_t^k represents the influence of the hidden layer state at the k-th layer on the DO prediction at the current moment. By weighted summing the β_t^k with the corresponding hidden layer state h_k, the comprehensive information of the predicted time series features could be obtained:. The calculation process is shown in Equations 7 and 9:

$$l_t^k = \boldsymbol{V}_d^T \tanh(\boldsymbol{W}_d d_{t-1} + \boldsymbol{U}_d h_k + \boldsymbol{b}_d) \tag{7}$$

$$\beta_t^k = \mathrm{softmax}(l_t^k) = \frac{\exp(l_t^k)}{\sum_{i=1}^{T} \exp(l_t^i)} \tag{8}$$

$$c_t = \sum_{k=1}^{t} \beta_t^k h_k \tag{9}$$

where $\boldsymbol{V}_d^T$, $\boldsymbol{W}_d$ and $\boldsymbol{U}_d$ represents the network feature weights that need to be learned, and $\boldsymbol{b}_d$ is the bias parameters. The softmax function is applied for normalization, ensuring that the sum of all weights equals 1.

Fuse the dissolved oxygen yt with ct as the input to the GRU network:

$$\tilde{y}_t = \tilde{\boldsymbol{W}}^T [y_t, c_t] + \tilde{b} \tag{10}$$

where $\tilde{\boldsymbol{W}}^T$ and $\tilde{b}$ represents the weights and biases for the fused input to the GRU neural network.

The hidden state after incorporating the temporal attention mechanism is updated using Equation 11:

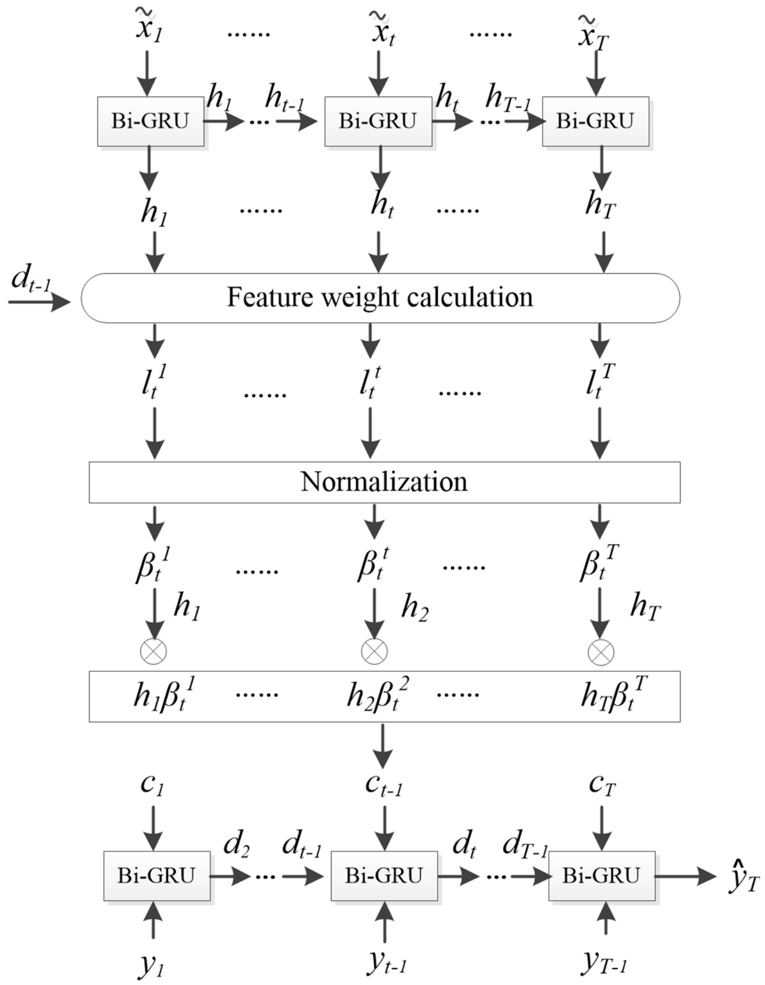


FIGURE 6
Structural diagram of the temporal attention mechanism.

$$d_t = f_1(d_{t-1}, \tilde{y}_{t-1}) \tag{11}$$

The predicted value of the dissolved oxygen to be predicted is:

$$\begin{aligned}\tilde{y}_{T+1} &= F(y_1, y_2, \ldots y_T, x_1, x_2, \ldots x_T) \\ &= V_y^T(W_y[d_T, c_T] + \boldsymbol{b}_w) + b_y\end{aligned} \tag{12}$$

where W_y and b_w are the weights and biases of the GRU network, respectively; while V_y^T and b_y are the weights and biases of the entire network, respectively.

2.2.4 Improved sparrow search algorithm

The hyperparameters of neural network models affect the structure, topology, and details of the training process, which in turn impact the learning process and performance of the models. Traditionally, the setting of Bi-GRU hyperparameters often relies on trial and error based on experience, leading to poor stability, susceptibility to overfitting and underfitting, and time-consuming processes. Existing research has demonstrated the importance of hyperparameter optimization in enhancing the robustness, generalization, stability, and accuracy of models (Sun et al., 2021; Yang and Liu, 2022; Jiange et al., 2023). There are numerous hyperparameter optimization algorithms, among which the sparrow search algorithm (SSA) proposed in 2020 is a novel swarm intelligence optimization algorithm inspired by bird foraging behavior (Xue and Shen, 2020). By simulating the foraging process of sparrows to search for optimal solutions, SSA boasts high search accuracy, fast convergence speed, and strong robustness, making it widely applicable to various optimization problems. This study proposes an improved sparrow search algorithm (ISSA) that integrates multiple strategies to search and optimize the hyperparameters of the Bi-GRU model, thereby enhancing the model's optimal learning capabilities.

SSA is a discoverer-follower model which superimposes detection and early warning mechanism. The individual who finds the best food in the sparrow acts as the discoverer, and the other individuals act as followers, and compete with the discoverer for food when the discoverer finds the better food. Additionally, a certain proportion of individuals within the population are selected as scouts to conduct reconnaissance and warning, abandoning food sources if danger is detected. Addressing the issues of insufficient population diversity, poor convergence performance, and the imbalance between global exploration and local exploitation capabilities in the standard SSA, this study proposes improvements to the algorithm from the following aspects.

2.2.4.1 Incorporating gauss chaotic sequence into population initialization

The standard SSA randomly generates the initial population, and once the population gathers, it will affect the breadth of the search space. Additionally, if a "super sparrow" (an individual with a fitness value significantly higher than the average) emerges prematurely during the iteration process, a large number of participants may converge towards it, drastically reducing the diversity of the population. To address these issues, the gauss chaotic sequence is introduced into the initialization phase of the SSA algorithm. The gauss chaotic mapping possesses properties such as regularity, randomness, and ergodicity, which can help ensure a uniform distribution of the initial population, enhancing both the diversity of the population and the global search performance of the model. The mathematical expression for the gauss chaotic mapping is given as:

$$x_{k+1} = \begin{cases} 0 & x_k = 0 \\ \frac{1}{x_k \bmod (1)}, & x_k \neq 0 \end{cases} \tag{13}$$

where "mod" represents the modulo operation.

2.2.4.2 Improving the discoverer's position update strategy by borrowing from the salp group algorithm

The position update strategy for discoverers in the standard SSA is:

$$x_{i,d}^{t+1} = \begin{cases} x_{i,d}^{t} \cdot \exp\left(-\frac{i}{\beta_1 T_{\max}}\right) & R_2 < S_T \\ x_{i,d}^{t} + \beta_2 \cdot \boldsymbol{L} & R_2 \geq S_T \end{cases} \tag{14}$$

where t represents the current iteration number; $T_{\max}$ represents the maximum number of iterations; β_1 and β_2 are random numbers, $\beta_1 \in (0,1]$ and β_2 follows a normal distribution; L is a 1×d matrix filled with 1; $R_2 \in [0,1]$,which represents the warning value; and $S_T \in [0.5,1]$ represents the safe value.

According to the Equation 14, when $R_2 < S_T$, each dimension of the position converges towards zero, leading the algorithm to easily become trapped in local optima near zero and potentially miss optimal solutions located away from zero. In order to improve the global search ability of the algorithm, this study draws on the leader's update strategy in the Salp Group Algorithm (Mirjalili et al., 2017), and modified the position update formula for the discoverer as follows:

$$x_{i,d}^{t+1} = \begin{cases} x_{i,d}^{t} \cdot \frac{c_1[(u_b - l_b)c_2 + l_b)]}{(1+c_3)u_b} & R_2 < S_T \\ x_{i,d}^{t} + \beta_2 \cdot \boldsymbol{L} & R_2 \geq S_T \end{cases} \tag{15}$$

$$c_1 = 2\exp -\left(\frac{4t}{T_{\max}}\right)^2 \tag{16}$$

In Equation 15, u_b and l_b represents the lower and upper bounds of the current dimension's search space, respectively. $c_2, c_3 \in (0,1)$ are random variables that follow a uniform distribution, and c_1 serves as a balancing parameter that regulates the trade-off between the algorithm's global search and local search capabilities. With these modifications, the SSA discoverer's position does not necessarily decrease in each dimension at the early stage of iteration, which improved the search range and global search ability of the population. Meanwhile, it also maintains a balance with the convergence speed and local search capabilities during the later iterations of the algorithm.

2.2.4.3 Improving the follower's position update strategy inspired by chicken swarm optimization

In the standard SSA, the follower's position update strategy is typically defined as follows:

$$x_{i,d}^{t+1} = \begin{cases} \beta_2 \cdot \exp\left(\frac{x_{worst}^{t} - x_{i,d}^{t}}{i^2}\right) & i > \frac{N}{2} \\ x_{p,d}^{t+1} + \left| x_{i,d}^{t} - x_{p,d}^{t+1} \right| \cdot \boldsymbol{A}^{+} \cdot \boldsymbol{L} & i \leq \frac{N}{2} \end{cases} \tag{17}$$

where $x_{p,d}^{t+1}$ refers to the best position found by the discoverer (or leader) of the swarm during the t+1-st iteration of the algorithm, and x_{worst}^{t} represents the worst position found by any individual (including both followers and the discoverer) in the current iteration or across all iterations so far. $\boldsymbol{A}^{+} = \boldsymbol{A}^{T}(\boldsymbol{A}\boldsymbol{A}^{T})^{-1}$, where A is a 1-by-$d$ matrix whose elements are randomly chosen from the set $\{1, -1\}$.

According to Equations 17, when $i \leq \frac{N}{2}$, the follower's position update is primarily guided by the leader $x_{p,d}^{t+1}$. It is prone to rapid aggregation of the population within a short period, leading to a sharp decline in population diversity and significantly increasing the probability of the algorithm falling into a local optimum. Drawing inspiration from the random following strategy in the chicken swarm algorithm (Osamy et al., 2020), where hens converge towards roosters with a certain probability, the follower's position update strategy is improved as follows:

$$x_{i,d}^{t+1} = \begin{cases} \beta_2 \cdot \exp\left(\frac{x_{worst}^{t} - x_{i,d}^{t}}{i^2}\right) & i > \frac{N}{2} \\ x_{i,d}^{t} + S\mathrm{rand}(0,1)(x_{k,d}^{t} - x_{i,d}^{t}) & i \leq \frac{N}{2} \end{cases} \tag{18}$$

$$S = \exp(f_s - f_i) \tag{19}$$

where $k \in [1, N]$ represents the fitness of any k-th sparrow, and $k \neq i$. The improved SSA ensures both convergence and population diversity, balancing local exploitation and global search capabilities.

2.2.4.4 Introduction of Cauchy-Gaussian mutation strategy

The standard SSA is prone to falling into local optima and stagnation in the later stages of iteration due to the decrease in population diversity. Therefore, the Cauchy-Gaussian mutation strategy (Wang et al., 2020) is adopted in this study to ensure population diversity and resistance to stagnation, thereby avoiding premature convergence of the algorithm. The specific formula is as follows:

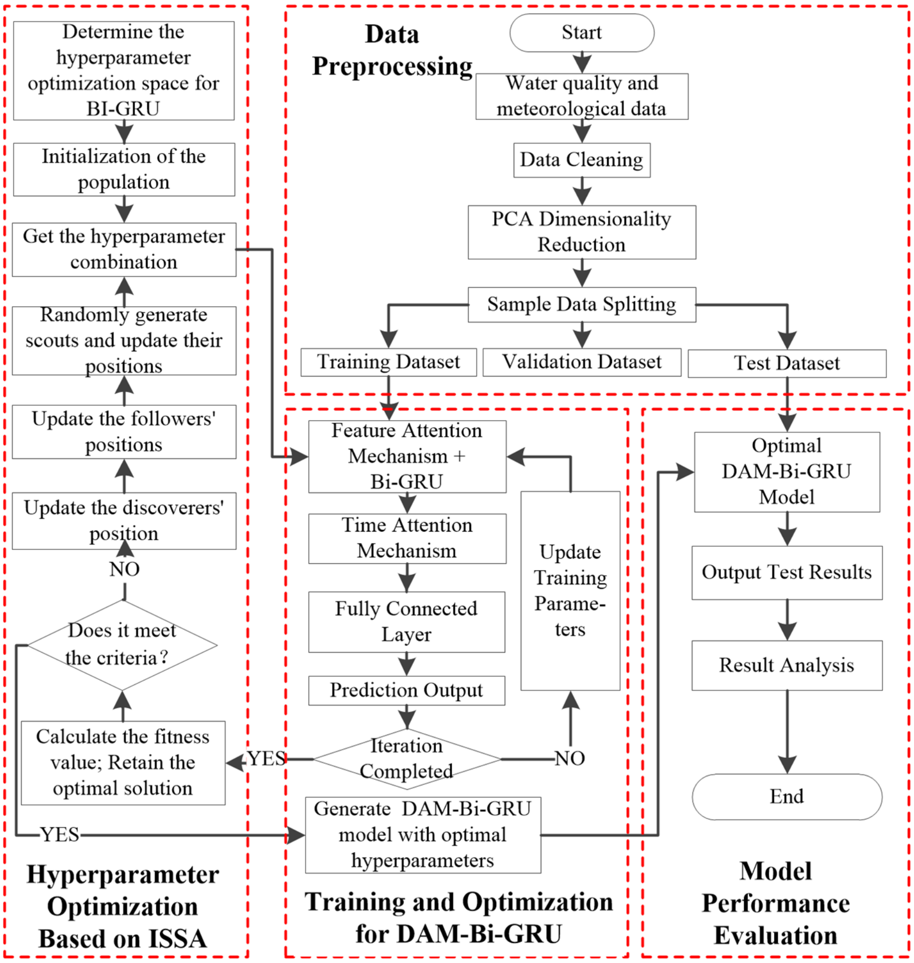

Determine the hyperparameter optimization space for BI-GRU
Initialization of the population
Get the hyperparameter combination
Randomly generate scouts and update their positions
Update the followers' positions
Update the discoverers' position
NO
Does it meet the criteria?
Calculate the fitness value; Retain the optimal solution
YES
YES
Hyperparameter Optimization Based on ISSA
Data Preprocessing
Start
Water quality and meteorological data
Data Cleaning
PCA Dimensionality Reduction
Sample Data Splitting
Training Dataset
Validation Dataset
Test Dataset
Feature Attention Mechanism + Bi-GRU
Time Attention Mechanism
Fully Connected Layer
Prediction Output
Iteration Completed
NO
Update Training Parame-ters
Generate DAM-Bi-GRU model with optimal hyperparameters
Training and Optimization for DAM-Bi-GRU
Optimal DAM-Bi-GRU Model
Output Test Results
Result Analysis
End
Model Performance Evaluation

FIGURE 7
Flowchart of DO prediction algorithm PCA-ISSA-DAM-Bi-GRU.

$$u_{best} = x_{best}[1 + \lambda_1 \mathrm{Cauchy}(0, \sigma^2) + \lambda_2 \mathrm{Gauss}(0, \sigma^2)] \tag{20}$$

$$\sigma = \begin{cases} 1, & f(x_{best}) < f(x_i) \\ \exp\left(\frac{f(x_{best})-f(x_i)}{|f(x_{best})|}\right) & f(x_{best}) \geq f(x_i) \end{cases} \tag{21}$$

In Equations 20 and 21, u_{best} represents the position of the optimal individual after mutation; σ denotes the standard deviation of the Cauchy-Gaussian mutation strategy; $\mathrm{Cauchy}(0, \sigma^2)$ is a random variable that follows a Cauchy distribution; $\mathrm{Gauss}(0, \sigma^2)$ is a random variable that follows a Gaussian distribution; $\lambda_1 = 1 - \frac{t^2}{T_{\max}^2}$ and $\lambda_2 = \frac{t^2}{T_{\max}^2}$ are dynamic parameters adaptively adjust with the number of iterations.

2.2.5 Dissolved oxygen prediction model fuse DAM and Bi-GRU optimized by ISSA

The flowchart of the ISSA-optimized DO prediction model integrating DAM and Bi-GRU proposed in this study is shown in Figure 7. The main processes include data the preprocessing based on PCA, the hyperparameter optimization conducted by ISSA, the training and optimization of the DAM-Bi-GRU model, and the evaluation of model performance.

3 Results

3.1 Data processing

To validate the performance of the proposed model in this article, data from the study area spanning 86 days from June 1st 2023 to August 25th 2023 were collected, with each data point recorded every 30 minutes, resulting in a total of 4,184 data sets for every given monitor point. The first 60 days' data were used as the training set, the next 13 days' data as the validation set, and the final 13 days' data as the test set, following a 7:1.5:1.5 ratio. For any given time *t*, the model's input comprised the aquaculture environmental parameters from the preceding 24 hours, and its output predicted the dissolved oxygen levels for the following 2 hours. This resulted in 2,832 training samples, 624 validation samples, and 624 test samples. Due to space limitations, a portion of the raw data collected on June 20th 2023 is presented in Table 1. Furthermore, taking monitor point A9 as an example, after removing outliers and filling in missing values through linear interpolation, statistical analysis was conducted on the data, as shown in Table 2. Subsequently, the PCA algorithm was applied to reduce the data's

TABLE 1A Water quality data collected by monitoring station A9 on June 20, 2023.

Time	Water quality parameters					
	Dissolved oxygen/ (mg·L^{-1})	Water temperature/°C	Conductivity/ (μS·cm^{-1})	pH value	Ammonia nitrogen/ (mg·L^{-1})	Turbidity/ NTU
06:00	5.35	27.14	2980.52	7.72	0.27	18.27
06:30	5.39	27.14	3080.74	7.72	0.27	18.29
07:00	5.47	27.14	3220.28	7.75	0.27	19.11
07:30	5.58	27.24	3170.19	7.76	0.28	19.63
08:00	5.77	27.24	3586.48	7.77	0.28	19.92
⋮	⋮	⋮		⋮	⋮	⋮
14:00	8.23	29.52	3800.46	7.82	0.38	20.35
14:30	8.41	29.64	3740.74	7.87	0.38	20.77
15:00	8.65	29.02	3826.92	7.95	0.39	21.06
15:30	8.92	30.18	3780.36	8.02	0.39	21.95
16:00	8.78	30.15	3776.62	8.11	0.41	21.84

TABLE 1B Meteorological parameter data collected by monitoring station A9 on June 20, 2023.

Time	Meteorological parameters						
	Temperature/ °C	Relative humidity/%	Pressure/KPa	Wind Speed/ (km·h^{-1})	Wind direction/°	Solar radiation/ (W·m^{-2})	Rainfall/ mm
06:00	28.46	87.38	101.42	12.25	117.75	68.45	0
06:30	28.64	87.24	101.42	14.37	127.36	60.24	0
07:00	28.91	87.41	101.42	13.96	123.95	88.90	0
07:30	29.32	86.95	101.41	16.75	131.24	120.37	0
08:00	29.75	86.23	101.41	15.33	135.78	135.66	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14:00	33.72	84.66	101.26	16.82	130.25	458.36	0
14:30	33.48	83.35	101.26	14.29	123.74	520.59	0
15:00	32.95	84.71	101.27	12.88	119.55	330.47	0
15:30	32.53	83.29	101.26	15.26	121.57	220.69	0
16:00	31.47	83.04	101.25	16.88	120.49	392.21	0

dimensionality, eliminating redundant information and noise. Finally the processed data was input into the neural network model for feature extraction. The PCA of the aquaculture environmental parameters is presented in Table 3. As can be seen, the cumulative contribution rate of the first seven components reaches 86.27%, representing the majority of environmental information. Therefore, this study selected seven principal components, utilizing PCA to reduce the original 13-dimensional data to seven dimensions.

3.2 Hyperparameter optimization and training of the model

The data, after being processed through outlier removal, linear interpolation for missing values, and principal component analysis, was input into the neural network model for hyperparameter optimization and training.

Step 1: Initialize the hyperparameters of the ISSA. The number of sparrows was set to 50, the maximum number of iterations T was

TABLE 2 Statistical results of data collected by monitoring station A9.

Category	Indicators	Mean ± SD	Range
Water quality parameters	Dissolved oxygen/(mg·L^{-1})	7.534 ± 2.175	3.29~11.64
	Water temperature/°C	27.422 ± 3.210	18.58~33.36
	Conductivity/μS·cm^{-1}	3450.463 ± 400.675	2240.45~5300.60
	pH value	7.920 ± 0.218	7.24~8.91
	Ammonia nitrogen/(mg·L^{-1})	0.324 ± 0.112	0.06~0.58
	Turbidity/NTU	20.301 ± 2.430	15.4~30.5
Meteorological Parameters	Temperature/°C	28.512 ± 5.351	22.32~34.05
	Relative humidity/%	85.638 ± 6.250	73.45~94.68
	Pressure/KPa	101.512 ± 0.782	99.25~102.07
	Wind speed/(km·h^{-1})	16.578 ± 5.530	7.00~52.00
	Wind direction/(°)	173.539 ± 56.821	22.5~360
	Solar radiation (W·m$^{-2)}$	625.537 ± 568.248	0.0~1915.00
	Rainfall/mm	2.350~8.852	0.0~38.8

100, with the proportions of producers, followers, and scouts being 70%, 10%, and 20% respectively. The safety threshold was set to 0.6, and the search space was 5-dimensional. For the two-layer Bi-GRU, the optimization range for the number of hidden neurons was [8, 128], the optimization range for the maximum number of iterations was [10, 100], the optimization range for the batch size was [16, 128], and the optimization range for the learning rate was [0.001, 0.1].

Step 2: Train the DAM-Bi-GRU model using the hyperparameter combinations provided by ISSA. Each sparrow corresponds to a set of hyperparameter combinations. The model was trained using supervised learning, with the root mean square error (RMSE) function serving as the loss function. The mathematical definition of RMSE is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{22}$$

where y_i and $\hat{y}_i$ represents the actual value and the predicted value by the model respectively, and N is the number of training samples in a batch. An end-to-end learning approach was adopted, where the neural network's weights were continuously adjusted through forward propagation and backward propagation of gradients. The iteration stops once the preset number of iterations is reached or the training objective is achieved, completing the neural network training. Ultimately, each hyperparameter combination corresponds to a trained DAM-Bi-GRU model.

Step 3: Validate the DAM-Bi-GRU models trained in Step 2 using the pre-divided validation dataset. The validation result of each trained DAM-Bi-GRU model was measured by RMSE, and the fitness of the sparrow corresponding to the set of hyperparameter combinations for that model is also evaluated using the same RMSE value.

Step 4: Determining whether the model training has concluded based on the fitness value. If it has reaches the maximum number of the presented iterations of ISSA or the optimal fitness value of the sparrow population has met the training objective, end the training and output the DAM-Bi-GRU model with the optimal parameter combination. Otherwise, update the positions of producers, followers, and scouts based on the fitness values of the sparrow population, and generate new hyperparameter combinations. Repeat Steps 2 to 4 until the training is completed.

Following the above optimization and training steps, the final results of DAM-Bi-GRU hyperparameter optimization were obtained, with the hidden neuron counts for the two-layer Bi-GRU being 46 and 72 respectively; the maximum number of iterations being 86; the batch size being 66; and the learning rate being 0.004. Furthermore, the proposed ISSA was compared with the original SSA, PSO, and GA in terms of optimization performance. The convergence of the algorithms during the iterative optimization process is illustrated in Figure 8. It can be seen that the fitness value of ISSA converges to around 0.21 after approximately 35 iterations, while SSA converges to around 0.23 after about 45 iterations, PSO converges to around 0.26 after approximately 55 iterations, and GA converges to around 0.28 after approximately 70 iterations. This indicates that the optimization ability and convergence speed of ISSA are significantly higher than those of SSA, GA, and PSO. Additionally, the fluctuating downward trend of the fitness value of ISSA in Figure 8 suggests its ability to quickly escape local optima. In contrast, the other three optimization algorithms exhibit varying degrees of stagnation.

3.3 Testing and evaluation of the model

In this study, the root mean squared error (RMSE), mean absolute percentage error (MAPE), and Nash-Sutcliffe efficient (NSE) were adopted to evaluate the predictive performance of the model. The calculation formulas are as follows:

TABLE 3 Principal component coefficient matrix of aquaculture environment parameters.

Indicators	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6	Component 7
Water temperature	0.467	−0.184	−0.114	0.147	−0.052	−0.030	−0.124
Conductivity	−0.352	−0.038	0.294	−0.132	−0.160	−0.084	−0.131
pH value	−0.278	−0.241	0.469	0.153	−0.078	−0.232	−0.094
Ammonia nitrogen	0.314	−0.296	−0.370	0.055	0.079	0.140	−0.207
Turbidity	0.114	−0.385	−0.208	−0.258	0.1411	−0.273	0.776
Temperature	0.452	0.242	−0.213	0.147	−0.037	−0.063	−0.097
Relative humidity	0.135	0.644	−0.187	−0.037	−0.087	−0.197	0.226
Pressure	−0.378	−0.229	−0.340	0.068	0.063	0.025	−0.075
Wind speed	−0.060	0.161	−0.305	−0.218	0.581	0.688	0.050
Wind direction	−0.086	0.032	−0.027	0.844	−0.055	0.280	0.416
Solar radiation	−0.306	0.197	−0.458	0.136	0.084	0.024	−0.258
Rainfall	0.018	0.281	0.031	0.249	0.761	−0.498	−0.065
eigenvalue	3.632	1.585	1.402	1.050	0.976	0.903	0.804
Contribution rate/%	30.266	13.208	11.686	8.750	8.136	7.524	6.701
Cumulative contribution rate/%	30.266	43.474	55.160	63.910	72.046	79.570	86.271

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{23}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{24}$$

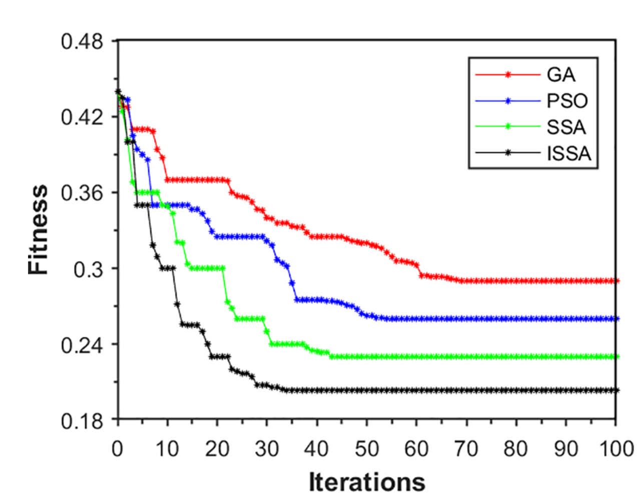


FIGURE 8
Iterative optimization and convergence curves for different optimization algorithm.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{25}$$

where y_i is the actual value, $\bar{y}_i$ is the mean of the actual values, $\hat{y}_i$ is the predicted value by the model, and N is the number of data points in the data set used for evaluating the model's performance. A lower RMSE indicates better predictive performance. MAPE measures the average magnitude of the percentage errors in a set of predictions, without considering their direction. A lower MAPE indicates better predictive accuracy. NSC ranges from negative infinity to 1, with 1 indicating a perfect match between observed and predicted values. Higher NSE values indicate better predictive performance. In summary, a lower RMSE and MAPE, and a higher NSC, all suggest better predictive performance of the model.

The 624 test data set samples were inputted one by one into the trained DAM-Bi-GRU model with the optimal combination of hyperparameters, the prediction results were obtained sequentially. The model's performance parameters on the test set were calculated by Equations 23–25, namely RMSE, MAPE, and NSE which found to be 0.2136, 0.0232, and 0.9427, respectively. Additionally, Figures 9A–D sequentially present the comparison curves of predicted and actual values for the test samples, the prediction errors, the distribution of prediction errors, and the linear fitting between predicted and actual values. From Figure 9A, it can be observed that the proposed PCA-ISSA-DAM-Bi-GRU model is capable of capturing the changing trends of real dissolved oxygen data, sensitively identifying subtle fluctuations

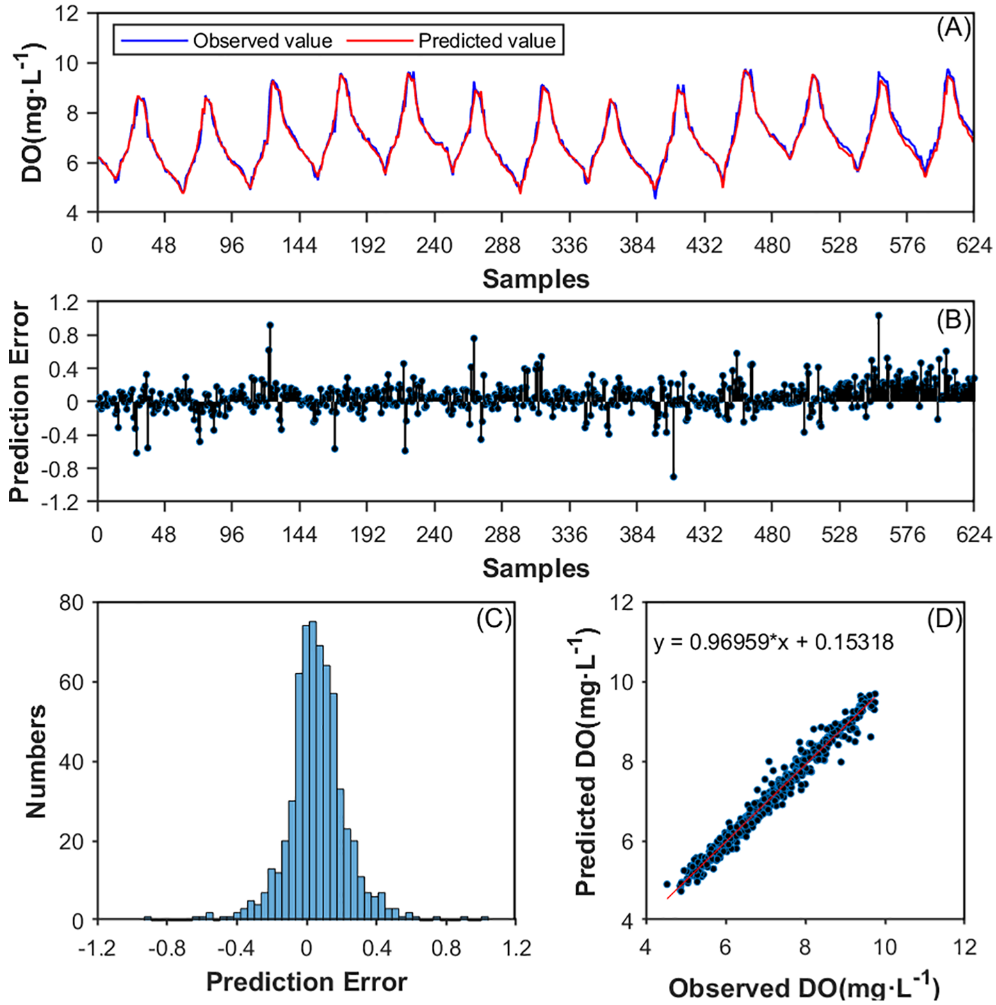


FIGURE 9
(A) DO prediction of the proposed model on the test data set. **(B)** Prediction error of the test data set; **(C)** Histogram of the prediction error distribution on the test data set; **(D)** Linear fitting between predicted and observed values.

in the data, and maintaining a high prediction accuracy. Figures 9B–D demonstrate that there is a small discrepancy between the predicted and actual values.

3.4 Comparison and analysis of the models

To analyze and evaluate the competitiveness and superiority of the proposed model, this article designed ablation experiments and comparative experiments, selecting different models to compare their predictive performance.

3.4.1 Ablation experiments

The ablation experiments were conducted in two groups, A and B. The models in Group A do not incorporate the hyperparameter optimization module ISSA, with the baseline model being Bi-GRU. The models in Group B all include the ISSA, with the baseline model being ISSA-Bi-GRU. Each group include three models: one with PCA added alone to the baseline module, one with DAM added alone, and one with both PCA and DAM added simultaneously. For experiments in Group A, the random search method was used to determine the model's hyperparameters with the number of random searches setted to be 100, which is equivalent to the maximum number of iterations for the ISSA module.

The prediction performance of each model on the test data set is shown in Table 4. In Group A, the prediction performance indicators RMSE, MAPE, and NSE of the baseline model Bi-GRU are 0.4077, 0.0527, and 0.8358, respectively. Compared with it, the PCA-Bi-GRU model shows a 9.22% decrease in RMSE, a 11.76% decrease in MAPE, and a 1.99% increase in NSE. The DAM-Bi-GRU model exhibits a 18.42% reduction in RMSE, a 28.27% reduction in MAPE, and a 5.56% increase in NSE. The PCA-DAM-Bi-GRU model, on the other hand, demonstrates a 24.63% decrease in RMSE, a 40.23% decrease in MAPE, and a 8.91% increase in NSE compared to the baseline. In Group B, the prediction performance indicators RMSE, MAPE, and NSE of the base model ISSA-Bi-GRU are 0.3424, 0.0392, and 0.8682, respectively. The PCA-ISSA-Bi-GRU model shows a 4.35% decrease in RMSE, a 8.16% decrease in MAPE, and a 3.44% increase in NSC compared to it. The ISSA-DAM-GRU model

TABLE 4 Predictive performance of different models for the ablation experiments.

Group	Model	RMSE/(mg·L^{-1})	MAPE	NSE
A (Model with-out ISSA)	Bi_GRU	0.4077	0.0527	0.8358
	PCA_Bi_GRU	0.3701	0.0465	0.8524
	DAM_Bi_GRU	0.3326	0.0378	0.8823
	PCA_DAM_Bi_GRU	0.3073	0.0315	0.9103
B (Model with ISSA)	ISSA_Bi_GRU	0.3424	0.0392	0.8682
	PCA_ISSA_Bi_GRU	0.3275	0.0360	0.8981
	ISSA_DAM_Bi_GRU	0.2780	0.0323	0.9266
	PCA_ISSA_DAM_Bi_GRU	0.2136	0.0232	0.9427

exhibits an 18.81% reduction in RMSE, a 17.6% reduction in MAPE, and a 6.73% increase in NSC. The PCA-ISSA-DAM-Bi-GRU model, however, demonstrates a 37.62% decrease in RMSE, a 40.82% decrease in MAPE, and an 8.85% increase in NSE compared to the base model. This indicates that both the DAM module and the PCA module can enhance the prediction performance of the models, with the DAM module showing a more significant improvement than PCA, and their fusion being even more effective. Figures 10A–C represent the three evaluation indicators (RMSE, MAPE, and NSE) for the models in Groups A and B, respectively. It can be observed that optimizing the hyperparameters of the Bi-GRU module through the ISSA module indeed enhances the prediction performance of the models.

3.4.2 Comparative experiments

3.4.2.1 Comparison with baseline modules

To evaluate the superiority of PCA, ISSA, and Bi-GRU in enhancing prediction accuracy within the proposed model, the following comparative experiments were also conducted in this study: 1) Pearson correlation coefficient analysis was used to replace PCA, resulting in the comparative model P-ISSA-DAM-Bi-GRU; 2) ISSA was replaced with SSA, GA, and PSO, respectively, generating comparative models PCA-SSA-DAM-Bi-GRU, PCA-GA-DAM-Bi-GRU, and PCA-PSO-DAM-Bi-GRU; 3) Bi-GRU was replaced with Bi-LSTM, LSTM, and CNN, respectively, resulting in comparative models PCA-ISSA-DAM-Bi-LSTM, PCA-ISSA-DAM-LSTM, and PCA-ISSA-DAM-CNN. Eight comparative models were evaluated in total corresponding to serial numbers 1 to 8. The experimental results are presented in Table 5 and Figure 11, revealing the following: 1) The prediction performance metrics of the PCA-ISSA-DAM-Bi-GRU model are superior to those of P-ISSA-DAM-Bi-GRU, indicating that PCA outperforms the Pearson correlation coefficient analysis method in dimensionality reduction for data input in terms of dissolved oxygen prediction performance; 2) The prediction performance metrics of PCA-ISSA-DAM-Bi-GRU are superior to those of PCA-SSA-DAM-Bi-GRU, PCA-GA-DAM-Bi-GRU, and PCA-PSO-DAM-Bi-GRU, with the NSE value reaching 0.9807, demonstrating that compared to baseline approaches such as SSA, GA, and PSO, the optimization of Bi-GRU hyperparameters by ISSA results in better model fitting; 3) The prediction performance metrics of PCA-ISSA-DAM-Bi-GRU are slightly higher than those of PCA-ISSA-DAM-Bi-LSTM and significantly higher than those of PCA-ISSA-DAM-LSTM and PCA-ISSA-DAM-CNN, indicating that bidirectional neural networks enhance temporal feature extraction for contextually related time series prediction.

3.4.2.2 Comparison with existing models

Furthermore, in order to test the overall predictive performance of the proposed hybrid model PCA-ISSA-DAM-Bi-GRU, this paper also selected dissolved oxygen prediction models proposed in the

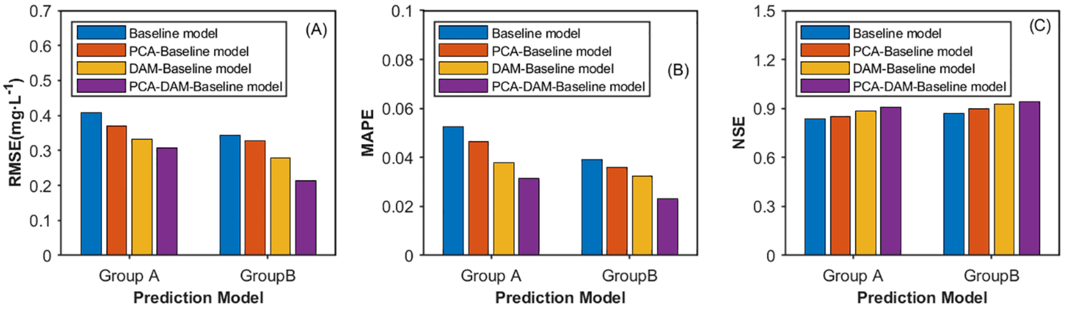


FIGURE 10
Prediction performance presented by **(A)** RMSE, **(B)** MARE and **(C)** NSE for various models in the ablation study. Group A do not incorporate attention mechanism and Group B incorporate attention Mechanism.

TABLE 5 Predictive performance of different models for the comparative experiments.

Model number	Prediction model	RMSE/(mg·L^{-1})	MAPE	NSE
1	PCA-ISSA-DAM-Bi-GRU	0.2136	0.0232	0.9427
2	P-ISSA-DAM-Bi-GRU	0.2742	0.0306	0.9294
3	PCA-SSA-DAM-Bi-GRU	0.2821	0.0317	0.9316
4	PCA-GA-DAM-Bi-GRU	0.2933	0.0346	0.9358
5	PCA-PSO-DAM-Bi-GRU	0.2928	0.0336	0.9346
6	PCA-ISSA-DAM-Bi-LSTM	0.2178	0.0292	0.9401
7	PCA-ISSA-DAM-LSTM	0.2558	0.0287	0.9395
8	PCA-ISSA-DAM-CNN	0.2931	0.0358	0.9162

past three years, namely IPSO-LSTM (Cao et al., 2021b), IBAS-LSTM (Sun et al., 2021), CNN-LSTM (Tan et al., 2022) and IDA-GRU (Zhang et al., 2023) for comparison. The results in Table 6 show that the model proposed in this paper outperforms those 4 models, indicating the effectiveness and superiority of the individual modules and their fusion in enhancing the prediction accuracy of dissolved oxygen.

3.5 Application of the model

To evaluate the practical effectiveness of the proposed model, the dissolved oxygen prediction for August 26, 2023, at the A9 monitoring station was selected as the experimental case. The prediction results and prediction error curves from the proposed PAC-ISSA-DAM-Bi-GRU model, along with the PCA-ISSA-Bi-GRU, PCA-DAM-Bi-GRU, IBAS-LSTM (Sun et al., 2021), and IDA-GRU (Zhang et al., 2023) models discussed in the previous section, are presented in Figure 12. The error value curves visually reflect the differences between the predicted curves and the actual curves, with smaller fluctuations and closer proximity to the zero-value line indicating better prediction performance. The analysis is as follows: 1) The prediction curve of the PCA-ISSA-DAM-Bi-GRU model proposed in this paper (Figures 12A, B) is closest to the actual observed values; 2) The prediction accuracy of PCA-ISSA-Bi-GRU without the dual attention mechanism (Figures 12E, F) and the IBAS-LSTM (Sun et al., 2021)model (Figures 12G, H) is significantly lower than that of the other three models, especially during the daytime. This maybe due to the factor that the dissolved oxygen is greatly affected by light intensity, and the introduced attention mechanism increases the weight of light intensity to improve the prediction accuracy; 3) The PCA-DAM-Bi-GRU model (Figures 12C, D), which do not incorporate hyperparameter optimization module ISSA, performs slightly better than IDA-GRU (Zhang et al., 2023) (Figures 12I, J), but both are significantly inferior to the PCA-ISSA-DAM-Bi-GRU model (Figures 12A, B) that incorporates ISSA for hyperparameter optimization.

As can be seen from Figure 11, the daily dissolved oxygen reaches the peak at around 15:00 and reaches the valley value at around 06:00, which can reflect the health state of the water environment to a large extent. Figure 13 presents the predicted dissolved oxygen distribution at various depths and monitoring stations at 06:00 and 15:00 on August 26, 2023, using the proposed PCA-ISSA-DAM-Bi-GRU model. The analysis of this distribution provides valuable insights into the health status of the aquatic environment. The key observations are: 1) Vertical dissolved oxygen gradient: compared Figures 13A–D, it could be concluded that within the same vertical profile, the dissolved oxygen levels at 1.6 meters depth are consistently lower than those at 0.8 meters, with this difference being more pronounced during the day compared to night. This vertical gradient is a common phenomenon in aquatic systems, where oxygen solubility decreases with depth due to factors such as temperature and pressure. 2) Spatial variations during daytime: it could be seen

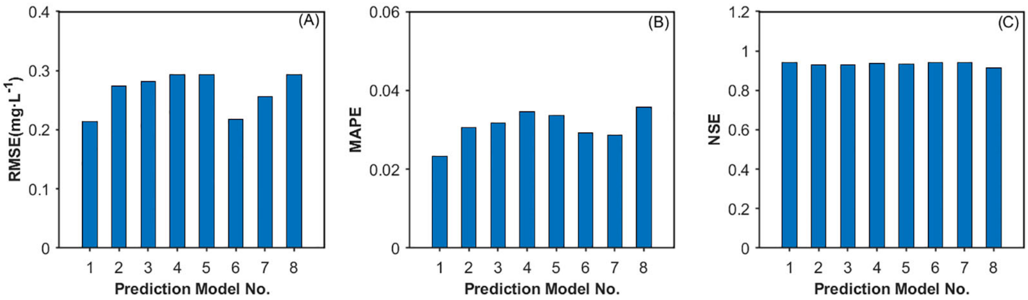


FIGURE 11
Prediction performance presented by **(A)** RMSE, **(B)** MARE and **(C)** NSE for various models in the comparative experiments.

TABLE 6 Predictive performance of existing models.

Model	RMSE/ (mg·L^{-1})	MAPE	NSE
PCA-ISSA-DAM-Bi-GRU	0.2136	0.0232	0.9427
IPSO-LSTM (Cao et al., 2021b)	0.3861	0.0492	0.8635
IBAS-LSTM (Sun et al., 2021)	0.3528	0.0426	0.8724
CNN-LSTM (Tan et al., 2022)	0.3495	0.0358	0.8631
IDA-GRU (Zhang et al., 2023)	0.3128	0.0327	0.9084

from Figure 13A that during the day, the dissolved oxygen concentration in regions A1, A2, and A4 is higher than in A3 and A7. This can be attributed to various factors, including wind direction, water temperature, and the photosynthetic activity of aquatic plants (e.g., phytoplankton). Favorable wind conditions can enhance mixing and oxygenation, while increased photosynthetic activity during daylight hours releases oxygen into the water. 3) Spatial variations during nighttime: it could be seen from Figure 13C that The distribution of dissolved oxygen at night is influenced by different factors, such as the aggregation patterns of fish schools, wind direction, and the location of feeding devices. Notably, the dissolved oxygen levels in regions A1 and A9 are higher, while those in A7 and A8 are lower. This can be explained by the possible concentration of fish schools or the efficiency of oxygen replenishment mechanisms in these areas. Additionally, the reduced photosynthetic activity at night leads to a general decrease in dissolved oxygen levels across all regions.

The observed diurnal and spatial variations in dissolved oxygen concentrations highlight the complexity of aquatic ecosystems and the importance of accurate monitoring and prediction. The PCA-ISSA-DAM-Bi-GRU model, by capturing these dynamic changes, provides a powerful tool for assessing the health of aquaculture systems and informing management decisions aimed at optimizing conditions for fish growth and welfare.

4 Discussion

4.1 Optimization mechanism of ISSA

As can be observed from Figure 8, the proposed ISSA in this study exhibits superior capability in optimizing model hyperparameters and convergence speed compared to the original SSA, GA, and PSO. Table 5 further indicates that the DO prediction performance of Bi-GRU optimized by ISSA is superior to that optimized by SSA, GA, and

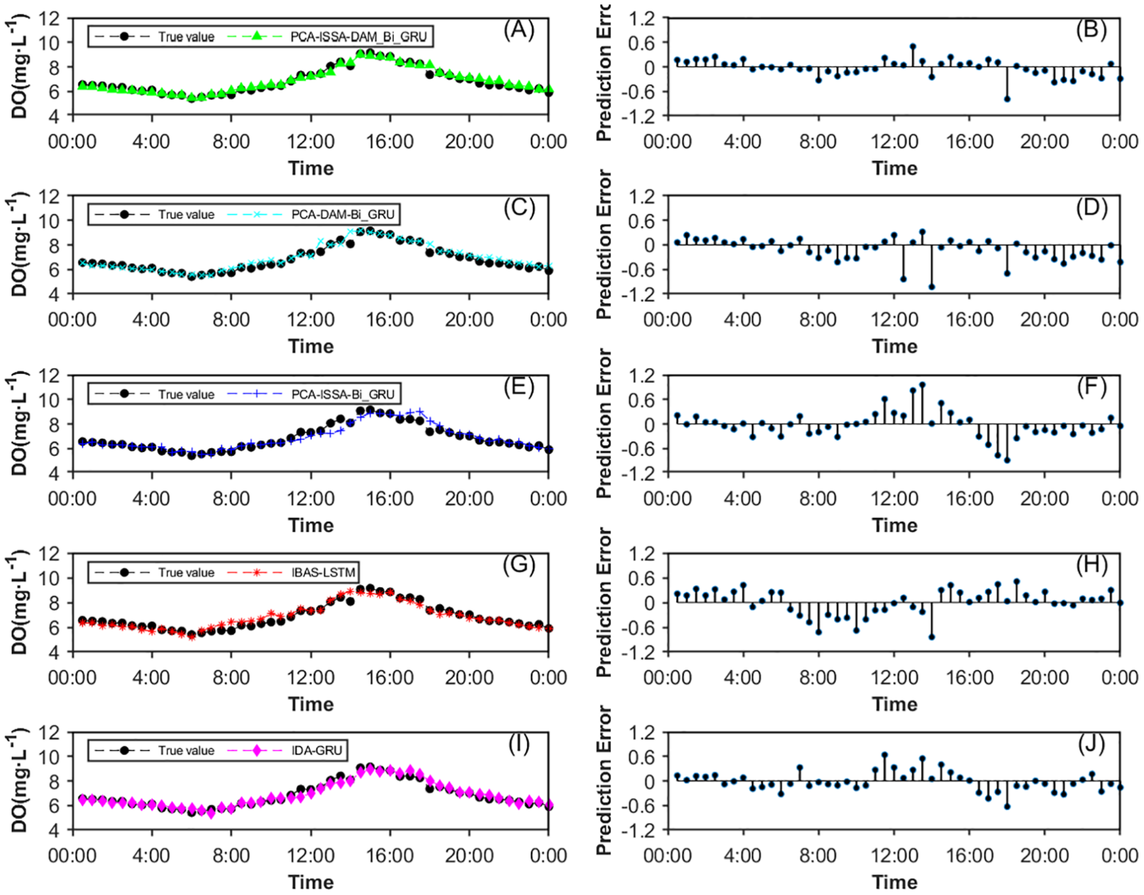


FIGURE 12
The prediction results and prediction error curves from five models on August 26, 2023. **(A, B)** PCA-ISSA-DAM-Bi-GRU model; **(C, D)** PCA-ISSA-Bi-GRU; **(E, F)** PCA-DAM-Bi-GRU; **(G, H)** IBAS-LSTM (Sun et al., 2021); **(I, J)** IDA-GRU (Zhang et al., 2023).

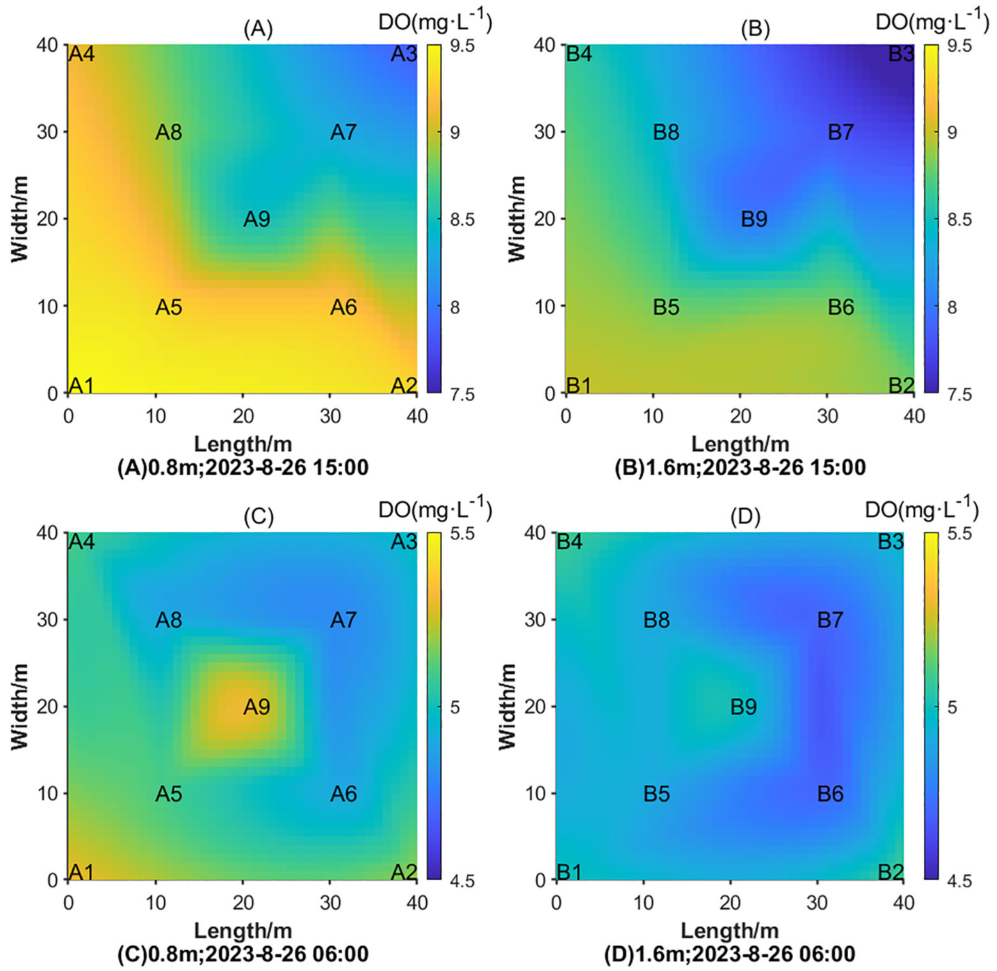


FIGURE 13
Dissolved oxygen distribution on different time at different water layers on August 26th 2023. **(A)** Dissolved oxygen distribution at a depth of 0.8 meters on 15:00; **(B)** Dissolved oxygen distribution at a depth of 1.6 meters on 15:00; **(C)** Dissolved oxygen distribution at a depth of 0.8 meters on 06:00; **(D)** Dissolved oxygen distribution at a depth of 1.6 meters on 06:00.

PSO. The optimization capability and convergence speed of SSA are primarily influenced by factors such as population diversity, global search performance, and local search ability. ISSA employs a multi-strategy fusion approach for improvement, which not only enhances the diversity and quality of the initial population but also fully utilizes information exchange among sparrow individuals to achieve a balance between local exploitation and global search in the algorithm. Additionally, it improves the algorithm's ability to escape from local extrema. Firstly, the introduction of Gauss chaotic sequence into the population initialization process ensured a uniform distribution of the initial population, thereby enhancing population diversity and the global search performance of the model. Secondly, the improvement of the position update strategy for discoverers by drawing inspiration from the Salp Swarm Algorithm allowing the discoverers to not necessarily decrease in every dimension during the early iterations, enhancing the search range and global search capability of the population while also maintaining the convergence speed and local search ability during the later iterations of the algorithm. Furthermore, the improvement of the position update process for followers by adopting the random following strategy from the Chicken Swarm Optimization (CSO) algorithm, where hens converge towards roosters with a certain probability. This ensures both convergence and population diversity, balancing local exploitation and global search. Lastly, the introduction of the Cauchy-Gaussian mutation strategy maintains population diversity and resistance to stagnation, preventing premature convergence of the algorithm.

4.2 Optimization effects of each module in the proposed PAC-ISSA-DAM-Bi-GRU

Based on ablation and comparison experiments, the analysis of the optimization effects of each module in the PCA-ISSA-DAM-Bi-GRU model on DO prediction is as follows: 1) ISSA can optimize the hyperparameters of the neural network model, thereby enhancing its prediction performance for the factor that hyperparameters control the structure, topology, and training process of the network, directly impacting the model's fitting degree, generalization ability, and stability during training. 2) Dimensionality reduction of data using PCA can improve model performance, and the effect is superior to that of the Pearson correlation coefficient analysis method. This is because the Pearson correlation coefficient analysis method only selects factors with high correlation coefficients with dissolved oxygen as inputs, completely ignoring factors weakly correlated with dissolved oxygen. In contrast, the PCA analysis method used in this study can capture 86.27% of water quality and meteorological information with only 7 dimensions of data. While reducing the dimensionality, it ensures that the input information is more complete and comprehensive, facilitating subsequent feature extraction. 3) The DAM module introduces a dual attention mechanism combining feature and temporal attention. The feature attention mechanism adaptively assigns weights to different environmental factors at each time point, while the temporal attention mechanism dynamically adjusts the weights of different time steps on the current DO concentration. This enables the neural network to better capture critical information in time series data. 4) The prediction performance of Bi-GRU is significantly higher than that of LSTM and CNN. This is because the dissolved oxygen concentration at a particular moment is correlated with environmental factors both before and after it. Bi-GRU can simultaneously explore the sequential and inverse correlations in time series, comprehensively extracting temporal features.

4.3 Competitiveness and superiority compared to existing models

4.3.1 Comparison with IPSO-LSTM and IBAS-LSTM

Both the IPSO-LSTM (Cao et al., 2021b) and IBAS-LSTM (Sun et al., 2021) models employed modified optimization algorithms, IPSO and IBAS, respectively, to optimize the hyperparameters of LSTM networks. In contrast to the PCA-ISSA-DAM-Bi-GRU model proposed in this paper, neither of these models performed PCA dimensionality reduction nor incorporates the feature and temporal attention mechanism DAM. Firstly, an ISSA-Bi-GRU model was constructed, and experiments revealed that its prediction performance was slightly higher than that of IPSO-LSTM (Cao et al., 2021b) and IBAS-LSTM (Sun et al., 2021), as shown in Table 7. This demonstrates the superiority of the ISSA and Bi-GRU modules proposed in this paper. Therefore, the optimization capabilities and convergence speeds of ISSA, IPSO, and IBAS were compared in this paper. As shown in Figure 14 and significantly higher than those of IPSO. This demonstrates that the ISSA, with its enhanced search mechanisms and adaptive parameter adjustments, exhibits superior performance in finding optimal solutions and converging towards them efficiently, compared to the other two algorithms. Furthermore, PCA-IPSO-DAM-LSTM and PCA-IBAS-DAM-LSTM were constructed based on IPSO-LSTM (Cao et al., 2021b) and IBAS-LSTM (Sun et al., 2021), respectively. Significant improvements in prediction performance were observed as shown in Table 7, thoroughly validating the effectiveness of PCA and DAM proposed in this paper in enhancing the predictive capabilities of the models.

TABLE 7 Predictive performance of various models.

Model	RMSE/ (mg·L^{-1})	MAPE	NSE
ISSA-Bi-GRU	0.3424	0.0392	0.8682
IPSO-LSTM (Cao et al., 2021b)	0.3861	0.0492	0.8635
IBAS-LSTM (Sun et al., 2021)	0.3528	0.0426	0.8724
PCA-IPSO-DAM-LSTM	0.3082	0.0397	0.8963
PCA-IBAS-DAM-LSTM	0.2762	0.0324	0.9178

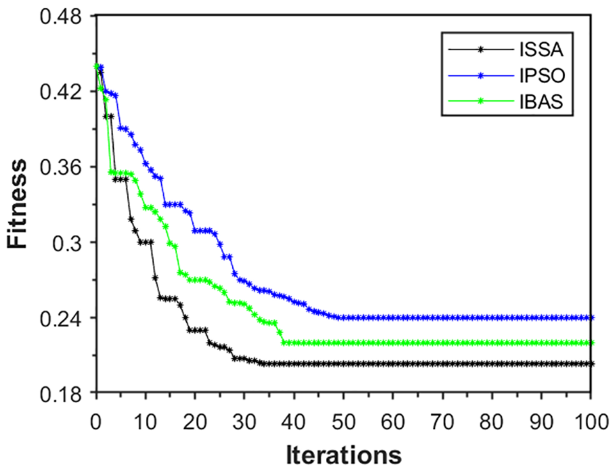


FIGURE 14
Iterative optimization and convergence curve for different optimization algorithm.

4.3.2 Comparison with CNN-LSTM

CNN-LSTM (Tan et al., 2022) employed CNN to extract local features from the data before feeding them into the LSTM network. Compared to the PCA-ISSA-DAM-Bi-GRU model proposed in this paper, CNN-LSTM functionally lacks the integration of the feature and temporal attention mechanism DAM, as well as the utilization of ISSA for optimizing the hyperparameters of the neural network. Firstly, PCA-Bi-GRU model was constructed for comparative experiments, with hyperparameter optimized through random search. Experimental results in Table 8 indicated that its predictive performance was slightly inferior to CNN-LSTM (Tan et al., 2022), suggesting that the combination of CNN and LSTM indeed enhances the feature extraction capability of the data. Furthermore, CNN-ISSA-DAM-LSTM model was built upon CNN-LSTM (Tan et al., 2022). Experiments revealed significant improvement in predictive performance as shown in Table 8, which reaffirms the effectiveness of the ISSA and DAM proposed in this paper in enhancing the predictive functionality of the model.

TABLE 8 Predictive performance of existing models.

Model	RMSE/ (mg·L^{-1})	MAPE	NSE
PCA-Bi-GRU	0.3701	0.0465	0.8524
CNN-LSTM (Tan et al., 2022)	0.3495	0.0358	0.8631
CNN-ISSA-DAM-LSTM	0.2474	0.0256	0.9397

4.3.3 Comparison with IDA-GRU

IDA-GRU (Zhang et al., 2023) employed a dual attention mechanism similar to this paper to optimize the hyperparameters of GRU, incorporating both feature and temporal attention at the input ends of the GRU encoder and decoder. However, its optimization effect is inferior to the model presented in this paper. Firstly, IDA-GRU (Zhang et al., 2023) utilized the Pearson correlation coefficient method to select environmental factors with high correlation coefficients with DO as input variables, whereas this paper adopts PCA, preserving approximately 86.27% of the information from all environmental factors. Secondly, IDA-GRU (Zhang et al., 2023) did not employ an intelligent optimization algorithm for hyperparameter tuning. In the comparison experiment, random search method was used to determine its hyperparameters, but its predictive performance still lags behind the model in this paper. This underscores the effectiveness of the ISSA proposed in this paper in enhancing the predictive performance of the model.

4.4 Practical application significance, limitations, and future research prospects of the model

This model utilized historical data from the past 24 hours to make real-time predictions of dissolved oxygen concentration 2 hours ahead, combined with LoRa+5G-based sensor deployment, enabling simultaneous prediction of dissolved oxygen concentrations at multiple points, thereby effectively forecasting the dissolved oxygen distribution in aquaculture areas. The engineering application analysis of the model reveals that it achieves good prediction results, effectively guiding water quality early warning and regulation, reducing aquaculture risks in marine ranching, and enhancing aquaculture efficiency. However, this study has limitations in spatial dimension prediction. The spatial distribution of dissolved oxygen was achieved through joint multi-point prediction, and the prediction accuracy of dissolved oxygen between points is related to the density of sensor deployment. Moreover, due to the limited availability of observed data, this study does not discuss the prediction performance of the model under different weather conditions. In future research, we will add more monitoring points in depth and attempt to employ a 3D convolutional neural network (3D-CNN) to capture the spatiotemporal characteristics of the data, providing more accurate prediction results. Additionally, we will further extend the experimental period to accumulate more data, which will be clustered according to weather conditions before predictive modeling for different categories, thereby enhancing the applicability and accuracy of the model.

5 Conclusion

To enhance the accuracy, generalization, and robustness of the dissolved oxygen prediction model in aquaculture water, this paper constructed a data-driven dissolved oxygen prediction model that integrates principal component analysis (PCA), dual attention mechanism (DAM), and bi-directional gated recurrent unit (Bi-GRU) neural network. Furthermore, an improved sparrow search algorithm with multi-strategy fusion (ISSA) is introduced for hyperparameter optimization. The main conclusions are as follows:

1. By applying PCA, the 13-dimensional input is reduced to 7 dimensions, eliminating redundancy and correlation among variables. This enhances the feature representation power of the input data for the prediction model and reduces its complexity. The fusion of DAM and Bi-GRU strengthens the feature extraction capability of the prediction model. The introduction of the feature attention mechanism in the encoder stage adaptively assigns weights to different environmental factors at each time step, while the time attention mechanism in the decoder stage dynamically adjusts the weights of the influence of different time steps on the current dissolved oxygen concentration. This enables the model to better capture the key information in the time series data. Combined with Bi-GRU, it simultaneously mines the sequential and inverse sequential correlations in the time series, comprehensively extracting temporal features.
2. The hyperparameters of the Bi-GRU model are searched and optimized using ISSA to enhance the model's optimal learning capability. The Gauss chaotic sequence is introduced into the population initialization, and the updating strategy of the discoverer's position is improved by referencing the salp swarm algorithm. Meanwhile, the updating strategy of the follower's position is optimized by drawing inspiration from the chicken swarm algorithm, and the Cauchy-Gaussian mutation strategy is incorporated to enhance the convergence performance of the SSA algorithm, balancing its global search and local exploitation capabilities.
3. The root mean square error (RMSE), mean absolute percentage error (MAPE), and Nash-Sutcliffe efficiency (NSE) of the proposed PCA-ISSA-DAM-Bi-GRU model for predicting dissolved oxygen are 0.2136, 0.0232, and 0.9427, respectively. The ablation study demonstrates that each component of the hybrid model contributes to enhancing the predictive performance of the model. By comparing the results with traditional baseline approaches, it is evident that each module in the hybrid model provides a more significant optimization effect on prediction accuracy.
4. By combining the proposed model with wireless sensor deployment, it can effectively predict the spatio-temporal distribution characteristics of dissolved oxygen in aquaculture water, enabling dynamic monitoring of water quality in marine ranching and intelligent analysis of the aquaculture environment, thereby facilitating the construction of modern marine ranching.

In summary, the model proposed in this paper, combined with wireless sensor deployment, can effectively predict the spatio-temporal distribution characteristics of dissolved oxygen in aquaculture water bodies. This enables dynamic monitoring of water quality in marine ranching and intelligent analysis of the aquaculture environment, thereby contributing to the modernization of marine ranching construction. The model provides a powerful tool for managing and optimizing aquaculture operations, ensuring sustainable development and improved productivity in marine ranching systems.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

WL: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JW: Conceptualization, Methodology, Writing – review & editing. ZL: Software, Visualization, Writing – review & editing. QL: Data curation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Key R&D Program of Shaanxi Province (2023-ZDLGY-15); General Project of National Natural Science Foundation of China (51979045); New Generation Information Technology Special Project in Key Fields of Ordinary Universities in Guangdong Province (2020ZDZX3008); Key Special Project in the Field of Artificial Intelligence in Guangdong Province (2019KZDZX1046); University level doctoral initiation project (060302112309); Guangdong Youth Fund Project (2023A15151110770); Zhanjiang Marine Youth Talent Innovation Project (2023E0010).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdel-Tawwab, M., Monier, M. N., Hoseinifar, S. H., and Faggio, C. (2019). Fish response to hypoxia stress: growth, physiological, and immunological biomarkers. *Fish Physiol. Biochem.* 45, 997. doi: 10.1007/s10695-019-00614-9

Arora, S., and Keshari, A. K. (2021). Dissolved oxygen modelling of the Yamuna river using different ANFIS models. *Water Sci. Technol.* 84, 3359–3371. doi: 10.2166/wst.2021.466

Cao, S., Zhou, L., and Zhang, Z. (2021a). Prediction of dissolved oxygen content in aquaculture based on clustering and improved ELM. *IEEE Access* PP, 1–1. doi: 10.1109/access.2021.3064029

Cao, S., Zhou, L., and Zhang, Z. (2021b). Prediction model of dissolved oxygen in aquaculture based on improved long short-term memory neural network. *Trans. Chin. Soc. Agric. Eng. (Transactions CSAE)* 37, 235–242. doi: 10.11975/j.issn.1002-6819.2021.14.027

Chen, H., Yang, J., Fu, X., Zheng, Q., Song, X., Fu, Z., et al. (2022). Water quality prediction based on LSTM and attention mechanism: A case study of the Burnett River, Australia. *Sustainability* 14, 13231. doi: 10.3390/su142013231

Choi, H., Suh, S.-I., Kim, S.-H., Han, E. J., and Ki, S. J. (2021). Assessing the performance of deep learning algorithms for short-term surface water quality prediction. *Sustainability* 13, 10690. doi: 10.3390/su131910690

Cuenco, M. L., Stickney, R. R., and Grant, W. E. (1985). Fish bioenergetics and growth in aquaculture ponds: ii. effects of interactions among, size, temperature, dissolved oxygen, unionized ammonia and food on growth of individual fish. *Ecol. Modelling* 27, 191–206. doi: 10.1016/0304-3800(85)90002-X

Guo, Y., Chen, S., Li, X., Cunha, M., Jayavelu, S., Cammarano, D., et al. (2022). Machine learning-based approaches for predicting SPAD values of maize using multi-spectral images. *Remote Sens.* 14, 1337. doi: 10.3390/rs14061337

Guo, Y., Xiao, Y., Hao, F., Zhang, X., Chen, J., de Beurs, K., et al. (2023). Comparison of different machine learning algorithms for predicting maize grain yield using UAV-based hyperspectral images. *Int. J. Appl. Earth Observation Geoinf.* 124, 103528. doi: 10.1016/j.jag.2023.103528

Huan, J., Li, M., Xu, X., Zhang, H., Yang, B., Jiang, J., et al. (2022). Multi-step prediction of dissolved oxygen in rivers based on random forest missing value imputation and attention mechanism coupled with recurrent neural network. *Water Supply* 22, 5480–5493. doi: 10.2166/ws.2022.154

Jiange, J., Liqin, Z., and Senjun, H. (2023). Water quality prediction based on IGRA-ISSA-LSTM model. *Water Air Soil pollut.* 234, 172. doi: 10.1007/s11270-023-06117-x

Jiang, X., Dong, S., Liu, R., Huang, M., Dong, K., Ge, J., et al. (2021). Effects of temperature, dissolved oxygen, and their interaction on the growth performance and condition of rainbow trout (Oncorhynchus mykiss). *J. Of Thermal Biol.* 98, 102928. doi: 10.1016/j.jtherbio.2021.102928

Kuang, L., Shi, P., Hua, C., Chen, B., and Zhu, H. (2020). An enhanced extreme learning machine for dissolved oxygen prediction in wireless sensor networks. *IEEE Access* 8, 198730–198739. doi: 10.1109/ACCESS.2020.3033455

Li, W., Wu, H., Zhu, N., Jiang, Y., Tan, J., and Guo, Y. (2021). Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). *Inf. Process. Agric.* 8, 185–193. doi: 10.1016/j.inpa.2020.02.002

Li, Y., Li, X., Xu, C., and Tang, X. (2023). Dissolved oxygen prediction model for the Yangtze River estuary basin using IPSO-LSSVM. *Water* 15, 2206. doi: 10.3390/w15122206

Lipizer, M., Partescano, E., Rabitti, A., Giorgetti, A., and Crise, A. (2014). Qualified temperature, salinity and dissolved oxygen climatologies in a changing Adriatic sea. *Ocean Sci. Discussions* 11, 331–390. doi: 10.5194/os-10-771-2014

Liu, P., Wang, J., Sangaiah, A., Xie, Y., and Yin, X. (2019). Analysis and prediction of water quality using LSTM deep neural networks in ioT environment. *Sustainability* 11, 2058. doi: 10.3390/su11072058

Liu, Y., Zhang, Q., Song, L., and Chen, Y. (2019). Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction. *Comput. Electron. Agriculture* 165, 104964. doi: 10.1016/j.compag.2019.104964

Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., and Mirjalili, S. M. (2017). Salp swarm algorithm: a bio-inspired optimizer for engineering design problems. *Adv. Eng. Software* 114, 163–191. doi: 10.1016/j.advengsoft.2017.07.002

Neilan, R. M., and Rose, K. (2014). Simulating the effects of fluctuating dissolved oxygen on growth, reproduction, and survival of fish and shrimp. *J. Theor. Biol.* 343, 54–68. doi: 10.1016/j.jtbi.2013.11.004

Osamy, W., El-Sawy, A. A., and Salim, A. (2020). CSOCA: Chickenswarm optimization based clustering algorithm forwireless sensor networks. *IEEE Access* 8, 60676–60688. doi: 10.1109/ACCESS.2020.2983483

Sharad, T., Richa, B., and Gagandeep, K. (2018). Performance evaluation of two ANFIS models for predicting water quality index of river Satluj (India). *Adv. Civil Eng.* 2018, 1–10. doi: 10.1155/2018/8971079

Sun, L., Wu, Y., Sun, X., and Zhang, S. (2021). Dissolved oxygen prediction model in ponds based on improved beetle antennae search and LSTM network. *Trans. Chin. Soc. Agric. Machinery* 52, 252–260. doi: 10.6041/j.issn.1000-1298.2021.S0.031

Tan, W., Zhang, J., Wu, J., Lan, H., Liu, X., Xiao, K., et al. (2022). Application of CNN and long short-term memory network in water quality predicting. *Intell. Autom. Soft Comput.* 34, 1943–1958. doi: 10.32604/iasc.2022.029660

Than, N. H., Ly, C. D., and Van Tat, P. (2021). The performance of classification and forecasting Dong Nai River water quality for sustainable water resources management using neural network techniques. *J. Hydrol.* 596, 126099. doi: 10.1016/j.jhydrol.2021.126099

Wang, W. C., Xu, L., Chau, K. W., and Xu, D. M. (2020). Yin-Yang firefly algorithm based on dimensionally Cauchy mutation. *Expert Syst. With Applications* 150, 113216. doi: 10.1016/j.eswa.2020.113216

Wang, J., Xie, Z., and Mo, C. (2023). Research progress in accurate prediction of aquaculture water quality by neural network. *J. Fish China*. 47(8), 089502. doi: 10.11964/jfc.20220913689

Wu, J., Li, Z., Zhu, L., Li, G., Niu, B., and Peng, F. (2018). Optimized BP neural network for Dissolved Oxygen prediction. *IFAC-PapersOnLine* 51, 596–601. doi: 10.1016/j.ifacol.2018.08.132

Xue, J., and Shen, B. (2020). A novel swarm intelligence optimization approach: Sparrow search algorithm. *Syst. Sci. Control Engineering* 8, 22–34. doi: 10.1080/21642583.2019.1708830

Yang, H., and Liu, S. (2022). Water quality prediction in sea cucumber farming based on a GRU neural network optimized by an improved whale optimization algorithm. *PeerJ Comput. Sci.* 8, e1000. doi: 10.7717/peerj-cs.1000

Zhang, Y.-F., Fitch, P., and Thorburn, P. J. (2020). Predicting the trend of dissolved oxygen based on the kPCA-RNN model. *Water* 12, 585. doi: 10.3390/w12020585

Zhang, Z., Jia, X., Zhang, Z., and Cao, S. (2023). Spatiotemporal prediction model of dissolved oxygen in aquaculture intergrating IDA-GRU and IIDW. *Trans. Chin. Soc. Agric. Engineering* 39, 161–171. doi: 10.11975/j.issn.1002-6819.202307067

Zhu, C., Liu, X., and Ding, W. (2017). “Prediction model of dissolved oxygen based on FOA-LSSVR,” in *IEEE 2017 36th Chinese Control Conference (CCC)*, Dalian, China, July 26-28, 2017. IEEE, 9819–9823. doi: 10.23919/ChiCC.2017.8028922

Zhu, N., Xia, Q., Tan, J., Jiang, Y., Xu, G., Chu, D., et al. (2019). Model-based prediction of dissolved oxygen content in fish production. *Trans. ASABE* 62, 1417–1425. doi: 10.13031/trans.13263

Frontiers in Marine Science

TYPE Original Research
PUBLISHED 23 October 2024
DOI 10.3389/fmars.2024.1466037

OPEN ACCESS

EDITED BY
Jianchuan Yin,
Guangdong Ocean University, China

REVIEWED BY
Sarat Chandra Tripathy,
National Centre for Polar and Ocean Research (NCPOR), India
Guifen Wang,
Hohai University, China

*CORRESPONDENCE
David Antoine
david.antoine@curtin.edu.au

RECEIVED 17 July 2024
ACCEPTED 24 September 2024
PUBLISHED 23 October 2024

CITATION
Li J, Antoine D and Huot Y (2024) Bio-optical variability of particulate matter in the Southern Ocean.
Front. Mar. Sci. 11:1466037.
doi: 10.3389/fmars.2024.1466037

Bio-optical variability of particulate matter in the Southern Ocean

Juan Li[1,2], David Antoine[1,2]* and Yannick Huot[3]

[1]Remote Sensing and Satellite Research Group, School of Earth and Planetary Sciences, Curtin University, Perth, WA, Australia, [2]Australian Research Council (ARC) Australian Centre for Excellence in Antarctic Science (ACEAS), University of Tasmania, Hobart, TAS, Australia, [3]Centre d'Applications et de Recherches en Télédétection, Département de géomatique appliquée, Université de Sherbrooke, Sherbrooke, QC, Canada

The composition and size distribution of particles in the ocean control their optical (scattering and absorption) properties, as well as a range of biogeochemical and ecological processes. Therefore, they provide important information about the pelagic ocean ecosystem's structure and functioning, which can be used to assess primary production, particle sinking, and carbon sequestration. Due to its harsh environment and remoteness, the particulate bio-optical properties of the Southern Ocean (SO) remain poorly observed and understood. Here, we combined field measurements from hydrographic casts from two research voyages and from autonomous profiling floats (BGC-Argo) to examine particulate bio-optical properties and relationships among several ecologically and optically important variables, namely the phytoplankton chlorophyll *a* concentration (Chl), the particulate absorption coefficient (a_p), the particulate backscattering coefficient (b_{bp}), and the particulate organic carbon (POC) concentration. In the clearest waters of the SO (Chl < 0.2 mg m^{-3}), we found a significant contribution to absorption by non-algal particles (NAP) at 442 nm, which was up to 10 times greater than the absorption by phytoplankton. This makes the particulate bio-optical properties there remarkably different from typical oceanic case 1 water. A matchup analysis confirms the impact of this larger NAP absorption on the retrieval of Chl from satellite ocean colour observations. For waters with Chl > 0.2 mg m^{-3}, no significant differences are observed between the SO and temperate waters. Our findings also demonstrate consistency in predicting phytoplankton carbon from either Chl or b_{bp}, suggesting that both methods are applicable in the SO.

KEYWORDS

Southern Ocean, particles, bio-optical properties, bio-optical relationships, BGC-Argo floats

1 Introduction

In the sunlit upper ocean, autotrophic organisms take up CO_2 and utilise inorganic nutrients via photosynthesis to produce organic matter, packaged in the form of phytoplankton cells, that accumulates in the water column as suspended particles (Falkowski et al., 1998). These phytoplankton cells provide energy for essentially the entire pelagic ecosystem and are, thereby, transformed into a large variety of living and nonliving particles through a myriad processes: viral infection, shedding of vesicles and other cellular parts, grazing by zooplankton (Jackson, 1980; Steinberg and Landry, 2017; Karakuş et al., 2022), remineralisation by microbes (Boyd et al., 2015; Belcher et al., 2016; Cavan et al., 2017), and (dis)aggregation by a series of biogeophysical and biogeochemical processes (Jackson and Burd, 1998; Slade et al., 2011; Briggs et al., 2020). A fraction of their accumulated carbon and nutrients eventually sinks (or is advected) into deep unlit layers as part of the so-called biological pump (Buesseler et al., 2007; Turner, 2015; Boyd et al., 2019). The particle flux and its composition in the water column represent a dynamic balance between ecosystem-driven processes that generate large sinking particles in the upper ocean and particle recycling processes within the ocean interior that consume, modify, and produce new sinking particles (Clements et al., 2022). Therefore, marine particles are critical in the characterisation of pelagic ecosystems, as they control a range of biogeochemical and ecological processes, and influence the ability of the ocean to sequester carbon.

The Southern Ocean (SO) is responsible for ~ 40% of the global oceanic CO_2 uptake (Gruber et al., 2009) and is a key driver of global ocean circulation and climate (Stark et al., 2019). Characterising and understanding particle dynamics in the surface layer is particularly important for assessing the strength of the biological pump under the pressure of climate change. However, the remoteness and difficult field conditions limit the opportunities for *in situ* studies of the SO. In this context, ocean colour remote sensing (OCRS) can provide a powerful tool to monitor it and obtain spatially resolved information. However, given the importance of the SO in oceanic carbon uptake and productivity, estimates must be accurate, as any error will have a large impact on our ability to obtain global estimates. In turn, for remote sensing to be accurate, we must determine whether global relationships derived elsewhere between satellite-measured quantities—such as spectral remote sensing reflectance or normalised water-leaving radiance—and *in situ* variables are applicable in the SO. This need has spurred studies examining the particulate bio-optical properties and relationships in the SO.

Allison et al. (2010a) found a different relationship between particulate organic carbon (POC) and the blue-to-green band ratios of reflectance in the SO compared to other oceans, such as the North Polar Atlantic (Stramska et al., 2003) and the eastern South Pacific and eastern Atlantic oceans (Stramski et al., 2008). Their new relationship has been applied to satellite observations to characterise the seasonal and interannual variability of POC in the SO (Allison et al., 2010b). Johnson et al. (2013) have reported three improved satellite chlorophyll algorithms for the SO to better monitor phytoplankton dynamics. These improved ocean colour products and relationships would lead to better estimation of primary production using bio-optical productivity models (Arrigo et al., 2008; Hirawake et al., 2011). However, due to the lack of contemporaneous *in situ* measurements in the SO, these particulate bio-optical relationships obtained from space have not been thoroughly evaluated and validated.

More observations are now available due to the BGC-Argo program, which deploys autonomous profiling floats worldwide, particularly in the SO, following the initial deployments by the Southern Ocean Carbon and Climate Observations and Modelling (SOCCOM) program (Sarmiento et al., 2023). To maximise the use of the BGC-Argo data, continuing efforts have been made to accurately convert chlorophyll fluorescence signals into chlorophyll *a* concentrations (Johnson et al., 2017; Roesler et al., 2017; Schallenberg et al., 2022). Particulate backscattering coefficients (b_{bp}) at 700 nm have been used to estimate POC in the SO by Johnson et al. (2017). In addition, Schallenberg et al. (2019) used 6 years of mooring data collected at the Southern Ocean Time Series (SOTS) site in the Subantarctic Zone south of Australia to estimate carbon-to-Chl ratios and interpret their seasonal dynamics. However, these particulate bio-optical relationships applied to float data in these studies were empirically developed based on limited concurrent measurements from hydrological casts on cruises. It is still unknown whether they are suitable for waters other than those where the relationships were developed.

In addition to pigment concentration (a desirable proxy for phytoplankton biomass) and the b_{bp}, hydrological casts during cruises can also provide data on mass concentrations (e.g., POC, particulate organic nitrogen (PON), macronutrients, and trace metal) and particle size distribution. These variables are essential for a better understanding of the particulate bio-optical properties and relationships in the SO. Based on ~ 280 samples collected during the Antarctic Circumpolar Expedition (ACE), Robinson et al. (2021) found that high-latitude SO phytoplankton have distinctive absorption properties compared to lower-latitude populations. However, other particulate bio-optical properties of the SO remain poorly observed and understood, leaving the question of whether they conform to or diverge from global relationships unanswered. Both the correct interpretation of satellite ocean colour observations and the appropriate parameterisation of bio-optical properties in biogeochemical models rely on this answer.

To address this gap, we collated field measurements from both hydrological casts and BGC-Argo floats to derive relationships among variables that are commonly used to describe the particulate pool, namely the particulate backscattering coefficient, b_{bp} (m^{-1}), the particulate absorption coefficient, a_p (m^{-1}), Chl (mg m^{-3}), POC (mg m^{-3}), and phytoplankton carbon (C_{phyto}, mg C m^{-3}). We specifically aim to elucidate large-scale distribution patterns of particle-related properties and evaluate the applicability of particulate bio-optical relationships developed for temperate oceanic waters to the SO.

2 Materials and methods

2.1 Datasets

The field data used in this study were acquired during the ACE aboard the RV Akademik Tryoshnikov during the austral summer from 20 December 2016 to 19 March 2017 (Walton and Thomas, 2018), and during the Southern Ocean Large Areal Carbon Export (SOLACE) research voyage aboard the RV Investigator (voyage IN2020_V08) from 05 December 2020 to 16 January 2021 (Figure 1). The ACE cruise travelled eastward around the Southern Ocean, starting from Cape Town, South Africa, to Hobart, Australia (leg 1), then proceeding via the Pacific Ocean to Punta Arenas, Chile (leg 2), and finally returning through the Atlantic Ocean back to Cape Town (leg 3). The SOLACE cruise investigated three sites: a subpolar site—SOTS (47°05′S, 141°22′E, Wynn-Edwards et al., 2019), and two polar possible phytoplankton bloom sites—Southern Site 1 (SS1, 55°49′S, 138°40′E) and Southern Site 2 (SS2, 57°54′S, 141°32′E), along with several stations during the transit (Figure 1, inset).

Measurements from both cruises and BGC-Argo floats were classified based on the seven bioregions defined by Ardyna et al. (2017) (Figure 2). This split aims to examine whether our dataset evenly samples various oceanographic regimes and to assess whether there are regional differences in the particulate bio-optical properties in addition to their large-scale patterns. The data from floats mostly fall into bioregions 3, 4, and 5, with only one-third belonging to the other bioregions. The distribution of cruise data is similar to that of float data, although high-latitude regions 6 and 7 are more prominently represented. Only average values from the top 10 m of the BGC-Argo data profiles are used in this work, combined with cruise data from underway sampling (depth ~5 m) and the top 10 m of data from CTD casts, as described in the following sections.

The BGC-Argo profiling floats (https://biogeochemical-argo.org) used in this study, equipped with CTD and bio-optical sensors, measured temperature, salinity, pressure, chlorophyll fluorescence, and volume scattering (used to derive backscattering coefficient). Figure 1 shows the geographical location of all profiles collected by 254 floats from January 2016 to June 2023. The data points are colour-coded based on the bioregions in which the floats operated, following the regionalisation outlined by Ardyna et al. (2017).

2.2 Phytoplankton pigments and particulate absorption

Phytoplankton pigment concentrations from ACE and SOLACE were determined using high-performance liquid chromatography (HPLC, see details in Ras et al. (2008); Antoine et al. (2020) and references therein). On both cruises, 2.2 L water samples were collected either 3 hourly from the underway seawater supply (sampling depth ~ 5 m) or from the shallowest depth of the

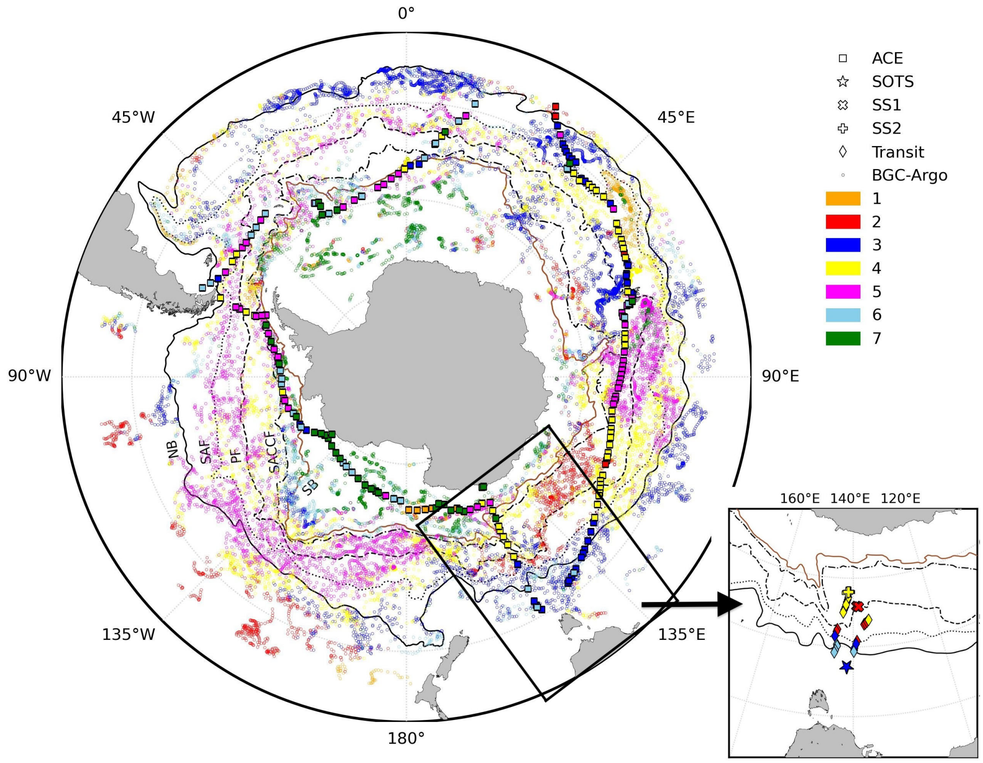


FIGURE 1
Sampling locations during ACE (squares) and SOLACE (SOTS, SS1, SS2, and transit sites, inset), as well as surface sampling points from BGC-Argo floats, colour-mapped according to the bioregions described in Ardyna et al. (2017). The Northern boundary (NB), Subantarctic Front (SF), Polar Front (PF), Southern Antarctic Circumpolar Current Front (SACCF), and Southern boundary (SB) from Park et al. (2019) are also displayed (black lines).

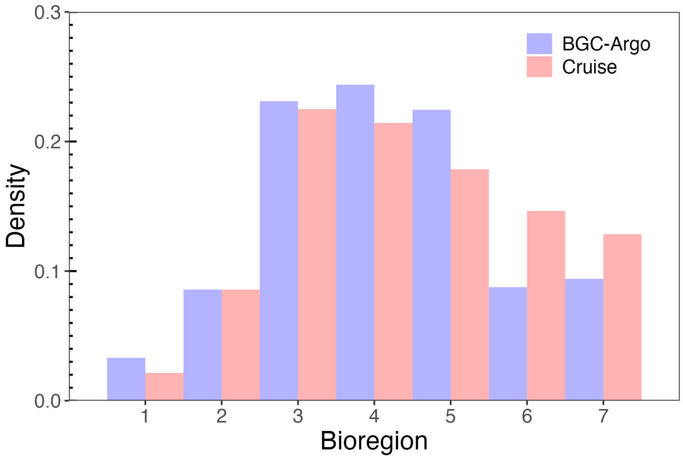


FIGURE 2
Distribution of cruise and float measurements among the seven bioregions defined by Ardyna et al. (2017).

conductivity, temperature, and depth (CTD) rosette casts (see Table 1 for a summary of the number of samples). Total chlorophyll *a* concentration was defined as the sum of mono- and divinyl chlorophyll *a* concentrations, chlorophyllide *a*, and the allomeric and epimeric forms of chlorophyll *a* (Hooker and Zibordi, 2005; Reynolds et al., 2016).

The pigments are used here to determine the relative contributions of micro- (f_{micro}, > 20 µm), nano- (f_{nano}, 2–20 µm), and picophytoplankton (f_{pico}, < 2 µm) to the total population, following Brewin et al. (2015).

For the BGC-Argo floats, the calibrated fluorescence profiles were adjusted for nonzero deep values (below 600 dbar) and corrected for spikes using a five-point median filter. Subsequently, they were divided by 3.79, as recommended by Schallenberg et al. (2022) for the SO. The surface Chl was obtained by averaging the values within the top 10 dbar.

A full description of the determination of the absorption coefficient of phytoplankton and non-algal particles (NAP) from the total particulate absorption coefficient can be found in Robinson et al. (2021), which is not repeated here.

2.3 Backscattering measurements

The particulate backscattering coefficient, $b_{\text{bp}}(\lambda)$ (m^{-1}), was determined on ACE and SOLACE using HOBI Labs HydroScat-6 sensors, which provide a measurement of the total spectral volume scattering function $\beta(\psi)$ (m^{-1} sr^{-1}) at an effective scattering angle $\psi = 140°$. The following equation (Maffione and Dana, 1997) was used to convert $\beta(140°, \lambda)$ to b_{bp} (λ),

$$b_{bp}(\lambda) = 2\pi\chi[\beta(140°, \lambda) - \beta_w(140°, \lambda)] \tag{1}$$

Where the subscripts *p* and *w* indicate the contributions from particles and seawater to scattering, respectively. χ is the coefficient of proportionality between β and b_b for particles, set to 1.13 for the HydroScat. Pure water values for $\beta_w(140°, \lambda)$ at given temperature and salinity were calculated following Zhang et al. (2009). Finally, vertical profile data of $b_{\text{bp}}(\lambda)$ were filtered to remove spikes and averaged into 0.5 m depth bins for analysis and correlation with discrete water samples. The Hydroscat channels were 420 nm, 488 nm, 550 nm, 620 nm, and 700 nm for ACE (the 442 nm channel failed), and 420 nm, 442 nm, 470 nm, 510 nm, 590 nm, and 700 nm for SOLACE. For ACE, only 13 b_{bp} spectra were obtained, primarily during leg 2 near the Antarctic continent.

To examine the wavelength dependency of particle backscattering, discrete spectral measurements were fitted to a power function of the following form:

$$b_{bp}(\lambda) = b_{bp}(\lambda_0)\left(\frac{\lambda}{\lambda_0}\right)^{\eta} \tag{2}$$

Where λ_0 represents a reference wavelength, and η denotes the dimensionless spectral slope of b_{bp}. Nonlinear least-squares fitting was applied to account for b_{bp} at all channels to derive η. Since the 442-nm spectral channel failed during the ACE cruise, this wavelength was excluded from the fitting of spectral relationships for the SOLACE cruise to ensure consistency between cruises.

Since our focus is on the total particle pool, we did not apply the nonzero deep value correction to the backscattering profile of BGC-Argo floats, as proposed by Uchida et al. (2019). This correction is meant to isolate the part of b_{bp} attributed to phytoplankton by removing the average of deep values (below 600 dbar), which are assumed to represent the NAP contribution. We applied a five-point median filter to remove spikes (Carranza et al., 2018; Mignot et al., 2018). Surface values were obtained by averaging the data within the top 10 dbar of the profile.

TABLE 1 Summary of the in situ dataset.

	Chl	$b_{\text{bp}}(\lambda)$	$a_{\text{p}}(\lambda)$	POC	PSD
ACE	221	13	274	355	264
SOLACE	31	3	31	3	3
BGC-Argo floats	21,872	21,872	–	–	–

2.4 Particulate organic carbon concentration

For the ACE dataset, the concentration of POC was obtained from the underway seawater supply every 3 h, as well as from several depths during the CTD rosette casts, and processed at the University of Cape Town (Fawcett and Forrer, 2020). Up to 2 L water samples were filtered through precombusted 25 mm GF/F filters. The filters were then dried at 40°C for 24 h, acidified to remove inorganic carbon, and stored until elemental analysis. Finally, POC for each sample was obtained by subtracting the average concentration of carbon in the dry blanks and was expressed in units of milligrammes per cubic metre. Samples outside the detection limits were eliminated. The same methodology was used for the SOLACE POC data, although it was processed at the University of Tasmania by Cathryn Wynn-Edwards.

2.5 Particle size distributions

The particle size distributions (PSD) determine their optical properties along with the particle composition. During ACE, PSD was measured with a Beckmann Coulter Counter Multisizer, which measures particle sizes by quantifying changes in electrical resistance produced by particles suspended in seawater as they pass through an aperture (Kinsman, 2018). In this study, 0.2 µm filtered seawater was used as the blank to detect particles in the range of 2–60 µm across 400 bins at each underway station. Twenty replicate measurements of 2 ml subsamples were made by the Counter for each sample and summed up to provide larger sample volumes, thereby improving statistical accuracy. Each discrete Coulter measurement included a set of values representing the particle concentrations (m^{-3}) within a size bin D, $N(D)$. The bin diameters were restricted to 2–30 µm, as no particles were observed in larger bins. Plots of particle concentration versus bin diameter were visually inspected, and samples with high noise levels or particle concentrations constrained to just a few bins were flagged as poor quality and removed (Robinson et al., 2021). Finally, 264 records of PSD were retained for further analysis.

During SOLACE, the PSD of large (> 100 µm) particles was measured using an Underwater Video Profiler 5 (UVP-5, Picheral et al., 2010) mounted on the rosette. Particle cross-sectional areas are quantified by assessing the contiguous pixels for a given image brightness level, which were then used to estimate the equivalent circular diameter. Finally, PSDs were determined for 24 bins, with centre bin sizes ranging from 115 to 23,300 µm. An upper limit of 2,315 µm was chosen to avoid regions of low particle counts and high statistical noise at large particle diameters. All data were binned vertically into 5-m intervals, and only surface values were used in this study.

The particle size distribution was fitted to a power law model (Bader, 1970; Jonasz and Fournier, 2011):

$$N(D) = N_0 \left(\frac{D}{D_0}\right)^{\zeta} \tag{3}$$

Where D_0 is a reference diameter, N_0 is the particle number at D_0, and ζ is the dimensionless slope of the distribution. For the Coulter measurements, these metrics were calculated to characterize the samples and were computed over the size range of 4–20 µm. Since ζ is very sensitive to the range of effective diameters used and can be biased when an abundant phytoplankton cell size is present, leading to a bump on the PSD, we visually inspected the spectra and removed 15 spectra where clear bumps were present to avoid these cases.

2.6 Phytoplankton carbon estimations

C_{phyto} is a key parameter in estimating primary production using various models (Sathyendranath et al., 2007). It also allows an understanding of phytoplankton physiology, as reflected in variations of cellular chlorophyll-to-carbon ratios generated by changes in light, temperature, and nutrients (Behrenfeld et al., 2005). However, C_{phyto} is difficult to distinguish experimentally *in situ* or in laboratory studies from the total carbon included in phytoplankton plus zooplankton, detritus, and bacteria. Consequently, direct C_{phyto} estimations are scarce, and essentially proxy measurements have been used to quantify it, such as Chl, cell biovolume, POC, and b_{bp}.

Here, in the absence of direct C_{phyto} measurements, it was estimated using either the POC vs. Chl relationship or from b_{bp}. The former approach assumes that at any given Chl, the lowest POC observed represents the phytoplankton fraction, C_{phyto} (Sathyendranath et al., 2009). In this approach, a 1% quantile regression is applied to the fit between POC and Chl to obtain C_{phyto}, hereafter denoted as C_{phyto}-S09. Since there will always be some contribution to POC from material other than phytoplankton, such as heterotrophs and detritus, this C_{phyto} estimate likely represents an upper limit for a given Chl. In addition, this approach does not allow for scenarios where C_{phyto} increases or decreases without a corresponding change in Chl (Thomalla et al., 2017). However, it is unlikely to be influenced by phenomena such as coccolith blooms or bubbles, which can significantly increase backscattering or the attenuation coefficient without increasing Chl.

Backscattering-based approaches allow C_{phyto} to vary independently of Chl, making them less susceptible to the package effect or photoacclimation. As a result, they are able to detect the high temporal variability in Chl:C_{phyto} ratios. These methods assume C_{phyto} is linearly related to b_{bp}. Behrenfeld et al. (2005) established such a relationship by fitting satellite-derived b_{bp} (440) to which a background value of 3.5×10^{-4} m^{-1} is subtracted to laboratory C_{phyto} values:

$$C_{phyto} = 13{,}000 \times (b_{bp}(440) - 3.5 \times 10^{-4}) \tag{4}$$

which is denoted as C_{phyto}-B05 hereafter. The subtraction of the background value accounts for the portion of backscattering attributed to a background of NAP that does not covary with phytoplankton.

Based on direct measurements of both C_{phyto} and b_{bp} in the Atlantic Ocean, Martinez-Vicente et al. (2013) found a significant linear relationship between C_{phyto} and $b_{bp}(470)$

(denoted as C_{phyto}-M13):

$$C_{phyto} = 30{,}100 \times (b_{bp}(470) - 7.6 \times 10^{-4}) \tag{5}$$

This linear regression was initially limited to b_{bp} (470) < 0.003 m^{-1} or Chl < 0.4 mg m^{-3}; however, in this study, we extended it for the larger Chl range as well.

Using data from the Equatorial Pacific Ocean and from the 22nd Atlantic Meridional Transect cruise, Graff et al. (2015) established yet another relationship:

$$C_{phyto} = 12{,}128 \times b_{bp}(470) + 0.59 \tag{6}$$

Hereafter denoted as C_{phyto}-G15.

Note that, for backscatter measurements lacking a 440- or 470-nm channel, the values at 700 nm were converted to these other wavelengths using Equation 2, with η equal to 1.08 (mean of the measured values).

3 Results and discussions

3.1 General latitudinal distribution of properties

The latitudinal distribution of average values of major environment parameters and inherent optical properties (IOPs) is presented in Figure 3. These values are calculated from all data found in 2° latitude bands centred on latitudes from 40°S to 74°S. Hereafter, Chl and b_{bp} are measured both through ship-based hydrographic casts and BGC-Argo floats are denoted separately as Chl-*Cruise* and Chl-*Float* (and b_{bp}-*Cruise* and b_{bp}-*Float*).

Temperature decreases toward the south (Figure 3A), from about 15°C at 40°S to − 1°C close to the Antarctic continent. Salinity also shows a general decreasing trend toward the south (Figure 3B), with two relative maxima observed around 65°S and near the continent.

Minima of Chl-*Cruise* are found around 60°S~68°S and maxima around 45°S and 72°S (Figure 3C). The Chl-*Cruise* and Chl-*Float* are quite consistent, with the differences mainly due to the uneven cruise sampling. In the 41°S~45°S latitudinal belt, Chl-*Cruise* measurements were constrained in the area south of the African continent and are higher than Chl-*Float*. In the 59°S to 67°S latitudinal belt, Chl-*Cruise* was at the lowest level (< 0.2 mg m^{-3}) at the Drake Passage and the Dumont d'Urville Sea. As for the 69°S~76°S latitudinal belt, Chl-*Float* measurements are higher than Chl-*Cruise* and have larger variance because they were collected in more varied environments. Among the SOLACE data, Chl-*Cruise* is the highest at SOTS with a mean value of 0.64 mg m^{-3}, followed by SS2 (0.31 mg m^{-3}) and SS1 (0.15 mg m^{-3}).

The latitudinal distribution of $a_{\mathrm{ph}}(442)$ (Figure 3D) reflects that of Chl-*Cruise*, yet shows a smaller relative increase toward high latitudes, leading to a decrease of the chlorophyll-specific absorption coefficient at 442 nm, $a^{*}_{ph}(442)$ (Figure 3E).

The fraction of larger phytoplankton, f_{micro}, increases from about 0.3 to 0.9 toward the south (Figure 3F), which is consistent with the findings by Robinson et al. (2021).

Generally, b_{bp} (700) (Figure 3G) is quite stable in the 40 to 60°S belt, with a mean value of ~ 0.0012 m^{-1}. South of 60°S, the mean values and associated variance both increase. The 16 b_{bp}-*Cruise* all fall within the range of the b_{bp}-*Float* values. Due to the very limited b_{bp} measurements on cruises, they were excluded from further analyses with respect to the large-scale latitudinal analyses. The mean b_{bp}:Chl ratios across latitudes (Figure 3H) fluctuate between 0.005 and 0.008 m^{2} mg^{-1}.

The POC (Figure 3I) varies between 30 and 200 mg m^{-3}, with minimum values around 60°S and an average value of 105 mg m^{-3} across the dataset. The POC:Chl ratio (Figure 3J) varies over one order of magnitude, from 100 to 1,000, and shows relative maxima around 54°S and 64°S, with a regular decrease for latitude south of about 64°S.

It is worth noting that BGC-Argo data from all seasons have been pooled together here, whereas the cruise data are for the summer months only (December to February, plus early March for ACE). If Figure 3 was to include BGC-Argo data for only the 3 summer months, the only two notable differences would be the slightly higher $b_{\mathrm{bp}}(700)$ values (0.002 instead of 0.0015 m^{-1} on average for latitudes above 70°S) and, similarly, the slightly higher Chl for latitudes above 60°S (with an average of approximately 0.4 instead of 0.25 mg m^{-3}). The discrepancy between Chl-*Cruise* and Chl-*Float* would be reinforced in the 60°S–68°S band, primarily due to the ship sampling being restricted to the Drake Passage and Dumont d'Urville Sea.

When the zonal averages displayed in Figure 3 are restricted to the Atlantic, Indian, and Pacific sectors, the latter displays the lowest Chl (average 0.2 mg m^{-3}). In contrast, the subtropical latitudes of the Indian Ocean are saltier (salinity ~ 35 psu), warmer (SST up to 17°C), and have the largest POC concentrations (around 100 mg m^{-3}). No other major differences are observed among the three sectors and between the results for each sector and those for the entire SO.

3.2 Bio-optical relationships

The various bio-optical relationships we have explored are illustrated in Figure 4. The ratio of NAP to phytoplankton absorption, a_{NAP}: a_{ph} (Figure 4A), shows an upward tail in low Chl waters (Chl < 0.2 mg m^{-3}), with values larger than 1 and as large as 10. For larger Chl values (> 0.2 mg m^{-3}), the ratio slowly decreases as Chl increases. This result not only suggests a high contribution of NAP, such as heterotrophs and detritus, to the particle pool in clear waters of the SO, but also that this contribution is highly variable.

This large contribution of NAP to the particle pool in the SO seems corroborated by the POC vs. Chl relationship (Figure 4B). Here, Chl-*Cruise* varies over nearly three orders of magnitude, and POC over two, and their relationship shows the generally expected increasing trend but only for Chl > 0.2 mg m^{-3}. Below this concentration, POC fluctuates around 80 mg m^{-3}, independent of Chl levels, again suggesting that NAP significantly contributes to POC in the clear waters of the SO.

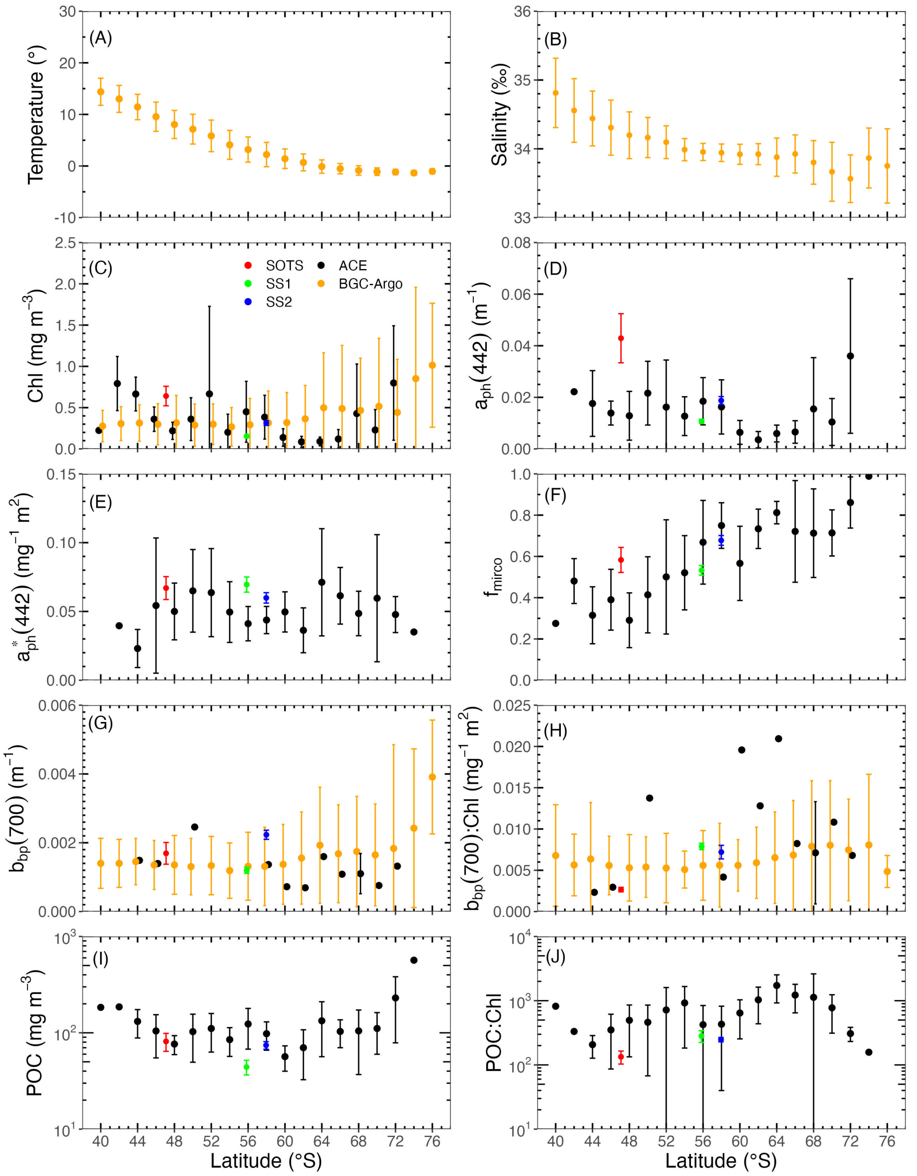


FIGURE 3

Latitudinal distribution of **(A)** temperature, **(B)** salinity, **(C)** Chl, **(D)** a_{ph}(442), **(E)** a^*_{ph}(442), **(F)** f_{micro}, **(G)** b_{bp}(700), **(H)** b_{bp}(700) to Chl ratio, **(I)** POC, and **(J)** POC to Chl ratio. Black symbols represent measurements made from hydrographic casts on ships, while orange symbols refer to BGC-Argo float measurements. The red, blue, and green dots refer to the SOTS, SS1, and SS2 sites, respectively. Data were grouped into 2° latitude intervals, within which the mean values and standard deviations were calculated.

The POC vs. Chl relationship is generally expressed through a linear fit on log-transformed data. For instance, Sathyendranath et al. (2009) found such a linear relationship (r^2 = 0.58) using data from the equatorial Pacific and Atlantic Oceans, denoted POC-S09 hereafter (solid purple line in Figure 4B). In our study, a linear regression in log space is not appropriate to describe the POC vs. Chl relationship because of the relatively constant POC in waters where Chl is less than 0.2 mg m^{-3}, attributed to the contribution of NAP to POC. Therefore, we added a constant background POC in our linear regression. When a linear fit without constant background value is applied to data for Chl > 0.2 mg m^{-3} only, the obtained POC vs. Chl relationship shows no significant difference with POC-S09.

We also applied 1% quantile regression on the data where Chl > 0.2 mg m^{-3} to derive C_{phyto}, following the approach of Sathyendranath et al. (2009), resulting in a remarkably similar relationship (dashed and solid purple lines in Figure 4B). By converting b_{bp}-*Float* to C_{phyto} according to Behrenfeld et al. (2005); Martinez-Vicente et al. (2013), and Graff et al. (2015), C_{phyto} vs. Chl can be obtained as well, denoted as B05, M13 and G15, respectively. Their comparison with S09 is illustrated in Figure 4C. The slope of M13 is significantly higher than the others, resulting in a difference up to 160 mg m^{-3} in C_{phyto} at Chl = 3 mg m^{-3}. B05 and G15 have similar intercepts, although the slope of B05 is slightly higher. Their largest difference is about 50 mg m^{-3} in C_{phyto} at Chl = 6 mg m^{-3}. S09 generally derives higher C_{phyto} per Chl, while G15 intersects with B05

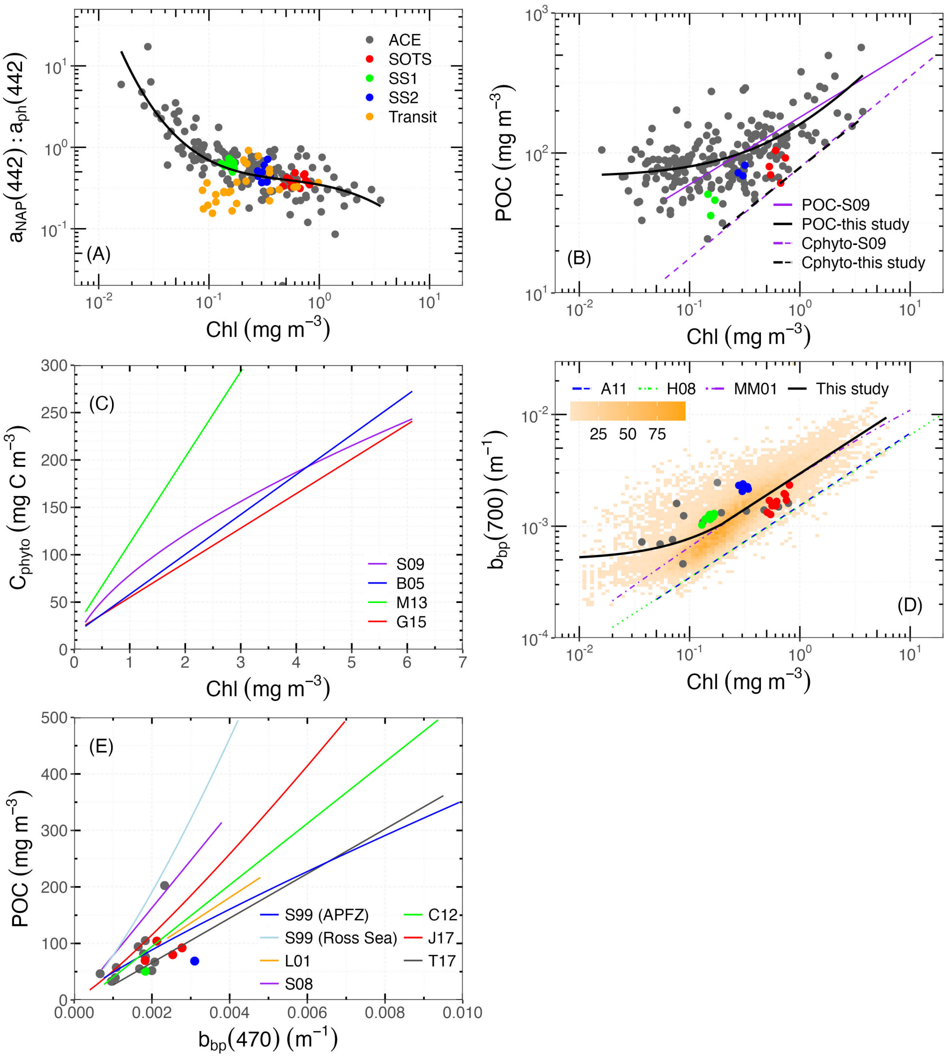


FIGURE 4

Bio-optical relationships. **(A)** a_{NAP}:a_{ph} ratio at 442 nm vs. Chl. The black solid line represents the fourth-degree polynomial regression in log space for our dataset, $y = -0.26 - 3.0x + 1.36x^2 - 0.90x^3 + 0.08x^4$ ($y = \log_{10}(a_{NAP}(442)/a_{ph}(442))$, $x = \log_{10}$Chl, n = 201, r^2 = 0.57). **(B)** POC vs. Chl. The black solid curve represents POC = 67.4 + 93.3 Chl$^{0.87}$ (n = 229, r^2 = 0.42). The dashed black line represents the 1% quantile regression on the same data, representing the relationship between C_{phyto} and Chl (C_{phyto} = 78.5 Chl$^{0.63}$), following Sathyendranath et al. (2009). The purple solid and dashed lines refer to the POC vs. Chl and C_{phyto} vs. Chl relationships from Sathyendranath et al. (2009), respectively. **(C)** C_{phyto} vs. Chl. See Section 2.6 for S09, B05, M13, and G15. **(D)** b_{bp}(700) vs. Chl. The background orange symbols refer to the density of float measurements. The solid black line represents the regression line between b_{bp}(700) and Chl, using both ship and float data from this study: $b_{bp}(700) = 0.0005 + 0.0028$ Chl (Chl ≤ 0.2 mg m^{-3}); $b_{bp}(700) = 0.0031$ Chl$^{0.67}$ (Chl > 0.2 mg m^{-3}). The blue, green, and purple lines refer to the relationships from Antoine et al. (2011) in the Northwestern Mediterranean Sea and Santa Barbara Channel, Huot et al. (2008) in the Eastern South Pacific Ocean, and Morel and Maritorena (2001) for global oceanic waters, respectively. **(E)** POC vs. b_{bp}(470). S99, L01, S08, C12, J17, and T17 refer to the relationships from Stramski et al. (1999) in the Antarctic Polar Front Zone (APFZ) and the Ross Sea, Loisel et al. (2001) in the Mediterranean Sea, Stramski et al. (2008) in the Pacific and Atlantic Oceans, Cetinić et al. (2012) in the North Atlantic Ocean, Johnson et al. (2017) in the SO, and Thomalla et al. (2017) in the South Atlantic and SO, respectively. Note that for models without b_{bp} at 470 nm, values were propagated from other available bands according to Equation 2 using η = 1.08.

at Chl = 4.3 mg m^{-3}, with their differences in C_{phyto} remaining within 50 mg m^{-3} across all Chl ranges.

B05 and M13 assume a constant background b_{bp} due to NAP (denoted bbp-NAP) that does not covary with phytoplankton. However, Bellacicco et al. (2016) found that b_{bp}-NAP varies both seasonally and regionally by more than one order of magnitude, which might result in significant errors in the C_{phyto} estimates. G15 is the only one using analytical field-measured C_{phyto}, which has been found to have a significant correlation with b_{bp} in the Equatorial Pacific Ocean and Atlantic Ocean.

None of these methods (either Chl or b_{bp}-based) have been derived using data collected in the SO, so their applicability here cannot be ascertained. Their consistent C_{phyto} prediction is however encouraging. It is still worth noting that the aforementioned approaches do not seem applicable to waters with Chl < 0.2 mg m^{-3}, due to the high NAP contribution to the particle pool. The substantial contribution of NAP to POC and b_{bp} further complicates the accurate estimation of C_{phyto} in such waters. Therefore, sufficient concurrent measurements of phytoplankton community composition and their specific chlorophyll and carbon content are needed to evaluate and validate optical methods of determining C_{phyto} concentrations, and then to assess Chl: C_{phyto} ratios for better understanding phytoplankton physiology under environmental forcing.

The b_{bp}-*Cruise* and b_{bp}-*Float* data are displayed as a function of Chl in Figure 4D. The former varies between 0.0004 and 0.004 m^{-1} in the combined ACE and SOLACE dataset, and all fall within the larger range (0.0002~0.1 m^{-1}) measured by floats over the entire SO. Values obtained during SOLACE were generally higher than those observed during ACE leg 2 near the Antarctic continent. The highest values were obtained at SS2 with the smallest variation, followed by those at SOTS and SS1. The spectral slope of $b_{bp}(\lambda)$, calculated using all available wavebands, fluctuates from 0.5 to 1.6, with a mean value of 1.08. The b_{bp}-*Float* values generally increase with Chl, although they remain relatively constant in the low Chl range (again, Chl < 0.2 mg m^{-3}). For Chl > 0.2 mg m^{-3}, b_{bp} covaries with Chl, which is consistent with previous observations (Antoine et al., 2011; Bellacicco et al., 2019). Therefore, we here combined a linear regression for Chl < 0.2 mg m^{-3} with a power law for Chl > 0.2 mg m^{-3} to fit the data (solid black curve in Figure 4D).

For comparison, other relationships obtained from *in situ* data collected in the Northwestern Mediterranean Sea and Santa Barbara Channel by Antoine et al. (2011) (denoted A11) and in the Eastern South Pacific Ocean by Huot et al. (2008) (denoted M08) are also shown. For A11, their $b_{bp}(560)$ was converted to $b_{bp}(700)$ according to Equation 2. The $b_{bp}(700)$ vs. Chl relationship from Morel and Maritorena (2001) (denoted MM01) developed for the oceanic case I waters is shown as well. Generally, the $b_{bp}(700)$ of MM01 is significantly higher than H08 and A11 across the Chl range, which might be due to the difference in study regions. For Chl > 0.2 mg m^{-3}, our fit generally coincides with MM01, but with a slightly larger slope. For Chl < 0.2 mg m^{-3}, $b_{bp}(700)$ is higher than predicted by MM01, with differences that can reach up to one order of magnitude. The high contribution of NAP is likely responsible for these larger b_{bp} values in clear waters as compared to what global models predict from Chl. Such relative constant b_{bp} is likely a consequence of the combination of photoacclimation and the high proportion of NAP. The former is typical for polar waters, where Chl variation is driven by photoacclimation to low light and thus uncoupled with biomass, leading to an increase of Chl without a corresponding increase in b_{bp} (Behrenfeld et al., 2005; Brewin et al., 2012; Bellacicco et al., 2019).

It is also worth noting that we did not find significant differences among b_{bp}-*Float* vs. Chl-*Float* relationships as separately derived for each Ardyna's bioregions.

The ACE and SOLACE datasets have only 16 concurrent measurements of POC and $b_{bp}(470)$. They are shown in Figure 4E on top of relationships obtained from both temperate oceans and subregion of the SO (see figure caption for details). There is a large spread of POC vs. $b_{bp}(470)$ relationships across different regions. The highest POC: $b_{bp}(470)$ statistical mean ratio is noticed in the Ross Sea, while the lowest is also found in the SO from 20°W ~ 20°E by Thomalla et al. (2017). In addition, their difference in POC is about 350 mg m^{-3} at $b_{bp}(470) = 0.004$ m^{-1} and continues to increase as b_{bp} increases. Thus, there is no clear POC vs. b_{bp} relationship in the SO, especially for high-scattering waters. This lack of correlation is due to the contributions of both POC and particulate inorganic carbon (PIC) to backscattering.

It appears that in the SO, we cannot use a single linear regression to describe the relationships of bio-optical properties and chlorophyll over the full concentration range. For Chl < 0.2 mg m^{-3}, the large NAP contribution tends to mask any possible phytoplankton-related changes in bio-optical properties.

3.3 Particle size distributions

PSD of surface waters measured by the Coulter Counter and UVP-5 are illustrated in Figure 5. Particle concentrations (m^{-3} μm^{-1}) decrease when the equivalent spherical diameter increases and generally follow a Junge-type size distribution. This is expected, yet exceptions occur, with peaks appearing at certain diameters, e.g., 10 μm, which may indicate phytoplankton blooms dominated by a particular size group. Although the Coulter and UVP-5 use different approaches to determine particle size, cross-sectional area for the UVP-5, and particle volume for the Coulter

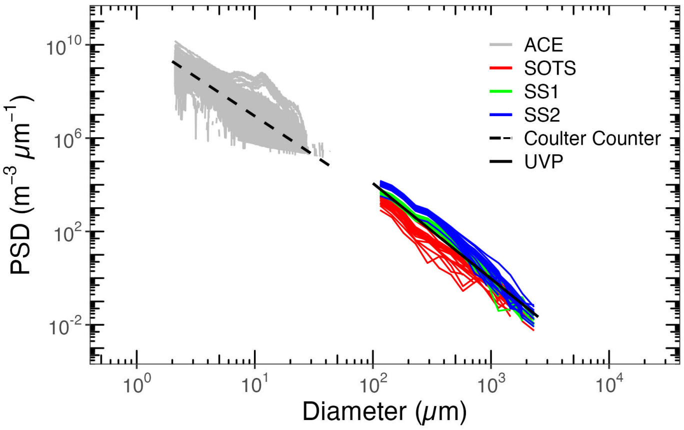


FIGURE 5
Particle size distributions derived from Coulter Counter and UVP-5.

Counter, the slopes determined over different size ranges are quite similar (a mean value of 3.99 for Coulter Counter data and 4.29 for UVP-5 data). Among the SOLACE data, the number of particles is overall larger at SS1 and SS2 than at SOTS, although the latter displays a larger Chl. This observation of more large particles (UVP-5 data start at 100 µm) at the two clearer southern sites seems consistent with the larger NAP contribution already identified for the domain of smaller particles.

PSD slopes (ζ, unitless) are displayed as a function of Chl, colour-mapped as a function of the fraction of microphytoplankton (f_{micro}) derived according to Brewin et al. (2015) (Figure 6). The slope ζ varies between 3 and 5. A somewhat decreasing trend can be discerned for Chl > 0.2 mg m^{-3}, similar to what Buonassissi and Dierssen (2010) have found in temperate regions, yet there is no significant relationship when Chl appears below 0.2 mg m^{-3}, suggesting the particle distribution is heterogenous within the SO. In addition, we did not notice a clear impact of f_{micro} on ζ. The expectation would be that populations dominated by larger cells would have a lower ζ, which is not clearly observed here.

3.4 Implications for ocean colour remote sensing

The absorption of NAP follows an exponential decay from the blue to the red parts of the spectrum. In the clear waters we have analysed here (Chl < 0.2 mg m^{-3}), the large NAP contribution leads to significant non-chlorophyll absorption in the blue part of the

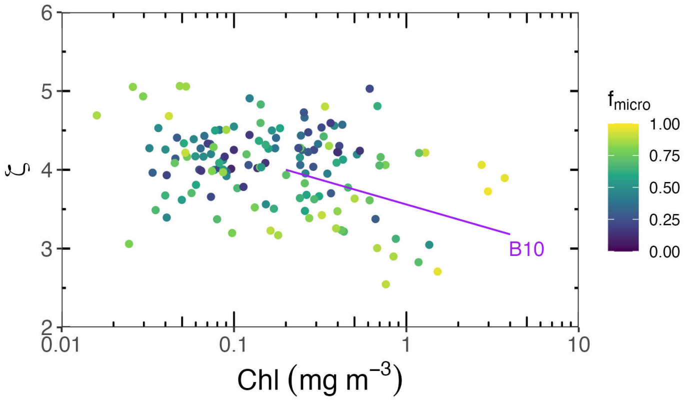


FIGURE 6
PSD slope ζ vs. Chl relationship. The dots are colour-mapped as a function of the fraction of micro-phytoplankton derived from Brewin et al. (2015). The purple line represents the relationship obtained by Buonassissi and Dierssen (2010) in the North Atlantic, where $\zeta = -0.63\,\log_{10}\mathrm{Chl} + 3.56$, $n = 25$, $r^2 = 0.45$.

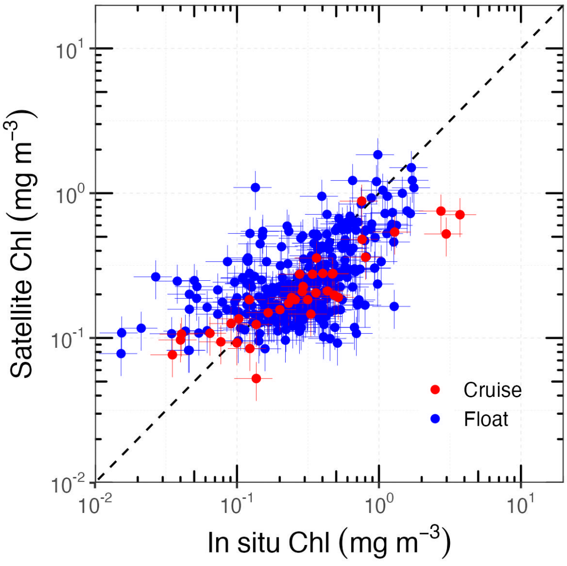


FIGURE 7
Comparisons between in situ and satellite-derived Chl. The black dashed line represents the 1:1 relationship. Error bars indicate the typical 30% uncertainty for both the HPLC- and satellite-derived Chl.

spectrum (λ < ~ 500 nm). A likely consequence is a lower blue-to-green remote-sensing reflectance band ratio than would otherwise exist for the same Chl concentration but with a lower NAP contribution. This would result in a significant overestimation of Chl when using global empirical ocean colour algorithms to derive Chl from the reflectance ratio, as the NAP absorption would be wrongly interpreted as phytoplankton absorption.

To verify this, we compared *in situ* Chl from both cruise and float measurements with satellite-derived Chl estimates. For this purpose, we used the Moderate Resolution Imaging Spectroradiometer (MODIS) L3b binned chlorophyll products, which use a sinusoidal projection so that each grid cell covers the same area, regardless of latitude. For each *in situ* measurement covered by a product, a 3 × 3 window centred on the *in situ* location was extracted. In total, we found 311 *in situ* and satellite-derived Chl matchups, which are displayed in Figure 7.

Previous similar matchup studies generally reported an underestimation of Chl by satellite products in the SO (Garcia et al., 2005; Marrari et al., 2006; Kahru and Mitchell, 2010; Szeto et al., 2011; Johnson et al., 2013; Jena, 2017; Pereira and Garcia, 2018; Moutier et al., 2019). This is confirmed here, but only for Chl > 0.2 mg m^{-3}. For lower values, we conversely observe an overestimation. This is consistent with the excess NAP absorption reported here. It cannot be ruled out, however, that larger absorption by coloured dissolved organic matter (CDOM) would also contribute to this overestimation of Chl (Morel and Gentili, 2009). The absence of reflectance measurements from the ACE and SOLACE voyages did not allow these hypotheses to be further tested here.

4 Conclusions

By combining ship-based measurements from the ACE and SOLACE research voyages and profiling-float-based measurements from over 20,000 profiles collected by the BGC-Argo network, we were able to analyse the general latitudinal distribution patterns of particle-related bio-optical and biogeochemical variables, as well as the associated bio-optical relationships.

At latitudes beyond 60°S, Chl, a_{ph}, b_{bp}, POC, and f_{micro} increase toward the Antarctic continent. In parallel, the chlorophyll-normalised values of a_{ph} and POC decrease, while the chlorophyll-specific b_{bp} remains stable across latitudes.

The absorption data showed a high proportion of NAP (a_{NAP}: a_{ph} up to 10) in the clear waters (Chl < 0.2 mg m^{-3}) of the SO. This substantial NAP contribution leads to higher POC and b_{bp} values, making particulate bio-optical properties significantly different from what is expected for temperate areas. In contrast, this divergence is not evident in waters with Chl > 0.2 mg m^{-3}. Therefore, using bio-optical relationships developed in temperate waters to study the SO is probably acceptable outside the domain of low Chl. The specific relationships we propose here (Figure 4) are presumably better adapted to the SO. Nonetheless, caution is warranted, as even minor alterations in these relationships can result in notable absolute differences due to the substantial variability present.

The implication for satellite ocean colour applications seems to be an overestimation of Chl in clear waters when using standard algorithms (again, in areas with Chl levels below 0.2 mg m^{-3}). Deriving better SO-adapted Chl retrieval algorithms that account for this peculiarity would, however, require more comprehensive datasets of bio-optical properties and radiometry measurements, which still do not exist at the required scale. This is definitely where more effort should be directed if we are to expect significant improvements in our ability to monitor SO ecosystems from space.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: BGC-Argo float data is available at https://biogeochemical-argo.org. Data collected during cruises is available at https://zenodo.org/records/3993096, https://doi.org/10.5281/zenodo.3816726, https://doi.org/10.5281/zenodo.3660852, https://doi.org/10.5281/zenodo.3706710.

Author contributions

JL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. DA: Conceptualization, Data curation, Funding acquisition, Resources, Supervision, Writing – review & editing. YH: Conceptualization, Methodology, Supervision, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. JL was supported by the Australian Research Council Special Research Initiative, Australian Centre for Excellence in Antarctic Science (ACEAS; Project Number SR200100008). ACE was funded by Ferring Pharmaceuticals with additional support from the Swiss Polar Institute. Funding from the Australian Research Council Discovery Program (DP160103387) contributed to the exploitation of ACE data.

Acknowledgments

We would like to thank all members of the ACE Project, including Charlotte Robinson, William Moutier, David Berliner, Alexandra Olivier, and Hazel Little, as well as staff from the Swiss Polar Institute and the crew from the RV Akademik Trynoshnikov. SOLACE received support from the Australian Government through ship time at the Marine National Facility (MNF; R/V Investigator).

We thank the scientific staff and crew of the R/V Investigator for facilitating work at the three SOLACE sites in the Southern Ocean.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allison, D. B., Stramski, D., and Mitchell, B. G. (2010a). Empirical ocean color algorithms for estimating particulate organic carbon in the Southern Ocean. *J. Geophysical Research: Oceans* 115. doi: 10.1029/2009JC006040

Allison, D. B., Stramski, D., and Mitchell, B. G. (2010b). Seasonal and interannual variability of particulate organic carbon within the Southern Ocean from satellite ocean color observations. *J. Geophysical Research: Oceans* 115. doi: 10.1029/2009JC005347

Antoine, D., Siegel, D. A., Kostadinov, T., Maritorena, S., Nelson, N. B., Gentili, B., et al. (2011). Variability in optical particle backscattering in contrasting bio-optical oceanic regimes. *Limnology Oceanography* 56, 955–973. doi: 10.4319/lo.2011.56.3.0955

Antoine, D., Thomalla, S., Berliner, D., Little, H., Moutier, W., Olivier-Morgan, A., et al. (2020). *Phytoplankton pigment concentrations of seawater sampled during the Antarctic Circumnavigation Expedition (ACE) during the Austral Summer of 2016/2017.* doi: 10.5281/zenodo.3816726

Ardyna, M., Claustre, H., Sallée, J.-B., D'Ovidio, F., Gentili, B., van Dijken, G., et al. (2017). Delineating environmental control of phytoplankton biomass and phenology in the Southern Ocean. *Geophysical Res. Lett.* 44, 5016–5024. doi: 10.1002/2016GL072428

Arrigo, K. R., van Dijken, G. L., and Bushinsky, S. (2008). Primary production in the southern ocean 1997–2006. *J. Geophysical Research: Oceans* 113. doi: 10.1029/2007JC004551

Bader, H. (1970). The hyperbolic distribution of particle sizes. *J. Geophysical Res.* 75, 2822–2830. doi: 10.1029/JC075i015p02822

Behrenfeld, M. J., Boss, E., Siegel, D. A., and Shea, D. M. (2005). Carbon-based ocean productivity and phytoplankton physiology from space. *Global Biogeochem. Cycles.* 19. doi: 10.1029/2004GB002299

Belcher, A., Iversen, M., Manno, C., Henson, S. A., Tarling, G. A., and Sanders, R. (2016). The role of particle associated microbes in remineralization of fecal pellets in the upper mesopelagic of the Scotia Sea, Antarctica. *Limnology Oceanography* 61, 1049–1064. doi: 10.1002/lno.10269

Bellacicco, M., Cornec, M., Organelli, E., Brewin, R. J. W., Neukermans, G., Volpe, G., et al. (2019). Global variability of optical backscattering by non-algal particles from a biogeochemical-argo data set. *Geophysical Res. Lett.* 46, 9767–9776. doi: 10.1029/2019GL084078

Bellacicco, M., Volpe, G., Colella, S., Pitarch, J., and Santoleri, R. (2016). Influence of photoacclimation on the phytoplankton seasonal cycle in the Mediterranean Sea as seen by satellite. *Remote Sens. Environ.* 184, 595–604. doi: 10.1016/j.rse.2016.08.004

Boyd, P. W., Claustre, H., Levy, M., Siegel, D. A., and Weber, T. (2019). Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature* 568, 327–335. doi: 10.1038/s41586-019-1098-2

Boyd, P. W., McDonnell, A., Valdez, J., LeFevre, D., and Gall, M. P. (2015). RESPIRE: An in situ particle interceptor to conduct particle remineralization and microbial dynamics studies in the oceans' Twilight Zone. *Limnology Oceanography: Methods* 13, 494–508. doi: 10.1002/lom3.10043

Brewin, R. J. W., Dall'Olmo, G., Sathyendranath, S., and Hardman-Mountford, N. J. (2012). Particle backscattering as a function of chlorophyll and phytoplankton size structure in the open-ocean. *Opt. Express* 20, 17632–17652. doi: 10.1364/OE.20.017632

Brewin, R. J. W., Sathyendranath, S., Jackson, T., Barlow, R., Brotas, V., Airs, R., et al. (2015). Influence of light in the mixed-layer on the parameters of a three-component model of phytoplankton size class. *Remote Sens. Environ.* 168, 437–450. doi: 10.1016/j.rse.2015.07.004

Briggs, N., Dall'Olmo, G., and Claustre, H. (2020). Major role of particle fragmentation in regulating biological sequestration of CO $_2$ by the oceans. *Science* 367, 791–793. doi: 10.1126/science.aay1790

Buesseler, K. O., Lamborg, C. H., Boyd, P. W., Lam, P. J., Trull, T. W., Bidigare, R. R., et al. (2007). Revisiting carbon flux through the ocean's twilight zone. *Science* 316, 567–570. doi: 10.1126/science.1137959

Buonassissi, C. J., and Dierssen, H. M. (2010). A regional comparison of particle size distributions and the power law approximation in oceanic and estuarine surface waters. *J. Geophys. Res.* 115, C10028. doi: 10.1029/2010JC006256

Carranza, M. M., Gille, S. T., Franks, P. J. S., Johnson, K. S., Pinkel, R., and Girton, J. B. (2018). When mixed layers are not mixed. Storm-driven mixing and bio-optical vertical gradients in mixed layers of the southern ocean. *J. Geophysical Research: Oceans* 123, 7264–7289. doi: 10.1029/2018JC014416

Cavan, E. L., Trimmer, M., Shelley, F., and Sanders, R. (2017). Remineralization of particulate organic carbon in an ocean oxygen minimum zone. *Nat. Commun.* 8, 14847. doi: 10.1038/ncomms14847

Cetinić, I., Perry, M. J., Briggs, N. T., Kallin, E., D'Asaro, E. A., and Lee, C. M. (2012). Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment. *J. Geophysical Research: Oceans* 117. doi: 10.1029/2011JC007771

Clements, D. J., Yang, S., Weber, T., McDonnell, A. M. P., Kiko, R., Stemmann, L., et al. (2022). Constraining the particle size distribution of large marine particles in the global ocean with in situ optical observations and supervised learning. *Global Biogeochemical Cycles* 36, e2021GB007276. doi: 10.1029/2021GB007276

Falkowski, P. G., Barber, R. T., and Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science* 281, 200–206. doi: 10.1126/science.281.5374.200

Fawcett, S., and Forrer, H. (2020). Particulate organic carbon concentration in seawater profiles collected on board the R/V Akademik Tryoshnikov in the Southern Ocean during the austral summer of 2016/2017 as part of the Antarctic Circumnavigation Expedition (ACE). *Zenodo.* doi: 10.5281/zenodo.3706710

Garcia, C. A. E., Garcia, V. M. T., and McClain, C. R. (2005). Evaluation of SeaWiFS chlorophyll algorithms in the Southwestern Atlantic and Southern Oceans. *Remote Sens. Environ.* 95, 125–137. doi: 10.1016/j.rse.2004.12.006

Graff, J. R., Westberry, T. K., Milligan, A. J., Brown, M. B., Dall'Olmo, G., Dongen-Vogels, V., et al. (2015). Analytical phytoplankton carbon measurements spanning diverse ecosystems. *Deep Sea Res. Part I: Oceanographic Res. Papers* 102, 16–25. doi: 10.1016/j.dsr.2015.04.006

Gruber, N., Gloor, M., Mikaloff Fletcher, S. E., Doney, S. C., Dutkiewicz, S., Follows, M. J., et al. (2009). Oceanic sources, sinks, and transport of atmospheric CO2. *Global Biogeochemical Cycles* 23. doi: 10.1029/2008GB003349

Hirawake, T., Takao, S., Horimoto, N., Ishimaru, T., Yamaguchi, Y., and Fukuchi, M. (2011). A phytoplankton absorption-based primary productivity model for remote sensing in the Southern Ocean. *Polar Biol.* 34, 291–302. doi: 10.1007/s00300-010-0949-y

Hooker, S. B., and Zibordi, G. (2005). Advanced methods for characterizing the immersion factor of irradiance sensors. *J. Atmos. Oceanic Technol.* 22, 757–770. doi: 10.1175/JTECH1736.1

Huot, Y., Morel, A., Twardowski, M. S., Stramski, D., and Reynolds, R. A. (2008). Particle optical backscattering along a chlorophyll gradient in the upper layer of the eastern South Pacific Ocean. *Biogeosciences* 5, 495–507. doi: 10.5194/bg-5-495-2008

Jackson, G. A. (1980). Phytoplankton growth and Zooplankton grazing in oligotrophic oceans. *Nature* 284, 439–441. doi: 10.1038/284439a0

Jackson, G. A., and Burd, A. B. (1998). Aggregation in the marine environment. *Environ. Sci. Technol.* 32, 2805–2814. doi: 10.1021/es980251w

Jena, B. (2017). The effect of phytoplankton pigment composition and packaging on the retrieval of chlorophyll-a concentration from satellite observations in the Southern Ocean. *Int. J. Remote Sens.* 38, 3763–3784. doi: 10.1080/01431161.2017.1308034

Johnson, K. S., Plant, J. N., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Riser, S. C., et al. (2017). Biogeochemical sensor performance in the SOCCOM profiling float array. *J. Geophysical Research: Oceans* 122, 6416–6436. doi: 10.1002/2017JC012838

Johnson, R., Strutton, P. G., Wright, S. W., McMinn, A., and Meiners, K. M. (2013). Three improved satellite chlorophyll algorithms for the Southern Ocean. *J. Geophysical Research: Oceans* 118, 3694–3703. doi: 10.1002/jgrc.20270

Jonasz, M., and Fournier, G. R. (2011). *Light Scattering by Particles in Water: Theoretical and Experimental Foundations.* (Elsevier). doi: 10.1016/B978-0-12-388751-1.X5000-5

Kahru, M., and Mitchell, B. G. (2010). Blending of ocean colour algorithms applied to the Southern Ocean. *Remote Sens. Lett.* 1, 119–124. doi: 10.1080/01431160903547940

Karakuş, O., Völker, C., Iversen, M., Hagen, W., and Hauck, J. (2022). The role of zooplankton grazing and nutrient recycling for global ocean biogeochemistry and phytoplankton phenology. *J. Geophysical Research: Biogeosciences* 127, e2022JG006798. doi: 10.1029/2022JG006798

Kinsman, S. (2018). "Particle Size Instrumentation — Coulter® Counter," in *Particle Characterization in Technology*. Ed. J. K. Beddow (Boca Raton: CRC Press), 183–186. doi: 10.1201/9781351075350-9

Loisel, H., Bosc, E., Stramski, D., Oubelkheir, K., and Deschamps, P.-Y. (2001). Seasonal variability of the backscattering coefficient in the Mediterranean Sea based on satellite SeaWiFS imagery. *Geophysical Res. Lett.* 28, 4203–4206. doi: 10.1029/2001GL013863

Maffione, R. A., and Dana, D. R. (1997). Instruments and methods for measuring the backward-scattering coefficient of ocean waters. *Appl. Optics* 36, 6057–6067. doi: 10.1364/AO.36.006057

Marrari, M., Hu, C., and Daly, K. (2006). Validation of SeaWiFS chlorophyll a concentrations in the Southern Ocean: A revisit. *Remote Sens. Environ.* 105, 367–375. doi: 10.1016/j.rse.2006.07.008

Martinez-Vicente, V., Dall'Olmo, G., Tarran, G., Boss, E., and Sathyendranath, S. (2013). Optical backscattering is correlated with phytoplankton carbon across the Atlantic Ocean. *Geophys. Res. Lett.* 40, 1154–1158. doi: 10.1002/grl.50252

Mignot, A., Ferrari, R., and Claustre, H. (2018). Floats with bio-optical sensors reveal what processes trigger the North Atlantic bloom. *Nat. Commun.* 9, 190. doi: 10.1038/s41467-017-02143-6

Morel, A., and Gentili, B. (2009). A simple band ratio technique to quantify the colored dissolved and detrital organic material from ocean color remotely sensed data. *Remote Sens. Environ.* 113, 998–1011. doi: 10.1016/j.rse.2009.01.008

Morel, A., and Maritorena, S. (2001). Bio-optical properties of oceanic waters: A reappraisal. *J. Geophys. Res.* 106, 7163–7180. doi: 10.1029/2000JC000319

Moutier, W., Thomalla, S., Bernard, S., Wind, G., Ryan-Keogh, T., and Smith, M. (2019). Evaluation of chlorophyll-a and POC MODIS aqua products in the southern ocean. *Remote Sens.* 11, 1793. doi: 10.3390/rs11151793

Park, Y.-H., Park, T., Kim, T.-W., Lee, S.-H., Hong, C.-S., Lee, J.-H., et al. (2019). Observations of the antarctic circumpolar current over the udintsev fracture zone, the narrowest choke point in the southern ocean. *J. Geophysical Research: Oceans* 124, 4511–4528. doi: 10.1029/2019JC015024

Pereira, E. S., and Garcia, C. A. E. (2018). Evaluation of satellite-derived MODIS chlorophyll algorithms in the northern Antarctic Peninsula. *Deep Sea Res. Part II: Topical Stud. Oceanography* 149, 124–137. doi: 10.1016/j.dsr2.2017.12.018

Picheral, M., Guidi, L., Stemmann, L., Karl, D. M., Iddaoud, G., and Gorsky, G. (2010). The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology Oceanography: Methods* 8, 462–473. doi: 10.4319/lom.2010.8.462

Ras, J., Claustre, H., and Uitz, J. (2008). Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: comparison between in situ and predicted data. *Biogeosciences* 5, 353–369. doi: 10.5194/bg-5-353-2008

Reynolds, R. A., Stramski, D., and Neukermans, G. (2016). Optical backscattering by particles in Arctic seawater and relationships to particle mass concentration, size distribution, and bulk composition: Particle backscattering in Arctic seawater. *Limnol. Oceanogr* 61, 1869–1890. doi: 10.1002/lno.10341

Robinson, C. M., Huot, Y., Schuback, N., Ryan-Keogh, T. J., Thomalla, S. J., and Antoine, D. (2021). High latitude Southern Ocean phytoplankton have distinctive bio-optical properties. *Opt. Express* 29, 21084–21112. doi: 10.1364/OE.426737

Roesler, C., Uitz, J., Claustre, H., Boss, E., Xing, X., Organelli, E., et al. (2017). Recommendations for obtaining unbiased chlorophyll estimates from in situ chlorophyll fluorometers: A global analysis of WET Labs ECO sensors. *Limnology Oceanography: Methods* 15, 572–585. doi: 10.1002/lom3.10185

Sarmiento, J. L., Johnson, K. S., Arteaga, L. A., Bushinsky, S. M., Cullen, H. M., Gray, A. R., et al. (2023). The Southern Ocean carbon and climate observations and modeling (SOCCOM) project: a review. *Prog. Oceanogr.* 219, 103130. doi: 10.1016/j.pocean.2023.103130

Sathyendranath, S., Platt, T., and Forget, M.-H. (2007). "Oceanic Primary Production: Comparison of Models," in *OCEANS 2007*. (Europe, Aberdeen, UK. (IEEE), 1–3. doi: 10.1109/OCEANSE.2007.4302468

Sathyendranath, S., Stuart, V., Nair, A., Oka, K., Nakane, T., Bouman, H., et al. (2009). Carbon-to-chlorophyll ratio and growth rate of phytoplankton in the sea. *Mar. Ecol. Prog. Ser.* 383, 73–84. doi: 10.3354/meps07998

Schallenberg, C., Harley, J. W., Jansen, P., Davies, D. M., and Trull, T. W. (2019). Multi-year observations of fluorescence and backscatter at the southern ocean time series (SOTS) shed light on two distinct seasonal bio-optical regimes. *Front. Mar. Sci.* 6. doi: 10.3389/fmars.2019.00595

Schallenberg, C., Strzepek, R. F., Bestley, S., Wojtasiewicz, B., and Trull, T. W. (2022). Iron limitation drives the globally extreme fluorescence/chlorophyll ratios of the southern ocean. *Geophysical Res. Lett.* 49, e2021GL097616. doi: 10.1029/2021GL097616

Slade, W. H., Boss, E., and Russo, C. (2011). Effects of particle aggregation and disaggregation on their inherent optical properties. *Opt. Express* 19, 7945–7959. doi: 10.1364/OE.19.007945

Stark, J. S., Raymond, T., Deppeler, S. L., and Morrison, A. K. (2019). "Chapter 1 - Antarctic Seas," in *World Seas: an Environmental Evaluation (Second Edition)*. Ed. C. Sheppard (Cambridge, Massachusetts: Academic Press), 1–44. doi: 10.1016/B978-0-12-805068-2.00002-4

Steinberg, D. K., and Landry, M. R. (2017). Zooplankton and the ocean carbon cycle. *Annu. Rev. Mar. Sci.* 9, 413–444. doi: 10.1146/annurev-marine-010814-015924

Stramska, M., Stramski, D., Hapter, R., Kaczmarek, S., and Stoń-Egiert, J. (2003). Bio-optical relationships and ocean color algorithms for the north polar region of the Atlantic. *J. Geophys. Res.* 108, 3143. doi: 10.1029/2001JC001195

Stramski, D., Reynolds, R. A., Babin, M., Kaczmarek, S., Lewis, M. R., Röttgers, R., et al. (2008). Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans. *Biogeosciences* 5, 171–201. doi: 10.5194/bg-5-171-2008

Stramski, D., Reynolds, R. A., Kahru, M., and Mitchell, B. G. (1999). Estimation of particulate organic carbon in the ocean from satellite remote sensing. *Science* 285, 239–242. doi: 10.1126/science.285.5425.239

Szeto, M., Werdell, P. J., Moore, T. S., and Campbell, J. W. (2011). Are the world's oceans optically different? *J. Geophysical Research: Oceans* 116. doi: 10.1029/2011JC007230

Thomalla, S. J., Ogunkoya, A. G., Vichi, M., and Swart, S. (2017). Using optical sensors on gliders to estimate phytoplankton carbon concentrations and chlorophyll-to-carbon ratios in the southern ocean. *Front. Mar. Sci.* 4. doi: 10.3389/fmars.2017.00034

Turner, J. T. (2015). Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Prog. Oceanography* 130, 205–248. doi: 10.1016/j.pocean.2014.08.005

Uchida, T., Balwada, D., Abernathey, R., Prend, C. J., Boss, E., and Gille, S. T. (2019). Southern ocean phytoplankton blooms observed by biogeochemical floats. *J. Geophysical Research: Oceans* 124, 7328–7343. doi: 10.1029/2019JC015355

Walton, D. W. H., and Thomas, J. (2018). *Cruise report - antarctic circumnavigation expedition (ACE) 20th december 2016 - 19th march 2017*. doi: 10.5281/zenodo.1443511

Wynn-Edwards, C., Davies, D. M., Jansen, P., Bray, S. G., Eriksen, R., and Trull, T. W. (2019). *IMOS-Southern Ocean Time Series (SOTS)-Annual Reports: 2012/2013*. (Australian Integrated Marine Observing System, University of Tasmania).

Zhang, Y. L., Liu, M. L., Wang, X., Zhu, G. W., and Chen, W. M. (2009). Bio-optical properties and estimation of the optically active substances in Lake Tianmuhu in summer. *Int. J. Remote Sens.* 30, 2837–2857. doi: 10.1080/01431160802558592

TYPE Original Research
PUBLISHED 07 November 2024
DOI 10.3389/fmars.2024.1456480

Optimizing data-driven arctic marine forecasting: a comparative analysis of MariNet, FourCastNet, and PhyDNet

Aleksei V. Buinyi[1,2]*, Dias A. Irishev[1], Edvard E. Nikulin[1], Aleksandr A. Evdokimov[3], Polina G. Ilyushina[4] and Natalia A. Sukhikh[1,2]

[1]Research and Development Department, Marine Information Technologies LLC, Moscow, Russia, [2]Department of Hydrometeorological Modeling, Lomonosov Moscow State University Marine Research Center (LMSU MRC), Moscow, Russia, [3]Department of Hydrometeorological Research, Lomonosov Moscow State University Marine Research Center (LMSU MRC), Moscow, Russia, [4]Geoinformation Technologies Department, Lomonosov Moscow State University Marine Research Center (LMSU MRC), Moscow, Russia

OPEN ACCESS

EDITED BY
Jianchuan Yin,
Guangdong Ocean University, China

REVIEWED BY
Richard Arthur Allard,
Naval Research Laboratory, United States
Qiannan Zhang,
Shanghai Maritime University, China

*CORRESPONDENCE
Aleksei V. Buinyi
a.buinyi@marine.tech

RECEIVED 28 June 2024
ACCEPTED 10 October 2024
PUBLISHED 07 November 2024

CITATION
Buinyi AV, Irishev DA, Nikulin EE, Evdokimov AA, Ilyushina PG and Sukhikh NA (2024) Optimizing data-driven arctic marine forecasting: a comparative analysis of MariNet, FourCastNet, and PhyDNet. *Front. Mar. Sci.* 11:1456480. doi: 10.3389/fmars.2024.1456480

Introduction: Marine forecasts play a crucial role in ensuring safe navigation, efficient offshore operations, coastal management, and research, particularly in regions with challenging conditions like the Arctic Ocean. These forecasts necessitate precise predictions of ocean currents, wind-driven waves, and various other oceanic parameters. Although physics-based numerical models are highly accurate, they come with significant computational requirements. Therefore, data-driven approaches, which are less computationally intensive, may present a more effective solution for predicting sea conditions.

Methods: This study introduces a detailed analysis and comparison of three data-driven models: the newly developed convLSTM-based MariNet, FourCastNet, and PhydNet, a physics-informed model designed for video prediction. Through the utilization of metrics such as RMSE, Bias, and Correlation, we illustrate the areas in which our model outperforms well-known prediction models.

Results: Our model demonstrates enhanced accuracy in forecasting ocean dynamics when compared to FourCastNet and PhyDNet. Additionally, our findings reveal that our model demands significantly less training data and computational resources, ultimately resulting in lower carbon emissions.

Discussion: These findings indicate the potential for further exploration of data-driven models as a supplement to physics-based models in operational marine forecasting, as they have the capability to improve prediction accuracy and efficiency, thereby facilitating more responsive and cost-effective forecasting systems.

KEYWORDS

Arctic, machine learning, ocean prediction, LSTM, short-term forecast

1 Introduction

Machine Learning is the process of making computer systems learn without explicit instructions by analyzing and drawing inferences from data patterns using algorithms and statistical models. One of the major limitations of Artificial Intelligence and Machine Learning has always been computational power, which has been a cause of concern for researchers. CPUs were not as powerful and efficient a few decades ago when it came to running large computations for machine learning. Hardware manufacturers have worked hard to create a processing unit capable of performing any AI operation.

Though CPUs are no longer viable sources of computational power, they were the pioneers. Today, those CPUs are rightfully replaced by GPUs and AI accelerators, specifically designed for large computing. The main features considered while purchasing an AI accelerator are cost, energy consumption, and processing speed.

The study of ocean circulation is crucial for many reasons, including the climate research, determining marine life distribution, shaping human activity, and more. Accurate prediction of currents can help forecast weather, estimate energy transfer rates in the ocean, predict the spread of oil spills and drift of the sea ice and icebergs. Sediment transport is another important correlated aspect correlated with the water circulation, affecting marine economic activities such as fishing, transport, logistics, and tourism. Therefore, in the seas, especially in the high latitudes, the prediction of currents is crucial for port, pipeline, and logistics development, as well as for the analysis of sea ice drift for safe logistics. In this context, the development of a machine learning model for the prediction of sea water movement and sea level variations is essential.

Sea currents and sea surface level prediction have a long history of development, starting with traditional empirical methods and evolving into modern AI methodologies. The early efforts held in the 17th -19th centuries (e.g. Halley, 1686; Maury, 1855) and relied on accidental *in situ* observations. With the transition from single observations to systematic measurements, the emergence of scientists specializing in hydrodynamics and ocean studies, the development of a network of observation stations and scientific equipment, analytical methods of describing observed phenomena were formed in (Navier, 1822; Stokes, 1845) and numerically solved in (Bjerknes, 1903, 2023). In the early 20th century, V. Walfrid Ekman's research on wind-driven surface currents laid important groundwork for understanding ocean transport mechanisms. It laid the foundation of geophysical fluid dynamics and led to the pioneering work of numerical weather forecasting of (Richardson, 1922). The first numerical forecasts in oceanography were developed for the wind-driven waves by (Sverdrup and Munk, 1947). Development of numerical methods based on solving the Navier-Stokes equations continued in the ocean simulations with the first models (Bryan, 1969) and succeeded in mesoscale ocean circulation forecasting by 1983 (Robinson, 1983). Over time, increased computational power and improved mathematical representations of ocean processes have enabled more sophisticated forecasting models. The satellite remote sensing era, that began nearly at the same time, provided massive volume of data for observing and assimilating sea surface height data into models.

All it made the global ocean reanalysis and forecasting projects available. Operational forecasting centers like the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Oceanic and Atmospheric Administration (NOAA) began running global ocean prediction systems to support weather, climate and marine applications (Storto et al., 2019), while regional models with finer resolutions also emerged for areas like the Arctic (Chen et al., 2009).

With the availability of petabytes of oceanographic and remote sensing observations, with the outputs of numerical model simulations, with the growth of computational power, artificial intelligence (AI) tools are increasingly being leveraged in a variety of applications in oceanography (Dong et al., 2022). The high energy efficiency of the AI models (e.g. (Pathak et al., 2022) also contributes to their spreading.

Various AI algorithms are now being used for the identification of mesoscale eddies (Franz et al., 2018; Lguensat et al., 2018; Du et al., 2019; Duo et al., 2019; Xu et al., 2019, 2021; Santana et al., 2020), forecasting surface waves (Mandal and Prabaharan, 2006; Fan et al., 2020; Gao et al., 2021; Zhou et al., 2021; Buinyi et al., 2022), prediction of features, like the Indian Ocean Dipole, with a multi-task deep learning model in (Ling et al., 2022), that outperformed traditional numerical multiseasonal prediction.

The topics of sea surface heights and currents forecasting are also covered with the AI methods. One approach is the use of deep learning methods such as ConvLSTMP3, which extracts spatial-temporal features of sea surface heights using convolutional operations and long short-term memory (LSTM) (Song et al., 2021). One more paper (Zulfa et al., 2021) uses LSTM to predict sea surface velocity and direction, achieving good results with low Mean Absolute Percentage Error (MAPE) values in Labuan Bajo waters. In the paper (Ning et al., 2021) an optimized Simple Recurrent Unit (SRU) deep network was developed for short-to-medium-term sea surface height prediction with AVISO data.

There are a lot of promising results in geosciences now. We have created MariNet, the ML architecture, and compared its output with two state-of-art ML models of different architectures to test their ability in the Arctic region forecasting. In the current work, we test the algorithms on the surface currents data and sea surface heights.

2 Materials and methods

In the initial stages of our research, we harnessed PhyDNet and FourCastNet, two of the most promising machine learning architectures applicable to the ocean state forecasting available at the time, for the comparison with MariNet, our Neural Network. The neural networks are described below.

2.1 MariNet neural network

MariNet is an artificial neural network (ANN) based on the parallel encoder-decoder architecture within which ConvLSTM modules are embedded in latent space (Buinyi et al., 2023). The

ConvLSTM itself is introduced by (Shi et al., 2015) and described as a type of neural network architecture that combines convolutional and LSTM layers. Because of its successful design, it has been used for spatiotemporal data analysis and prediction in various applications, including precipitation nowcasting (Shi et al., 2015), air temperature forecasting (Lin et al., 2019), flood forecasting (Moishin et al., 2021), arctic sea ice concentration prediction (Liu et al., 2021), and seismic events prediction (Fuentes et al., 2021), with relatively high reliability.

The architecture of our model is shown on the Figure 1. MariNet consists of several interconnected encoder-decoder blocks, within which ConvLSTM modules are embedded between the encoder and the decoder. For this study, we employed four encoder layers and four decoder layers. Each ConvLSTM module contains several parallel ConvLSTM cells connected in a manner that the sum of their outputs forms the resulting forecast of time series in the latent space. This design enables the neural network to simultaneously detect temporal dependencies at various frequencies without assuming any specific frequency distribution and *a priori* defined data distributions.

The encoder-decoder blocks are interconnected in such a way that the input to each subsequent block is the result of subtracting the original data from the original data passed through the first convolutional layer in the block, which produces average pooling. Moreover, the size of the convolution in the first layer of the block varies for each block. This solution facilitates hierarchical pattern highlighting in images: first, the neural network is trained to work with larger patterns. Then it analyzes smaller patterns and their conditional dependencies on larger ones.

During this research, we employed three encoder-decoder architectures. We settled on using three parallel architectures because they enabled us to capture different scales of spatial and temporal variability. The first block learns and captures the largest features, which can be considered as general sea state variability. The second block focuses on large-scale patterns, such as global circulation dynamics. Meanwhile, the third block captures the finest details.

A key feature of the model's operation is the forecasting algorithm. Unlike typical recursive algorithms, where the forecast from the previous step is cyclically fed into the neural network to form a forecast for the next steps, our neural network sequentially receives several previous values for the water velocity and sea surface heights. When predicting the sea state 3 days ahead with a temporal resolution of 6 hours (i.e., 12 timesteps), we provide the model with 12 consecutive input sequences. Therefore, instead of getting a single array for one time point, our neural network is initialized by the dynamics of such arrays, which allows for a more accurate assessment of the state of the forecasted values, and consequently, ensures a more precise forecast. To train MariNet, we used a learning rate of 0.001, a sigmoid activation function, and the Adam optimizer. The number of training epochs turned out to be 300.

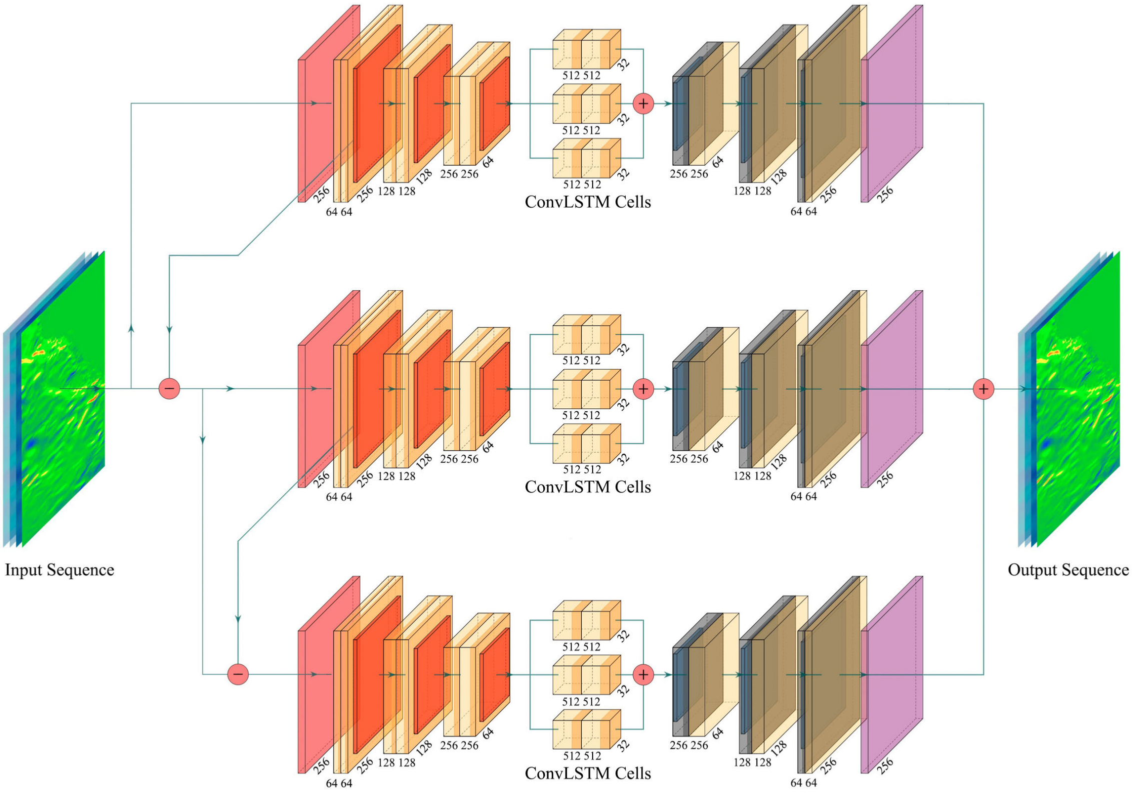


FIGURE 1
The architecture of MariNet.

2.2 PhyDNet neural network

PhyDNet is a deep learning model introduced in (Le Guen and Thome, 2020) and designed for unsupervised video prediction. Due to its architecture, the model integrates physical knowledge into the learning process, making it effective for tasks such as weather forecasting, fluid dynamics, and other physical phenomena prediction. The model leverages physical knowledge on dynamics and disentangles it from other unknown factors. To achieve this goal, the authors introduced a PhyDNet disentangling architecture, and PhyCell physically-constrained recurrent cell. The recurrent block projects input video frames into a latent space. This projection is achieved through a deep convolutional encoder, which transforms the input video into a lower-dimensional representation. The latent space is where the disentanglement of physical dynamics and residual information occurs. Two parallel neural networks are responsible for it: PhyCell and ConvLSTM. PhyCell is a recurrent cell that models and solves Partial Differential Equations (PDE) with internal physical predictor computing and combining partial derivatives with convolutions. PhyCell allows exploiting prior physical knowledge to improve prediction of a model, add explainability and leverages physical constraints to limit the number of model parameters. The ConvLSTM network is trained to learn the residuals, or errors, of the physical model's predictions. By learning these residuals, the network can correct the physical model's predictions and improve the overall accuracy of the system. Learned physical and residual representations are summed before decoding to predict the future video frame. As a result, PhyDNet generates one-step-ahead prediction that can be extended by recursive feeding predicted frame into the model. It's important to note that predictions are reinjected as the next input only for the ConvLSTM branch, and not for PhyCell. This is because the PhyCell is designed to capture the deterministic physical dynamics, which should not be influenced by the predictions.

In (Le Guen and Thome, 2020) PhyDNet has been compared with PredRNN, ConvLSTM, Causal LSTM, Memory in Memory (MIM), outperformed them and showed itself as one of the state-of-the-art model of its time. Therefore, we have chosen PhyDNet to compare with our model.

2.3 FourCastNet neural network

FourCastNet, or Fourier ForeCasting Neural Network is first described in (Pathak et al., 2022). It is a data-driven global weather forecasting model that provides short to medium range predictions. It is trained with an ERA5 reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF), which has hourly estimates of atmospheric variables at a 0.25° resolution. FourCastNet utilizes a Fourier transform-based token-mixing scheme (Guibas et al., 2021) which is complemented with a vision transformer (ViT) backbone (Dosovitskiy et al., 2021). This method is grounded in the recent advancements in the Fourier neural operator, or Adaptive Fourier Neural Operator (AFNO) that has demonstrated success in modeling challenging partial differential equations (PDE), including fluid dynamics, in a resolution-invariant manner (Li et al., 2020).

According to (Pathak et al., 2022) the use of ViT backbone is preferred due to its ability to effectively model long-range dependencies. The combination of ViT and Fourier-based token mixing produces a model that effectively resolves fine-grained features and scales well with the size and resolution of the dataset, leading to the training of high-fidelity data-driven models at an unprecedented resolution.

The original version of FourCastNet models 20 variables at five vertical levels, that are: surface air pressure, mean sea level pressure, air temperature at 2m from the surface, zonal and meridional wind velocity 10m from the surface; zonal and meridional wind velocity at 1000, 850, and 500 hPa; air temperature at 850 and 500 hPa; geopotential at 1000, 850, 500, and 50hPa; relative humidity at 850 and 500hPa, and Total Column Water Vapor. The authors use the model to predict such variables as the surface wind speed, precipitation, and atmospheric water vapor. They propose FourCastNet to be used for planning wind energy resources, predicting extreme weather events such as tropical cyclones, extra-tropical cyclones, and atmospheric rivers. FourCastNet matches the forecasting accuracy of the ECMWF Integrated Forecasting System (IFS), a state-of-the-art Numerical Weather Prediction (NWP) model, at short lead times for large-scale variables, while outperforming IFS for small-scale variables, including precipitation.

According to (Pathak et al., 2022), the FourCastNet uses such metrics as Root Mean Squared Error (RMSE), Anomaly Correlation Coefficient (ACC) at lead times of up to three days and gives results comparable to the ECMWF Integrated Forecasting System (IFS), considered one of the best classical numerical model used by ECMWF to construct reanalyzes and make weather forecasts.

2.4 Metrics for the model output quality estimation

We trained all three networks with the data on the surface water currents and the sea surface heights, started the inferences and compared their outputs with several metrics: Root Mean Squares Error (RMSE), Bias and Correlation. They are defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2},$$

$$Bias = \frac{1}{N}\sum_{i=1}^{N}y_i - \frac{1}{N}\sum_{i=1}^{N}x_{i,}$$

$$Correlation = \frac{Sxy}{\sqrt{SxxSyy}}, \text{with}$$

$$S_{xy} = \sum_{i=1}^{N}x_i\, y_i - \frac{\sum_{i=1}^{N}x_i \sum_{i=1}^{N}y_i}{N},$$

$$S_{xx} = \sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N},$$

$$S_{yy} = \sum_{i=1}^{N} y_i^2 - \frac{(\sum_{i=1}^{N} y_i)^2}{N},$$

where x_i is the original data value for a given timestep, y_i is a predicted value for a given timestep, N – the length of the timeseries.

In scholarly terms, the Root Mean Square Error (RMSE) quantifies the divergence in magnitude between the model's predictions and the actual observations. It is preferable for the RMSE to be smaller as this signifies a better alignment between predicted and actual values.

Bias, on the other hand, signifies the systematic deviation of the approximated quantifier from the real value and can be interpreted as a consistent overestimation or underestimation of an output. It is desirable for the bias to be closer to zero, indicating that the estimates are nearer to the actual data.

The correlation, in contrast, is a statistical measure that sheds light on the degree to which two variables share a linear relationship. This relationship is frequently deployed to depict the linear association between two contingent factors. Greater values of correlation denote a stronger relationship between the two variables.

3 Data

The Copernicus Marine Environment Monitoring Service (CMEMS) offers a comprehensive global ocean analysis and forecast system through its Global Ocean Physics Analysis and Forecast (CMEMS-GLO-PUM-001-024) product. The system operates at a resolution of 1/12°, updated daily, and provides global ocean forecasts for a 10-day period (Operational Mercator Global Ocean System). The dataset employs a combination of the numerical ocean model NEMO 3.6 with LIM3 Multi-categories sea ice model, ECMWF IFS HRES atmospheric forcing, and several data assimilation techniques, like SAM2 (SEEK Kernel) 4D, allowing for seamless integration of *in-situ* and satellite observations.

For our needs we choose the region bounded by 60°N-90°N and 5°E-150°W and obtain the hourly surface data of zonal sea water velocity (u), meridional sea water velocity (v), and sea surface height above geoid (zos) for 2019-2022. We interpolate them to the 6-hour temporal resolution and 0.25°x0.25° spatial resolution with and feed the data to the ML models.

4 Results and discussion

The MariNet model demonstrates promising performance. Notably, the figures representing metrics for the FourCastNet model display artifacts. The average metrics are presented in Table 1. As shown in the table, the MariNet model demonstrates minimal RMSE values for sea surface heights and components of surface sea water velocities. Furthermore, the bias of the MariNet model is closest to zero among the mentioned models. Although the mean correlation between models is not significantly high, PhyDNet and MariNet display the highest correlation, approximately 0.5 for sea surface velocities and 0.4 for sea surface heights.

Figure 2 illustrates the temporal evolution of RMSE for sea surface velocity and sea surface heights for the selected models throughout the prediction period. As observed, all RMSE values monotonically increase over time. Notably, MariNet and PhyDNet demonstrate comparable results, with their RMSE values growing from approximately 0.01 m/s and 0.01 m to about 0.04 m/s and 0.045 m for sea water velocity and sea surface heights, respectively.

Figures 3–8 depict maps of RMSE for zonal and meridional components of surface water velocity within the research area. All three models exhibit similar spatial distributions of RMSE values, with high values observed in the Barents Sea, Kara Sea, and coastal areas of other seas, as well as low values in eastern offshore regions and the region to the north of the continental shelf. The high RMSE values may be attributed to two primary factors: (1) the model's poor learning quality or (2) the high standard deviation and variability of the original data. Conversely, the low RMSE values observed in the eastern and northern parts may be due to the relatively low variability of the original water velocity data, with the latter potentially being exacerbated by the presence of sea ice cover in these areas for a significant portion of the year. On the other hand, the high RMSE values in coastal areas could be attributed to the active hydrodynamics in these regions, characterized by larger values and greater variability of the water velocity data.

Notably, the general patterns of spatial variability for RMSEs are consistently present across all models' results; however, MariNet outperforms the other two neural networks in terms of absolute error values. Furthermore, FourCastNet is observed to exhibit artifacts both in zonal and meridional components of surface water velocity. This could indicate that the FourCastNet model fails to accurately capture sea water dynamics patterns.

TABLE 1 Metrics of MariNet, FourCastNet and PhyDNet for zonal and meridional components of surface water velocities and Sea Surface Heights.

	Model	RMSE (m/s)	Bias (m/s)	Correlation
u	MariNet	**0.027**	**-0.001**	0.507
	FourCastNet	0.051	0.003	0.432
	PhyDNet	0.043	0.004	**0.519**
v	MariNet	**0.028**	**0**	0.515
	FourCastNet	0.051	0.002	0.428
	PhyDNet	0.044	**0**	**0.524**
ssh	MariNet	**0.027**	**-0.001**	0.430
	FourCastNet	0.082	-0.050	0.367
	PhyDNet	0.046	0.003	**0.451**

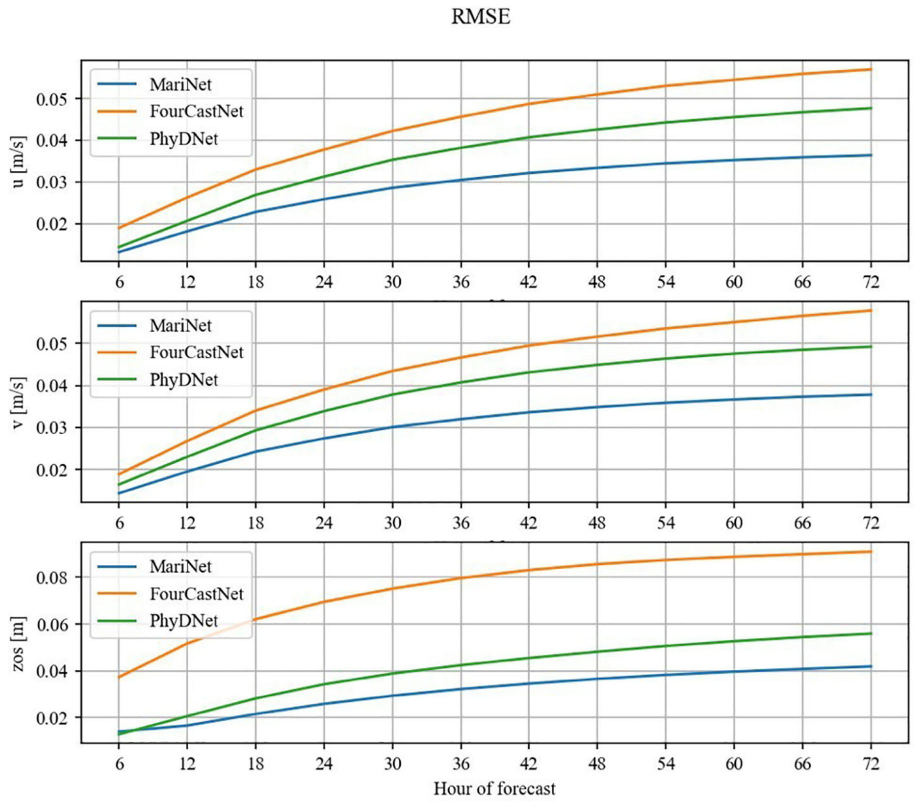


FIGURE 2
Plots of temporal evolution of RMSE for zonal (upper image), meridional (middle image) components of surface water velocity (m/s), and the sea surface heights (m) above geoid for MariNet (blue line), PhyDNet (green line), and FourCastNet (orange line).

4.1 Computational cost of MariNet

With the CodeCarbon software package, we have calculated the carbon emissions and the energy consumption of the MariNet, FourCastNet and the PhyDNet for our calculations. Results are shown in the Table 2. Training of the MariNet model has the least carbon emission rate, but, due to the relatively large time of training, it takes the most energy. At the same time, PhyDNet wins the energy consumption and the emission rate challenges.

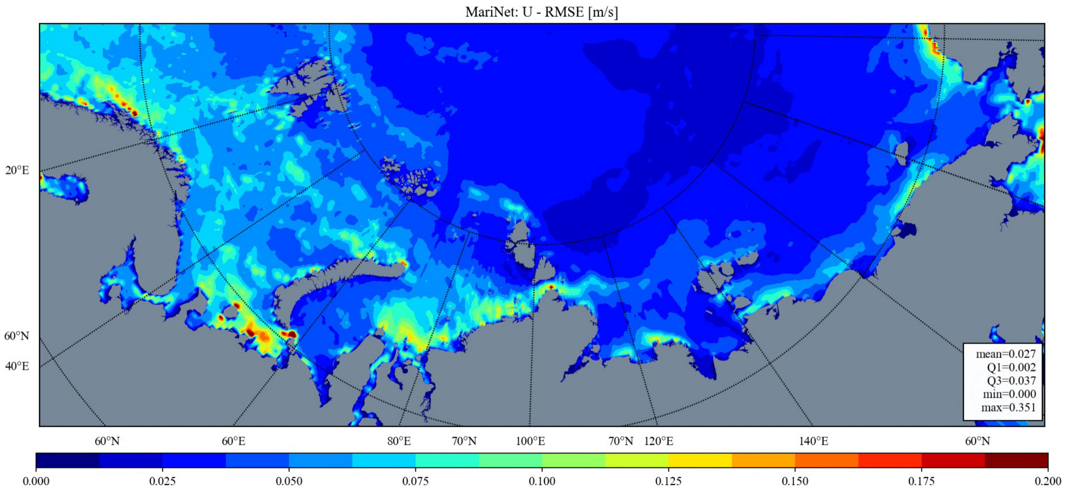


FIGURE 3
RMSE (in m/s) for zonal component of the surface water velocity for MariNet model.

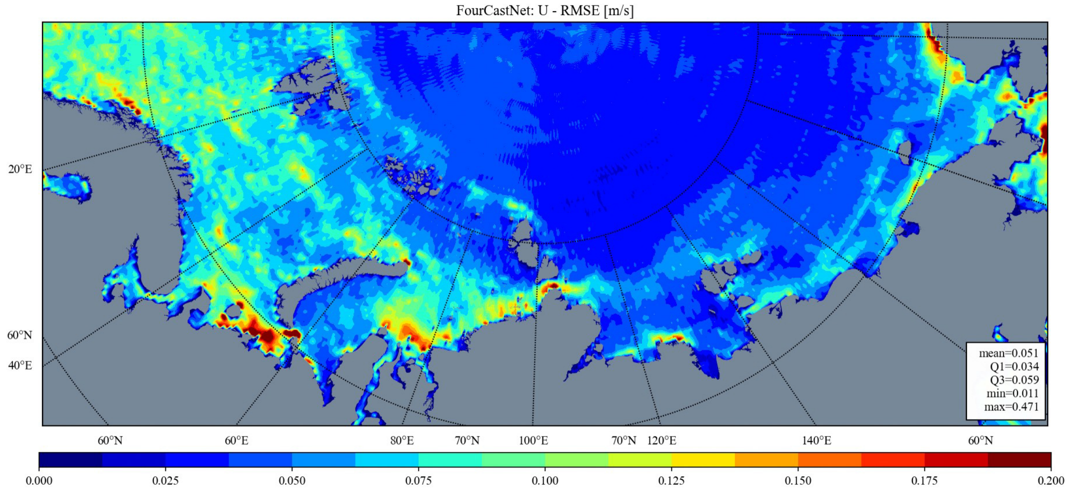


FIGURE 4
RMSE (in m/s) for zonal component of the surface water velocity for FourCastNet model.

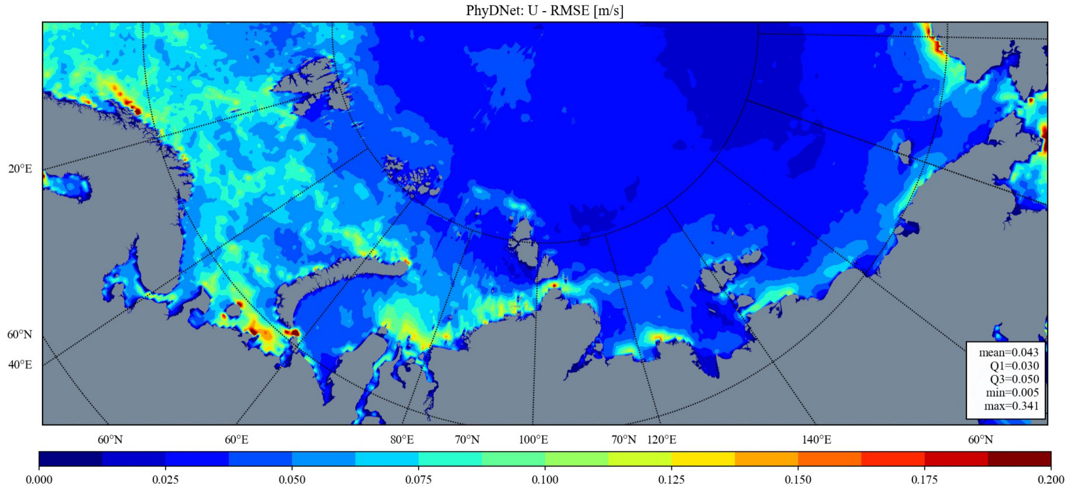


FIGURE 5
RMSE (in m/s) for zonal component of the surface water velocity for PhyDNet model.

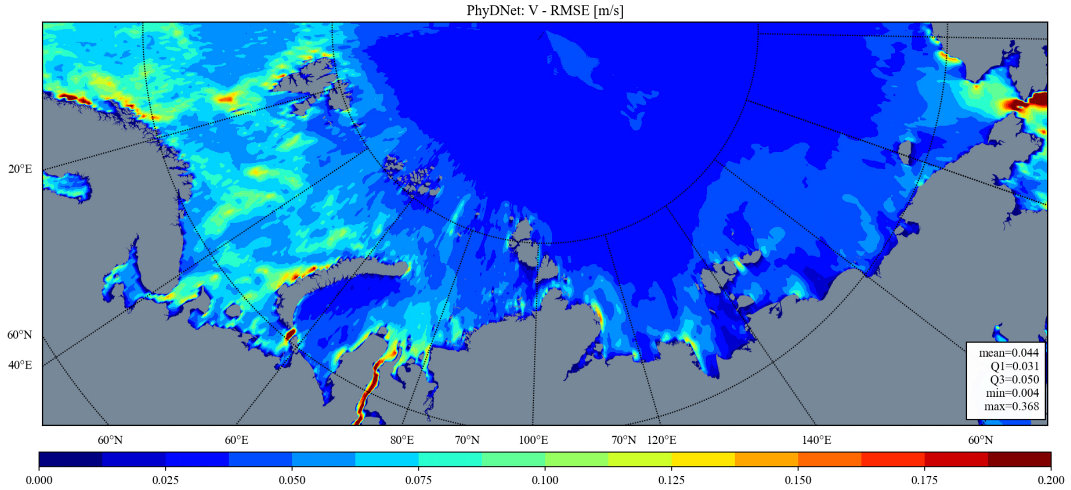


FIGURE 6
RMSE (in m/s) for meridional component of the surface water velocity for MariNet model.

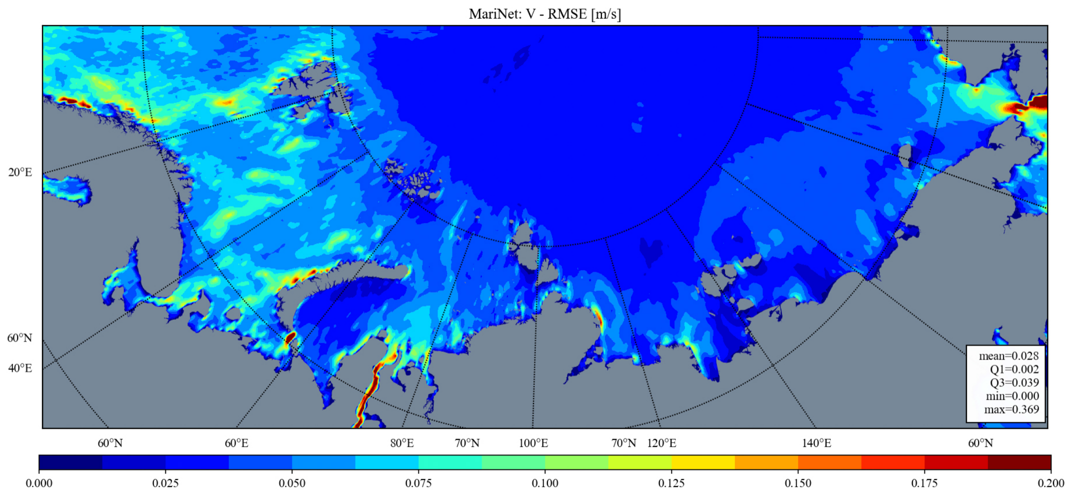


FIGURE 7
RMSE (in m/s) for meridional component of the surface water velocity for FourCastNet model.

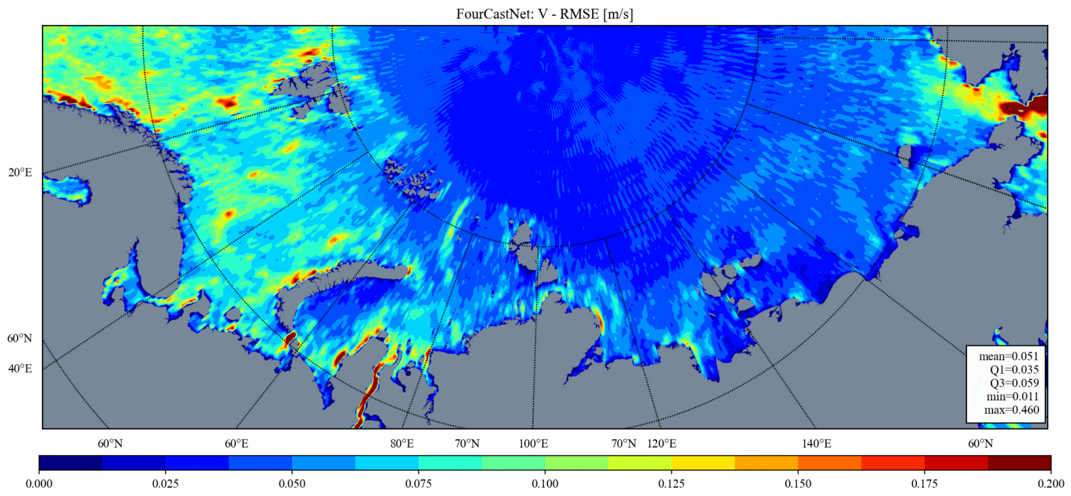


FIGURE 8
RMSE (in m/s) for meridional component of the surface water velocity for PhyDNet model.

5 Conclusions

In the study, we proposed a forecast model MariNet model, based on the encoder-decoder architecture, and compared it with FourCastNet and PhyDNet, the most promising ML models in the field weather prediction of their time. We have chosen the Arctic region, one of the hottest spots of the modern climate science research and obtained the hourly data on zonal and meridional

TABLE 2 Comparison of the carbon emissions and energy consumption during the models training and inference.

	Model Training			Model Inference		
	Emissions Rate (g/s)	Energy Consumed (kW)	Time (hrs)	Emissions Rate (g/s)	Energy Consumed (W)	Time (sec)
FourCastNet	0.100	**103.356**	**103.30**	0.104	0.431	1.997
PhyDNet	0.116	103.734	119.06	**0.02353**	**0.001097**	**0.0224**
MariNet	**0.083**	214.788	257.90	0.0908	0.0973	0.515

velocities of the surface sea water and sea surface heights above geoid from the Copernicus Marine Data Store. We switched temporal resolution from 1 hour to 6 hours and fed the datasets to the MariNet model, PhyDNet and FourCastNet.

In comparison with the other mentioned ML models, the RMSE and bias of the MariNet model are significantly lower. At the same time, the mean correlations of all three models with the original data are moderate and located between 0.4-0.5.

The above experimental results all show that the MariNet model has great potential in the mid-term predictions of the ocean dynamics. The further development of the model incudes improving the efficiency of computational operations, expanding the number of parallel running modules of our model to capture more temporal and spatial features of data variability, and increase the number of variables used in training.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

AB: Conceptualization, Data curation, Formal Analysis, Investigation, Writing – original draft. DI: Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing. EN: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – review & editing. AE: Funding acquisition, Project administration, Supervision, Writing – review & editing. PI: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. NS: Conceptualization, Formal Analysis, Investigation, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The research was funded under a grant for implementing a pilot project aimed at testing artificial intelligence technologies in priority industries as a part of the federal project 'Artificial Intelligence' within the national program 'Digital Economy of the Russian Federation'.

Acknowledgments

The study has been conducted using E.U. Copernicus Marine Service Information; https://doi.org/10.48670/moi-00016. We would like to acknowledge the use of the GitHub repository https://github.com/nvlabs/fourcastnet for providing the FourCastNet code, https://github.com/vincent-leguen/PhyDNet for providing the PhyDNet code used in our experiments, and https://github.com/mlco2/codecarbon for the CodeCarbon package.

Conflict of interest

Authors AB, DI, EN, and NS were employed by the company Marine Information Technologies LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bjerknes, V. (1904). *Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. Meterologische Z. Wien*, Vol. 21, 1-7.

Bjerknes, V. (1914). Meteorology as an exact science. *Monthly Weather Review* 42 (1), 11–14. Available online at: https://journals.ametsoc.org/view/journals/mwre/42/1/1520-0493_1914_42_11_maaes_2_0_co_2.xml (Accessed October 18, 2024).

Bryan, K. (1969). Climate and the ocean circulation: III. The ocean model. *Mon. Weather Rev.* 97, 806–827. doi: 10.1175/1520-0493(1969)097<0806:CATOC>2.3.CO;2

Buinyi, A., Irishev, D., Nikulin, E., Evdokimov, A., and Sukhikh, N. (2023). On the machine learning experience for the ocean circulation modeling and forecast in the arctic ocean. *Conference Poster.* doi: 10.13140/RG.2.2.23386.31686

Buinyi, A., Zhdanova, E., Evdokimov, A., and Sukhikh, N. (2022). On the artificial intelligence for the sea state forecast along the northern sea route. *Int. J. Offshore Polar Eng.* 32, 411–417. doi: 10.17736/ijope.2022.jc870

Chen, C., Gao, G., Qi, J., Proshutinsky, A., Beardsley, R. C., Kowalik, Z., et al. (2009). A new high-resolution unstructured grid finite volume Arctic Ocean model (AO-FVCOM): An application for tidal studies. *J. Geophys. Res.* 114, C08017. doi: 10.1029/2008JC004941

Dong, C., Xu, G., Han, G., Bethel, B. J., Xie, W., and Zhou, S. (2022). Recent developments in artificial intelligence in oceanography. *Ocean-Land-Atmosphere Res.* doi: 10.34133/2022/9870950

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint. doi: 10.48550/arXiv.2010.11929

Du, Y., Song, W., He, Q., Huang, D., Liotta, A., and Su, C. (2019). Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection. *Inf. Fusion* 49, 89–99. doi: 10.1016/j.inffus.2018.09.006

Duo, Z., Wang, W., and Wang, H. (2019). Oceanic mesoscale eddy detection method based on deep learning. *Remote Sens.* 11, 1921. doi: 10.3390/rs11161921

Fan, S., Xiao, N., and Dong, S. (2020). A novel model to predict significant wave height based on long short-term memory network. *Ocean Eng.* 205, 107298. doi: 10.1016/j.oceaneng.2020.107298

Franz, K., Roscher, R., Milioto, A., Wenzel, S., and Kusche, J. (2018). "Ocean eddy identification and tracking using neural networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia, Spain. 6887–6890 (New York, NY: IEEE). doi: 10.1109/IGARSS.2018.8519261

Fuentes, A. G., Nicolis, O., Peralta, B., and Chiodi, M. (2021). "ConvLSTM Neural Networks for seismic event prediction in Chile," in *2021 IEEE XXVIII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, Lima, Peru. 1–4 (New York, NY: IEEE). doi: 10.1109/INTERCON52678.2021.9532946

Gao, S., Huang, J., Li, Y., Liu, G., Bi, F., and Bai, Z. (2021). A forecasting model for wave heights based on a long short-term memory neural network. *Acta Oceanol. Sin.* 40, 62–69. doi: 10.1007/s13131-020-1680-3

Guibas, J., Mardani, M., Li, Z.-Y., Tao, A., Anandkumar, A., and Catanzaro, B. (2021). Adaptive fourier neural operators: efficient token mixers for transformers. *ArXiv*. doi: 10.48550/arXiv.2111.13587

Halley, E. (1686). An historical account of the trade winds, and monsoons, observable in the seas between and near the Tropicks, with an attempt to assign the physical cause of the said winds. *Philos. Trans. R. Soc Lond.* 16, 153–168. doi: 10.1098/rstl.1686.0026

Le Guen, V., and Thome, N. (2020). "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA. 11471–11481 (New York, NY: IEEE). doi: 10.1109/CVPR42600.2020.01149

Lguensat, R., Sun, M., Fablet, R., Tandeo, P., Mason, E., and Chen, G. (2018). "EddyNet: A deep neural network for pixel-wise classification of oceanic eddies," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia. 1764–1767 (New York, NY: IEEE). doi: 10.1109/IGARSS.2018.8518411

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., et al. (2020). Fourier neural operator for parametric partial differential equations. arXiv preprint. doi: 10.48550/ARXIV.2010.08895

Lin, H., Hua, Y., Ma, L., and Chen, L. (2019). "Application of convLSTM network in numerical temperature prediction interpretation," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, New York, NY, USA. 109–113 (New York, NY, USA: Association for Computing Machinery). doi: 10.1145/3318299.3318381

Ling, F., Luo, J.-J., Li, Y., Tang, T., Bai, L., Ouyang, W., et al. (2022). Multi-task machine learning improves multi-seasonal prediction of the Indian Ocean Dipole. *Nat. Commun.* 13, 7681. doi: 10.1038/s41467-022-35412-0

Liu, Q., Zhang, R., Wang, Y., Yan, H., and Hong, M. (2021). Daily prediction of the arctic sea ice concentration using reanalysis data based on a convolutional lstm network. *J. Mar. Sci. Eng.* 9. doi: 10.3390/jmse9030330

Mandal, S., and Prabaharan, N. (2006). Ocean wave forecasting using recurrent neural networks. *Ocean Eng.* 33, 1401–1410. doi: 10.1016/j.oceaneng.2005.08.007

Maury, M. F. (1855). *The physical geography of the sea* (London: Sampson, Low, Son & Co). doi: 10.5962/bhl.title.102148

Moishin, M., Deo, R. C., Prasad, R., Raj, N., and Abdulla, S. (2021). Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm. *IEEE Access* 9, 50982-50993. doi: 10.1109/ACCESS.2021.3065939

Navier, C. L. M. H. (1822). Mémoire sur les lois mouvement des fluides. *Mem Acad. Sci. Inst Fr* 6, 389-440. Available online at: https://fr.wikisource.org/wiki/Page:M%C3%A9moires_de_l%E2%80%99Acad%C3%A9mie_des_sciences,_Tome_6.djvu/577 (accessed October 18, 2024).

Ning, P., Zhang, C., Zhang, X., and Jiang, X. (2021). Short- to medium-term sea surface height prediction in the bohai sea using an optimized simple recurrent unit deep network. *Front. Mar. Sci.* 8. doi: 10.3389/fmars.2021.672280

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint. doi: 10.48550/ARXIV.2202.11214

Richardson, L. (1922). *Weather Prediction by Numerical Process*. Cambridge: Cambridge University Press.

Robinson, A. R. (1983). "Overview and summary of eddy science," in *Eddies in Marine Science*. Ed. A. R. Robinson (Springer Berlin Heidelberg, Berlin, Heidelberg), 3–15. doi: 10.1007/978-3-642-69003-7_1

Santana, O., Hernández-Sosa, D., Martz, J., and Smith, R. (2020). Neural network training for the detection and classification of oceanic mesoscale eddies. *Remote Sens.* 12, 2625. doi: 10.3390/rs12162625

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W.-K., and Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *NIPS*. 28, 802-810. Available online at: https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf (accessed October 18, 2024).

Song, T., Han, N., Zhu, Y., Li, Z., Li, Y., Li, S., et al. (2021). Application of deep learning technique to the sea surface height prediction in the South China Sea. *Acta Oceanol. Sin.* 40, 68–76. doi: 10.1007/s13131-021-1735-0

Stokes, G. (1845). On the theories of the internal friction of fluids in motion. *Trans. Camb. Philos. Soc* 8, 287-319.

Storto, A., Alvera-Azcárate, A., Balmaseda, M. A., Barth, A., Chevallier, M., Counillon, F., et al. (2019). Ocean reanalyses: recent advances and unsolved challenges. *Front. Mar. Sci.* 6. doi: 10.3389/fmars.2019.00418

Sverdrup, H. U., and Munk, W. H. (1947). *Wind, sea and swell: Theory of relations for forecasting (No. 601)*. Hydrographic Office. Available online at: https://api.semanticscholar.org/CorpusID:132860343. (accessed October 18, 2024).

Xu, G., Cheng, C., Yang, W., Xie, W., Kong, L., Hang, R., et al. (2019). Oceanic eddy identification using an AI scheme. *Remote Sens.* 11, 1349. doi: 10.3390/rs11111349

Xu, G., Xie, W., Dong, C., and Gao, X. (2021). Application of three deep learning schemes into oceanic eddy detection. *Front. Mar. Sci.* 8. doi: 10.3389/fmars.2021.672334

Zhou, S., Bethel, B. J., Sun, W., Zhao, Y., Xie, W., and Dong, C. (2021). Improving significant wave height forecasts using a joint empirical mode decomposition–long short-term memory network. *J. Mar. Sci. Eng.* 9, 744. doi: 10.3390/jmse9070744

Zulfa, I. I., Novitasari, D. C. R., Setiawan, F., Fanani, A., and Hafiyusholeh, M. (2021). Prediction of sea surface current velocity and direction using LSTM. *IJEIS Indones. J. Electron. Instrum. Syst.* 11, 93. doi: 10.22146/ijeis.63669

frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 14 November 2024
DOI 10.3389/fmars.2024.1485161

OPEN ACCESS

EDITED BY
Jianchuan Yin,
Guangdong Ocean University, China

REVIEWED BY
Mbaye Tine,
Gaston Berger University, Senegal
Lara Veylit,
SINTEF Ocean, Norway

*CORRESPONDENCE
Gert Everaert
gert.everaert@vliz.be

RECEIVED 23 August 2024
ACCEPTED 22 October 2024
PUBLISHED 14 November 2024

CITATION
Standaert W, Musimwa R, Stevens M, Guerra JA, Muñiz C, Debusschere E, Pint S and Everaert G (2024) Modeling Atlantic herring distribution in the Northeast Atlantic for informed decision-making towards sustainable fisheries.
Front. Mar. Sci. 11:1485161.
doi: 10.3389/fmars.2024.1485161

Modeling Atlantic herring distribution in the Northeast Atlantic for informed decision-making towards sustainable fisheries

Ward Standaert[1], Rutendo Musimwa[1], Martha Stevens[1], Jesus Alonso Guerra[1], Carlota Muñiz[1], Elisabeth Debusschere[1], Steven Pint[1,2] and Gert Everaert[1*]

[1]Research Department, Flanders Marine Institute, Ostend, Belgium, [2]Marine Biology Research Group, Ghent University, Ghent, Belgium

The withdrawal of the United Kingdom from the European Union will likely result in reduced fishing grounds for the Belgian fishing fleet. This fleet now targets demersal fish, but there used to be a tradition of catching Atlantic herring (*Clupea harengus*). After the stock collapse of Atlantic herring in the 1970s, fishing on herring by the Belgian fleet did not recover and herring quotas are now exchanged with the Netherlands and Germany. To assess the feasibility of reintroducing herring fisheries for the Belgian fishing fleet, our study created spatiotemporal species distribution models for Atlantic herring in the Northeast Atlantic Ocean, focusing results on the Belgian Part of the North Sea. In total 30078 occurrence records were derived and processed to fit species-environmental relationships with temperature, salinity, seabed characteristics and plankton concentration using Maximum entropy (Maxent) models. The Area Under the Curve of the Receiver Operating Characteristic plot (AUC) and the True Skill Statistic (TSS) were used to assess model fit. Models performed well (AUC > 0.7 and TSS > 0.6). While a broad spatiotemporal distribution of Atlantic herring in the Northeast Atlantic Ocean was inferred, regional differences show that herring habitat is most suitable during winter months in the Belgian Part of the North Sea for both adult and larval herring (habitat suitability index > 75%). This regional trend in the Belgian Part of the North Sea was negatively correlated (R = -0.8) with the North Atlantic Oscillation (NAO). We anticipate that these findings will provide valuable insights for policymakers to implement sustainable fisheries management practices.

KEYWORDS

Atlantic herring, species distribution modeling, machine learning, fishery, sustainable blue economy, Northeast Atlantic, North Atlantic Oscillation, Belgian Part of the North Sea

1 Introduction

Since the start of the Common fisheries policy in 1970, European Union (EU) members have been allowed equal access to fish in their shared waters, including those of the United Kingdom (UK). With the withdrawal of the UK from the EU on the 1st of January 2021, it was decided that this will change (Regulation 1380/2013, 2013). As a transition period, EU vessels are allowed to access UK waters until 30 June 2026 and afterwards, access will require annual negotiation. Furthermore, fishing quotas with a value of 25% of the EU landings in UK waters will be gradually transferred from the EU to the UK from 2021 to 2025 (Popescu and Scholaert, 2022).

For the Belgian fishing fleet, the loss by transfer of fishing quota to the UK is estimated at 3.7 million euros in 2023 (Popescu and Scholaert, 2022) and estimated to increase up to 6.8 million euros in 2026 (Coudyser, 2021). Recently, the Belgian fishing fleet has experienced a steady decline in catches in the southern North Sea and a decrease in the number of vessels (data up to 2021, Maertens, 2022). The Belgian fleet targets demersal fish, mainly sole (*Solea solea*) and plaice (*Pleuronectes platessa*) using beam trawlers (Regulation No 1380/2013). Historically, pelagic fish species, particularly Atlantic herring (*Clupea harengus*), were also targeted with annual yields reaching up to 58 000 tons in 1943 (Lescrauwaet et al., 2010). One of the initiatives to overcome the loss of fishing grounds and quota following Brexit is to provide information about alternative fishing grounds and niche fisheries (European Commission, 2020). One of those alternatives for the Belgian fishing fleet could be the restoration of pelagic fisheries in Belgian waters. Because the current Belgian fishing fleet mainly targets bottom-dwelling fish (Regulation No 1380/2013, n.d), the whereabouts of pelagic fish are often unknown, anecdotical, or expert-based.

Atlantic herring is a valuable pelagic fish species in the North Sea in terms of both economy but also in terms of ecology. It plays a key role in the ecosystem as a regulator for seabird abundance (through bottom-up control; Fauchald et al., 2011) and zooplankton (through top-down control; Fauchald et al., 2011). During early larval stages, Atlantic herring feeds on phytoplankton and unicellular organisms (Marshall et al., 1937; Joly et al., 2021), subsequently larval diet shifts towards zooplankton (Van Ginderdeuren et al., 2014). For adults, the main prey is copepods, but feeding on other zooplankton and fish larvae is also common (Van Ginderdeuren et al., 2014). Atlantic herring can tolerate a wide range of temperatures (1 – 19°C; de Groot, 1980; Whitehead, 1985) and salinities (2 – 35 PSU; Brevé et al., 2007; de Groot, 1980), which allows them to migrate between feeding, spawning, and nursing grounds. Typical for this species is their natal homing behavior or the return to the same area where they hatched for spawning (Geffen, 2009).

In the Northeast Atlantic, Atlantic herring populations are divided into several distinct stocks, based primarily on their spawning grounds and migration behaviors. These stocks include the North Sea autumn spawners, the West of Scotland stock, the Irish and Celtic Sea stocks, and the Downs stock. Each stock shows unique migration and spawning behaviors. Based on the spawning period, Atlantic herring populations are divided into two groups in the Northeast Atlantic: spring and autumn spawners (Heath, 1993). Herring in the Belgian Part of the North Sea (BPNS) belong to the Downs stock which are autumn spawners.

Currently, the stock of autumn-spawning Atlantic herring is stable in the North Sea (ICES, 2023a). For the Irish and Celtic Sea, the International Council for the Exploration of the Sea (ICES) advises zero catches of herring to sustain maximum sustainable yield (ICES, 2023b, ICES, 2023c, ICES, 2023d). Atlantic herring stocks are prone to collapse due to natural fluctuations in abundance, aggravated by poorly managed fishing pressure with slow recovery rates (Stephenson et al., 2001). In addition to maintaining the overall stock biomass, effective management also requires the preservation of their spatial and temporal spawning distribution (Frost and Diele, 2022; Stephenson et al., 2001). During spawning, herring lay their eggs on the seabed or aquatic vegetation creating dense egg carpets that are vulnerable to bottom trawling (Morrison et al., 1991; Watling and Norse, 1998). Since visual observations of spawning grounds *in situ* are scarce, spawning areas have been allocated using the position of young larvae (Frost and Diele, 2022). Following a stock collapse of the Down's stock of Atlantic herring in 1955, which saw recovery only after a fishing ban was implemented from 1977 to 1980, commercial pelagic fisheries disappeared completely in the Belgian fishing fleet (Cushing, 1992; Lescrauwaet et al., 2010). Currently, Belgium exchanges its Atlantic herring fishing quota with the Netherlands and Germany in return for quotas on sole and Atlantic cod (*Gadus morhua*) (Departement Landbouw en Visserij, 2021).

Species distribution models use species field observations and environmental information to create species-environment relationships and infer the spatial distribution and ecological niche of a species. Outcomes quantify the habitat suitability for the species at each location, given the local environmental conditions. Previous studies by Turner et al. (2016) and Wang et al. (2018) modeled the distribution of adult Atlantic herring, while Aires et al. (2014) and Maravelias et al. (2000) modeled their spawning distribution. Of these studies, they either looked at different areas (Shetland Islands, Maravelias et al., 2000; Northeast US continental shelf, Turner et al., 2016 and Northwest Atlantic Shelf, Wang et al., 2018) or did not consider the monthly variation of their distribution (Aires et al., 2014; Maravelias et al., 2000). In addition to studying the monthly distributional variation (Turner et al., 2016), the current abundance of available data enables analysis of distributional variation across different years (Wang et al., 2018). Notably, comparable models for both adults and larvae are lacking. Generating two models with identical settings for these two life stages of herring facilitates the comparison of their seasonal distribution and their ecological tolerance to various environmental gradients within the research area.

The objective of this study is to model the spatiotemporal distribution of Atlantic herring in the Northeast Atlantic towards informing sustainable fisheries. Due to our focus on the Belgian fishing fleet, outcomes for the BPNS will be highlighted throughout the study. Furthermore, since Atlantic herring is prone to stock collapse, we aim to model their spawning distribution as well and compare the ecological needs of two life stages: adults and larvae. It is expected that this study will contribute to sustainable fisheries management by providing the time and location of Atlantic herring occurrences and by giving insight into the ecological needs of Atlantic herring. In the Northeast Atlantic, we hypothesize that both adults and larvae will have recurring annual distributional patterns due to their natal homing behavior (Geffen, 2009). Furthermore, we expect that larvae and adults will occur in the BPNS during November – January since they are known to spawn in the English Channel during this period (Limborg et al., 2012).

2 Materials and methods

2.1 Occurrence data

A total of 22176 occurrences of adult Atlantic herring were retrieved from the Database of Trawl Surveys (DATRAS) for the Northeast Atlantic (http://marineregions.org/mrgid/5664), restricted to 48°N – 62°N and 12°W – 10°E for all months for the years 2000 to 2020 (Figure 1, Supplementary Table 1; ICES, 2023e). This region spans about 1 367 600 km^2 of ocean and includes the English Channel, the North Sea, the Scottish Sea, the Irish Sea, the Celtic Sea and part of the North Atlantic Ocean. Following a general overview of the Northeast Atlantic, we specifically looked at the model outcomes for the BPNS, situated between 51°N – 52°N and 2°E – 4°E and spanning 3454 km^2 or 0.25% of the total area of the study area (https://marineregions.org/gazetteer.php?p=details&id=3293, Figure 1). Herring observations with a length below 20 cm, the approximate length when Atlantic herring become mature (Brevé et al., 2007), were discarded from the adult set. This way, in total 13112 occurrences of adult herring were available for the Northeast Atlantic in all months excluding April, May and June and this for a time frame of 21 years (i.e. years 2000 – 2020; Figure 1). Larval occurrence data of Atlantic herring was retrieved from the International Herring Larvae Surveys (IHLS) within the same spatiotemporal frame as adult occurrences (ICES, 2023f). In total 7902 larval observations were available for the North Sea and the English Channel in September, October, December, and January from 2000 to 2020 inclusive (Figure 1) with larval size ranging from 5 – 24 mm in length. Since these are the months when spawning occurs in the North Sea and the English Channel (Geffen, 2009), larval model outcomes were restricted to these months.

Four steps were taken to transform the raw occurrence data into the final pre-processed occurrence dataset ready for analysis, being 1) conducting outlier analysis, 2) eliminating duplicated occurrences, 3) applying geographical filtering to address spatial bias and 4) applying environmental filtering to address spatial autocorrelation (SAC).

Outliers were defined as being farther away from other observations than 1.5 times the interquartile range of geographical distances (Yang et al., 2019) and were flagged using the cc_outl function from R-package *CoordinateCleaner* (Zizka

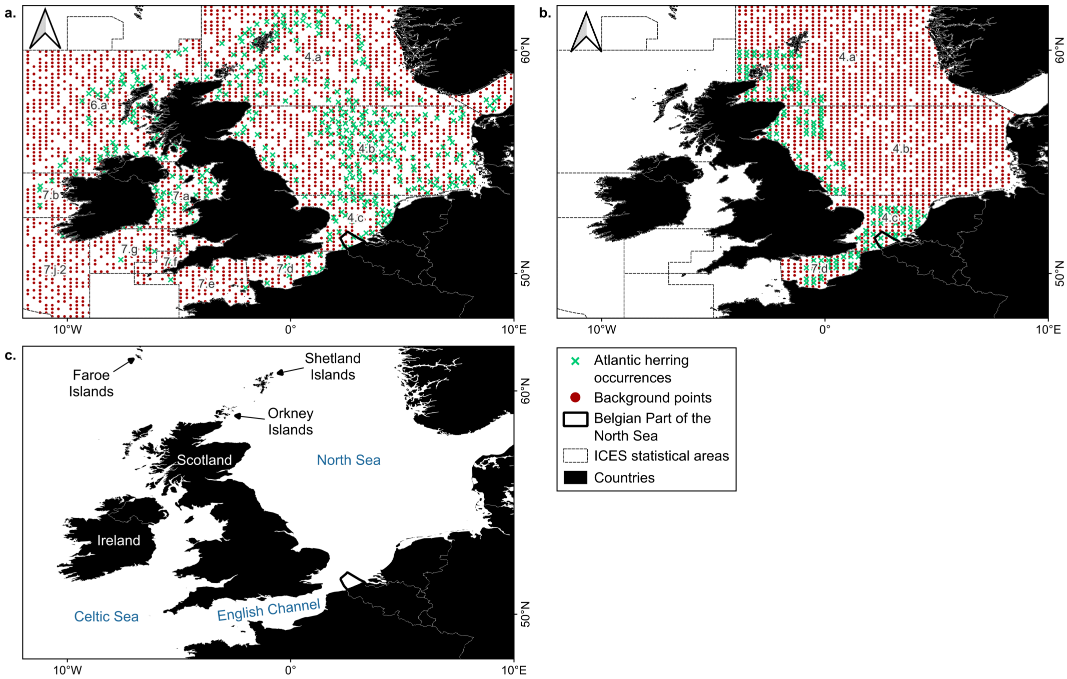


FIGURE 1
Study area with occurrences in green and background points in red of adult (**A**; ICES, 2023e) and larval herring (**B**; ICES, 2023f). Occurrences were compiled from the period 2000 – 2020 inclusive. In this period, larvae were present in September, October, December and January only, while adults were present in all months excluding April, May and June. Background points were restricted to the International Council for the Exploration of the Sea (ICES) statistical areas where occurrences are present. The location of geographic features that are used throughout the text are indicated in panel **(C)**. Here, names of water bodies are denoted in blue and names of terrestrial areas in black or white. Projection: EPSG 4326/WGS 84.

et al., 2019). For the response variable, no outliers were identified, and all observations were kept. Next, duplicated occurrences per location (grid cell, 10 x 10 NM, see section 2.2) and time steps (months) were removed (Phillips, 2021).

To account for sampling bias in the compiled datasets of DATRAS and IHLS (Loiselle et al., 2008; Merow et al., 2013), we applied a filtering technique in geographical space as per Vollering et al. (2019). Occurrences were removed randomly until a dataset was retained where each pair of occurrences had a minimum distance of 10 nautical miles (NM, or 18.5 km), which is the recommended distance between valid hauls in the DATRAS trawl surveys (ICES, 2020). This filtering was done on the projected datum ETRS89-extended/LCC Europe (EPSG 3034), which covers the entire study area, using the R-package *spThin* (Aiello-Lammens et al., 2015).

Spatial autocorrelation (SAC), where locations close to each other are more similar than locations further apart, is common in spatial data. Spatial autocorrelation in model residuals violates the assumption that model residuals are independent and identically distributed (Legendre, 1993). Since initial results revealed SAC in the model residuals (tested using Moran's I statistic from the *ape* package; Paradis and Schliep, 2019), SAC was reduced by an additional filtering technique in environmental space (de Oliveira et al., 2014). First, the environmental Mahalanobis distance was calculated between all observations using eight environmental variables (see section 2.2). Based on these environmental distances, the two most distant observations were selected and added to a new dataset. Subsequently, we iteratively added the observation that is most distant to this new dataset until we retained a new dataset of 400 occurrences. The objectively selected 400 occurrences minimized SAC while keeping enough observations to construct robust models. We applied this procedure to both adult and larval datasets, resulting in a total of 800 occurrences for calibrating and validating the two species distribution models (Figure 1, section 0).

2.2 Environmental variables

Relevant environmental variables were selected through a literature review on the ecology of Atlantic herring and include bathymetry, sea surface temperature (SST) (Turner et al., 2016; Wang et al., 2018), sea surface salinity (SSS) (Aires et al., 2014), seabed substrate and energy (Brevé et al., 2007; Maravelias et al., 2000), sea surface phytoplankton concentration (Marshall et al., 1937), zooplankton concentration in the epipelagic layer (Maravelias et al., 2000; Van Ginderdeuren et al., 2014), and windfarm presence to include a measure of artificial coarse substrate (Frost and Diele, 2022). Substrate energy is a measure of the average hydrodynamics of the seabed by European Marine Observation and Data Network (EMODnet Seabed Habitats product). It has been found that a highly energetic seabed is important for Atlantic herring egg development (Haegele and Schweigert, 1985). Of all eight variables, bathymetry, seabed substrate and energy were considered static variables over time for each location, while the remaining variables were dynamic and derived monthly for 2000 – 2020. All environmental variables used were obtained from the European Marine Observation and Data Network (EMODnet) and the Copernicus Marine Service (CMEMS, Table 1). Zooplankton concentration was derived from CMEMS, with the units g C/m^2, representing the average biomass (expressed in carbon content) over the depth of the epipelagic layer (see also Global ocean low and mid trophic levels biomass content hindcast | Copernicus Marine Service).

Preprocessing of the environmental variables involved aggregation by averaging to match the coarser spatiotemporal resolution of the occurrence data (10 NM x 10 NM, monthly for 2000 – 2020 inclusive) (Sillero and Barbosa, 2021). Additionally, we calculated a measure of windfarm presence using a buffer of 200 m around active windfarms to indicate nearby windfarm presence since Atlantic herring are known to spawn on coarse substrate (Frost and Diele, 2022). To avoid adding highly correlated variables in the models, the Variance Inflation Factor (VIF) was calculated for each combination of variable pairs and a VIF larger than 10 was considered as a threshold for collinearity (Zuur et al., 2010). No correlated variable pairs were found and hence all variables were kept.

2.3 Model settings

Since Atlantic herring is a mobile and migratory species, the absence of the fish at a location during a sampling event does not necessarily indicate that environmental conditions are not suitable at this location and time (Lobo et al., 2010). For this reason, we based our models on presence-background data instead of presence-absence data (Fernandez et al., 2022). Background points are defined as a general sample of the environmental conditions of the entire study area or a sample of all sites that are available for the species to occupy (Phillips et al., 2009). We developed Maximum entropy (Maxent) presence-background models, a machine learning model that is commonly used in species distribution modeling because it is flexible, simple to use and performs well (using R-package *dismo*; Hijmans et al., 2023) (Barber et al., 2022; Phillips et al., 2006; Valavi et al., 2023). Background points were sampled randomly in the study area, restricted to the ICES areas of the occurrences (Figure 1). The number of background points was set at 10 times the number of presences (Hysen et al., 2022). Maxent can be tailored by employing combinations of feature classes and regularization multipliers (Phillips et al., 2006). Feature classes are transformations that can be applied to each predictor variable by the model, for example, linear and quadratic transformations, while the regularization multiplier is a penalty to avoid overfitting (Merow et al., 2013). Fifteen combinations were tested using the corrected Akaike's Information Criterion (AICc) as a selection criterion (R-package *ENMeval*; Kass et al., 2021) (Table 2A; Zeng et al., 2016). Following the recommendations of Merow et al. (2013), we included all combinations of the feature classes L, LQ and LQH (with L linear, Q quadratic and H hinge) and the regularization multipliers 1, 2, 4, 8 and 32. Finally, one model was retained for adult herring and one model for larvae. Next, habitat suitability

TABLE 1 Selected environmental variables for modeling with the corresponding source, value range, data type, uni, original resolution, and DOI/URL.

Variable	Source name	Value range in study area	Value range in BPNS	Data type/ unit	Original resolution		DOI/URL
					Spatial	Temporal	
Bathymetry	EMODnet Digital Bathymetry (DTM)- 2022	5 - 4866	12 - 37	m	0.063’ x 0.063’	/	https://emodnet.ec.europa.eu/en/bathymetry
Seabed substrate	EUSeaMap 2023 Broad-Scale Predictive Habitat Map for Europe	/	/	Categorical	Polygon	/	https://emodnet.ec.europa.eu/en/seabed-habitats
Seabed energy	EUSeaMap 2023 Broad-Scale Predictive Habitat Map for Europe	/	/	Categorical	Polygon	/	https://emodnet.ec.europa.eu/en/seabed-habitats
Windfarm presence	EMODnet Human Activities	/	/	Binary data	Polygon	/	https://emodnet.ec.europa.eu/en/human-activities
Sea surface temperature	Global Ocean Physics Reanalysis	0 - 22	5 – 21	°C	0.083° x 0.083°	Monthly	https://doi.org/10.48670/moi-00021
Sea surface salinity	Global Ocean Physics Reanalysis	24 – 36	30 – 35	PSU	0.083° x0.083°	Monthly	https://doi.org/10.48670/moi-00021
Sea surface phytoplankton concentration	Atlantic- European North West Shelf- Ocean Biogeochemistry Reanalysis	0 – 47	0 – 24	mmol C m-3	0.111° ×0.067°	Monthly	https://doi.org/10.48670/moi-00058
Zooplankton concentration in the epipelagic layer	Global ocean low and mid trophic levels biomass content hindcast	0 - 81	1 – 25	g C m-2	0.083° ×0.083°	Daily	https://doi.org/10.48670/moi-00020

Before modeling, all variables were resampled from their original resolution towards a resolution of 10 NM x 10 NM and monthly. C, Carbon; EMODnet, European Marine Observation and Data Network.

maps were calculated per month and year in 2000 - 2020, by applying the models to each monthly map of the environmental variables in 2000 - 2020. Average and standard deviation maps were calculated per month. Model outcomes are shown in terms of habitat suitability indices (HSI). To enhance clarity, habitat suitability indices above 50% will be addressed as suitable and indices below 50% as unsuitable (Manel et al., 1999).

The importance of each environmental variable in the model was evaluated using a bootstrapping method adopted by Thuiller et al. (2009). Hereby, the correlation was calculated between the original model prediction and a model prediction where one variable was randomly permuted. Following the approach of Thuiller et al. (2009), this calculation was repeated 50 times for each variable. The variable importance score was calculated as the mean correlation coefficient for each variable, normalized across all variables to collectively contribute to a total variable importance of 100%. Response plots were created that depict the modeled relationship between each environmental variable and the HSI. For each plot, HSI was simulated across 100 values over the range of the environmental variable, with other variables constant at their mean value (response.plot function from R-package *Maxnet*, Phillips, 2021).

Model performance was evaluated using the Area Under the Curve of the Receiver Operating Characteristic plot (AUC) and the True Skill Statistic (TSS) metrics (Báez et al., 2020; Liu et al., 2013). The AUC ranges from 0 to 1, whereby an AUC of 0.5 or lower indicates that the model is no better than random and 1 indicates perfect model performance; the TSS ranges from -1 to 1, whereby a random model would have a TSS of 0 or less, and a perfect model a TSS of 1. To have a full view of the various aspects of the model performance (Grimmett et al., 2020), the model sensitivity (ability to accurately predict presences) and specificity (ability to accurately predict background points) were also included as performance metrics. All four metrics (AUC, TSS, sensitivity and specificity) were calculated using k-fold cross-validation as follows: (1) the complete dataset was divided randomly into a training and a test set (training-test ratio of 80-20%), (2) a model was built on the training set and (3) the model performance was tested based on its ability to predict the test set. This process was repeated ten times.

2.4 North Atlantic Oscillation

The North Atlantic Oscillation (NAO) is an important climate process in the North Atlantic and affects ocean dynamics (Hurrell and Deser, 2010). Variations in the NAO index can have direct effects on the biology in the ocean (e.g., Alheit et al., 2005). For example, Corten (1999) found a link between the NAO and the occurrence of Atlantic herring in the Norwegian trench and Gröger et al. (2010) between the NAO and the number of Atlantic herring recruits in the North Sea. We correlated the seasonal and inter-annual variability of our models’ spatiotemporal HSI with NAO indices. The impact of the NAO varies across regions of the Northeast Atlantic (van der Molen and Pätsch, 2022). Therefore, a regional assessment was made, particularly for the focus area of this study: the BPNS.

To analyze the effect of the NAO on the HSI in the BPNS, the following method was adapted from the one used by Corten (1999): Monthly NAO indices from 2000 – 2020 were retrieved from the National Oceanic and Atmospheric Administration (NOAA, https://www.ncei.noaa.gov/access/monitoring/nao/) and seasonal averages were calculated. For each season, a separate time series was created. For example, for the winter, NAO indices of January, February and March were averaged for each year from 2000 until 2020. Next, based on our model outcomes, an average HSI was calculated for adults and larvae in the BPNS for each year in 2000 – 2020, using the months during which habitat was calculated to be most suitable in the BPNS (winter, specifically January – March and December – January for adults and larvae respectively, see section 3.4). As per Corten (1999), these NAO and HSI time series were smoothed using a running average. We used three smoothing windows: (1) non-smoothed indices, (2) a three-year average window and (3) a five-year average window. Finally, the autocorrelation was calculated by crossing NAO and HSI time series using the base R function *ccf* (R Core Team, 2023). This autocorrelation was calculated for each of the three smoothing windows and for each of the four seasonal NAO time series. Besides calculating autocorrelation, the *ccf* function was also used to detect if there is a lag between two correlated time series. The significance of the correlation coefficients was tested by calculating 99% confidence intervals using Fisher's Z transformation for correlation coefficients (CorCI function from R-package *DescTools*; Signorell, 2024) (Zou, 2007).

3 Results

3.1 Data exploration

After filtering, 400 adult and 400 larval Atlantic herring occurrences were retained for modeling. Of these, two adult occurrences and four larval occurrences were located in the BPNS. After filtering in geographical and environmental space, adult occurrences were present at a broad range of bathymetry (8 – 700 m), sea surface temperature (3 – 20°C) and sea surface salinity (29 – 35 PSU) and at zooplankton concentrations in the epipelagic layer of 0 – 12 g C m^{-2}, sea surface phytoplankton concentrations of 0 – 11 mmol C m^{-3}, all seabed energy classes (low, moderate and high) and above multiple seabed substrate classes (including both coarse and muddy classes and several categories in between). Larval occurrences for modeling were present at a narrower range of bathymetry (12 – 134 m), sea surface temperature (5 – 17°C), sea surface salinity (31 – 35 PSU), zooplankton concentrations in the epipelagic layer (1 – 10 g C m^{-2}) and sea surface phytoplankton concentrations (0 – 8 mmol C m^{-3}). Additionally, these occurrences were situated above all seabed energy levels and sandy and coarse substrate types.

3.2 Model performance

The larval model performed best with an AUC of 0.9 and a TSS of 0.7 (Table 2). The adult model had lower performance metrics with an AUC of 0.7 and a TSS of 0.6. The adult model scored better at accurately predicting presences (model sensitivity) than background points (model specificity). Optimal Maxent model settings (minimal AICc) were obtained using feature classes linear, quadratic and hinge and a regularization multiplier of one for both adult and larval models. Filtering in environmental space was successful in reducing spatial autocorrelation but did not remove it completely (from I = 0.18 to I = 0.09 and from I = 0.16 to I = 0.14 for the adult and larval models, respectively). However, bootstrapping methods for variable importance ensured that selected variables were not selected due to type I errors derived from SAC in the residuals of the models.

3.3 Variable importance and response curves

Bathymetry was the most influential variables for both adults (63%) and larvae (37%) (Table 3). The dynamic variables SST, SSS, phyto- and zooplankton concentrations were important in both models but have a higher summed importance in the larval model (46%) compared to the adult model (35% in total). Seabed characteristics were important (18% in total) in the larval model only, while windfarm presence did not influence any model.

Response curves displayed distinct patterns for the adult and larval life stages (Figure 2). In general, adult Atlantic herring were inferred to have a capacity to withstand a broader range of environmental conditions than their larvae. For both larvae and adults, the habitat was most suitable at shallow depths with a decreasing HSI towards deeper depths. Adults were able to tolerate a wider bathymetrical range compared to larvae (HSI drops to 25% at 660 and 81 m respectively). Sea surface temperature was optimal at 5 and 7°C for adults and larvae respectively, situated at the lower end of the temperature range observed in the study area (Northeast Atlantic 2000 – 2020, 3 – 20°C). Adult herring tolerate a wide temperature range (3 – 15°C, HSI > 50%), while larvae favor a narrower temperature range (5 – 8°C). The response of suitability to salinity is low in the study area. The highest HSI (38%) were reached for adults at high salinity values around 35 PSU and at both ends of the salinity range for larvae (HSI > 50% at 29 – 35 PSU). Adult herring

TABLE 2 Model evaluation using the area under the curve (AUC), true skill statistic (TSS), sensitivity and specificity.

	Adult	Larva
AUC	0.73 ± 0.02	0.89 ± 0.02
TSS	0.61 ± 0.03	0.71 ± 0.05
Sensitivity	0.78 ± 0.14	0.71 ± 0.01
Specificity	0.68 ± 0.01	0.72 ± 0.01

Numbers depict averages and standard deviations after ten cross-validation repetitions.

TABLE 3 Variable importance (%) of the environmental variables in the adult and larval models.

	Adult	Larvae
Bathymetry	62.6	36.7
Sea surface temperature	12.8	14.7
Sea surface salinity	11.1	5.3
Zooplankton concentration	6.2	12.6
Phytoplankton concentration	4.9	13.3
Seabed substrate	2.4	13.6
Seabed energy	0.0	3.8
Windfarm presence	0.0	0.0

was more likely to be found in low concentrations of zooplankton in the pelagic layer (< 2.5 g C m^{-2}), although other concentrations in the study area were not restrictive (HSI around 50%). For larvae, optimal zooplankton concentrations were quantified at 4.5 – 6.5 g C m^{-2}. On top of these environmental variables that were important in both adult and larval models, larval models were also influenced by seabed substrate and sea surface phytoplankton concentration (Table 3). Models suggested that larvae are more likely to be found above coarse substrate, sandy mud, and sand (HSI > 45%) than the unclassified group 'seabed' (unclassified in EMODnet Predictive Habitat Map for Europe), fine mud, and rock or other hard substrata (HSI < 20%). Finally, larvae were simulated to be found in the highest concentrations of phytoplankton in the study area (8 mmol C m^{-3}).

3.4 Spatiotemporal distribution maps

On average, adult Atlantic herring was projected to have a wide spatial distribution across the Northeast Atlantic throughout the entire year (Figure 3, left; see Supplementary Figure 1 for distribution maps for all months). Early in the year, habitat was suitable in the North Sea and around the Faroe Islands (HSI > 50%, Figure 3). No observations were present around the Faroe Islands, so this is an extrapolation of the model and should be interpreted with caution (Elith et al., 2010). In July, waters surrounding Ireland become suitable and in October, the English Channel was included as suitable waters for adult herring (HSI > 50%). Year-to-year variability of the habitat suitability was highest in the Celtic Sea in July and October (Figure 3, right). For example, the standard deviation peaked at 18% in October in the Celtic Sea.

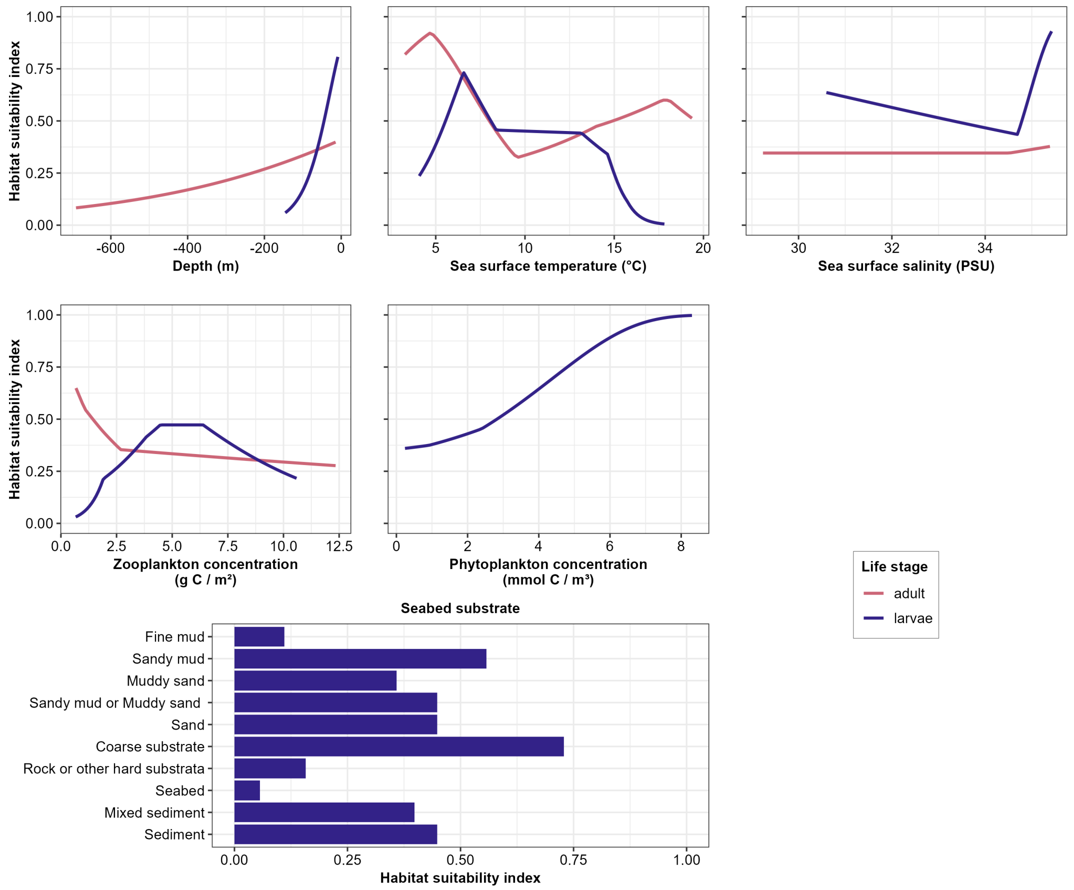


FIGURE 2
Response curves relating the modeled habitat suitability index for Atlantic herring to the environmental variables for adult and larval life stages. Only environmental variables with a variable importance larger than 5% are shown. The range of the environmental values shown was restricted to the range of values of the variable where occurrences were present.

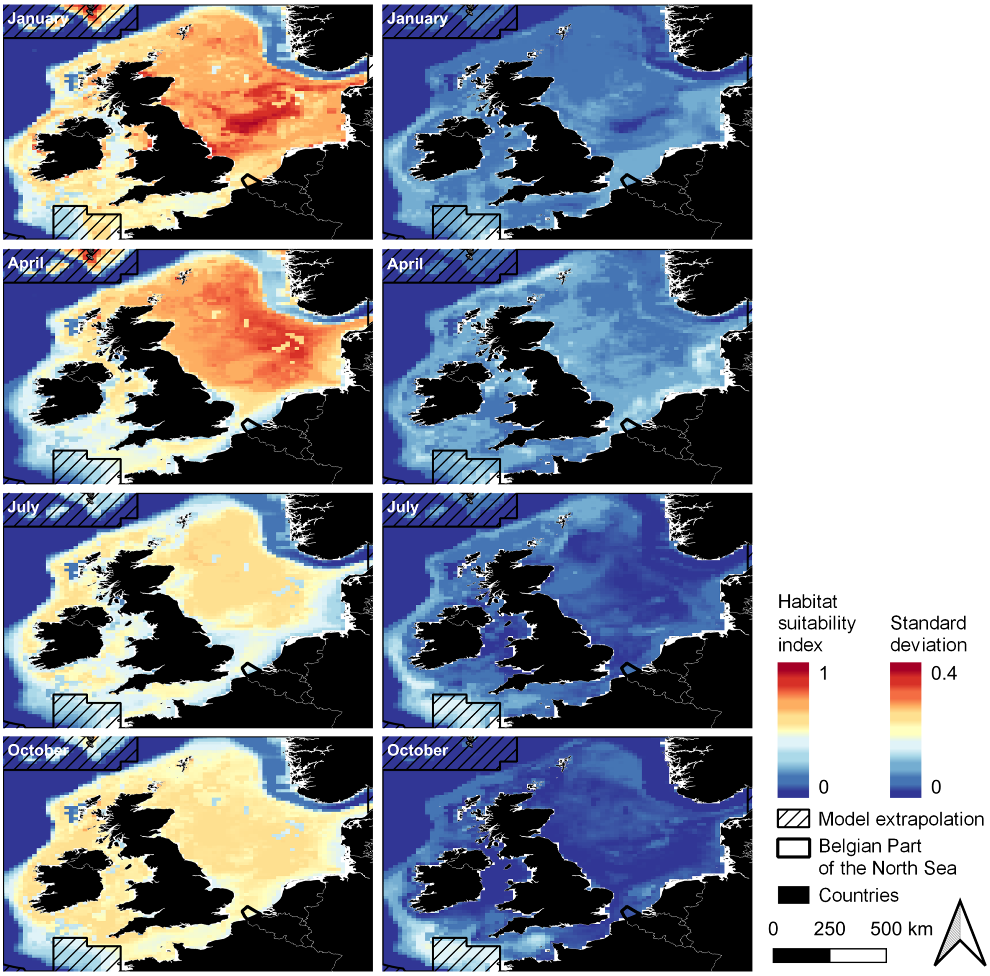


FIGURE 3
Average HSI (left) and average year-to-year variability (given as standard deviation, right) of the habitat suitability for adult Atlantic herring for January, April, July and October (representing the four seasons) in 2000 – 2020. High values are indicated in red, and low values in blue. Areas where the model extrapolated, i.e., where no occurrences were present, are hatched. Projection: EPSG 4326/WGS 84. Distribution maps for all months are shown in Supplementary Figure 1.

Zooming in on the BPNS specifically, habitat was calculated to be suitable for adult Atlantic herring throughout the entire year with increasing values through autumn and peaking in winter (January, February, and March, Figure 4). More specifically, monthly averages never dropped below 25% HSI and the highest average HSI was 75% in February. The variability of the HSI was highest in March and April with HSI varying between 21 and 85% and 12 and 64%, respectively.

For larvae of Atlantic herring, a gradual southward movement of suitable areas (HSI > 50%) can be seen in the distribution maps from September to January in the Northeast Atlantic (Figure 5). In September and October, high habitat suitability was calculated surrounding Ireland and the United Kingdom. Later, in December and January, areas with high habitat suitability occurred in the Celtic Sea and the English Channel (including the BPNS, HSI up to 85%, Figure 5). Year-to-year variability was higher in larval HSI than adult HSI (Figure 3 right vs. Figure 5 right) and reached up to a standard deviation of 30% in all months (Figure 5, right). In the BPNS, this variation showed an oscillating pattern, as visualized using boxplots in Figure 6.

3.5 North Atlantic Oscillation

The presence of Atlantic herring did not only show seasonal patterns (section 3.4) but also patterns across years. The HSI in the BPNS was negatively correlated with the NAO. More specifically and using a confidence interval of 99%, a significant correlation coefficient was found between winter NAO and winter HSI of adults (correlation of -0.57) at a moving average window of one year. At a moving average window of three years, significant correlation coefficients were found between autumn NAO and the winter HSI (-0.83 and -0.63 for adults and larvae respectively) and winter NAO and winter HSI (-0.83 and -0.76 for adults and larvae respectively). At a moving average of five years, both the autumn NAO and winter HSI (-0.86 and -0.63 for adults and larvae

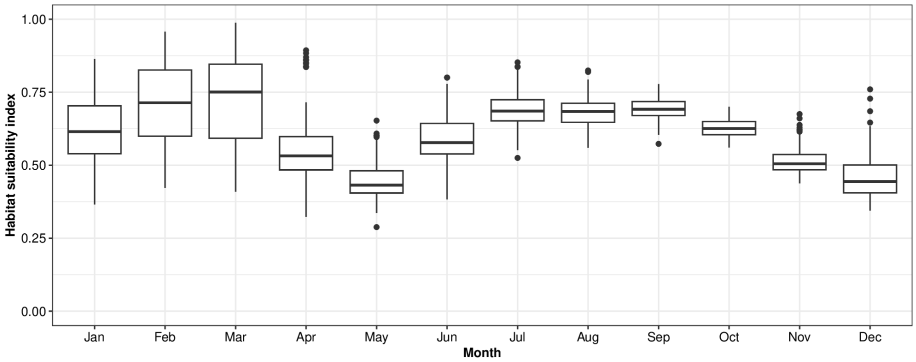


FIGURE 4
Monthly variability of the habitat suitability index for adult Atlantic herring in the Belgian Part of the North Sea, averaged over 2000 – 2020. For each month, the horizontal lines in the rectangular part of the boxplot represent, from low to high respectively, the 25th percentile, the 50th percentile and the 75th percentile. Points that fall outside of these ranges are shown by whiskers (vertical lines) that extend up to 1.5 times the interquartile range. Points falling outside of 1.5 times the interquartile range are shown as dots.

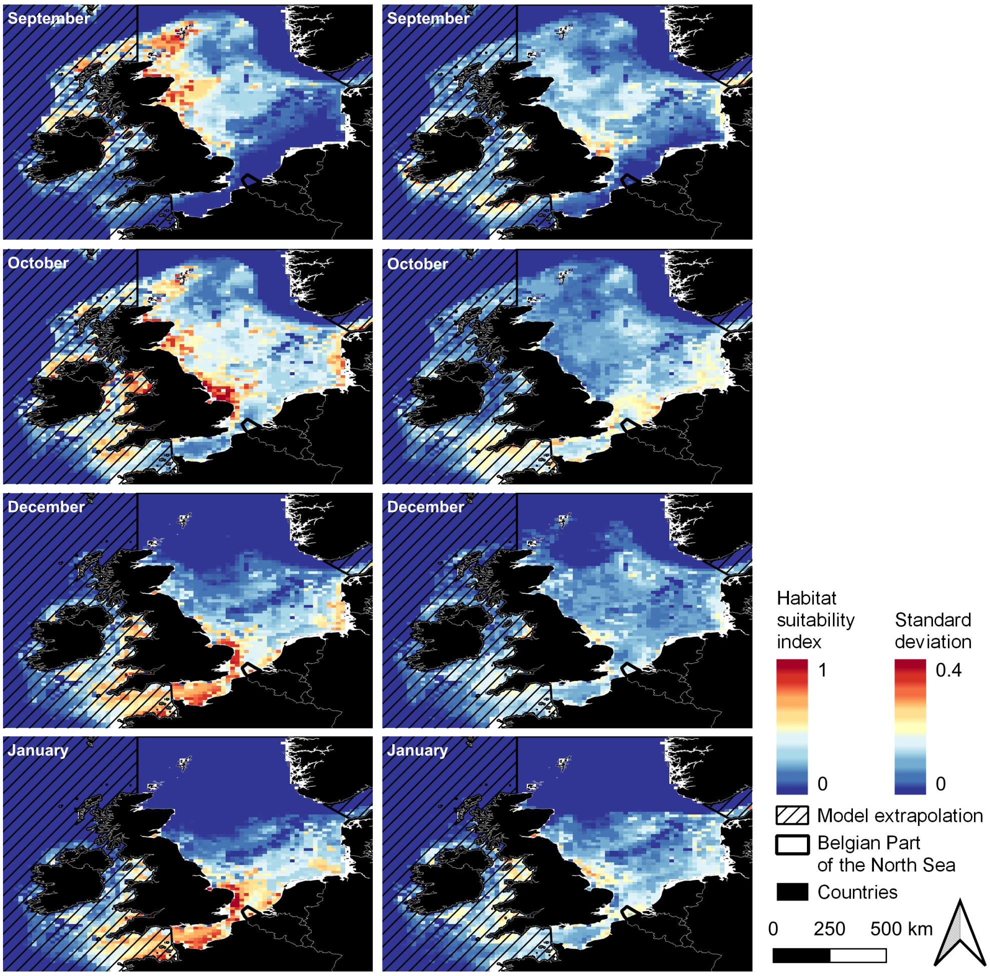


FIGURE 5
Average HSI (left) and average year-to-year variability (given as standard deviation, right) of the habitat suitability for larvae of Atlantic herring for September, October, December and January (months where data was available) in 2000 – 2020. High values are indicated in red, and low values in blue. Areas where the model extrapolated, i.e., where no occurrences were present, are hatched. Projection: EPSG 4326/WGS 84.

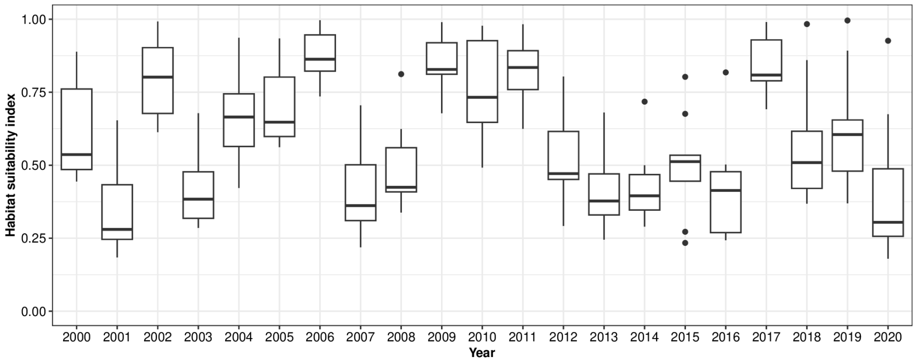


FIGURE 6
Yearly variability of habitat suitability for larvae of Atlantic herring in January in the Belgian Part of the North Sea. For each month, the horizontal lines in the rectangular part of the boxplot represent, from low to high respectively, the 25th percentile, the 50th percentile and the 75th percentile. Points that fall outside of these ranges are shown by whiskers (vertical lines) that extend up to 1.5 times the interquartile range. Points falling outside of 1.5 times the interquartile range are shown as dots.

respectively) and the winter NAO and winter HSI were significantly correlated (-0.88 and -0.78 for adults and larvae respectively). For all correlations, the time lag was either zero (winter NAO – winter HSI) or one season (autumn NAO – winter HSI), but no autocorrelations with a time lag of (over) a year were seen. For example, the effect of the winter NAO on the winter HSI for adults using a three-year averaging window is shown in Figure 7. In the winter of 2010 in the BPNS, the HSI for adults was simulated high (0.8) coinciding with low average NAO indices (-0.7). In contrast in the winter of 2015, the HSI was simulated relatively low (0.6) during high average NAO indices (0.8). The same pattern can be seen for the other significant combinations from above between autumn/ winter NAO and modeled HSI for adults/larvae (Supplementary Figures 5–7).

4 Discussion

This is one of the first studies making dynamic species distribution models for both adults and larvae of Atlantic herring in the Northeast Atlantic. Model outcomes suggest that suitable habitat for adult herring is widely spread over the Northeast Atlantic throughout the entire year (Figure 3). Suitable habitat for larvae occurs first in the North of the UK in September and moves gradually southward towards the English Channel throughout the spawning season (Figure 5). Focusing on the BPNS provided valuable insights for the development of spatiotemporally specific management strategies for fisheries on Atlantic herring by the Belgian fishing fleet. In this area, both adults and larvae are most likely to occur during winter months (adults: January – March; larvae: December – January).

4.1 Modeling outcomes in an ecological context

Bathymetry was the main explanatory variable in the adult model, followed by sea surface temperature and salinity (variable importance of 63, 13 and 11% respectively, Table 3). Response curves showed that habitat was suitable at depths shallower than 200 m, aligning with the region of the European continental shelf of which the edge is situated at approximately 200 m depth (Figure 2,

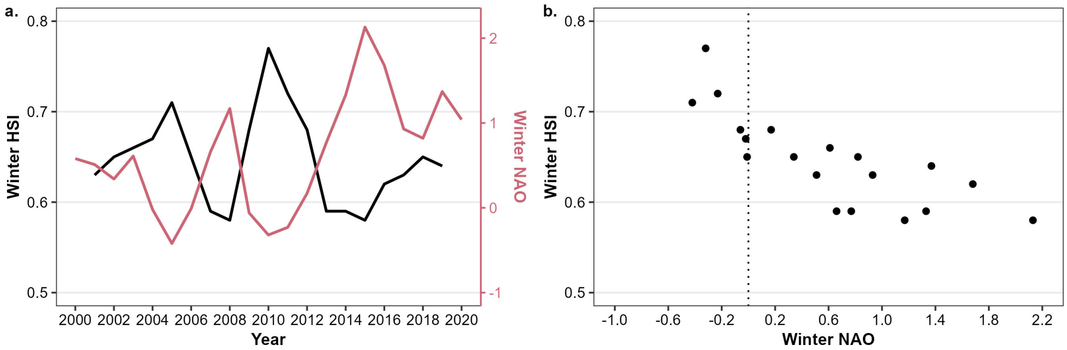


FIGURE 7
Effect of the North Atlantic Oscillation (NAO) on modeled habitat suitability indices (HSI) of adults in the Belgian Part of the North Sea during winter. **(A)** time series of 3-year averaged winter HSI (left y-axis) and winter NAO (in red, right y-axis). **(B)** Correlation plot between three-year averaged winter NAO and winter HSI.

Ricker and Stanev, 2020). The importance of bathymetry is assumed to be related to the Atlantic herring's adaptability to a wide range of temperatures and salinities (de Groot, 1980; Whitehead, 1985). This was also reflected in the response curves for sea surface temperature and salinity, being able to tolerate the entire range of values over 21 years in the study area (Figure 2).

To model the distribution of larvae, apart from bathymetry, sea surface temperature and salinity, also seabed substrate, zooplankton and phytoplankton concentration were important variables (37, 15, 5, 14, 13 and 13% variable importance respectively, Table 3). In terms of abiotic factors, response curves matched well with known characteristics of Atlantic herring spawning sites for the bathymetric range (7 – 150 m, Figure 2; Brevé et al., 2007; Dickey-Collas et al., 2004; Frost and Diele, 2022), substrate (sandy and course; de Groot, 1980) and salinity (both marine and brackish; Frost and Diele, 2022). Modeled temperature ranges (optimum at 7°C) also agree well with historical catch data from the North Sea (8°C; Hay et al., 2000). The inferred inclination of Atlantic herring towards colder water temperatures suggests that global warming may disrupt their current migration patterns, potentially leading to a shift towards more northern spawning grounds. However, the visual feeding behavior of herring, dependent on light for hunting (Blaxter, 1968), may hinder a northward shift during the short daylengths in northern winters, as suggested by Hufnagl and Peck (2011). Spawning sites are often classified as high-energy environments, by wave or tidal movement, which is important for egg development (Haegele and Schweigert, 1985). Seabed energy was included for modeling, but no effect was found.

A remarkable difference between the adult and the larval model was that prey concentration was important to model the distribution of larvae (26% variable importance in total including both phyto- and zooplankton concentration) while being less important to model the distribution of adults (11% variable importance in total, Table 3). An explanation could be that the adult model includes data on the complete cycle of their migration, including feeding and spawning grounds (de Groot, 1980; Coull et al., 1998). Therefore, the effect of feeding might be partly obscured. Furthermore, even though adult herring may be drawn to areas with high concentrations of zooplankton, their top-down influence could lead to lower concentrations of phyto- and zooplankton in those specific locations. Supporting both scenarios, Atlantic herring has previously been reported directly inside or at the edges of plankton patches (Maravelias, 2001).

For larvae, habitat suitability increased with increasing phytoplankton concentration and was optimal at zooplankton concentration in the epipelagic layer between 4.5 and 6.5 g C m^{-2}. This aligns with the fundamental physiological need of early life stages, as young herring prioritize somatic growth. Phytoplankton serves as the base of the food web, and the availability of zooplankton, which feeds on phytoplankton, becomes crucial for larval development. A lack of sufficient prey during these early stages can hinder growth and survival, as larvae rely on abundant food sources to sustain their rapid growth during early life (Fletcher et al., 2019). For example, for an epipelagic layer depth of 81 m (the average epipelagic layer depth in the Northeast Atlantic at noon, 01/01/2020, https://doi.org/10.48670/moi-00020), this would correspond to 55.6 – 80.2 mg C m^{-3}. Spawning grounds of Atlantic herring are chosen to promote larval retention (Sinclair and Power, 2015) and longer retention times near spawning grounds can lead to higher recruitment of Down's herring (Dickey-Collas et al., 2009). On top of this, Hufnagl and Peck (2011) reported that the duration of the hatching period is influenced by minimum prey concentration and prey size. Integrating these findings with our modeling results suggests that optimal spawning occurs in specific spawning grounds that promote retention and where an ample supply of prey is available. Our modeled outcomes suggest optimal prey concentrations for larval herring in the Northeast Atlantic.

The inclusion of nearby wind farm presence as a predictive variable to model the distribution of Atlantic herring is novel. During the construction of offshore wind farms, no spatial deterrence was found for free-ranging pelagic fish (Hubert et al., 2024). Offshore wind farms with scour protection introduce artificial hard substrates on the seabed and can attract demersal benthopelagic fish species by providing shelter, food sources and spawning sites (Degraer et al., 2018). To date, the spawning of Atlantic herring on windfarm substrate has not been observed, nor has our model detected any effect of nearby windfarm presence on the occurrence of Atlantic herring. However, the potential for wind farms to serve as artificial reefs could offer some benefits, particularly for larvae and early life stages. Artificial reefs have been shown to enhance local biodiversity by providing shelter, feeding opportunities, and protection from predators (Higgins et al., 2022). For herring larvae, the structural complexity of wind farms could increase prey availability and offer protection from predation, supporting their early development. However, the effect might be obscured by (1) inadequate sampling near windfarms due to fishing restrictions (Bonsu et al., 2024) or (2) the dispersal of larvae away from spawning sites following hatching (Sinclair and Power, 2015).

4.2 North Atlantic Oscillation

Adding on to the specific spawning requirements, larval habitat suitability indices had a larger year-to-year variability than adult HSI (up to 30 and 18% standard deviation respectively). This variability was found to be correlated with the NAO index. More specifically, positive autumn and winter NAO indices had a negative effect on adult and larval occurrence in the BPNS during winter. The NAO index represents an atmospheric sea level pressure difference between the Azores, Portugal and Iceland (Rogers, 1984). These pressure differences result in temporal variation of storms, precipitation, temperature, salinity, mixed-layer depth and circulation patterns (Hurrell and Deser, 2010). From the 1960s until the early 1990s a general positive trend of winter NAO indices has been observed and afterwards, the trend was less positive or even negative (Gulev et al., 2021). Climate models forecast a slight increase in the winter NAO in the future, with large natural variations (Lee et al., 2021).

Positive NAO indices and associated westerly winds lead to increased inflow of Atlantic waters in the North Sea which can increase nutrient concentrations and temperatures (van der Molen and Pätsch, 2022). Additionally, a stronger outflow of Baltic waters under positive NAO indices leads to a reduced exchange between the northern and southern North Sea (Salt et al., 2013). Considering the modeled response curves for temperature for both life stages of herring, the increase in temperature during positive NAO indices might have a direct impact on their physiology. Furthermore, the reduced exchange will affect the plankton distribution, which might affect herring. Investigating a correlation between temperature and Atlantic herring spawning stock biomass, Akimova et al. (2016) employed comparable reasoning. They suggested that the correlation was more likely attributed to fluctuations in the zooplankton composition than a direct impact of water temperature on larval growth rates. However, they were not able to find a match between the time series of the spawning stock biomass of Atlantic herring and zooplankton species in the North Sea. Finally, the impact of large-scale climate processes, including the NAO, on North Sea herring stock was modeled by Gröger et al. (2010). Gröger et al. (2010) did not find any correlation between the NAO and herring spawning stock biomass but did find a correlation with the number of recruits at a time lag of 5 years. Here they defined the number of recruits as the number of fish at 1 year of age (about 10 cm in length, Brevé et al., 2007), which is a different age group from the larvae used for modeling in our study (0.5 – 2.4 cm length). The age difference can partly explain the time lag seen in the study of Gröger et al. (2010), a phenomenon we did not observe in our study. On top of this, Gröger et al. (2010) looked at a different spatial scale than we did (North Sea vs. BPNS) and the effect of the NAO differs regionally (van der Molen and Pätsch, 2022). Since Atlantic herring is a key species in the North Sea food web (Fauchald et al., 2011) and is impacted by both top-down and bottom-up processes (Lynam et al., 2017), the net effect of NAO on the larval occurrence of herring in the BPNS is probably a combination of different effects. We recommend future ecological modeling work to focus on integrated approaches that include environmental variables, food web interactions, and climate processes and consider both the effects of space and time.

4.3 Model validation & limitations

Models performed well in terms of AUC and TSS values. Performance of the larval model was higher (AUC of 0.89 and TSS of 0.71) compared to the adult model (AUC of 0.73 and TSS of 0.61). To get a comprehensive understanding of the model's strengths and weaknesses, Grimmett et al. (2020) emphasize the importance of using multiple performance statistics alongside commonly used metrics like AUC. Specifically, the sensitivity and the specificity give information on the model's capability of predicting presences and background points, respectively. The adult model showed a better performance in predicting suitable habitat (sensitivity of 0.78), compared to unsuitable habitat (specificity of 0.68). The lower specificity suggests that the actual species range might be more confined than what the models projected.

The higher performance of the larval model compared to the adult model indicates that larvae exhibit more characteristics of habitat specialists than the generalist adults. Habitat specialists, characterized by narrow environmental tolerances, are often more straightforward to model compared to habitat generalists (Brotons et al., 2004; Elith et al., 2006). In modeling terms, this can be explained by the selection of background points, which are more likely to be true absences for a species with a narrow spatial distribution (habitat specialist) compared to a widely distributed species (habitat generalist; Grimmett et al., 2020; Lobo et al., 2008). Fernandez et al. (2022) suggest adopting the non-observation of highly mobile species in dynamic environments (such as the ocean) as part of the study area background, rather than treating it as an absence. Building upon this recommendation, we opted for Maxent models, which integrate non-observations as background points.

One of the limitations of this study is the reliance on occurrence data rather than abundance data to model Atlantic herring distribution. While occurrence-based models such as Maxent are valuable for capturing species-environment relationships, they do not account for the density or biomass of herring, which is essential for effective fisheries management. Species that are found in high numbers in specific locations can play a significant role in ecosystem dynamics, and focusing solely on occurrence data may miss capturing these ecological interactions. On the other hand, incorporating abundance data poses challenges of its own. Abundance data can be influenced by biases stemming from differences in sampling methods, spatial and temporal coverage, and variable sampling efforts, leading to difficulties in interpreting true species-environment relationships (Bonar et al., 2011). Factors like fishing gear types and tactics can strongly impact the perceived abundance of species in a particular area, introducing inaccuracies in data analysis (Mehdi et al., 2021; Moriarty et al., 2020). Additionally, the migratory nature and aggregative behavior of Atlantic herring makes it challenging to gather consistent and comprehensive abundance data.

While abundance data was available for adult herring, we opted not to use it. The ICES surveys are compiled of different surveys of different countries (Supplementary Table 1). On top of this, no abundance data was available for larvae. Using the same form of input data (occurrences data) for both life stages allowed for comparison of the ecological needs for the two life stages. While this approach may not fully capture the ecological dynamics of Atlantic herring, it provides a reliable and expansive dataset for evaluating habitat suitability over large spatial extents. Occurrence data is also noted to enhance the performance of species distribution models, particularly for migratory species or those with extensive ranges. Despite the advantages of using occurrence data, it is essential to recognize the limitations of this approach for fisheries management and consider integrating abundance data where feasible to gain a more nuanced insight into herring distribution and its implications for fisheries.

We acknowledge that the outcomes of our model may be partly biased due to the use of demersal input data to model pelagic adult herring (Brevé et al., 2007; ICES, 2023e). Demersal sampling is likely to miss some occurrences of herring when they are swimming in the upper water column. Since the Maxent model does not consider non-observations as true absences, but rather as a part of the background, this model could be more robust against such biases (Fernandez et al., 2022). Additionally, the DATRAS trawl surveys, which are compiled from different surveys organized by multiple countries, involve varying sampling depth ranges (Supplementary Table 1). This variation can introduce bias into the model output, leading to instances such as the unsuitability of deeper waters in the Norwegian Trench (Figure 3). The absence of occurrences at these depths is reflected in the model as being unsuitable, despite evidence that spring-spawning Atlantic herring are present in this region during winter (Corten, 2000). To address these biases, we have highlighted areas where the model extrapolates beyond the spatial range of observations in Figure 3 and Figure 5. Moreover, to the best of our knowledge, we have used the most reliable dataset currently available for our study area. Fisheries independent datasets, such as the DATRAS trawl surveys, are often preferred over fishery dependent data because they follow a recurring sampling scheme with sufficient spatiotemporal coverage (Hilborn and Walters, 2013). Finally, other studies such as Turner et al. (2016) and Wang et al. (2018) have also employed demersal surveys to develop species distribution models for Atlantic herring with good model prediction accuracies (AUC values > 0.75).

4.4 Spatiotemporal distribution maps

A visual comparison shows a good match of our spatiotemporal habitat suitability maps of larval herring with the location and timing of known spawning grounds (Figure 5 and Supplementary Figure 8). Spawning time is used to distinguish different autumn-spawning stocks (Heath et al., 1997). In the North Sea, spawning starts in August around the Shetland Islands, Orkney Islands and west of Scotland and ends in January in the southern North Sea (Figure 7; Coull et al., 1998; Gröger et al., 2010). On the east of the UK, model outcomes (Figure 5) correspond accurately with the Shetland stock in September, the Buchan stock in September – October, the Banks stock in October – December and the Downs stocks in December – January (Figure 7). No observations were found to include in the model from the west side of the UK, however, the model was able to extrapolate larval habitat preferences to these areas as well. These extrapolations accurately show the spawning grounds west of Scotland and Ireland during September and October. Spawning grounds in the east of Ireland (October – December – January) and in the English Channel (December – January) were predicted wider than the findings from Coull et al. (1998) (Figures 5 , 7). These authors stress that the location of spawning grounds should be under continuous revision. The precise location might be blurred in our model due to the lack of direct spawning ground observations since larvae can show some degree of dispersal from spawning grounds through local currents (Bauer et al., 2014; Funk et al., 2001; Sinclair and Power, 2015).

The adult model forecasts a wide distribution of Atlantic herring across the Northeast Atlantic throughout the entire year (Figure 3). Over different seasons, some regional differences could be seen. The most suitable areas were centered in the North Sea and around the Faroe Islands in the first half of the year to an even wider area including the east of the UK during the second half of the year.

4.5 Implications for fisheries, particularly the Belgian fishing fleet

The Belgian fishing fleet has witnessed a steady decline in number of catches and number of fishing vessels in the southern North Sea (Maertens, 2022) and this decline might be aggravated when the effects of Brexit come fully into force (Popescu and Scholaert, 2022). This study aimed to provide information on the location of Atlantic herring as a first assessment towards restoring pelagic fishing for the Belgian fishing fleet. Note that our models predict environmentally suitable areas where herring can be found during different seasons, but they do not provide any information on their biomass at those locations. The outcomes of the adult model show that Atlantic herring is likely to occur in the Greater North Sea throughout the entire year. For the Belgian fishing fleet, fishing directly in the BPNS would incur the lowest cost for ship operation. Here, areas of high habitat suitability for Atlantic herring were simulated in December – January for larvae and later, in January – February, for adults. The habitat in the BPNS being suitable for larvae first, before adults, could indicate that early spawning takes place outside of the BPNS in November – December and that, following hatching, larvae could be transported towards the BPNS through eastward local currents (Turrell, 1992). Sinclair and Power (2015) found that Atlantic herring choose their spawning sites to limit larval transport and hence spawning might occur nearby, likely in the Downs and Banks stocks. Later, in December – January adults arrive in the BPNS and spawning might occur on the sandbanks and gravel grounds of the BPNS itself.

Given these outcomes, pelagic fishing on adult Atlantic herring in the BPNS would be most suitable during winter months. However, due to the potential presence of spawning nearby and herring's susceptibility to collapse (Stephenson et al., 2001), fisheries must be managed effectively. Bottom trawling can have a direct negative impact on deposited eggs (Watling and Norse, 1998). Atlantic herring are caught using different types of gear including purse seine, mid-water trawl, pair trawl and otter trawl (ICES, 2005). Therefore, if fishing near spawning areas would be permitted, fisheries should at least consider employing non-bottom-stirring techniques.

5 Conclusion

Our study showed the widespread spatiotemporal distribution of Atlantic herring in the Northeast Atlantic, using species distribution models (AUC of 0.7). Models based on larval data were effective in deriving the Atlantic herring spawning distribution (AUC of 0.9). For the BPNS, outcomes show that Atlantic herring is likely to be present during winter months, both as adults and larvae. The year-to-year variability of habitat suitability during these months in the BPNS was negatively correlated (up to - 0.88) with the autumn and winter NAO indices. Positive NAO events might negatively impact spawning success through increased temperature and changes in prey composition.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://datras.ices.dk/home/descriptions.aspx#FRA; https://www.ices.dk/data/data-portals/Pages/DATRAS.aspx; https://emodnet.ec.europa.eu/en/bathymetry; https://emodnet.ec.europa.eu/en/seabed-habitats; https://emodnet.ec.europa.eu/en/human-activities; https://doi.org/10.48670/moi-00021; https://doi.org/10.48670/moi-00021; https://doi.org/10.48670/moi-00058; https://doi.org/10.48670/moi-00020. All datasets and scripts of this study can be found in the Marine Data Archive (MDA), https://doi.org/10.14284/657 (under Creative Commons Attribution 4.0 International License).

Author contributions

WS: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. RM: Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. MS: Writing – review & editing, Conceptualization. JG: Writing – review & editing, Resources, Data curation. CM: Writing – review & editing, Project administration, Funding acquisition. ED: Writing – review & editing, Project administration, Funding acquisition. SP: Writing – review & editing, Conceptualization. GE: Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the European Commission's Brexit Adjustment Reserve (BAR) – Veerkracht voor de kustvloot en klein vlootsegment post-Brexit (project No 1654) and the EU Public Infrastructure for the Digital Twin of the Ocean (EDITO-Infra) (project No 101101473).

Acknowledgments

The authors would like to thank Damian Villagra Villanueva (Flanders Research Institute for Agriculture, Fisheries and Food) and Ruben Perez Perez (Flanders Marine Institute) for their vital help with the data retrieval of the DATRAS and OBIS occurrence data. Furthermore, we would like to thank Dr. Leandro Ponsoni (Flanders Marine Institute) for his input on the decomposition of the NAO effect. This study has been conducted using E.U. Copernicus Marine Service Information; https://doi.org/10.48670/moi-00021, https://doi.org/10.48670/moi-00058, https://doi.org/10.48670/moi-00020. During the preparation of this work, the authors used ChatGPT 3.5 and Grammarly to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2024.1485161/full#supplementary-material

References

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* 38, 541–545. doi: 10.1111/ecog.01132

Aires, C., Manuel González-Irusta, J., and Watret, R. (2014). *Updating fisheries sensitivity maps in British waters*. UK: Scottish Marine and Freshwater Science.

Akimova, A., Núñez-Riboni, I., Kempf, A., and Taylor, M. H. (2016). Spatially-resolved influence of temperature and salinity on stock and recruitment variability of commercially important fishes in the North Sea. *PloS One* 11. doi: 10.1371/journal.pone.0161917

Alheit, J., Möllmann, C., Dutz, J., Kornilovs, G., Loewe, P., Mohrholz, V., et al. (2005). Synchronous ecological regime shifts in the central Baltic and the North Sea in the late 1980s., in. *ICES J. Mar. Sci.* 62, 1205–1215. doi: 10.1016/j.icesjms.2005.04.024

Báez, J. C., Barbosa, A. M., Pascual, P., Ramos, M. L., and Abascal, F. (2020). Ensemble modeling of the potential distribution of the whale shark in the Atlantic Ocean. *Ecol. Evol.* 10, 175–184. doi: 10.1002/ece3.5884

Barber, R. A., Ball, S. G., Morris, R. K. A., and Gilbert, F. (2022). Target-group backgrounds prove effective at correcting sampling bias in Maxent models. *Divers. Distrib* 28, 128–141. doi: 10.1111/ddi.13442

Bauer, R. K., Gräwe, U., Stepputtis, D., Zimmermann, C., and Hammer, C. (2014). Identifying the location and importance of spawning sites of Western Baltic herring using a particle backtracking model. *ICES J. Mar. Sci.* 71, 499–509. doi: 10.1093/icesjms/fst163

Blaxter, J. H. S. (1968). Visual thresholds and spectral sensitivity of herring larvae. *J. Exp. Biol.* 48, 39–53. doi: 10.1242/jeb.48.1.39

Bonar, S., Fehmi, J., and Mercado-Silva, N. (2011). "An onverview of sampling issues in species diversity and abundance surveys," in *Biological Diversity Frontiers in Measirement and Assessment*. (UK: Oxford University Press Oxford), 11–27.

Bonsu, P. O., Letschert, J., Yates, K. L., Svendsen, J. C., Berkenhagen, J., Rozemeijer, M. J. C., et al. (2024). Co-location of fisheries and offshore wind farms: Current practices and enabling conditions in the North Sea. *Mar. Policy* 159, 105941. doi: 10.1016/j.marpol.2023.105941

Brevé, N. P. W., van Emmerik, W. A. M., and van Beek, G. C. W. (2007). *Kennisdocument Atlantische haring*. Available online at: http://www.clupea.net (Accessed May 19, 2023).

Brotons, L., Thuiller, W., Araújo, M. B., and Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* 27, 437–448. doi: 10.1111/j.0906-7590.2004.03764.x

Corten, A. (1999). A proposed mechanism for the Bohuslän herring periods. *ICES J. Mar. Sci.* 56, 207–2020. doi: 10.1006/jmsc.1998.0429

Corten, A. (2000). A possible adaptation of herring feeding migrations to a change in timing of the Calanus finmarchicus season in the eastern North Sea. *ICES J. Mar. Sci.* 57, 1261–1270. doi: 10.1006/jmsc.2000.0812

Coudyser, C. (2021). *Eerste gevolgen van Brexit voor Vlaamse visserijsector worden duidelijk*. Available online at: https://www.cathycoudyser.be/nieuws/eerste-gevolgen-van-brexit-voor-de-vlaamse-visserijsector-worden-duidelijk (Accessed May 22, 2024).

Coull, K. A., Johnstone, R., and Rogers, S. I. (1998). *Fishery sensitivity maps in british waters*. (UK: UKOOA Ltd).

Cushing, D. H. C. (1992). *A short history of the Downs stock of herring*. Available online at: https://academic.oup.com/icesjms/article/49/4/437/653339 (Accessed January 10, 2024).

Degraer, S., Brabant, R., Rumes, B., and Vigin, L. (2018). *Environmental impacts of offshore wind farms in the belgian part of the north sea: assessing and managing effect spheres of influence*. Available online at: https://www.researchgate.net/publication/328095905 (Accessed June 1, 2023).

de Groot, S. J. (1980). The consequences of marine gravel extraction on the spawning of herring, Clupea harengus. *J. Fish Biol.* 16, 605–611. doi: 10.1111/j.1095-8649.1980.tb03739.x

de Oliveira, G., Rangel, T. F., Lima-Ribeiro, M. S., Terribile, L. C., and Diniz-Filho, J. (2014). Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: A new approach based on environmentally equidistant records. *A. F.Ecography* 37, 637–647. doi: 10.1111/j.1600-0587.2013.00564.x

Departement Landbouw en Visserij (2021). *Visserijrapport 2020* (Brussels: Departement Landbouw en Visserij).

Dickey-Collas, M., Bolle, L. J., Van Beek, J. K. L., and Erftemeijer, P. L. A. (2009). Variability in transport of fish eggs and larvae. II. effects of hydrodynamics on the transport of Downs herring larvae. *Mar. Ecol. Prog. Ser.* 390, 183–194. doi: 10.3354/meps08172

Dickey-Collas, M., Van Beek, D. F. A., Wot, H., and Voor Visserijonderzoek, C. (2004). *The current state of knowledge on the ecology and interactions of North Sea Herring within the North Sea ecosystem*. (Ijmuiden, Netherlands: Stichting DLO, Centre for Fishery Research).

Elith, J., Graham, H., C., P., Anderson, R., Dudík, M., Ferrier, S., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29, 129–151. doi: 10.1111/j.2006.0906-7590.04596.x

Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. *Methods Ecol. Evol.* 1, 330–342. doi: 10.1111/j.2041-210x.2010.00036.x

European Commission (2020). *Proposal for a regulation of the European Parliament and of the Council establishing the Brexit Adjustment Reserve*. Available online at: https://commission.europa.eu/publications/brexit-adjustment-reserve_enfiles (Accessed February 14, 2024).

Fauchald, P., Skov, H., Skern-Mauritzen, M., Johns, D., and Tveraa, T. (2011). Wasp-Waist interactions in the North Sea ecosystem. *PloS One* 6. doi: 10.1371/journal.pone.0022729

Fernandez, M., Sillero, N., and Yesson, C. (2022). To be or not to be: the role of absences in niche modelling for highly mobile species in dynamic marine environments. *Ecol. Modell* 471. doi: 10.1016/j.ecolmodel.2022.110040

Fletcher, C., Collins, S., Nannini, M., and Wahl, D. (2019). Competition during early ontogeny: Effects of native and invasive planktivores on the growth, survival, and habitat use of bluegill. *Freshw. Biol.* 64, 697–707. doi: 10.1111/FWB.13255

Frost, M., and Diele, K. (2022). Essential spawning grounds of Scottish herring: current knowledge and future challenges. *Rev. Fish Biol. Fish* 32, 721–744. doi: 10.1007/s11160-022-09703-0

Funk, F., Blackburn, J., Hay, D., Paul, A. J., Stephenson, R., Toresen, R., et al. (2001). *Herring: expectations for a new millennium* (Anchorage, Alaska, USA: University of Alaska Sea Grant).

Geffen, A. J. (2009). *Advances in herring biology: from simple to complex, coping with plasticity and adaptability*. Available online at: https://academic.oup.com/icesjms/article/66/8/1688/672963 (Accessed June 1, 2023).

Grimmett, L., Whitsed, R., and Horta, A. (2020). Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecol. Modell* 431. doi: 10.1016/j.ecolmodel.2020.109194

Gröger, J. P., Kruse, G. H., and Rohlf Gröger, N. (2010). *Slave to the rhythm: how large-scale climate cycles trigger herring (Clupea harengus) regeneration in the North Sea*. Available online at: https://academic.oup.com/icesjms/article/67/3/454/732742 (Accessed February 1, 2024).

Gulev, S. K., Thorne, P. W., Ahn, J., Dentener, F. J., Domingues, C. M., Gerland, S., et al. (2021). "Changing state of the climate system," in *Climate change 2021 – the physical science basis*. Eds. V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan and S. Berger (Cambridge, UK: Cambridge University Press), 287–422. doi: 10.1017/9781009157896.004

Haegele, C. W., and Schweigert, F. (1985). Distribution and characteristics of herring spawning grounds and description of spawning behavior. *Can. J. Fisheries Aquat. Sci.* 42, s39-s55. Available at: www.nrcresearchpress.com.

Hay, D. E., Toresen, R., Stephenson, R., Thompson, M., Claytor, R., Funk, F., et al. (2000). *Herring: expectations for a new millennium taking stock: an inventory and review of world herring stocks in 2000*. (USA: University of Alaska Sea Grant College Program).

Heath, M. (1993). An evaluation and review of the ICES Herring Larval Surveys in the North Sea and adjacent waters. *Bull. Mar. Sci.* 53, 993.

Heath, M., Scott, B., and Bryant, A. D. (1997). Modelling the growth of herring from four different stocks in the North Sea. *J. Sea Res.* 38, 413–436. doi: 10.1016/S1385-1101(97)00045-2

Higgins, E., Metaxas, A., and Scheibling, R. E. (2022). A systematic review of artificial reefs as platforms for coral reef research and conservation. *PloS One* 17, e0261964. doi: 10.1371/journal.pone.0261964

Hijmans, R. J., Phillips, S., Leathwick, J., and Elith, J. (2023). *dismo: species distribution modeling*. Available online at: https://CRAN.R-project.org/package=dismo (Accessed October 4, 2023).

Hilborn, R., and Walters, C. J. (2013). *Quantitative fisheries stock assessment: choice, dynamics and uncertainty* (Berlin/Heidelberg, Germany: Springer Science & Business Media).

Hubert, J., Demuynck, J. M., Remmelzwaal, M. R., Muñiz, C., Debusschere, E., Berges, B., et al. (2024). An experimental sound exposure study at sea: No spatial deterrence of free-ranging pelagic fish. *J. Acoust Soc. Am.* 155, 1151–1161. doi: 10.1121/10.0024720

Hufnagl, M., and Peck, M. A. (2011). Physiological individual-based modelling of larval Atlantic herring (Clupea harengus) foraging and growth: Insights on climate-driven life-history scheduling. *ICES J. Mar. Sci.* 68, 1170–1188. doi: 10.1093/icesjms/fsr078

Hurrell, J. W., and Deser, C. (2010). North Atlantic climate variability: The role of the North Atlantic Oscillation. *J. Mar. Syst.* 79, 231–244. doi: 10.1016/j.marsys.2008.11.026

Hysen, L., Nayeri, D., Cushman, S., and Wan, H. Y. (2022). Background sampling for multi-scale ensemble habitat selection modeling: Does the number of points matter? *Ecol. Inform* 72. doi: 10.1016/j.ecoinf.2022.101914

ICES (2005). ICES FishMap species factsheet: Herring. In: *ICES-fishMap*. Available online at: https://www.ices.dk/about-ICES/projects/EU-RFP/EU%20Repository/ICES%20FIshMap/ICES%20FishMap%20species%20factsheet-herring.pdf (Accessed January 23, 2024).

ICES (2020). "Manual for the north sea international bottom trawl surveys," in *Series of ICES survey protocols*, vol. 10. . doi: 10.17895/ices.pub.7562

ICES (2023a). "Herring (Clupea harengus) in division 7.a north of 52°30'N (Irish sea)," in *Report of the ICES advisory comitte. ICES advice 2023, her.27.nirs*. doi: 10.17895/ices.advice.23608098

ICES (2023b). "Herring (Clupea harengus) in divisions 6.a South of 56°00'N and West of 07°00'W and 7.b-c (Northwest and West of Ireland)," in *Report of the ICES advisory comitte. ICES advice 2023, her.27.6aS7bc*. doi: 10.17895/ices.advice.21907953

ICES (2023c). "Herring (Clupea harengus) in divisions 7.a South of 52°30'N, 7.g-h, and 7.j-k (Irish Sea, Celtic Sea, and Southwest of Ireland)," in *Report of the ICES advisory comitte. ICES advice 2023, her.27.irls*. doi: 10.17895/ices.advice.21907962

ICES (2023d). "Herring (Clupea harengus) in Subarea 4 and divisions 3.a and 7.d, autumn spawners (North Sea, Skagerrak and Kattegat, eastern English Channel)," in *Report of the ICES advisory committe. ICES advice 2023, her.27.3a47d*. doi: 10.17895/ices.advice.21907947

ICES (2023e). *ICES database on trawl surveys (DATRAS)*. Available online at: https://datras.ices.dk (Accessed October 2, 2023).

ICES (2023f). *The international herring larvae surveys*. Available online at: http://eggsandlarvae.ices.dk/ (Accessed November 23, 2023).

Joly, L. J., Loots, C., Meunier, C. L., Boersma, M., Collet, S., Lefebvre, V., et al. (2021). Maturation of the digestive system of Downs herring larvae (Clupea harengus, Linnaeu): identification of critical periods through ontogeny. *Mar. Biol.* 168. doi: 10.1007/s00227-021-03894-z

Kass, J. M., Muscarella, R., Galante, P. J., Bohl, C. L., Pinilla-Buitrago, G. E., Boria, R. A., et al. (2021). ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods Ecol. Evol.* 12, 1602–1608. doi: 10.1111/2041-210X.13628

Lee, J. Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J. P., et al. (2021). "Future global climate: scenario-based projections and near-term information," in *Climate change 2021 – the physical science basis*. Eds. V. ,. P. Masson-Delmotte, A. Zhai, S. L. Pirani, C. Connors, S. Péan, Berger,, et al (Cambridge, UK: Cambridge University Press), 553–672. doi: 10.1017/9781009157896.006

Legendre, P. (1993). Spatial autocorrelation trouble or new paradigm? *Ecology* 74, 1659–1673. doi: 10.2307/1939924

Lescrauwaet, A. K., Debergh, H., Vincx, M., and Mees, J. (2010). Fishing in the past: Historical data on sea fisheries landings in Belgium. *Mar. Policy* 34, 1279–1289. doi: 10.1016/j.marpol.2010.05.006

Limborg, M. T., Helyar, S. J., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., et al. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Mol. Ecol.* 21, 3686–3703. doi: 10.1111/j.1365-294X.2012.05639.x

Liu, C., White, M., and Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *J. Biogeogr* 40, 778–789. doi: 10.1111/jbi.12058

Lobo, J. M., Jiménez-Valverde, A., and Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography* 33, 103–114. doi: 10.1111/j.1600-0587.2009.06039.x

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeography* 17, 145–151. doi: 10.1111/J.1466-8238.2007.00358.X

Loiselle, B. A., Jørgensen, P. M., Consiglio, T., Jiménez, I., Blake, J. G., Lohmann, L. G., et al. (2008). Predicting species distributions from herbarium collections: Does climate bias in collection sampling influence model outcomes? *J. Biogeogr* 35, 105–116. doi: 10.1111/j.1365-2699.2007.01779.x

Lynam, C. P., Llope, M., Möllmann, C., Helaouët, P., Bayliss-Brown, G. A., and Stenseth, N (2017). Interaction between top-down and bottom-up control in marine food webs. *C.Proc. Natl. Acad. Sci. U.S.A.* 114, 1952–1957. doi: 10.1073/pnas.1621037114

Maertens, B. (2022). *Annual fleet report 2021* (Belgium: Departement Landbouw en Visserij).

Manel, S., Dias, J. M., and Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecol. Modell* 120, 337–347. doi: 10.1016/S0304-3800(99)00113-1

Maravelias, C. D. (2001). Habitat associations of Atlantic herring in the Shetland area: Influence of spatial scale and geographic segmentation. *Fish Oceanogr* 10, 259–267. doi: 10.1046/j.1365-2419.2001.00172.x

Maravelias, C. D., Reid, D. G., and Swartzman, G. (2000). Seabed substrate, water depth and zooplankton as determinants of the prespawning spatial aggregation of North Atlantic herring. *Mar. Ecol. Prog. Ser.* 195, 249–259. doi: 10.3354/meps195249

Marshall, S. M., Nicholls, A. G., and Orr, A. P. (1937). On the growth and feeding of the larval and post-larval stages of the Clyde herring. *J. Mar. Biol. Assoc. United Kingdom* 22, 245–267. doi: 10.1017/S002531540001198X

Mehdi, H., Lau, S. C., Synyshyn, C., Salena, M. G., Morphet, M. E., Hamilton, J., et al. (2021). A comparison of passive and active gear in fish community assessments in summer versus winter. *Fisheries Res.* 242, 106016. doi: 10.1016/j.fishres.2021.106016

Merow, C., Smith, M. J., and Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography* 36, 1058–1069. doi: 10.1111/j.1600-0587.2013.07872.x

Moriarty, M., Sethi, S. A., Pedreschi, D., Smeltz, T. S., McGonigle, C., Harris, B. P., et al. (2020). Combining fisheries surveys to inform marine species distribution modelling. *ICES J. Mar. Sci.* 77, 539–552. doi: 10.1093/icesjms/fsz254

Morrison, J. A., Napier, I. R., Gamble Morrison, J. C., and Napier, ,. J. A. (1991).Mass mortality of herring eggs associated with a sedimenting diatom bloom. Available online at: https://academic.oup.com/icesjms/article/48/2/237/644514 (Accessed January 12, 2024).

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Phillips, S. (2021). *maxnet: Fitting "Maxent" Species Distribution Models with "glmnet". R package version 0.1.4.* Available online at: https://CRAN.R-project.org/package=maxnet (Accessed February 6, 2024).

Phillips, S. B., Aneja, V. P., Kang, D., and Arya, S. P. (2006). "Modelling and analysis of the atmospheric nitrogen deposition in North Carolina," in *International journal of global environmental issues* (Geneva, Switzerland: Inderscience Publishers), 231–252. doi: 10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., et al. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, pp. 181 197. doi: 10.1890/07-2153.1

Popescu, I., and Scholaert, F. (2022). Brexit and the reduction in EU fishing quota share to 2023. *Eur. Parliamentary Res. Service.*

R Core Team (2023).R: A language and environment for statistical computing. In: *R foundation for statistical computing*. Available online at: https://www.R-project.org/ (Accessed August 11, 2023).

Regulation 1380/2013 (2013). *Regulation (EU) No 1380/2013 Of The European Parliament And Of The Council of 11 December 2013 on the Common Fisheries Policy, amending Council Regulations (EC) No 1954/2003 and (EC) No 1224/2009 and repealing Council Regulations (EC) No 2371/2002 and (EC) No 639/2004 and Council Decision 2004/585/EC.* Strasbourg: Official Journal of the European Union.

Ricker, M., and Stanev, E. V. (2020). Circulation of the European northwest shelf: A Lagrangian perspective. *Ocean Sci.* 16, 637–655. doi: 10.5194/os-16-637-2020

Rogers, J. C. (1984). The association between the North Atlantic Oscillation and the Southern Oscillation in the northern hemisphere. *Mon Weather Rev.* 112, 1999–2015. doi: 10.1175/1520-0493(1984)112<1999:TABTNA>2.0.CO;2

Salt, L. A., Thomas, H., Prowe, A. E. F., Borges, A. V., Bozec, Y., and De Baar, H. J. W. (2013). Variability of North Sea pH and CO2 in response to North Atlantic Oscillation forcing. *J. Geophys Res. Biogeosci* 118, 1584–1592. doi: 10.1002/2013JG002306

Signorell, A. (2024). *DescTools tools for descriptive statistics*. Available online at: https://CRAN.R-project.orgpackage=DescTools (Accessed March 18, 2024).

Sillero, N., and Barbosa, A. M. (2021). Common mistakes in ecological niche models. *Int. J. Geographical Inf. Sci.* 35, 213–226. doi: 10.1080/13658816.2020.1798968

Sinclair, M., and Power, M. (2015). The role of "larval retention" in life-cycle closure of Atlantic herring (Clupea harengus) populations. *Fish Res.* 172, 401–414. doi: 10.1016/j.fishres.2015.07.026

Stephenson, R., Clark, K. J., Power, M. J., Fife, F. J., and Melvin, G. D. (2001). "Herring stock structure, stock discreteness, and biodiversity," in *Herring: expectations for a new millennium* (Alaska, US: University of Alaska Sea Grant College Program), 559–571.

Thuiller, W., Lafourcade, B., Engler, R., and Araújo, M. B. (2009). BIOMOD - A platform for ensemble forecasting of species distributions. *Ecography* 32, 369–373. doi: 10.1111/j.1600-0587.2008.05742.x

Turner, S. M., Manderson, J. P., Richardson, D. E., Hoey, J. J., and Hare, J. A. (2016). Using habitat association models to predict Alewife and Blueback Herring marine distributions and overlap with Atlantic Herring and Atlantic Mackerel: Can incidental catches be reduced? *ICES J. Mar. Sci.* 73, 1912–1924. doi: 10.1093/icesjms/fsv166

Turrell, W. R. (1992). New hypotheses concerning the circulation of the northern North Sea and its relation to North Sea fish stock recruitment. *ICES J. Mar. Sci.* 49, 107–123. doi: 10.1093/icesjms/49.1.107

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2023). Flexible species distribution modelling methods perform well on spatially separated testing data. *Global Ecol. Biogeography* 32, 369–383. doi: 10.1111/geb.13639

van der Molen, J., and Pätsch, J. (2022). An overview of Atlantic forcing of the North Sea with focus on oceanography and biogeochemistry. *J. Sea Res.* 189. doi: 10.1016/j.seares.2022.102281

Van Ginderdeuren, K., Vandendriessche, S., Prössler, Y., Matola, H., Vincx, M., and Hostens, K. (2014). Selective feeding by pelagic fish in the Belgian part of the North Sea. *ICES J. Mar. Sci.* 71, 808–820. doi: 10.1093/icesjms/fst183

Vollering, J., Halvorsen, R., Auestad, I., and Rydgren, K. (2019). Bunching up the background betters bias in species distribution models. *Ecography* 42, 1717–1727. doi: 10.1111/ecog.04503

Wang, L., Kerr, L. A., Record, N. R., Bridger, E., Tupper, B., Mills, K. E., et al. (2018). Modeling marine pelagic fish species spatiotemporal distributions utilizing a maximum entropy approach. *Fish Oceanogr* 27, 571–586. doi: 10.1111/fog.12279

Watling, L., and Norse, E. A. (1998). Disturbance of the seabed by mobile fishing gear: A comparison with forest clear-cutting. *Conserv. Biol.* 12, 1180–1197. doi: 10.1046/j.1523-1739.1998.0120061180.x

Whitehead, P. J. P. (1985). "Clupeoid fishes of the world. An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, shads, anchovies and wolf-herrings," in *Part 1 chirocentridae, clupeidae and pristigasteridae*. (FAO fisheries synopsis).

Yang, J., Rahardja, S., and Fränti, P. (2019). "Outlier detection: How to threshold outlier scores?," in *ACM International Conference Proceeding Series* (Sanya, China: Association for Computing Machinery). doi: 10.1145/3371425.3371427

Zeng, Y., Low, B. W., and Yeo, D. C. J. (2016). Novel methods to select environmental variables in MaxEnt: A case study using invasive crayfish. *Ecol. Modell* 341, 5–13. doi: 10.1016/j.ecolmodel.2016.09.019

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., et al. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* 10, 744–751. doi: 10.1111/2041-210X.13152

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychol. Methods* 12, 399–413. doi: 10.1037/1082-989X.12.4.399

Zuur, A. F., Ieno, E. N., and Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* 1, 3–14. doi: 10.1111/j.2041-210x.2009.00001.x

TYPE Original Research
PUBLISHED 20 November 2024
DOI 10.3389/fmars.2024.1493275

OPEN ACCESS

EDITED BY
Yu Jiang,
Jilin University, China

REVIEWED BY
Alberto Alberello,
University of East Anglia, United Kingdom
Gui Gao,
Southwest Jiaotong University, China

*CORRESPONDENCE
Xiaolong Zhao
✉ 210321050349@stu.haust.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 08 September 2024
ACCEPTED 28 October 2024
PUBLISHED 20 November 2024

CITATION
Jiang H, Zhao X, Zhang Z and Ji J (2024) Research on adaptive dimming management methods for intelligent lighting systems in port traffic based on ocean weather perception. *Front. Mar. Sci.* 11:1493275. doi: 10.3389/fmars.2024.1493275

Research on adaptive dimming management methods for intelligent lighting systems in port traffic based on ocean weather perception

Haoyu Jiang[1†], Xiaolong Zhao[2*†], Zeguo Zhang[1] and Jiacheng Ji[3]

[1]Naval Architecture and Shipping College, Guangdong Ocean University, Zhanjiang, China, [2]School of Information Engineering, Henan University of Science and Technology, Luoyang, China, [3]Department of Electrical Engineering, Shanghai Maritime University, Shanghai, China

In challenging visibility conditions, the reliability of existing port lighting systems is significantly affected by abrupt changes in environmental factors (primarily stemming from ocean weather). This study proposes a cloud-edge collaborative dimming model that integrates a combined filter, enabling dynamic adaptation to these weather variations to ensure the stability of the lighting system. Additionally, the application of edge computing not only alleviates computational pressure but also facilitates the model's ability to achieve effective regional adaptive dimming in accordance with environmental regulations. Experimental results indicate that this method is suitable for scenarios with unknown mutations under extreme conditions, providing a more reliable and intelligent solution for port lighting systems within the Internet of Things (IoT) framework.

KEYWORDS

extreme weather, port streetlights, Internet of Things (IoT), combined filtering, adaptive dimming

1 Introduction

In recent years, global climate change has led to an increasing probability of extreme weather events (Clarke et al., 2022). Due to the complexity and variability of weather in coastal ports, various challenging visibility conditions (such as haze, overcast skies, and heavy rain) frequently occur, resulting in economic losses and casualties in several coastal cities and ports (Yang et al., 2021). Geographical factors contribute to the significant impact of extreme weather on coastal ports (Izaguirre et al., 2021). These weather conditions can rapidly alter the lighting environment of the port, causing dramatic fluctuations in natural light intensity and visibility, which directly impacts the safety of vehicle movements and cargo handling operations (Al-Behadili et al., 2023). Therefore, under these challenging visibility conditions, effective and reliable port lighting systems are crucial for ensuring traffic and personnel safety (Galbraith and Grosjean, 2019).

To ensure the safety and visual comfort of port personnel, lighting systems are among the highest energy-consuming components in port operations, sometimes accounting for over 70% of the port's total energy consumption (Sifakis et al., 2021). This has prompted many researchers to focus on the integration of the Internet of Things (IoT, refers to the interconnection of various physical devices via the internet, allowing them to communicate and exchange data with each other) with port lighting systems, exploring methods such as adjusting lighting schedules (Sun, 2019), introducing solar-assisted lighting (Muhamad and Ali, 2018), and optimizing energy management strategies (Prousalidis et al., 2019) to save energy required for lighting. However, despite the important role these lighting systems play in port safety, there is still a problem of insufficient intelligence (Pham, 2023). Under extreme conditions, existing lighting systems often require manual intervention and have long response times, lacking real-time monitoring and fine-tuning of environmental changes (Yau et al., 2020). Meanwhile, with the widespread adoption of intelligent assisted driving, safety hazards for logistics vehicles are becoming increasingly serious under the influence of extreme weather (He et al., 2021). Therefore, it is necessary to conduct further research on the perception capabilities and adaptive regulation capabilities of port lighting systems under extreme conditions.

In the face of extreme weather conditions, the effectiveness of environmental perception and dimming in port lighting systems relies not only on accurate localized weather data but also on overcoming the influence of urban structures on roadway monitoring (Bowden and Heinselman, 2016; Gao et al., 2024; Cao et al., 2024). However, existing methods struggle to meet these requirements. In recent years, although artificial intelligence technologies such as deep neural networks (e.g., MetNet, AI Earth) have gradually been applied to extreme weather forecasting and can achieve minute-level short-term predictions under ideal conditions with a resolution of up to 1 kilometer, these methods still have limitations in model interpretability and input sample quality (Bojesomo et al., 2021). Additionally, AI methods face challenges in integrating heterogeneous data and computational capabilities, making it difficult for existing port lighting systems to meet their computational demands (Zhang and Lu, 2021; Gao et al., 2023c, a, b; Zhang et al., 2024). Therefore, from a technical perspective, it is necessary to introduce a cloud-edge collaborative computing model to address the tracking and dimming issues of port lighting through edge computing methods. This approach not only enables real-time detection of environmental changes at the port but also ensures the precision of dimming adjustments on the edge, thereby improving the overall efficiency and reliability of the system while reducing computational pressure (Saeik et al., 2021).

Since 2015, countries such as China have gradually implemented smart streetlight infrastructure in major cities and published relevant standards (Wang et al., 2019). These standards define smart lighting, video capture, and mobile communication as standard configurations for urban roadways and require the deployment of weather monitoring functions at major roads, bridges, and intersections. The deployment of these functions enables cities to directly perceive weather changes based on edge computing capabilities, determine dimming targets, and achieve tracking and dimming of municipal lighting systems under extreme weather conditions (De Paz et al., 2016). Therefore, this paper will explore the application prospects of smart streetlights in port lighting, focusing on adaptive dimming management methods based on the perception of ocean weather conditions to enhance the intelligence level of port lighting systems and ensure safe operations. Specifically, it aims to clarify how to utilize the information collected from smart streetlight hardware systems, in conjunction with the physical state of extreme weather (primarily the impacts brought by ocean conditions) in the port environment, to improve combined tracking filters and achieve precise dimming of port lighting.

2 Problem description

2.1 Smart lighting system description

The system composition of intelligent street lighting is illustrated in Figure 1. It primarily consists of six components: the lighting module, video monitoring module, power supply module, environmental monitoring module, communication module, and information display module. The lighting module can be configured with either a light sensor or a photovoltaic panel. The brightness of the light source is regulated through the lighting controller. Currently, individual lamp control is primarily achieved through the DC intelligent control power supply, while centralized control of an entire street is accomplished by the centralized controller in the power distribution cabinet. The edge controller in the power distribution cabinet possesses enhanced computational capabilities, enabling smooth processing of video and image streams. It also offers a wide range of communication interfaces, such as Ethernet, RS485/232, CAN, HDMI, LVDS, USB2.0, line out, etc., facilitating the integration of diverse data sources and expanding various analytical functionalities. The video monitoring and environmental monitoring modules serve as the information foundation for intelligent light poles. Equipped with various sensor devices, the cameras primarily serve the recognition and tracking of specific targets for urban security, while also providing real-time monitoring of traffic flow and pedestrian movement. The environmental sensors encompass a variety of types, capable of measuring parameters such as temperature, humidity, particle concentration, wind speed, wind direction, air pressure, noise, and more. The information obtained or received by the aforementioned modules, including weather and traffic data, can be disseminated to pedestrians through LED display screens and speakers mounted on the light poles. Simultaneously, the communication module transmits this information to the big data cloud platform of the lighting system. This transmission trend is gradually shifting towards the development of 5G, facilitating distributed connections while serving as small base stations to provide external support for WIFI signals.

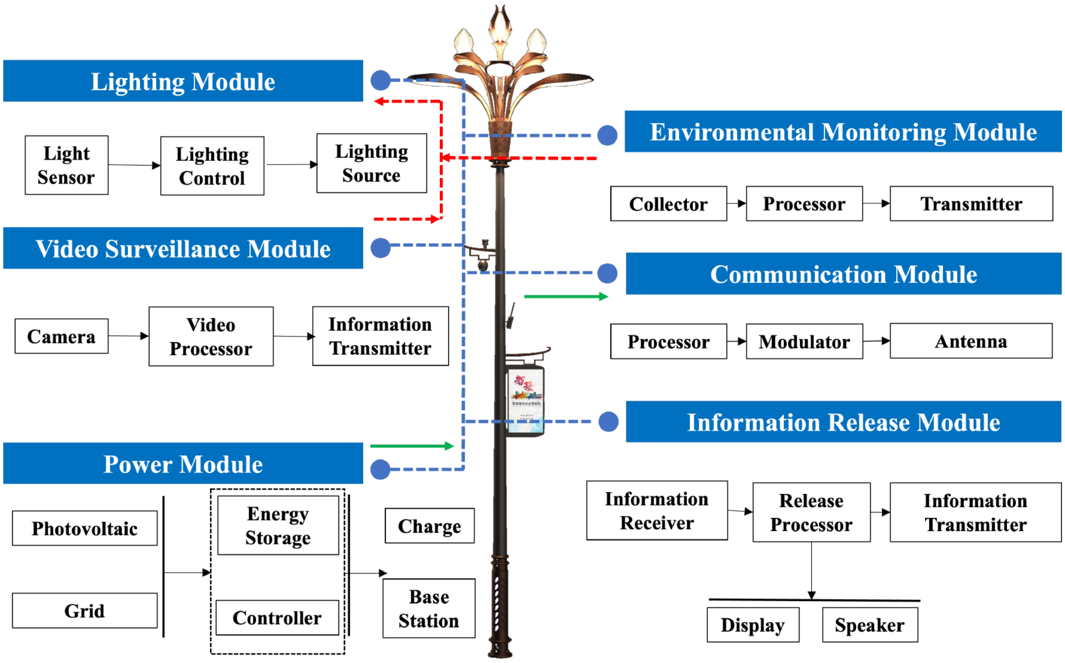


FIGURE 1
Description of smart streetlights system architecture.

Currently, the vast amount of data generated by intelligent light poles is primarily transmitted to dedicated management and operation platforms through optical fiber communication, as illustrated in Figure 2. Multiple communication protocols, including MoDBUS, DMX512, MQTT, GPRS/LTE, RPC, and HTTPS, are employed to enable application interactions at the Internet layer. To support the IoT information system implemented on light poles, the intelligent light pole system requires collaborative power supply from photovoltaic renewable energy and the grid. It is equipped with energy storage and control systems to provide energy assurance for electric vehicle charging and 5G services. Therefore, an intelligent street light, as indicated by the green arrows in Figure 1, can be regarded as a process that starts from the power supply module, delivers data to various information modules, and then transmits it externally through the communication (closed-loop) or information display (open-loop) modules. Tracking and dimming for extreme weather conditions deviate from the fixed path and enable information flow equivalent to the red arrows in Figure 2.

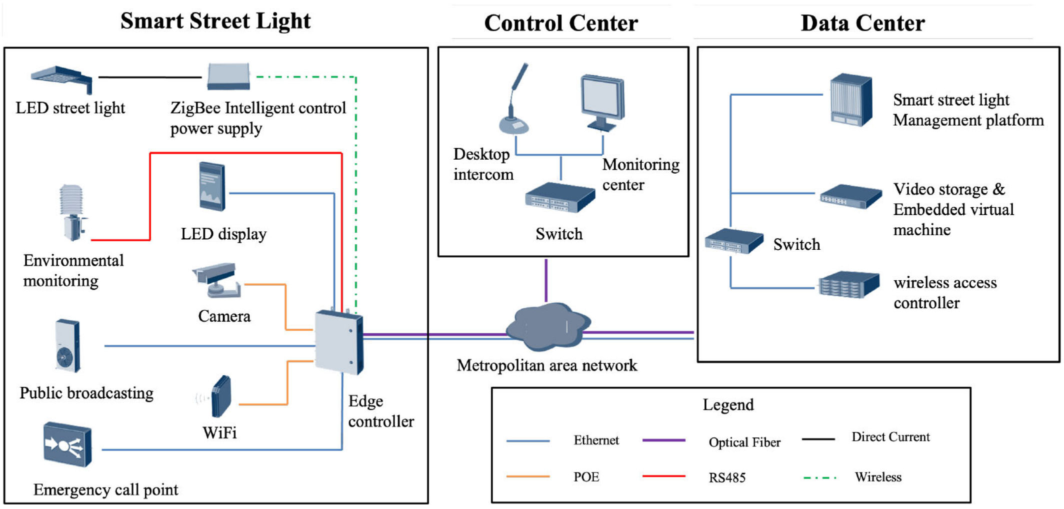
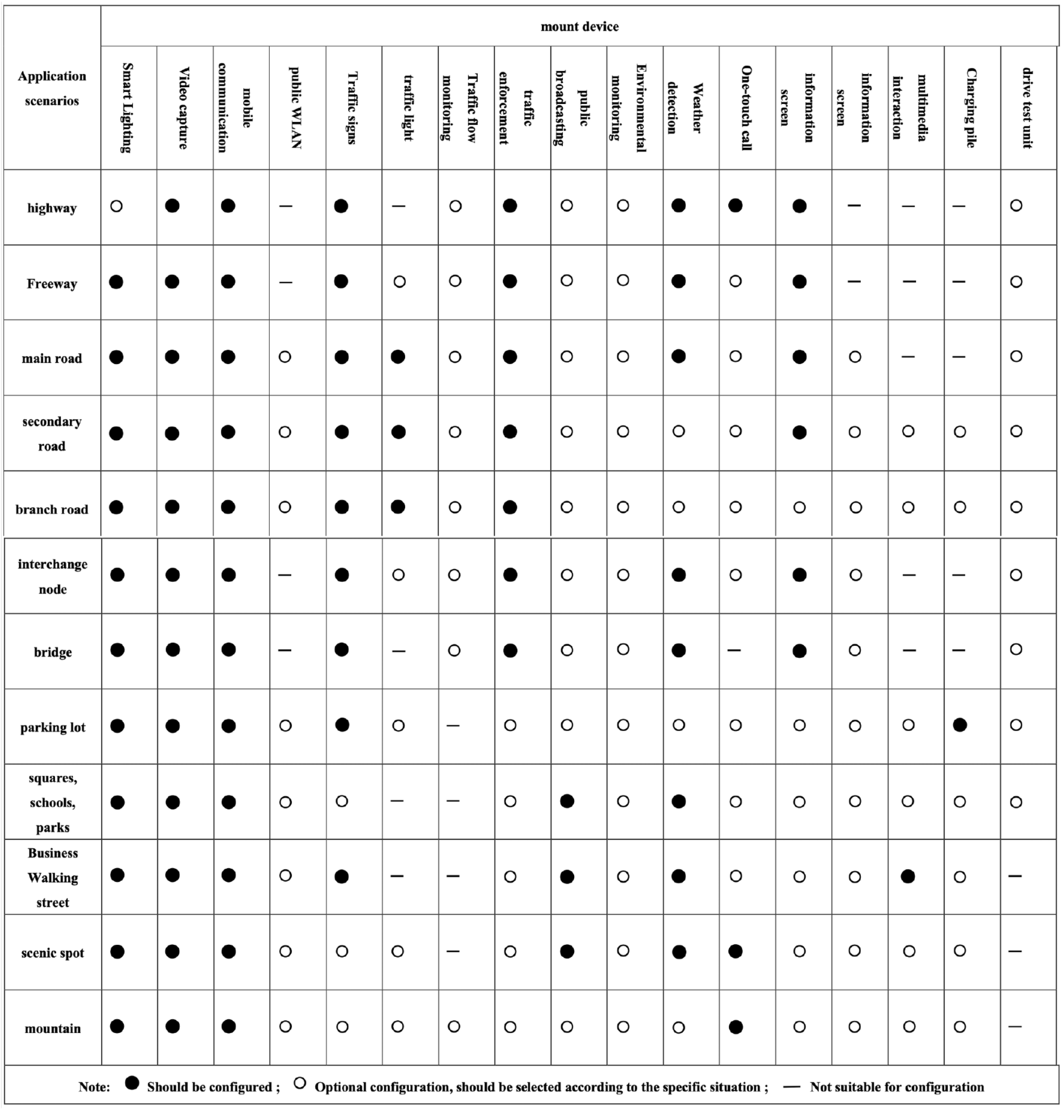


FIGURE 2
Smart streetlights control system.

Due to the lack of unified management entities for the operation of intelligent streetlights, different operators tend to emphasize different aspects based on their respective business characteristics. To ensure the general applicability of the research methodology (to address the dimming requirements of smart streetlights under various conditions in the port, thereby achieving a level of normal operation and safety assurance), it is necessary to consider the hardware configuration standards of smart streetlights. Figure 3 presents a reference specification indicating the configuration standards. It can be observed that smart lighting, video capture, and mobile communication are fundamental and commonly found configurations of smart streetlights. Additionally, meteorological monitoring is also required in urban road regulations. Therefore, this standard can serve as a hardware constraint reference for algorithm design, ensuring the consistency and compatibility of the proposed methods.

2.2 Description of the dimming problem

The impact of extreme weather in coastal port scenarios on port lighting systems primarily manifests in sudden changes in meteorological conditions such as rain and fog, posing threats to the safety of logistics vehicles and pedestrians. Extreme weather reduces visibility, thereby affecting traffic safety and logistics efficiency. In this context, the smart streetlights system at the port faces photometric issues, with the adjustment target being the luminous flux Φ_v. Given that existing streetlights are generally optimized through lens design, it is assumed that they possess directional uniformity within the specified emission angle (non-uniformity is considered an optimization problem of the luminaire hardware and is not included in the scope of this discussion). Therefore, the adjustment of the luminance L_v with respect to the emission angle Ω can be simplified as a problem of constant light



Application scenarios	mount device																
	Smart Lighting	Video capture	mobile communication	public WLAN	Traffic signs	traffic light	Traffic flow monitoring	traffic enforcement	public broadcasting	Environmental monitoring	Weather detection	One-touch call	information screen	information screen	multimedia interaction	Charging pile	drive test unit
highway	○	●	●	—	●	—	○	●	○	○	●	●	●	—	—	—	○
Freeway	●	●	●	—	●	○	○	●	○	○	●	○	●	—	—	—	○
main road	●	●	●	○	●	●	○	●	○	○	●	○	●	○	—	—	○
secondary road	●	●	●	○	●	●	○	●	○	○	○	○	●	○	○	○	○
branch road	●	●	●	○	●	●	○	●	○	○	○	○	○	○	○	○	○
interchange node	●	●	●	—	●	○	○	●	○	○	●	○	●	○	—	—	○
bridge	●	●	●	—	●	—	○	●	○	○	●	—	●	○	—	—	○
parking lot	●	●	●	○	●	○	—	○	○	○	○	○	○	○	○	●	○
squares, schools, parks	●	●	●	○	○	—	—	○	●	○	●	○	○	○	○	○	○
Business Walking street	●	●	●	○	●	—	—	○	●	○	●	○	○	○	●	○	—
scenic spot	●	●	●	○	○	○	—	○	●	○	●	●	○	○	○	○	—
mountain	●	●	●	○	○	○	○	○	○	○	○	●	○	○	○	○	—

Note: ● Should be configured ; ○ Optional configuration, should be selected according to the specific situation ; — Not suitable for configuration

FIGURE 3
Installation scenarios and configuration of smart streetlights (Zhou, 2018).

emission degree M_v. To further optimize the target illuminance E_v, two ideal assumptions are made: 1) assuming that the port road environment is fully diffusive, the illuminance can be considered uniformly consistent within a certain range of streets; 2) assuming that there is little difference between the light escaping at the boundaries and entering, or that the total amount of escaping light is small, thus considering the dynamic energy balance within the entire study space. Although these two ideal conditions may deviate to some extent in practical construction, they can be effectively approximated through engineering optimization of the lighting system. Therefore, under a fixed area conversion coefficient $K_{\mathbb{A}}$, E_v can be expressed as:

$$K_{\mathbb{A}}E_v = M_v = \int L_v d\Omega = d\Phi_v/d\mathbb{A} \tag{1}$$

where $\mathbb{A}$ is the illuminated area, measured in square meters.

Introducing extreme weather factors, it is considered that the illuminance from the environment undergoes attenuation or fluctuation, and environmental factors weaken the inherent illuminance of the lighting system. It is believed that this attenuation or fluctuation exhibits a significant dynamic range, during which both the cone and rod cells of the human eye are involved. The spectral luminous efficiency function for mesopic vision is denoted as V_m, and it is defined using the MES2-system model as V_m (Gao et al., 2018):

$$V_m(\lambda,p) = [pV(\lambda) + (1-p)v(\lambda)]/\mathbb{M}(p) \tag{2}$$

where λ represents the wavelength of light, p denotes a coefficient, V (λ) corresponds to the luminous efficiency function under photopic vision, $v(\lambda)$ represents the luminous efficiency function under scotopic vision, and $\mathbb{M}$ (p) stands for a normalization function that is influenced by photopic luminance and determined by visual adaptation conditions. Therefore, the appropriate mesopic luminance, L_m, can be defined as follows:

$$L_m = \mathbb{K}_m\int_{-\infty}^{\infty} E(\lambda)V_m(\lambda,p)d\lambda, \tag{3}$$

under the given light source, where $E(\lambda)$ represents the spectral radiance distribution of the light source, and $\mathbb{K}_m$ is the maximum spectral luminous efficiency, the luminance standard for mesopic vision can be obtained by measuring the standard photopic luminance and the standard scotopic luminance of the given light source. Consequently, L_m can be used as a reference for adjusting the system dimming based on mesopic vision.

Based on Equation 1 and its validity conditions, it is evident that the introduction of extreme weather conditions disrupts the energy balance of the existing lighting system, necessitating a reevaluation of the regulation behavior. However, the specific manner in which this balance is disrupted varies depending on the type of weather. For instance, cloudy conditions primarily lead to rapid changes (reductions) in natural illuminance, which can be addressed by directly adjusting the brightness or color (i.e., wavelength) of the light source based on the corresponding visual state. On the other hand, degradation of effective illuminance caused by rain (liquid droplets), haze (liquid-solid aerosols), or dust storms (solid particles) occurs due to the scattering and absorption of light by particulate matter, resulting in attenuation of light intensity after propagation through the medium. The extent of this attenuation is dependent on the size and concentration of the particles and can be described by the Lambert-Beer law (Swinehart, 1962):

$$I = I_0\exp(-\tau l), \tag{4}$$

where I_0 represents the initial intensity of light, and I denotes the intensity of light after extinction, which is equivalent to the integral of the corresponding luminance over the spherical degree. Here, l represents the optical path length, and τ signifies the turbidity of the medium. For a polydisperse particle system consisting of n particles with an average diameter of ϖ, τ can be quantitatively expressed as described by (Gledhill, 1962):

$$\tau = \pi/4\int_a^b n(\varpi)\varpi^2 k_{ext}(\lambda,\varpi,m)d\varpi, \tag{5}$$

where a and b represent the lower and upper limits, respectively, of the particle size distribution. The parameter m corresponds to the relative refractive index of the particles with respect to the surrounding medium, while k_{ext} denotes the extinction coefficient (Bruce et al., 1980). When both absorption k_{abs} and scattering k_{sca} processes occur simultaneously:

$$k_{ext} = k_{abs} + k_{sca} = 2/a^2\sum_{l=0}^{\infty}(2l+1)(|a_l| + |b_l|), \tag{6}$$

where

$$a_l = (\varphi_l(a)\varphi_l(ma) - m\varphi_l(a)\varphi_l(ma))/(\zeta_l(a)\varphi_l(ma) - m\zeta_l(a)\varphi_l(ma)), \tag{7}$$

$$b_l = (m\varphi_l(a)\varphi_l(ma) - \varphi_l(a)\varphi_l(ma))/(m\zeta_l(a)\varphi_l(ma) - \zeta_l(a)\varphi_l(ma)), \tag{8}$$

where

$$\varphi_l = \sqrt{\pi a/2}J_{1+1/2}(a), \tag{9}$$

$$\zeta_l = \sqrt{\pi a/2}H_{1+1/2}(a). \tag{10}$$

The functions $J_{1+1/2}(a)$ and $H_{1+1/2}(a)$ represent the Bessel functions of half-integer order and the Hankel functions of the first kind, respectively, both of which are series functions. It can be assumed that the absorption of particulate matter in general weather conditions is negligible, that is, the imaginary part of the complex refractive index m is zero. However, this calculation requires a substantial number of computational resources (the speed of convergence is directly proportional to the computational resources invested), which consequently increases the energy required for the entire port lighting system. Therefore, when designing tracking and dimming algorithms for extreme weather variations, it is advisable to avoid direct computation of the extinction coefficient or make necessary simplifications.

Based on the foundational discussions above, the regulation problem of the port illumination system in the face of extreme oceanic conditions can be transformed into a strong tracking problem by leveraging existing smart streetlight hardware standards. Guided by this approach, this paper proposes a method for regulating the port illumination system based on a combined tracking filter, consisting of three main components: the physical acquisition layer, the edge processing layer, and the platform processing layer. The physical acquisition layer is primarily responsible for providing observation data and system structural information. The edge processing layer is focused on tracking E_k^x (the actual value of the streetlight illumination state during the k-th tracking dimming) and adjusting the vector I_v^q (the dimming matrix received by the q streetlights at the edge). The platform processing layer is responsible for receiving the dimming matrix I_v, updating, and issuing macroscopic decisions T, as illustrated in Figure 4, outlining the basic framework. In a nutshell, the main contributions of this paper are outlined as follows:

- In response to the impact of extreme oceanic weather on port road illuminance, a cloud-edge collaborative dimming model is proposed, incorporating the hardware system of smart streetlights. The dimming model's cloud control risk items and decision items are expanded and described in detail, while optimization objectives for the target matrix are provided.
- To address the real-time dynamic changes of the dimming matrix, this paper presents state estimation and observation methods under static conditions. Specifically, for air turbidity, a calculation method based on video monitoring devices and neighboring streetlights is proposed, circumventing the direct computation of the extinction coefficient.
- A dynamic system model for discretized illuminance based on Kalman filtering theory is presented to address the dynamic adjustment problem of illuminance in response to time-varying solar input and air turbidity. The uncertainties and nonlinearity of the system are decoupled from the state vector, ensuring that the main iterative process achieves a convergence rate suitable for edge computing capabilities.
- Given the challenge of *a priori* judgment of state mutation resulting from the aforementioned operations, and considering the distinctions between the two strong tracking filtering methods, STF and STAKF, a strategy that combines the strengths of both approaches is proposed. Additionally, an optimized step size is adopted to account for the variability in tracking.

3 Cloud-edge collaborative dimming model

3.1 Model architecture

According to Section 2.1, it is evident that there are multiple approaches for controlling the luminous intensity of smart

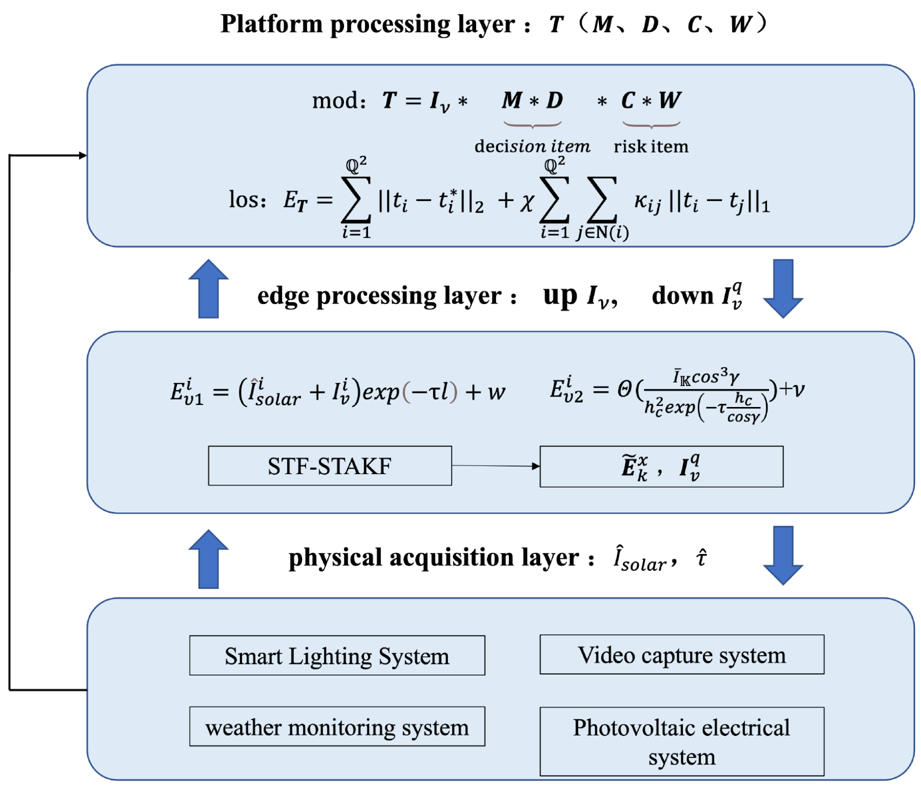

Platform processing layer： T（M、D、C、W）
mod： $T = I_v * \underbrace{M * D}_{\text{decision item}} * \underbrace{C * W}_{\text{risk item}}$
los： $E_T = \sum_{i=1}^{\mathbb{Q}^2} ||t_i - t_i^*||_2 + \chi \sum_{i=1}^{\mathbb{Q}^2} \sum_{j \in N(i)} \kappa_{ij} ||t_i - t_j||_1$
edge processing layer： up I_v, down I_v^q
$E_{v1}^i = (\hat{I}_{solar}^i + I_v^i) exp(-\tau l) + w$
$E_{v2}^i = \Theta(\frac{\bar{I}_{\mathbb{K}} cos^3\gamma}{h_c^2 exp(-\tau \frac{h_c}{cos\gamma})}) + \nu$
STF-STAKF
$\tilde{E}_k^x$, I_v^q
physical acquisition layer： $\hat{I}_{solar}$, $\hat{\tau}$
Smart Lighting System
Video capture system
weather monitoring system
Photovoltaic electrical system

FIGURE 4
Overview of research ideas in this paper.

streetlights. In this subsection, a cloud-edge collaborative dimming strategy model will be proposed, where the dimming decisions of all smart streetlights are based on the cloud-edge collaborative streetlight network model presented in this subsection, and the smart streetlights are interrelated while operating. The decision of whether to adopt a centralized or fixed strategy, which is cost-effective, requires prior decision-making at the cloud level. Based on this decision, concrete collaborative strategies can be formulated. The decision model can be expressed in the following form:

$$T = I_v * (M * D) * (C * W)\,. \tag{11}$$

Let W denote the risk matrix primarily based on meteorological observations. For convenience, let's assume that the streetlights scattered within the selected urban area for dimming can be projected into a square matrix of size $\mathbb{Q} \times \mathbb{Q}$ through an affine transformation. Here, $\mathbb{B}$ represents the smallest scale that the existing forecasting system (mainly based on meteorological satellites and radars) can discern in the projection onto W. The matrix W can be expressed as follows:

$$W = \begin{bmatrix} B_{11} & \cdots & B_{1\tilde{\mathbb{Q}}} \\ \vdots & \ddots & \vdots \\ B_{\tilde{\mathbb{Q}}1} & \cdots & B_{\tilde{\mathbb{Q}}\tilde{\mathbb{Q}}} \end{bmatrix}_{\mathbb{Q}\times\mathbb{Q}} \tag{12}$$

where $\tilde{\mathbb{Q}} \in \mathbb{N}^+$, the matrix B_{ii} represents a submatrix of size $\mathbb{B} \times \mathbb{B}$, which can be interpreted as a city block within the port area. Due to variations in port planning and infrastructure, different blocks may exhibit varying levels of response to extreme marine meteorological risks.

Therefore, based on meteorological forecasts of disaster types and severity from marine meteorological monitoring, the cloud platform can leverage historical data and the GIS+BIM system of the smart city to further refine and adjust B_{ii}, forming a transition matrix C:

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1\mathbb{Q}} \\ \vdots & \ddots & \vdots \\ C_{\mathbb{Q}1} & \cdots & C_{\mathbb{Q}\mathbb{Q}} \end{bmatrix}_{\mathbb{Q}\ \times\mathbb{Q}} \tag{13}$$

where, $C_{ii} \in \mathbb{R}^+$ is the adjustment factor. C and W together constitute the risk term in the collaborative dimming model T. Their purpose is to assign varying degrees of dimming based on evaluations of the individual impacts of extreme marine meteorological conditions on port streetlights.

The decision matrix D primarily serves to accommodate constraints from the power system and other aspects, including considerations of hardware controllability, grid dispatch, and economic factors. It can be represented as a binary matrix (1-0 matrix). If optimization operations on the D matrix are required in subsequent model applications, a sigmoid transformation can be applied to the matrix:

$$D = Sigmoid \begin{bmatrix} X_{11} & \cdots & X_{1\mathbb{Q}} \\ \vdots & \ddots & \vdots \\ X_{\mathbb{Q}1} & \cdots & X_{\mathbb{Q}\mathbb{Q}} \end{bmatrix}_{\mathbb{Q}\ \times\mathbb{Q}} \tag{14}$$

where $X \in \{0,1\}$.

In addition, the influence of extreme marine weather types needs to be considered. As discussed in Section 2.2, existing LED port lighting systems can adjust the color temperature based on marine meteorological conditions and environmental changes. Different color temperatures correspond to different S/P ratios, which in turn affect the intermediate visual brightness L_m. Since $I_v = \int L_v d\mathbb{A} \cos\theta$, the differences in brightness adjustment targets will impact the decision-making process for light intensity adjustment. To ensure a unified behavioral scale for the dimming matrix I_v in the model, it is necessary to normalize the influence in this aspect into a pattern matrix M, where the matrix elements $M_{ii} \in (0,1]$. Both W, C, D, and M can be determined based on the existing information and instructions from the cloud-based control system of smart streetlights. The matrix I_v needs to reflect the dynamic changes in extreme marine meteorological conditions on the edge side, thus requiring the adoption of a strong tracking algorithm combined with relevant sensor data for control. The evaluation target E_T of the entire control can be written as the target matrix T norm regularization form:

$$E_T = \sum_{i=1}^{\mathbb{Q}^2} \|t_i - t_i^*\|_2 + x \sum_{i=1}^{\mathbb{Q}^2} \sum_{j\in N(i)} k_{ij} \left\|t_i - t_j\right\|_1 \tag{15}$$

In the above equation, t_i represents the elements in T arranged according to certain geographic rules. We define the calibration matrix T^* as the reference values for evaluating T, where t_i^* corresponds to the elements in T^* that correspond to t_i. The determination of T^* is carried out by specialized instruments carried by engineering vehicles during road maintenance operations under specific conditions. It is based on standards (Jaskowski et al., 2022) that are related to road types, traffic flow, road morphology, and luminaire settings. Ideally, the dimming target T should closely resemble the standards and measured values in T^*. Hence, the evaluation objective E_T includes the 2 norms $\|\cdot\|_2$ of both T and T^*.

However, due to uncertainties in the model, standards, and measurement processes, including inaccuracies and imprecisions, as well as inherent biases in the control system transfer function, overfitting tendencies may arise when characterizing T with respect to T^*. To limit local flatness and encourage proximity, a sparse 1 norm $\|\cdot\|_1$ is introduced in the regularization term. This ensures that adjacent light intensities do not exhibit sudden changes and are as close as possible. χ represents a tunable hyperparameter of the evaluation model, which controls the tendency for proximity and can be freely set based on preferences. $N(i)$ denotes the local neighborhood of i, determined by the field of view of the smart streetlight's video surveillance module (as discussed in Section 2.2). t_j represents the light intensity of the streetlights within the field of view. κ_{ij} denotes the affinity coefficient, and its calculation method can be expressed as (Li et al., 2020):

$$K_{ij} = \exp(-\left\|t_i - t_j\right\|_2/(\sigma_1^2))\exp(-(\max(ST_i,\ ST_j)))/(\sigma_2^2)) \tag{16}$$

where the constants σ_1 and σ_2 are predefined constants used to control the model's attention to the differences in light intensity and structure. ST_i and ST_j represent the multi-scale structures based on *a priori* weighted strategies at points i and j, respectively. Taking ST_i as an example:

$$ST_i = \max\left(\sum_{t_j \in \Omega(t_i)} \left|\left(\sum_{t_j \in \Omega(t_i)} G_p(t_i, t_j) \nabla T(t_j)\right)\right| / (G_p(t_i, t_j))\right) \tag{17}$$

here, $G_\rho = \exp(-||t_i - t_j||^2/2\sigma^2)$ denotes the two-dimensional Gaussian kernel with multi-scale parameter σ, where $\sigma \in \{1,2,3\}$.

In summary, the main objective of evaluating and optimizing T is to adjust the decision and risk terms of the entire model. The parameters or strategies in this part are relatively fixed and can be allowed to be completed offline with a delay. It can be observed that the entire model, based on a large-scale data-driven smart streetlight operation and management cloud platform, is executable. Therefore, the key focus of this research lies in utilizing the edge hardware capabilities of smart streetlights to achieve tracking and adjustment of I_v.

3.2 Dimming matrix

As indicated in Section 3.1, the key challenge of the entire T model lies in handling the real-time dynamic variations of I_v. Following a data-driven approach, the main task in this regard is to establish a dataset comprising measurements from smart streetlight solar irradiance sensors, ground illuminance, and corresponding adjustment values of I_v. The real-time adjustment value of I_v can be directly obtained from the electrical system of the smart streetlight. On one hand, the solar irradiance intensity I_{solar} originating from solar radiation can be acquired or estimated through the environmental-meteorological sensing system integrated into the smart streetlight. On the other hand, in a more general scenario, I_{solar} can also be estimated from the solar panel mounted on the top of the streetlight, which processes the solar power P_{solar}:

$$\hat{I}_{solar} = \mathbb{K}(P_{solar}/\varrho\mathbb{A}_{pv} - I_r) \tag{18}$$

where P_{solar} represents the output power of the photovoltaic array, ϱ denotes the photoelectric conversion efficiency of the photovoltaic cells, and $\mathbb{A}_{pv}$ signifies the total area of the photovoltaic panel. The solar power P_{solar} is determined by the total solar irradiance received on the photovoltaic array, which includes the ground-reflected component multiplied by the spectral efficiency factor $\mathbb{K}$ to account for photometric considerations. Since the smart streetlight's photovoltaic panel is typically installed at the top of the pole and is nearly horizontal, $I_r = 0$ is negligible. The remaining term $I_{solar}/\mathbb{K}$ comprises the direct solar irradiance I_d and the sky-scattered radiation I_s :

$$I_{solar}/\mathbb{K} = I_d + I_s = I_{ba}\left[cos(h)cos(\Delta\varphi)sin(\theta) + sin(h)cos(\theta) + k_{sca}\left(\frac{1 - cos\theta}{2}\right)\right] \tag{19}$$

In the above equation, I_{ba} represents the total solar radiation at the location, primarily determined by factors such as solar declination angle, and is a known function of the date. h denotes the solar altitude angle, $\Delta\varphi$ represents the angle between the solar azimuth angle and the orientation of the photovoltaic array, and θ is the tilt angle of the photovoltaic array. Under non-extreme marine climatic conditions, atmospheric scattering due to particle sizes $\alpha \ll 1$ is mainly accounted for by Rayleigh scattering. In this case, the received solar radiation intensity I_{solar} is primarily composed of the direct solar radiation intensity I_d. Additionally, assuming that the tilt angle θ of the photovoltaic panel at the top of the smart streetlight tends towards zero, the ideal form of I^{*}_{solar}can be expressed primarily in terms of the local solar altitude angle:

$$I^{*}_{solar} \approx \mathbb{K} I_{ba}\, sin(h) \tag{20}$$

Clearly, the height of the streetlight pole can be disregarded compared to the atmospheric height, and the differential unit area can be approximated as a solid angle. Therefore, the illuminance E^i_{v1} of any streetlight i on the road surface below it is expressed as:

$$E^i_v = (\hat{I}^{\,i}_{solar} + I^i_v)\exp(-\tau I) + w \tag{21}$$

In this case, the range of $\hat{I}^{\,i}_{solar}$, is $[0, I(\mathrm{L_m})]$ where $I(L_m)$ represents the upper limit for intermediate vision [which can be defined according to relevant standards Ito et al. (2024)]. w denotes the uncertainty of the real-time state of E_{v1}, and its magnitude is mainly positively correlated with $\hat{I}_{solar}/I^{*}_{solar}$. Since E^i_v has a well-defined reference standard, the estimation of turbidity τ is required to compute the value of $\exp(-\tau l)$ in order to generate the dimming matrix I_v, where l is known as the height of the lamp post.

As discussed in Section 2.2, it is not feasible to estimate τ in real-time solely relying on the environmental monitoring devices at the edge of the smart streetlights. However, the emitted light intensities of adjacent centrally controlled streetlights within any solid angle Ω are known, and their relative positional relationships are determinate. Therefore, an estimation of τ can be achieved by observing nearby streetlights using a video surveillance system installed beneath the streetlight, obtaining the observed values of road surface illuminance E^i_{v2}. Consequently, Z neighboring streetlights within the field of view of the video surveillance equipment are selected to estimate τ, and the i-th estimation result τ_i is given by:

$$\tau_i = \frac{\cos\Omega_i}{d}\ln\frac{I_i}{I^0_i\Omega_i} \tag{22}$$

I_i represents the illumination intensity of the current streetlight, I^0_1 represents the illumination intensity of the nearest streetlight, and d denotes the vertical height between the observation device of the current streetlight and that of the nearest streetlight's light source. Then we have

$$\hat{\tau} = \omega^{\top}\tau\,. \tag{23}$$

In the case of data availability, the adjustment of the dynamic weights of vector ω can be achieved using shallow neural networks such as Extreme Learning Machines (ELM) (Liu et al., 2022) to solve for it. Alternatively, considering that the nearby streetlights have stronger light intensity and therefore a higher signal-to-noise ratio, an exponential weighted average can be employed to assign higher weights to the closest streetlights. Neglecting the influence of road surface materials, it is worth noting that different road surfaces (such as concrete or asphalt) have varying reflectance coefficients. For the sake of convenience in the discussion, it is assumed that the road surface undergoes complete diffuse reflection. The observed brightness of the road surface in the vertical direction below the nearest neighboring streetlight, as captured by the camera, is denoted as $I_{\mathbb{K}}$. To mitigate the impact of outliers, actual image processing can replace individual pixel values with the average value within a pixel region, denoted as $\bar{I}_{\mathbb{K}}$, corresponding to the direction angle γ. Therefore, under the condition of camera height h_c, the following relationship holds:

$$E^i_{v2} = \Theta\left(\frac{\bar{I}_{\mathbb{K}} cos^3\gamma}{h_c^2 \exp\left(-\tau\frac{h_c}{cos\gamma}\right)}\right) + v \tag{24}$$

Θ represents the transfer function, which is dependent on the specific parameter settings of the camera. Therefore, there exists a constant proportionality relationship between E^i_{v2} and $\bar{I}_{\mathbb{K}}$. Furthermore, as discussed in Section 3.1, the overall optimization strategy exhibits smoothness locally due to the constraint imposed by the 1 norm. Consequently, the illuminance of the nearest neighboring road surface captured by the camera can be regarded as an observation of the vertical illuminance of the *i*-th streetlight, with v representing the uncertainty associated with this observation.

Based on the above analysis, static computation of I_v can be achieved. However, extreme weather conditions are typically subject to dynamic changes. Therefore, the introduction of robust tracking filtering methods is necessary to enable adaptive adjustment of I_v.

4 Adaptive adjustment method based on STF-STAKF combination

4.1 Dynamic system model

Considering the variation of I_v with extreme marine weather conditions, I_{solar}, τ, and other parameters are functions of time. The estimates $\hat{I}_{solar}$, $\hat{\tau}$, and so on form time series with a certain interval (time step) Δt. By discretizing E_v according to Equations 21 and 24, and extending it to the illuminance vector E corresponding to q streetlights:

$$E^x_{k+1} = (I + \Delta t_k)E^x_k + U_k + w_k \tag{25}$$

$$E^z_k = HE^x_k + v_k \tag{26}$$

where $E^x_k \in R^{q\times 1}$ is the state vector, $I \in R^{q\times q}$ is the identity matrix, $U_k \in R^{q\times 1}$ is the control vector, and the process noise $w_k \in R^{q\times 1}$ satisfies the Gaussian distribution $N(0, \Delta t_k Q_k \Delta t_k^\top)$; $E^z_k \in R^{q\times 1}$ is the observation vector, $H \in R^{q\times q}$ is the observation matrix, and the observation noise $v_k \in R^{q\times 1}$ satisfies the Gaussian distribution $N(0, \Delta t_k R_k \Delta t_k^\top)$. Define:

$$U_k = \Delta((\hat{I}^q_{solar} + I^q_v)\exp(-\ \hat{\tau}^q l)) \tag{27}$$

$$H = I\theta/(h_c^2 \exp(-\hat{\tau}^q h_c))\,. \tag{28}$$

Referring to the standard linear Kalman filter theory (Shao et al., 2021), the recursive calculation formula can be listed as follows:

$$\hat{E}^x_{k|k-1} = (I + \Delta t_k)\hat{E}^x_{k-1} + U_k \tag{29}$$

$$P_{k|k-1} = (I + \Delta t_k)P_{k-1}(I + \Delta t_k)^\top + \Delta t_k Q_k \Delta t_k \tag{30}$$

$$K_k = P_{k|k-1}H^\top(HP_{k|k-1}H^\top + \Delta t_k R_k \Delta t_k)^{-1} \tag{31}$$

$$\hat{E}^x_k = \hat{E}^x_{k|k-1} + K_k\ (E^z_k - H\hat{E}^x_{k|k-1}) \tag{32}$$

$$P_k = (I - K_k H)P_{k|k-1} \tag{33}$$

where $\hat{E}^x_{k|k-1} \in R^{q\times 1}$ represents the *a priori* estimate of E^x_k, $P_{k|k-1} \in R^{q\times q}$ denotes the *a priori* error covariance, $I \in R^{q\times q}$ is the identity matrix, $K_k \in R^{q\times q}$ represents the Kalman gain, $\hat{E}^x_k \in R^{q\times 1}$ is the posterior estimate of $\hat{E}^x_k$, and $P_k \in R^{q\times q}$ corresponds to the updated error covariance. In the context of adaptive adjustment of the time step, the standard Kalman filter, as a non-closed-loop filter, faces challenges in adapting K_k to sudden changes caused by extreme ocean weather conditions and accumulated errors resulting from limited modeling accuracy. Consequently, there is room for improving the performance in practical light adjustment tracking and response.

To address the aforementioned issues, the algorithm needs to incorporate robust tracking filtering to tackle the challenges posed by inaccurate modeling and sudden environmental state changes. The core idea is to introduce a dynamically changing fading factor to adjust the covariance matrix of the prediction error. A computationally efficient approximation of this approach is given by:

$$\xi_k = \begin{cases} \xi_0,\ \xi_0 \geq 1 \\ 1,\ \ \xi_0 < 1 \end{cases}, \tag{34}$$

Where

$$\xi_0 = tr[N_k]/tr[A_k] \tag{35}$$

where

$$N_k = V_k - H\Delta t_k Q_k \Delta t_k H^\top - \beta\Delta t_k R_k \Delta t_k \tag{36}$$

$$A_k = (I + \Delta t_k)HP_{k-1}H^\top(I + \Delta t_k)^\top \tag{37}$$

In Equation 36, the parameter $\beta \in [1, \infty)$ is a user-defined damping factor that controls the smoothness of the state estimation. It plays a role in adjusting the level of smoothing in the estimated values. V_k represents the innovation covariance matrix (Zhou et al., 1991):

$$V_k = \begin{cases} \Upsilon_1\Upsilon_1^\top, & k = 0 \\ \frac{\rho V_{k-1} + \Upsilon_k\Upsilon_k^\top}{1+\rho}, & k \geq 1 \end{cases}, \tag{38}$$

where $\rho \in (0,1]$ is the forgetting factor, and Υ_k is the innovation sequence:

$$\Upsilon_k = E^z_k - H\hat{E}^x_{k|k-1} \tag{39}$$

If the fading factor ξ_k is applied to the error covariance matrix, the Strong Tracking Filter (STF) method is obtained (Han et al., 2006):

$$P^1_{k|k-1} = \xi_k(I + \Delta t_k)P_{k-1}(I + \Delta t_k)^\top + \Delta t_k Q_k \Delta t_k \tag{40}$$

Under the constraint of the orthogonality principle, adjusting the error covariance matrix is equivalent to a modification of process noise without differentiation. However, by directly applying the damping factor to the process noise, we can obtain the Strong Tracking Adaptive Kalman Filter (STAKF) method with multiple fading factors (Ge et al., 2016):

$$P^2_{k|k-1} = (I + \Delta t_k)P_{k-1}(I + \Delta t_k)^\top + \Delta t_k \Gamma_k Q_k \Delta t_k \tag{41}$$

Where

$$\Gamma_k = \begin{bmatrix} l_{k,1} & 0 & \cdots & 0 \\ 0 & l_{k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{k,q} \end{bmatrix} \tag{42}$$

To ensure the symmetry of $P_{k|k-1}$ when the diagonal elements of Γ_k are not equal, Equation 41 can be written as

$$P^2_{k|k-1} = (I + \Delta t_k)P_{k-1}(I + \Delta t_k)^\top + \Delta t_k \bar{\Gamma}_k Q_k \bar{\Gamma}_k^\top \Delta t_k \tag{43}$$

where $\bar{\Gamma}_k$ is obtained by performing Cholesky decomposition on Γ_k:

$$\Gamma_k = \bar{\Gamma}_k \times \bar{\Gamma}_k^\top \tag{44}$$

$$F_k = (H)^+(V_k - \Delta t_k R_k \Delta t_k - (1 + \Delta t_k)HP_{k-1}H^\top(1 + \Delta t_k)^\top)(H^\top)^+ \tag{45}$$

let F^{ii}_k represent the element in the i-th row and i-th column on the diagonal of F_k, and Q^{ii}_k denote the corresponding element in Q_k. Then, we have:

$$l_{k,i} = F^{ii}_k / Q^{ii}_k \tag{46}$$

As a result, the matrix Γ_k or $\bar{\Gamma}_k$with multiple fading factors can be determined. The difference in this approach is reflected in the tracking performance of transient variables. STF tends to assume the system model is reliable and focuses on modifying the estimation error from the previous time step. On the other hand, STAKF tends to attribute the transient changes to the inaccuracy of the system model, indicating a difference in their underlying processing principles. In this research problem, optical-electric measurement methods are frequently employed, which are susceptible to environmental disturbances. Moreover, the study focuses on extreme oceanic weather conditions where the parameters may undergo sudden changes within a processing interval. Therefore, an effective combination of both tracking filters is required.

4.2 Adaptive adjustment method based on STF-STAKF combination

Due to the time-varying nature of fitting functions such as I_{solar} and τ, which may exhibit non-stationary first and second-order differentials, it is crucial to emphasize the role of the discrete time step Δt_k in order to closely capture the trajectory of extreme oceanic meteorological variations. Moreover, the determination and updating of Δt_k need to be considered. To begin, we define the normalized distance of the error covariance matrix as $\mathbb{D}_k$:

$$\mathbb{D}_k = (P_k|_k{-}1 - P_k)(P_k|_k{-}1 + P_k)^{-1} \tag{47}$$

By introducing $\mathbb{D}_k$, we establish a criterion for adjusting Δt_k, such that as Δt_k approaches zero, $\mathbb{D}_k$ tends to zero. Referring to (Or et al., 2021), we can obtain the minimum value d_k of the diagonal elements of $\mathbb{D}_k$ and define a target threshold d_k. Consequently, the adjustment rule for Δt_k can be formulated as follows:

$$\Delta_{t_{k+1}} = \begin{cases} \Delta t_{k-\varepsilon}, & d_k - d^* > sd^* \\ \Delta t_k, & |d_k - d^*| < sd^* \\ \Delta t_{k+\varepsilon} & d_k - d^* < sd^* \end{cases} \tag{48}$$

the parameter s takes values within the range of 0.1 to 0.2, primarily serving as an auxiliary criterion for convergence determination. The fine-tuning quantity ε is a predefined parameter, and its range is constrained as follows:

$$0 < \frac{\varepsilon}{\Delta t_k} < 2sd^* \tag{49}$$

In order to meet the deployment requirements of the margin, the handling approach for the dynamic system model deliberately avoids nonlinearity in the state transition matrix. This strategy facilitates the real-time adaptability and tracking of various hyperparameters within the model. For states with unclear trends in extreme marine meteorological variations, a conservative adjustment effect is desired. Specifically, the outputs of the two filters under this condition, denoted as $\tilde{E}^x_k$ for Filter 1 and Filter 2, are collectively referred to as Y_1 and Y_2. Consequently, the final output result, denoted as $\tilde{E}^x_k$, is obtained by:

$$\tilde{E}^x_k = \eta_k Y_1 + (I - \eta_k)Y_2 \tag{50}$$

the fusion coefficient matrix $\eta_k \in R^{q\times q}$ is a diagonal matrix. In this example, P_k is also a diagonal matrix of the same size as η_k. As a result, there exists a correspondence between the diagonal elements $P^1_{k,\,i}$ and $P^2_{k,\,i}$ of the a priori error covariance P_k for the two filters and the diagonal elements η^i_k of η_k. By calculating η^i_k (Claser and Nascimento, 2021), the fusion coefficient matrix η_k is obtained:

$$\eta^i_k = \sqrt{P^1_{k,i}P^2_{k,i}} + \frac{(P^1_{k,i} - \sqrt{P^1_{k,i}P^2_{k,i}})(P^2_{k,i} - \sqrt{P^1_{k,i}P^2_{k,i}})}{P^1_{k,i} + P^2_{k,i} - 2\sqrt{P^1_{k,i}P^2_{k,i}}}. \tag{51}$$

To further prevent interference in the combined filtering under extreme oceanic weather conditions, the normalization of η^i_k is performed for $k \geq 2$ as follows:

$$\eta^i_k = (\eta^i_k - min(\eta^i_k))/(max(\eta^i_k) - min(\eta^i_k)). \tag{52}$$

A single filter can be regarded as η_k taking values of 0 or 1. Consequently, the difference between $\tilde{E}^x_k$ and the target illuminance

$\tilde{E}_k^*$ in the coordinate system can be used to set U_{k+1} and obtain I_v^q at time $k + 1$. The collection of I_v^q received by the cloud is arranged according to predefined rules, resulting in the overall dimming matrix I_v. The complete algorithm flow is illustrated in Figure 5.

5 Experiments and analysis

The effectiveness of the STF-STAKF (Strong Tracking Filter - Strong Tracking Adaptive Kalman Filter) approach will be validated through two aspects:

Computer data simulation will be employed to compare the tracking performance of STF, STAKF, and STF-STAKF under scenarios involving abrupt process noise mutations. This analysis aims to verify the effectiveness of STF-STAKF in the presence of process noise mutations. Observational data of port street lighting illuminance, influenced by oceanic meteorological factors, will be utilized to compare the tracking performance of STF, STAKF, and STF-STAKF. This evaluation will further validate the effectiveness of STF-STAKF in real-world scenarios where both process noise and state value mutations occur simultaneously.The experimental Root Mean Square Error (RMSE) formula is:

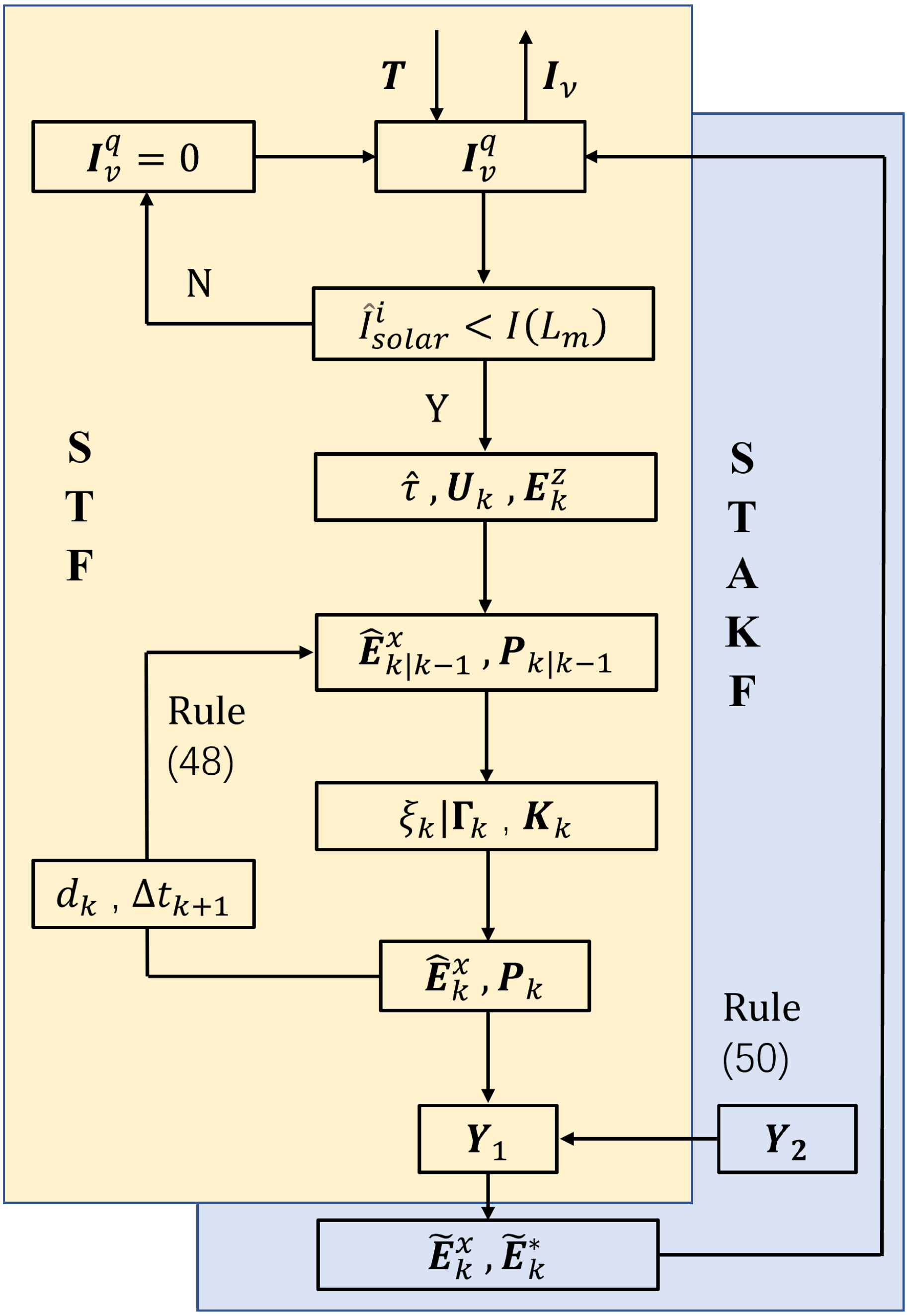


FIGURE 5
Flowchart of algorithm based on STF-STAKF combination.

$$RMSE = \sqrt{\frac{1}{T}\sum_{K=1}^{T}(\hat{E}^{x}_{k|k} - E^{x}_{k})^{2}} \tag{53}$$

The advantage of RMSE in measuring filter performance lies in its ability to effectively quantify the differences between predicted values and true values, providing an intuitive understanding of estimation errors and making it easier to identify situations of poor performance.

5.1 Computer data simulation

This part mainly uses computer numerical simulation examples to verify the effectiveness of combined filtering in the event of sudden changes in process noise. Since the research object of this study is extreme oceanic meteorological conditions, the relevant parameters may change suddenly in a short period of time, and the response scenarios mostly involve short-term parameter mutations and filter tracking. To analyze the tracking effect of the combined tracking filter when the process noise changes abruptly, the parameters and model of the simulation system are set as follows, and the filtering effect is analyzed.

The observation matrix and observed noise covariance are as follows:

$$H = \begin{bmatrix} 9 & 2 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix}, R = \begin{bmatrix} 5 & 8 & 6 \\ 8 & 5 & 6 \\ 6 & 6 & 5 \end{bmatrix} \tag{54}$$

define the process noise covariance as follows

$$Q = \begin{cases} \begin{bmatrix} 4 & 1 & 0 \\ 1 & 8 & 0 \\ 0 & 0 & 1 \end{bmatrix}, & 1 < t \le 15 \\ \begin{bmatrix} 20 & 5 & 0 \\ 5 & 30 & 0 \\ 0 & 0 & 5 \end{bmatrix}, & 15 < t \le 30 \\ \begin{bmatrix} 30 & 10 & 0 \\ 10 & 50 & 0 \\ 0 & 0 & 10 \end{bmatrix}, & 30 < t \le 45 \\ \begin{bmatrix} 50 & 20 & 0 \\ 20 & 80 & 0 \\ 0 & 0 & 20 \end{bmatrix}, & 45 < t \le 50 \end{cases} \tag{55}$$

The initial state vector and the initial state vector covariance are divided into:

$$\hat{E}^{x}_{0|0} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \ P_{0|0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{56}$$

In this subsection, computer simulation experiments are conducted by pre-setting parameters artificially, fixing certain model parameters (H, R and P), while configuring Q in a dynamic form to simulate an environment of abrupt process noise changes. The experimental interval is set to 1 second, and a total of 1000 Monte Carlo simulations are performed.

The following will compare the three filtering methods of the STF algorithm, the STAKF algorithm, and the fusion algorithm discussed in this paper, and analyze the filtering effects of the three.

In Figure 6, the root mean square error (RMSE) of both STAKF and combined filtering is lower than that of STF, indicating good tracking performance for STAKF, with the combined filtering curve being close to or slightly better than STAKF, demonstrating effective filtering.

In cases of sudden changes in process noise, STAKF shows good filtering effects, and combined filtering can closely approach the STAKF curve in real time, often providing better tracking performance. This experiment verifies that under such conditions, the estimation error of STAKF is smaller than that of STF, with the overall root mean square error of combined filtering (refer to Table 1) being better than STAKF, achieving improved tracking effects.

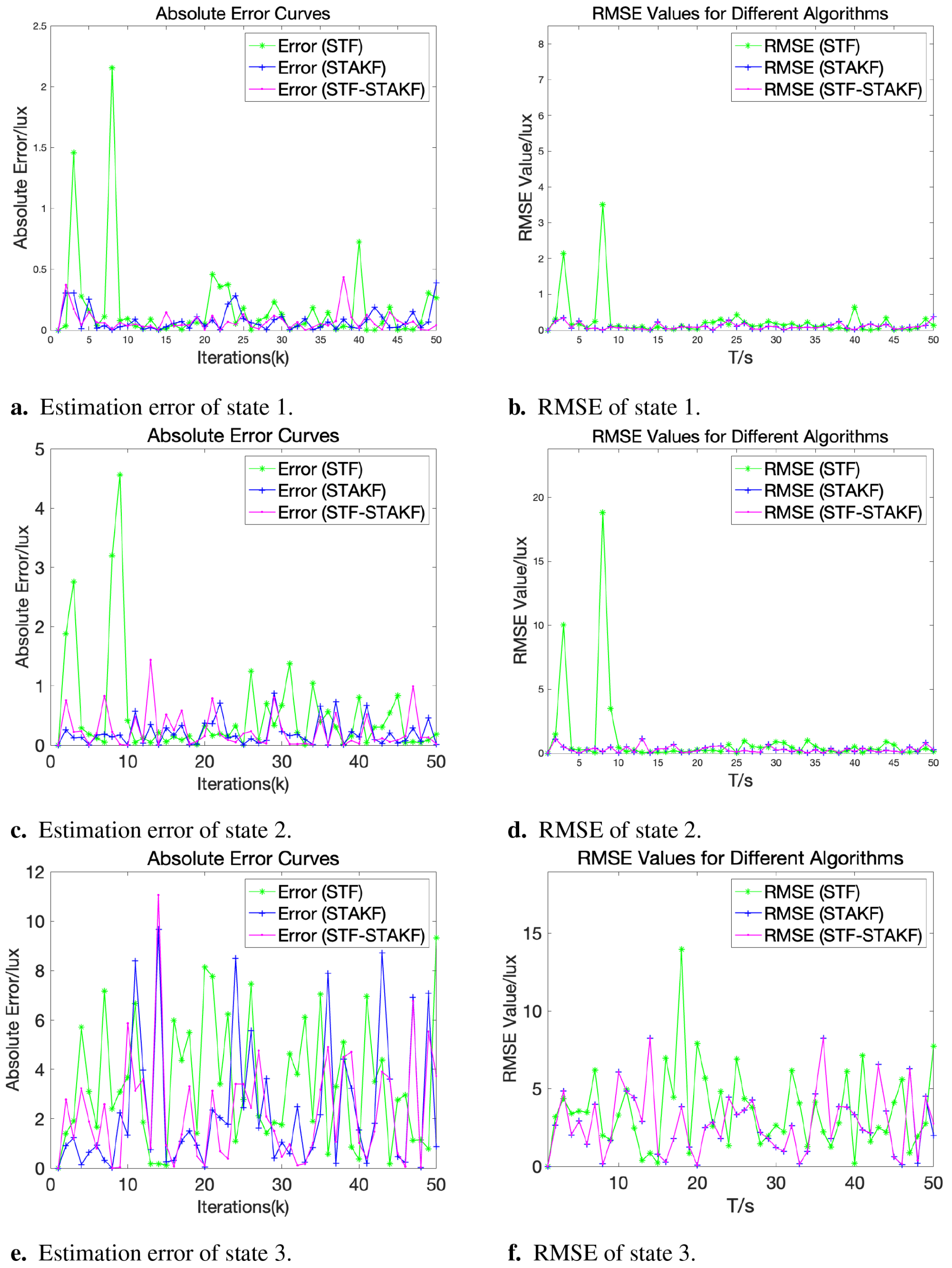
5.2 Experiment on actual observational data of port street lighting under rain and fog conditions

This subsection uses actual observational data of port street lighting under rain and fog conditions to simulate and verify the effectiveness of the combined filter for tracking dimming in extreme oceanic weather conditions. Most of the street lighting and ocean weather data are sourced from the Qiandao Lake Research Institute and Guangdong Ocean University.

5.2.1 Single head streetlight

During extreme oceanic weather conditions such as rain and fog at the port, there may also be sudden changes in the observation data itself, leading to some uncertainty. This section analyzes the tracking effect of the fusion tracking filter when the observed data changes abruptly. Here, the change curve of streetlight dimming illuminance in rainy and foggy weather is selected to experiment with the tracking effect of combined filtering and illuminance. This observational data can be directly collected by the weather perception and video monitoring modules of the smart streetlights in the actual project. The photovoltaic panel on top is calculated simply.

This part of the experimental principle obtains the observational data of sudden changes in rainy and foggy weather by monitoring the illuminance of the nearest single-head streetlight using the monitoring device (refer to Figure 7), which does not affect the verification of the effectiveness of the combined strong tracking filter for tracking dimming in this experiment. In the actual project, the observational data obtained by adjusting the dynamic weights of turbidity after observing multiple streetlights or multi-head streetlights in the current weather (such as ELM), according to



a. Estimation error of state 1.

b. RMSE of state 1.

c. Estimation error of state 2.

d. RMSE of state 2.

e. Estimation error of state 3.

f. RMSE of state 3.

FIGURE 6
Estimation error. **(A)** Estimation error of state 1. **(B)** RMSE of state 1. **(C)** Estimation error of state 2. **(D)** RMSE of state 2. **(E)** Estimation error of state 3. **(F)** RMSE of state 3.

the actual streetlight placement and the position of the observation camera, can be used as the final observational data.

In calm and clear oceanic meteorological conditions, the adjustment of port streetlight brightness is related to factors such as traffic flow and the speed of port vehicles from morning to night. The experiment collected relevant information through the camera to detect the lighting output values when different port vehicle flows and speeds were matched. The lighting output calculation program is developed based on PSO-FNN. As shown in Figure 8, the lighting trend meets the requirements for urban lighting energy

TABLE 1 Mean square error of three algorithms.

RMSE	STF	STAKF	STF-STAKF
$X1_{RMSE}$	0.6085	0.1332	0.1362
$X2_{RMSE}$	3.0825	0.3842	0.3840
$X3_{RMSE}$	4.3652	3.4802	3.4708
MEAN	2.6854	1.3325	1.3324

conservation control. To better reduce the residues between actual and predicted values, the residues were optimized using BLS. The results are shown in Figure 9. This method can effectively fit the actual port streetlights based on small experimental samples.

In rainy and foggy weather, the size of air particles changes, fog suddenly appears and disappears, natural illuminance suddenly increases and decreases, and the extinction coefficient changes abruptly. These factors lead to sudden changes in the observed values. Through the video surveillance system installed under the smart streetlight, the above factors and turbidity τ in rainy and foggy weather were estimated. Figure 10 shows the change curve of the observation data of streetlight illumination and the filtered curve during rainy and foggy conditions over a short period.

As shown in Figure 11, when the observation data is abrupt, the tracking effect of STF is better than that of STAKF, and the combined filtering has a conservative tracking effect, positioned between STF and STAKF at this time. According to Table 2, the overall estimation error of the combined filter is better than that of

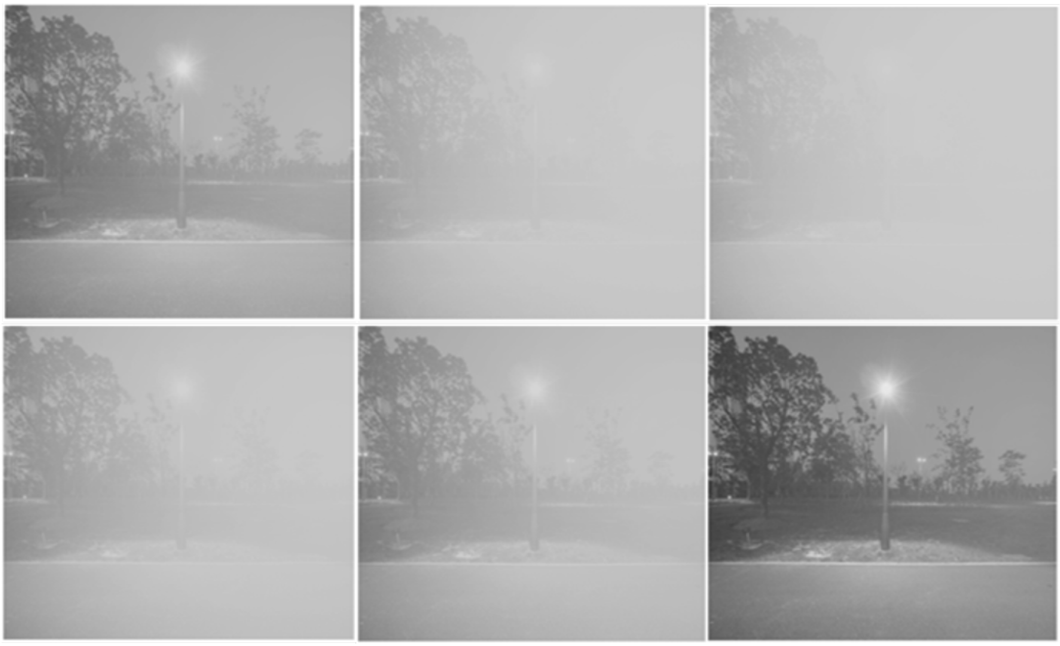


FIGURE 7
Observation and shooting map of smart streetlight in rainy and foggy weather.

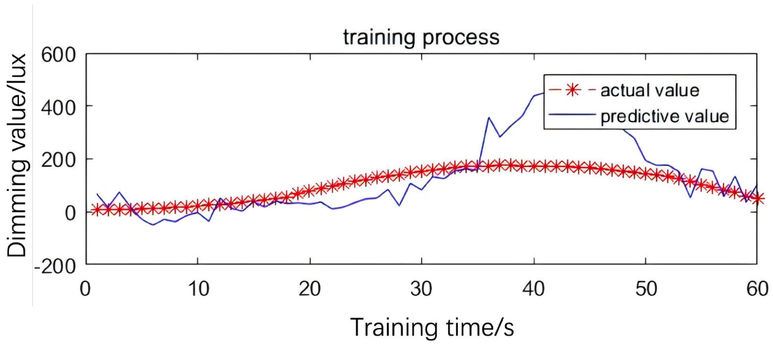


FIGURE 8
Lighting fitting results based on PSO-FNN.

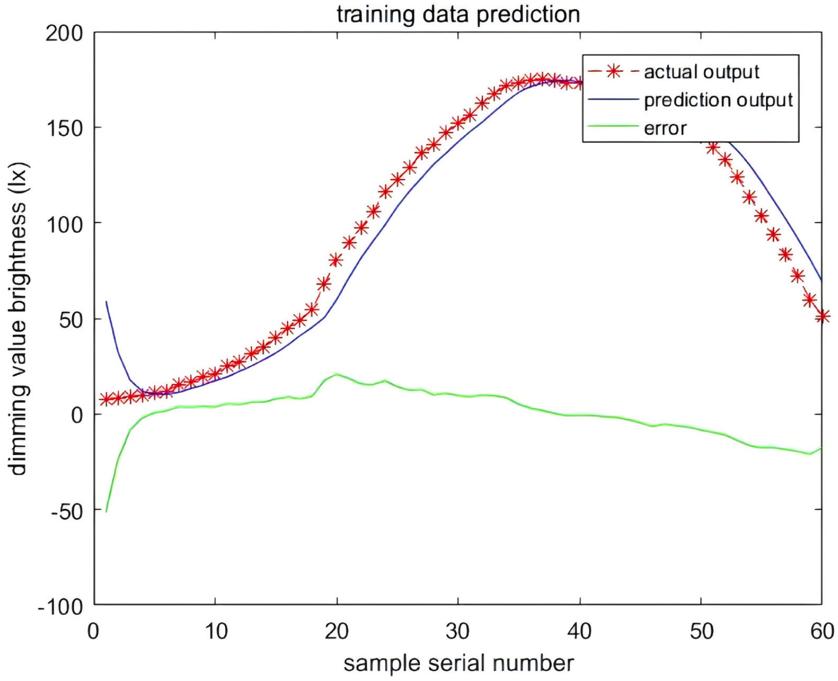


FIGURE 9
Output results after BLS optimization.

STF when observed mutations occur. From this, it can be predicted that when multiple observations undergo mutations, the estimation error of combined filtering will increasingly approach and exceed the current optimal filtering.

This experiment verifies that the estimation error of STF is smaller than that of STAKF in the case of sudden changes in extreme weather observation data, with STF showing better tracking performance at that time. The combined filtering curve

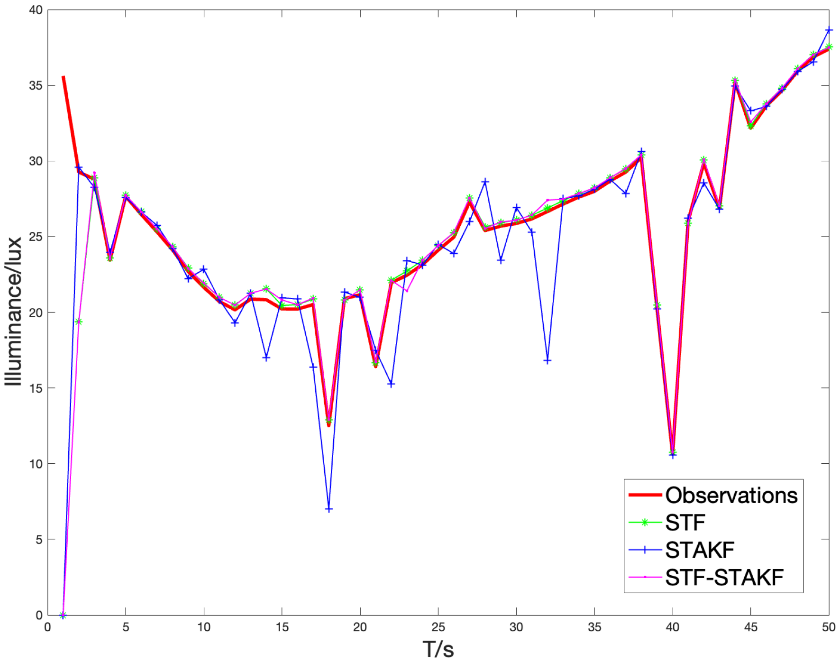
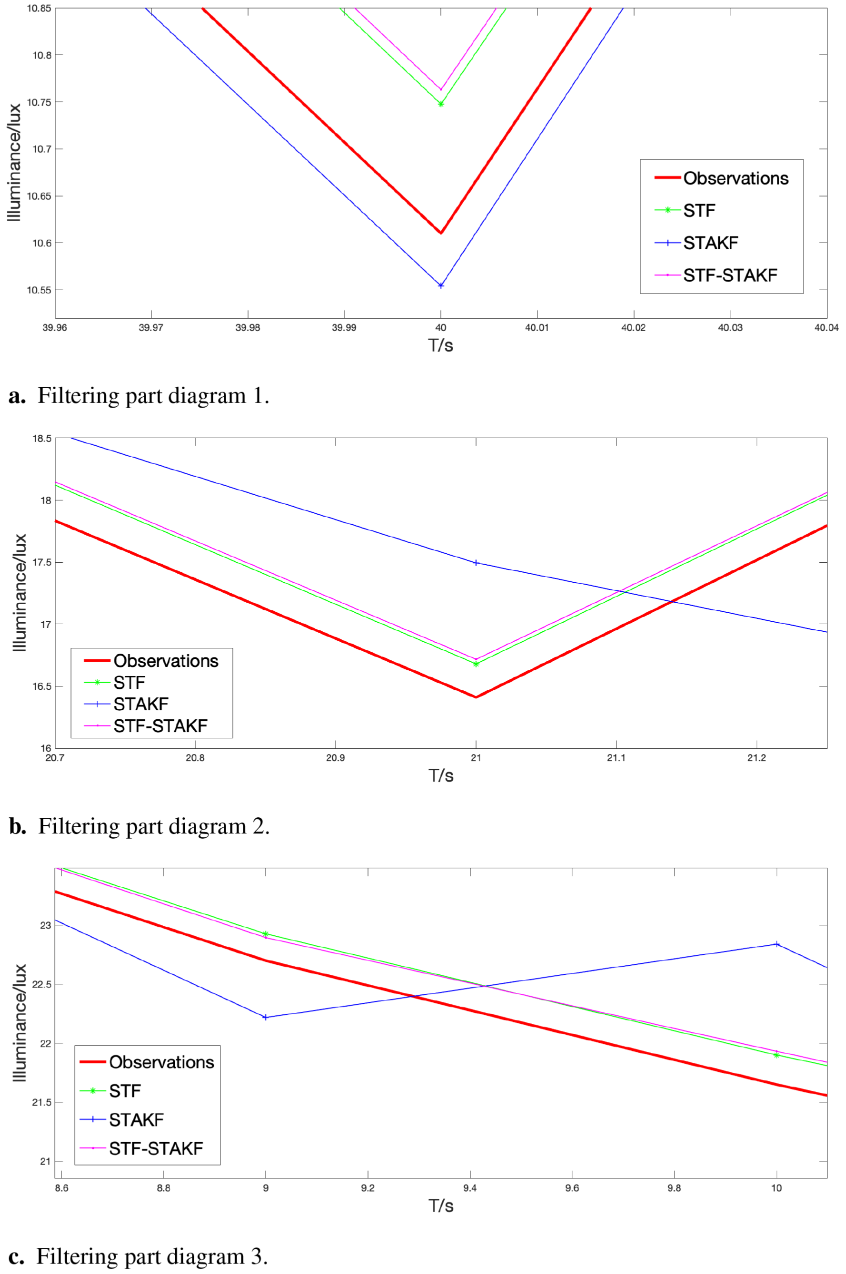


FIGURE 10
Illuminance observation curve of streetlight in rainy and foggy weather.



a. Filtering part diagram 1.

b. Filtering part diagram 2.

c. Filtering part diagram 3.

FIGURE 11
Filtering part diagram. **(A)** Filtering part diagram 1. **(B)** Filtering part diagram 2. **(C)** Filtering part diagram 3.

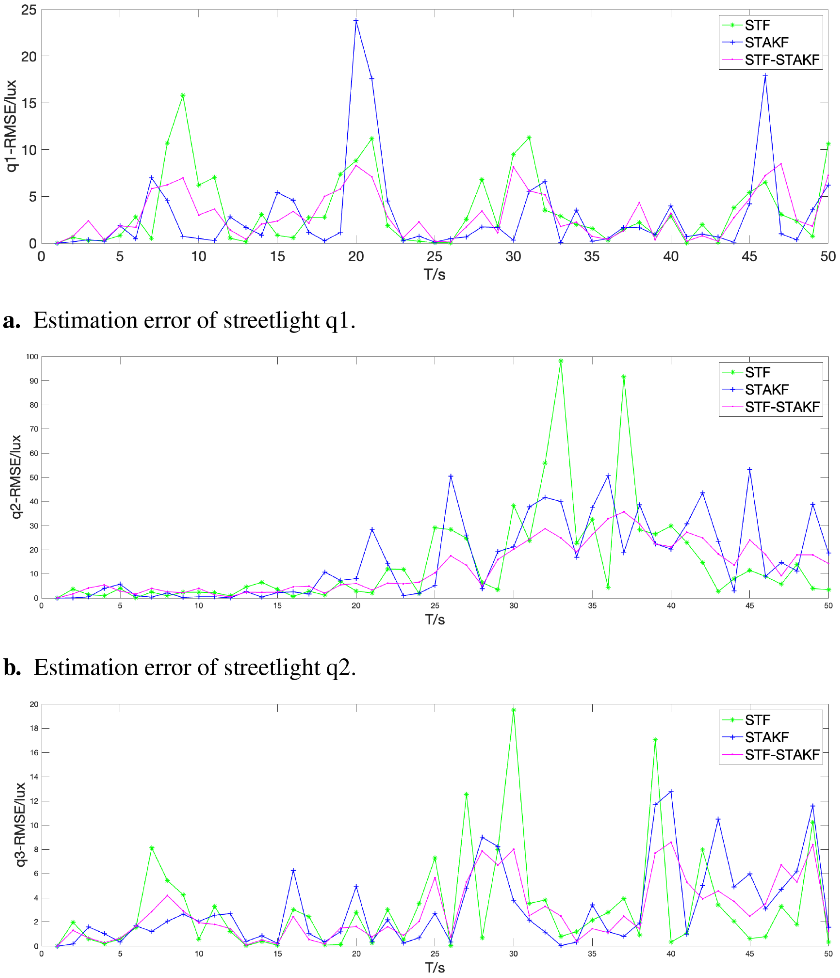
lies between the STF and STAKF curves during abrupt changes (as shown in Figure 11), and the overall root mean square error of the combined filtering (see Table 2) is better than that of STF, achieving improved tracking effects.

TABLE 2 Mean square error of three algorithms.

	STF	STAKF	STF-STAKF
RMSE	20.1759	20.7000	20.1680

5.2.2 Multiple streetlights

Since, in extreme oceanic weather conditions, any possible situation is unpredictable, data exchange between multiple port streetlights may sometimes fail to work for special reasons, preventing effective communication of the average illumination of the surrounding environment. Under this assumption, three adjacent streetlights are dimmed and tracked using the same method as described in section 5.2.1, and the noise parameters are as follows (This data is sourced from Guangdong Ocean University, and the relevant parameters were obtained by collecting hardware information from three adjacent smart streetlights):



a. Estimation error of streetlight q1.

b. Estimation error of streetlight q2.

c. Estimation error of streetlight q3.

FIGURE 12
Estimation error of streetlight. **(A)** Estimation error of streetlight q1. **(B)** Estimation error of streetlight q2. **(C)** Estimation error of streetlight q3.

$$Q=\begin{cases}\begin{bmatrix}4&1&0\\1&8&0\\0&0&1\end{bmatrix}, & 1<t\le 15\\ \begin{bmatrix}40&5&0\\5&25&0\\0&0&5\end{bmatrix}, & 15<t\le 30\\ \begin{bmatrix}15&10&0\\10&35&0\\0&0&10\end{bmatrix}, & 30<t\le 45\\ \begin{bmatrix}20&5&0\\5&30&0\\0&0&5\end{bmatrix}, & 45<t\le 50\end{cases},\quad R=\begin{cases}\begin{bmatrix}5&8&6\\8&5&5\\6&5&6\end{bmatrix}, & 1<t\le 15\\ \begin{bmatrix}5&10&0\\10&50&0\\0&0&25\end{bmatrix}, & 15<t\le 30\\ \begin{bmatrix}50&25&0\\25&80&80\\0&80&20\end{bmatrix}, & 30<t\le 45\\ \begin{bmatrix}30&40&0\\40&80&0\\0&0&80\end{bmatrix}, & 45<t\le 50\end{cases} \tag{57}$$

Through preprocessing and noise simulation of the actual data collected, each streetlight is independently estimated based on the processed data, and its observational data is dynamically affected by the dimming of other streetlights. When both the observational data and noise are abrupt, the dimming error of the three filters after 50 Monte Carlo simulations is shown in Figure 12 and Table 3, with the error of combined filtering being significantly smaller than that of STF and STAKF. This demonstrates that in extreme oceanic climate conditions, whether it is a single-head streetlight under data interconnection or multiple streetlights under data interconnection, combined filtering exhibits a certain effectiveness and versatility, effectively addressing the impacts of extreme weather.

TABLE 3 Mean square error of three algorithms.

RMSE	STF	STAKF	STF-STAKF
$q1_{RMSE}$	3.396	2.922	3.039
$q2_{RMSE}$	14.36	15.86	12.34
$q3_{RMSE}$	3.202	3.109	2.857
MEAN	6.986	7.297	6.078

6 Conclusions

Under challenging conditions influenced by various factors, the causes of visibility mutations in coastal ports typically stem from the weather's impact on the environment surrounding the port streetlights. When the mutation arises from process noise, STF-STAKF can closely approximate the current optimal STF in real-time and outperform STF in most states. When the mutation arises from the state value itself, observation noise, and process noise, the combined filtering approach can dynamically approach and surpass the current optimal tracking performance of STAKF. Experimental data from STF-STAKF demonstrate its overall real-time tracking performance, closely approximating and exceeding the current optimal filtering method. This tracking performance is highly suitable for scenarios with unknown mutations in extreme oceanic climate conditions. Moreover, due to limited computational resources at the edge of the port streetlight network, the proposed STF-STAKF approach can effectively utilize edge computing power to implement adaptive dimming at the edge. Considering that the dimming basis of port streetlights in actual projects will be based on regulations and standards stipulated by the state, the dimming standards should be set according to the environmental regulations of the streetlights. In practical engineering, the calculation methods for pavement materials, reflection coefficients, and brightness distribution have been studied in more detail, allowing for more accurate initial estimates and calibration judgments. However, these corrections do not affect the core idea of this paper, and it can be considered to further improve accuracy by combining and comparing more actual data in larger-scale application processes.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HJ: Funding acquisition, Investigation, Methodology, Resources, Writing – review & editing. XZ: Conceptualization, Data curation, Formal analysis, Project administration, Software, Writing – original draft, Writing – review & editing. ZZ: Supervision, Writing – review & editing. JJ: Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The work was supported in part by the National Natural Science Foundation of China under Grant 62272109.

Acknowledgments

We would like to thank Guangdong Ocean University and Qiandao Lake Research Institute for their support and contributions to this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al-Behadili, A. A., Al-Taai, O. T., and Al-Muhyi, A. H. A. (2023). "Impact of weather on marine vessel accidents in the Iraqi port of umm qasr, a case study of the salihiah tugboat accident," in *IOP Conference Series: Earth and Environmental Science*, Vol. 1158. 032008 (Bristol, UK: IOP Publishing).

Bojesomo, A., Al-Marzouqi, H., Liatsis, P., Cong, G., and Ramanath, M. (2021). "Spatiotemporal swin-transformer network for short time weather forecasting," in *Proceedings of the (CIKM) 2021 Workshops co-located with 30th (ACM) International Conference on Information and Knowledge Management (CIKM 2021)*, Gold Coast, Queensland, Australia, November 1-5, 2021. (Gold Coast, Australia: Central Europe (CEUR) Workshop).

Bowden, K. A., and Heinselman, P. L. (2016). A qualitative analysis of nws forecasters' use of phased-array radar data during severe hail and wind events. *Weather Forecasting* 31, 43–55. doi: 10.1175/WAF-D-15-0089.1

Bruce, C. W., Yee, Y. P., and Jennings, S. (1980). *In situ* measurement of the ratio of aerosol absorption to extinction coefficient. *Appl. Optics* 19, 1893–1894. doi: 10.1364/AO.19.001893

Cao, C., Bao, L., Gao, G., Liu, G., and Zhang, X. (2024). A novel method for ocean wave spectra retrieval using deep learning from sentinel-1 wave mode data. *IEEE Trans. Geosci. Remote Sens.* 62, 1–16. doi: 10.1109/TGRS.2024.3369080

Clarke, B., Otto, F., Stuart-Smith, R., and Harrington, L. (2022). Extreme weather impacts of climate change: an attribution perspective. *Environ. Res.: Climate* 1, 012001. doi: 10.1088/2752-5295/ac6e7d

Claser, R., and Nascimento, V. H. (2021). On the tracking performance of adaptive filters and their combinations. *IEEE Trans. Signal Process.* 69, 3104–3116. doi: 10.1109/TSP.2021.3081045

De Paz, J. F., Bajo, J., Rodr´ıguez, S., Villarrubia, G., and CorChado, J. M. (2016). Intelligent system for lighting control in smart cities. *Inf. Sci.* 372, 241–255. doi: 10.1016/j.ins.2016.08.045

Galbraith, D., and Grosjean, L. (2019). "Wind-alarm systems: Emerging observing technologies for port operations," in *Australasian Coasts and Ports 2019 Conference: Future directions from 40 [degrees] S and beyond, Hobart, 10-13 September 2019: Future directions from 40 [degrees] S and beyond*, Hobart, 10-13 September 2019 (Hobart: Engineers Australia), 418–423.

Gao, G., Bai, Q., Zhang, C., Zhang, L., and Yao, L. (2023a). Dualistic cascade convolutional neural network dedicated to fully polsar image ship detection. *ISPRS J. Photogrammetry Remote Sens.* 202, 663–681. doi: 10.1016/j.isprsjprs.2023.07.006

Gao, G., Dai, Y., Zhang, X., Duan, D., and Guo, F. (2023b). Adcg: A cross-modality domain transfer learning method for synthetic aperture radar in ship automatic target recognition. *IEEE Trans. Geosci. Remote Sens.* 61, 1–14. doi: 10.1109/TGRS.2023.3313204

Gao, G., Yao, B., Li, Z., Duan, D., and Zhang, X. (2024). Forecasting of sea surface temperature in eastern tropical pacific by a hybrid multiscale spatial–temporal model combining error correction map. *IEEE Trans. Geosci. Remote Sens.* 62, 1–22. doi: 10.1109/TGRS.2024.3353288

Gao, C., Zhang, X., Xu, Y., Wang, Z., Melgosa, M., Quesada-Molina, J. J., et al. (2018). Theoretical consideration on convergence of the fixed-point iteration method in cie mesopic photometry system mes2. *Optics Express* 26, 31351–31362. doi: 10.1364/OE.26.031351

Gao, G., Zhang, C., Zhang, L., and Duan, D. (2023c). Scattering characteristic-aware fully polarized sar ship detection network based on a four-component decomposition model. *IEEE Trans. Geosci. Remote Sens.* 61, 1–22. doi: 10.1109/TGRS.2023.3336300

Ge, Q., Shao, T., Duan, Z., and Wen, C. (2016). Performance analysis of the kalman filter with mismatched noise covariances. *IEEE Trans. Automatic Control* 61, 4014–4019. doi: 10.1109/TAC.2016.2535158

Gledhill, R. (1962). Particle-size distribution determination by turbidimetry. *J. Phys. Chem.* 66, 458–463. doi: 10.1021/j100809a021

Han, C., Zhu, H., and Duan, Z. (2006). *Multi-source Information Fusion* (Beijing, China: Tsinghua University Press).

He, R., Wan, C., and Jiang, X. (2021). “Risk management of port operations: A systematic literature review and future directions,” in *2021 6th International Conference on Transportation Information and Safety (ICTIS)*. (Wuhan, China: IEEE), 44–51.

Ito, K., Kang, Y., Zhang, Y., Zhang, F., and Biljecki, F. (2024). Understanding urban perception with visual data: A systematic review. *Cities* 152, 105169. doi: 10.1016/j.cities.2024.105169

Izaguirre, C., Losada, I. J., Camus, P., Vigh, J. L., and Stenek, V. (2021). Climate change risk to global port operations. *Nat. Climate Change* 11, 14–20. doi: 10.1038/s41558-020-00937-z

Jaskowski, P., Tomczuk, P., and Chrzanowicz, M. (2022). Construction of a measurement system with gps rtk for operational control of street lighting. *Energies* 15, 9106. doi: 10.3390/en15239106

Li, Z., Xu, G., Cheng, Y., Wang, Z., Wu, Q., and Yan, F. (2020). Spatially adaptive hybrid variational model for temperature-dependent nonuniformity correction of infrared images. *Optical Eng.* 59, 123103–123103. doi: 10.1117/1.OE.59.12.123103

Liu, R., Liang, Z., Yang, K., and Li, W. (2022). Machine learning based visible light indoor positioning with single-led and single rotatable photo detector. *IEEE Photonics J.* 14, 1–11. doi: 10.1109/JPHOT.2022.3163415

Muhamad, M., and Ali, M. M. (2018). “Iot based solar smart led street lighting system,” in *TENCON 2018 - 2018 IEEE Region 10 Conference*. (Jeju, Korea (South): IEEE), 1801–1806.

Or, B., Bobrovsky, B.-Z., and Klein, I. (2021). Kalman filtering with adaptive step size using a covariancebased criterion. *IEEE Trans. Instrumentation Measurement* 70, 1–10. doi: 10.1109/TIM.19

Pham, T. Y. (2023). A smart port development: Systematic literature and bibliometric analysis. *Asian J. Shipping Logistics* 39, 57–62. doi: 10.1016/j.ajsl.2023.06.005

Prousalidis, J., Kanellos, F., Lyridis, D., Dallas, S., Spathis, D., Georgiou, V., et al. (2019). “Optimizing the operation of port energy systems,” in *2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*. (Genova, Italy: IEEE), 1–6.

Saeik, F., Avgeris, M., Spatharakis, D., Santi, N., Dechouniotis, D., Violos, J., et al. (2021). Task offloading in edge and cloud computing: A survey on mathematical, artificial intelligence and control theory solutions. *Comput. Networks* 195, 108177. doi: 10.1016/j.comnet.2021.108177

Shao, T., Duan, Z., and Tian, Z. (2021). Performance ranking of kalman filter with pre-determined initial state prior. *IEEE Signal Process. Lett.* 28, 902–906. doi: 10.1109/LSP.2021.3071979

Sifakis, N., Kalaitzakis, K., and Tsoutsos, T. (2021). Integrating a novel smart control system for outdoor lighting infrastructures in ports. *Energy Conversion Manage.* 246, 114684. doi: 10.1016/j.enconman.2021.114684

Sun, Q. (2019). Opening time control method of port building lighting based on artificial intelligence. *J. Coast. Res.* 93, 335–340. doi: 10.2112/SI93-044.1

Swinehart, D. F. (1962). The beer-lambert law. *J. Chem. Educ.* 39, 333. doi: 10.1021/ed039p333

Wang, A., Xiang, M., Chen, W., and Chen, D. (2019). Exploration into the development of smart cities and the application of smart light poles (Zhao, Xiaolong, Trans). *Light Lighting* 43, 33–37. doi: CNKI:SUN:LAMP.0.2019-01-009

Yang, Z., Kagawa, S., and Li, J. (2021). Do greenhouse gas emissions drive extreme weather conditions at the city level in China? evidence from spatial effects analysis. *Urban Climate* 37, 100812. doi: 10.1016/j.uclim.2021.100812

Yau, K.-L. A., Peng, S., Qadir, J., Low, Y.-C., and Ling, M. H. (2020). Towards smart port infrastructures: Enhancing port activities using information and communications technology. *IEEE Access* 8, 83387–83404. doi: 10.1109/ACCESS.2020.2990961

Zhang, C., and Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *J. Ind. Inf. Integration* 23, 100224. doi: 10.1016/j.jii.2021.100224

Zhang, C., Zhang, X., Gao, G., Lang, H., Liu, G., Cao, C., et al. (2024). Development and application of ship detection and classification datasets: A review. *IEEE Geosci. Remote Sens. Magazine*, 2–36. doi: 10.1109/MGRS.2024.3450681

Zhou, Y. (2018). Urban management based on the internet of lights (Zhao, Xiaolong, Trans). *Shanghai Informatization* 05, 48–52. doi: CNKI:SUN:SHXX.0.2018-05-014

Zhou, D., Xi, Y., and Zhang, Z. (1991). An extended kalman filter with multiple suboptimal fading factors. *Chin. J. Automation* 17, 689–695.

Frontiers in Marine Science

TYPE Methods
PUBLISHED 25 November 2024
DOI 10.3389/fmars.2024.1503552

Significant wave height prediction in monsoon regions based on the VMD-CNN-BiLSTM model

Wengeng Shen[1], Zongquan Ying[1,2,3], Yiming Zhao[1] and Xuegang Wang[1,2,3]*

[1]China Communications Construction Company (CCCC) Fourth Harbor Engineering Institute Co., Ltd., Guangzhou, China, [2]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China, [3]Key Laboratory of Environment and Safety Technology of Transportation Infrastructure Engineering, China Communications Construction Company (CCCC), Guangzhou, China

OPEN ACCESS

EDITED BY
Jianchuan Yin,
Guangdong Ocean University, China

REVIEWED BY
Zeguo Zhang,
Guangdong Ocean University, China
Lianbo Li,
Dalian Maritime University, China

*CORRESPONDENCE
Xuegang Wang
510simon@163.com

RECEIVED 29 September 2024
ACCEPTED 31 October 2024
PUBLISHED 25 November 2024

CITATION
Shen W, Ying Z, Zhao Y and Wang X (2024) Significant wave height prediction in monsoon regions based on the VMD-CNN-BiLSTM model.
Front. Mar. Sci. 11:1503552.
doi: 10.3389/fmars.2024.1503552

A novel significant wave height prediction method for monsoon regions is proposed, utilizing the VMD-CNN-BiLSTM model to enhance prediction accuracy under complex meteorological conditions. Traditional numerical models exhibit limitations in managing extreme marine conditions and fail to fully integrate wind field information. Meanwhile, existing machine learning models demonstrate insufficient generalization and robustness for long-term predictions. To address these shortcomings, the predictive approach combines Variational Mode Decomposition (VMD) with a hybrid deep learning model (CNN-BiLSTM). VMD is employed to decompose the original wave height sequence and extract key features, while CNN captures the spatial features of wind field and wave height data. BiLSTM, in turn, models the temporal dependencies. Experimental results reveal that the VMD-CNN-BiLSTM model provides substantial advantages in prediction performance across all seasons, including the entire year. Compared to traditional models, the proposed method demonstrates significantly reduced Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), alongside an improved coefficient of determination (R^2). These findings confirm the effectiveness and reliability of the method under complex meteorological conditions such as monsoons and typhoons.

KEYWORDS

wave height, prediction, CNN-BiLSTM, VMD, monsoon

1 Introduction

Wave height prediction is a crucial issue in coastal and marine engineering. The larger the wave height, the worse the sea conditions, significantly impacting the safe operation of platform structures (Abed-Elmdoust and Kerachian, 2012). Therefore, forecasting wave height in advance allows for timely assessment of platform safety levels and risk mitigation.

However, due to the highly nonlinear and non-stationary statistical characteristics of waves, analyzing and predicting wave height is challenging.

Numerous efforts have been made in existing research on wave height prediction. Numerical wave models are widely applied in global sea state forecasting (Simmons et al., 2004). The principle of numerical wave models is to obtain information such as wave height and period by solving the wave spectrum equation of oceanic physical processes. Bottcher et al. (2012) compared the wave heights observed by buoys with the model predictions, concluding that numerical prediction is a reliable method for wave height forecasting. Advanced third-generation models, such as the Wave Model (WAM) (Mentaschi et al., 2015), WAVEWATCH-III (WW3) (Rogers et al., 2003), and Simulation Waves Nearshore (SWAN) (Swain et al., 2019), are currently among the most sophisticated numerical models. The WAM and WW3 models have a similar structure, but WW3 uses more complex dissipation source terms and wind input terms than WAM. Liu et al. (2019) compared the performance of WAM and WW3 using data from the South Indian Ocean, concluding that both methods can predict significant wave height well. The SWAN model was developed to address complex wave conditions in coastal areas. Liang et al. (2019) validated the performance of SWAN through buoy measurements in the northwest Pacific, northeast Pacific, and northwest Atlantic. The experimental results showed that, under accurate boundary conditions, the SWAN model could simulate coastal waves effectively. However, the fixed energy spectrum equations with fixed expressions used by these models may not fully represent the complex and variable ocean environment. Specifically, the accuracy of numerical wave predictions under extreme and highly variable ocean conditions still needs improvement.

Machine learning is a data-driven approach that has recently been successfully applied to wave height prediction (Yu and Wang, 2021). Based on long-term, accurate wave height measurements obtained from buoys, satellites, and scatterometers, machine learning methods predict future wave heights by learning the inherent variability in the data (Fan et al., 2019). Deo et al. (2001) explored a three-layer feedforward network to obtain significant wave height outputs. Berbic et al. (2017) used artificial neural networks (ANN) and support vector machines (SVM) to predict significant wave heights over 0.5–5.5 hours, demonstrating that ANN and SVM outperform numerical models in this range. Shen Lixiang et al. (2023) proposed an Attention-LSTM model based on attention mechanisms and multivariable inputs for short-term wave height prediction in the Longkou sea area of Shandong. Pradnya and Londhe (2016) used neural wavelet technology to predict extreme wave heights, showing that multi-level decomposition of wave data helps improve prediction accuracy. Recurrent neural networks (RNN) (Mikolov et al., 2021) and their variant long short-term memory networks (LSTM) (Gers et al., 2002) have unique advantages in solving prediction problems. Zhang et al. (2021) proposed the N-LSTM model, combining numerical forecasts with measured data, using LSTM and Gaussian approximation modules to improve the accuracy of numerical forecasts. Pushpam and Enigo V.S., 2020) applied RNN-LSTM to predict significant wave heights, showing good performance within 24 hours. Kaloop et al. (2020) integrated wavelet, particle swarm optimization (PSO), and extreme learning machine (ELM) methods into the wavelet PSO-ELM model for estimating coastal and deep-sea wave heights, with evaluation results showing high prediction accuracy. Hao et al. (2023) systematically analyzed the effects of input length, forecast length, and model complexity on wave height prediction using RNN/LSTM/GRU and other recurrent neural networks. Minghao et al. (2024) introduced Rayleigh parameters in wave height prediction, showing improvements in mid- to long-term prediction capabilities for BPNN and LSTM. Yifan et al. (2024) introduced Spearman correlation analysis into RNN/LSTM/GRU models and proposed the LSTM-Attention model. These studies achieved promising results using various neural network models for wave height prediction. However, they have not fully incorporated wind field information. As the key driver of wave formation and evolution, wind field data is crucial for wave height prediction. Ignoring wind field information may limit the model's ability to capture the complex relationships between wind and waves (Ahmed et al., 2024). Yin et al. (2023) proposed an adaptive tidal level prediction mechanism based on EMD and the Lipschitz quotients method, combining harmonic analysis with a variable structure neural network to automatically determine model parameters, thereby improving the accuracy and adaptability of tidal level prediction. Additionally, machine learning models often experience a decline in prediction accuracy over long-term forecasts, particularly when dealing with complex nonlinear time series wave data, limiting the model's generalization capability and robustness.

This study addresses the limitations in existing models, particularly their inability to fully incorporate wind field information for long-term wave height prediction, and proposes a hybrid model based on VMD-CNN-BiLSTM for a typical wind-wave region—the southeastern sea of China—aimed at improving wave height prediction accuracy by comprehensively considering wind field and significant wave height information. First, the model uses Variational Mode Decomposition (VMD) to decompose the wave height data, breaking down the complex non-stationary wave height sequence into multiple relatively stationary mode functions, facilitating subsequent feature extraction. Then, the decomposed wave height modes and wind field data are input into a Convolutional Neural Network (CNN) for feature extraction, where CNN extracts local spatial features of the wind field and wave height modes. Finally, the extracted features are fed into a Bidirectional Long Short-Term Memory (Bi-LSTM) network to capture the dependencies in the wave height time series, thereby better understanding the intrinsic relationship between wind and waves. Through this approach, the proposed model demonstrates greater robustness and generalization ability in long-term wave height prediction, providing a more reliable solution for significant wave height forecasting.

2 WW3-SWAN numerical simulation

2.1 Model settings

The WW3 model (Tolman, 2009) was developed based on the third-generation wave model WAM, with its governing equations modeled by solving the action balance equation over the wave number-direction spectrum. The model uses the global digital elevation model (DEM) dataset released by the General Bathymetric Chart of the Oceans (GEBCO), with a resolution of 15″×15″, and wind field data at a height of 10 meters from the ERA5 reanalysis data by the European Centre for Medium-Range Weather Forecasts (ECMWF), with a resolution of 0.25°×0.25°, from January 1, 2017, 00:00 to December 31, 2021, 23:00. The extent of the wind field should be greater than or equal to the extent of the WW3 and SWAN numerical simulations. No additional data is input into the boundary conditions of the WW3 model. The wave spectrum grid of the WW3 model is set to 32×24, with a frequency range from 0.0373 Hz to 0.7159 Hz, divided into 32 bands, and wave direction divided into 24 directions. The calculation area of the model covers the longitude range of 110°E to 130°E and the latitude range of 10°N to 30°N, with a spatial resolution of 0.25°×0.25°. The layout of the model region is shown in Figure 1.

The SWAN model was modified and improved by Booij et al. (1996) from Delft University of Technology based on the third-generation wave model WAM. The model discretizes the governing equations using an implicit method, taking into account wave-wave interactions and the breaking effects caused by depth changes during wave propagation, making it effective in simulating the evolution of nearshore waves. The computational range of the SWAN model is from 115.59°E to 117.71°E in longitude and from 21.78°N to 23.66°N in latitude, using an unstructured grid, as shown in Figure 1B. Bathymetric data comes from the GEBCO dataset, wind field data uses ERA5 reanalysis data, and the wave spectrum data at open boundary points is obtained from the wave spectrum output of the WW3 model. The simulation time range is from 00:00:00 on January 1, 2017, to 23:00:00 on December 31, 2021, with an output time interval of one hour.

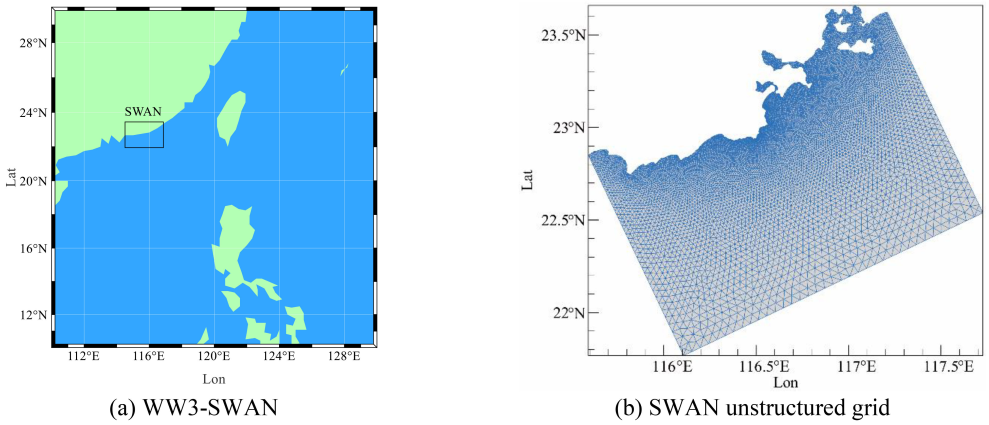
To verify the accuracy of the WW3-SWAN numerical model, a MARK III Wave Rider instrument was deployed in the waters off the Stone Tablet Mountain Cape, at the coordinate position (22° 55.7046′N, 116°31.4034′E), as shown in Figure 2. The Wave Rider instrument has a wave height measurement range of ±20m. The measured data were processed by the instrument's built-in software, which then statistically generated hourly wave height observation data. The observation period was from 00:00 on April 1, 2021, to 23:00 on November 18, 2021.

2.2 Data validation

Figure 3 shows a comparison between the significant wave heights from the numerical model and the measured values. The significant wave height values from the WW3-SWAN numerical simulation are consistent with the observed values in terms of the overall trend. However, due to the ERA5 reanalysis data underestimating the intensity of typhoons in the Northwest Pacific (Li and Hu, 2021), the numerical simulation slightly underestimates the peak values of the significant wave heights.

Figure 4 shows the situation of some typhoons in the Western Pacific in 2021, with longitude on the horizontal axis, latitude on the vertical axis, and wind speed represented by the color scale. As shown in Figure 4A, during the spring season, Typhoon Surigae formed on April 10, 2021, with wind speeds rapidly increasing from 28 m/s to 60 m/s, and was upgraded to a super typhoon on April 17-18, 2021. The typhoon's center was located approximately 1,280 km southeast of Manila, Philippines, in the Northwest Pacific Ocean (10.3°N, 131.9°E), with maximum winds near the center reaching 15 on the Beaufort scale (50 m/s). It transitioned into an extratropical cyclone on April 25. At 12:00 on April 18, 2021, the South China Sea was affected by the typhoon, with wind speeds around 10 m/s in the area of the wave monitoring site, leading to higher waves. Therefore, during the typhoon period, the average significant wave height measured in Figure 3A was 1.4m, slightly higher than the numerical simulation value. In May, with no typhoon influence, the average significant wave height at the wave monitoring site was 0.66m, with relatively calm sea conditions, and the numerical simulation values were closer to the measured values at this time.

As shown in Figure 4B, during the summer season, Typhoon Choi-Wan entered the South China Sea on June 3, 2021, with maximum sustained winds near the center reaching 65 km/h. At 9:00 on June 4, 2021, wind speeds at the wave monitoring site



(a) WW3-SWAN (b) SWAN unstructured grid

FIGURE 1
Calculation area of the WW3-SWAN model. **(A)** WW3-SWAN, **(B)** SWAN unstructured grid.

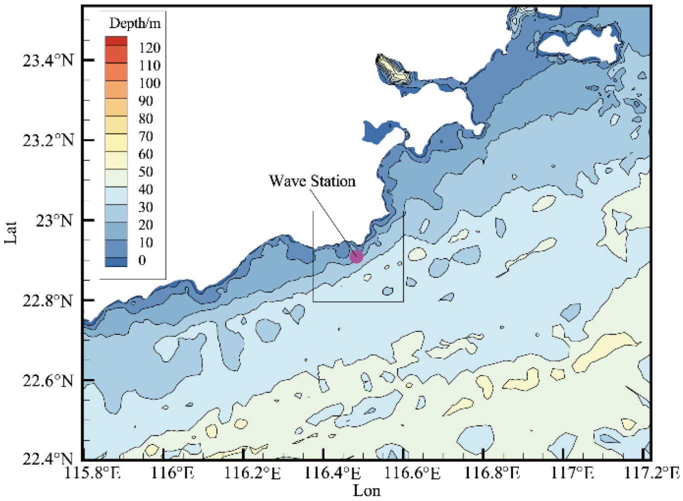


FIGURE 2
MARK III Wave instrument monitoring position.

reached 6-10 m/s, with a peak significant wave height of around 1.5 m. As shown in Figure 4C, Typhoon Lupit formed in Zhanjiang, Guangdong, on August 2, 2021, and gradually approached the coasts of Fujian and Guangdong. By 15:00 on August 6, 2021, wind speeds from Typhoon Lupit along the Fujian-Guangdong coast reached around 10 m/s, causing the wave height at the monitoring site to reach a maximum of approximately 2.5 m. Therefore, during the typhoon periods shown in Figure 4B, the

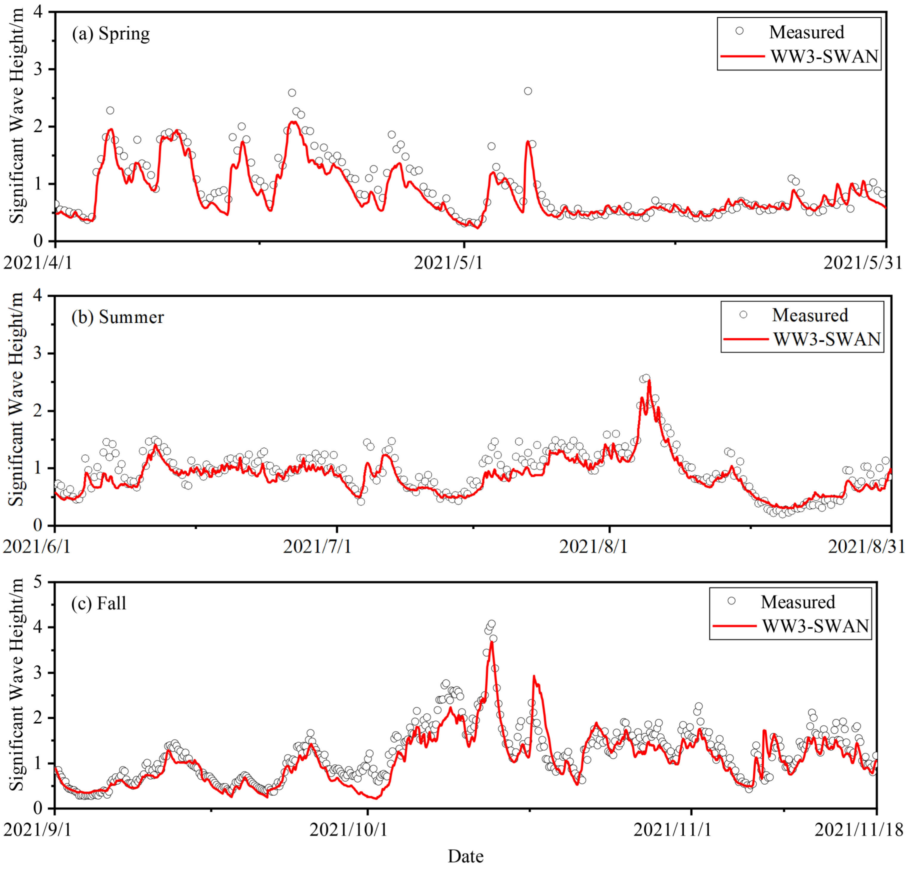


FIGURE 3
Comparison of WW3-SWAN SWH with the measured value. **(A)** Spring, **(B)** Summer, **(C)** Fall.

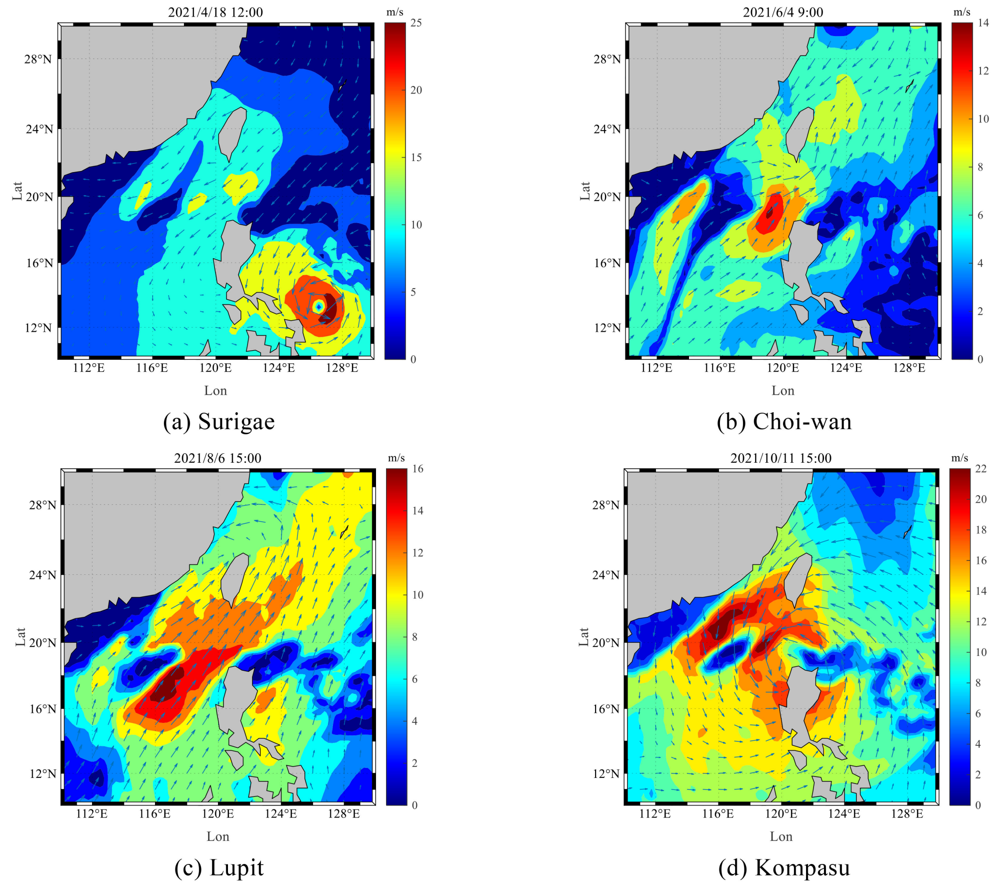


FIGURE 4
Typhoons in the Western Pacific during 2021. **(A)** Surigae, **(B)** Choi-wan, **(C)** Lupit, **(D)** Kompasu.

observed values of significant wave height were consistently higher than the values simulated by the numerical model.

As shown in Figure 4D, during the autumn season, Typhoon Kompasu formed in the Philippine Sea on October 8, 2021, and steadily moved westward after entering the South China Sea, with its center approaching the coastal areas of the South China Sea. Therefore, as seen in Figure 3C, the measured wave heights increased significantly during mid-October 2021, while the simulated wave heights were slightly lower.

To further validate the accuracy of the numerical simulation results in this study against the measured data, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2) were used to quantitatively evaluate the accuracy of the numerical results. The calculation formulas are shown in Equations 1–3.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{y})^2} \tag{3}$$

In the formulas, x_i represents the numerical simulation values, y_i represents the measured values, n is the total number of samples, and $\bar{x}$ and $\bar{y}$ are the mean values of the numerical simulation and measured values, respectively.

To evaluate the WW3-SWAN numerical simulation model, Table 1 uses MAE, RMSE, and R^2 for a quantitative assessment of model performance. Statistical analysis shows that the WW3-SWAN model performs well across different seasons. The MAE ranges from 0.1413 m to 0.2130 m, indicating that the average deviation between the simulated and observed values is quite small. RMSE, which is more sensitive to larger errors, is slightly higher, ranging from 0.1828 m to 0.2844 m. This is mainly due to the impact of extreme weather conditions like typhoons, which cause deviations in significant wave height at peak values. The R^2 values are notably high, between 0.7801 and 0.8493, indicating a strong linear relationship between the simulated and observed significant wave heights. The model performs best in the spring, with the

TABLE 1 Numerical simulation error.

Season	MAE/m	RMSE/m	R^2
Spring (April and May)	0.1540	0.2067	0.8493
Summer (June, July and August)	0.1413	0.1828	0.7910
Fall (September, October and November)	0.2130	0.2844	0.7801

highest R^2 value of 0.8493. In summer and autumn, frequent typhoons lead to reduced accuracy in the numerical simulation. Overall, the WW3-SWAN model reliably reflects the significant wave height in the study area, capturing the magnitude and temporal variation, and can serve as input data for the VMD-CNN-BiLSTM model.

3 Forecasting models

3.1 VMD model

VMD (Variational Mode Decomposition) is an adaptive, fully non-recursive signal processing technique that combines Wiener filtering, Hilbert transform, and the Alternating Direction Method of Multipliers (ADMM). As a non-stationary time series, significant wave height is well-suited for decomposition using VMD. The VMD decomposition process effectively transforms into an optimization process. The two main components of VMD are constructing the variational problem and solving it. Variational modes refer to the modes obtained by solving the variational problem. VMD iteratively searches for the optimal solution of the variational modes, adaptively updating the optimal center frequency and bandwidth for each Intrinsic Mode Function (IMF). VMD redefines the intrinsic mode function, as shown in Equation 4. Compared to other decomposition methods like Empirical Mode Decomposition (EMD) or Wavelet Transform, VMD was chosen for its superior ability to reduce mode mixing and provide more stable component separation under complex wave conditions.

$$u_k(t) = A_k(t)\cos(\phi_k(t)) \tag{4}$$

Where k represents the mode number, $A_k(t)$ is the amplitude of the k-th mode, $\varphi k(t)$ is the phase of the k-th mode, and $uk(t)$ is the k-th mode function.

At this point, the variational problem constructed by VMD is shown in Equation 5:

$$\begin{cases} \min\limits_{\{u_k,\bar{w}_k\}} \left\{ \sum\limits_{k=1}^{K} \| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} * u_k(t) \right) \right] e^{-j\bar{w}_k t} \|_2^2 \right\} \\ s.t.\cdots \sum\limits_{k=1}^{K} u_t(t) = f(t) \end{cases} \tag{5}$$

Where u_k represents the corresponding mode function, and $\bar{w}_k$ represents the center frequency of the corresponding mode.

By introducing Lagrange multipliers, the constrained optimization problem above is transformed into an unconstrained problem, as shown in Equation 6:

$$L(\{u_k\},\{\bar{w}_k\}) = \alpha \sum_{k=1}^{K} \| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} * u_k(t) \right) \right] e^{-j\bar{w}_k t} \|_2^2 + \| f(t) - \sum_{k=1}^{K} u_k(t) \|_2^2 + \left\langle \lambda(t), f(t) - \sum_{k=1}^{K} u_k(t) \right\rangle \tag{6}$$

Where α represents the variance regularization parameter, and λ represents the Lagrange multiplier.

To solve this problem, the Alternating Direction Method of Multipliers (ADMM) is used. The specific solving steps are as follows:

1. Initialize u_1^k, $\bar{w}_k$, λ_1^k and the iteration number n.
2. Increase the variable n to 1 and enter the loop.
3. Update the variables according to Equation 7 until the number of iterations exceeds k, then stop updating:

$$\begin{cases} \hat{u}_{n+1}^{k}(w) = \dfrac{\hat{f}(w) - \sum\limits_{i<k} \hat{u}_{n+1}^{i}(w) + \sum\limits_{i>k} \hat{u}_{n}^{i}(w) + \hat{\lambda}_n(w)/2}{1+2\alpha(w-\bar{w}_n^k)^2} \\ \bar{w}_{n+1}^{k} = \dfrac{\int_0^{\infty} w|\hat{u}_{n+1}^{k}(w)|^2 dw}{\int_0^{\infty} |\hat{u}_{n+1}^{k}(w)|^2 dw} \end{cases} \tag{7}$$

4. Update the Lagrange multipliers λ

$$\hat{\lambda}_{n+1}(w) = \hat{\lambda}_n(w) + \tau \left(\hat{f}(w) - \sum_{k=1}^{K} \hat{u}_{n+1}^{k}(w) \right) \tag{8}$$

5. If the condition of Equation 9 is met, the loop ends; if not, return to step 2.

$$\sum_{k=1}^{K} \frac{\| \hat{u}_{n+1}^{k}(w) - \hat{u}_{n}^{k}(w) \|_2^2}{\| \hat{u}_{n}^{k}(w) \|_2^2} < \in \tag{9}$$

By constructing and solving the variational problem, VMD can effectively decompose non-stationary data. However, the number of modes after VMD decomposition needs to be manually selected. Multiple tests are required to find the most appropriate number of modes.

3.2 CNN model

CNN are an effective deep learning model widely used for feature extraction in image processing and spatio-temporal data. Through mechanisms like local receptive fields and weight sharing, CNNs can effectively capture local spatial features in the data. In this forecasting model, CNN is used to extract the spatial features of wind fields and wave heights, which will serve as inputs for subsequent time series modeling. CNN architecture is constructed by stacking three main types of layers: convolutional layers, pooling layers, and fully connected (FC) layers. Each convolutional layer contains a set of learnable filters, which aim to automatically extract local features from the input matrix. These filters perform convolution operations based on two important concepts: weight sharing and local connections, which help reduce computational complexity and enhance model performance. The pooling layer follows the convolutional layer, performing down-sampling. A notable feature of the pooling layer is its ability to reduce the dimensionality of feature maps, thus preventing overfitting. Typically, FC layers are used in the final layers of CNN architecture, and their role is to learn nonlinear combinations of features extracted by convolutional layers, generating the final output. Since wave height and wind field data usually exhibit significant spatio-temporal dependencies, CNN can effectively extract local features and patterns from this data through its receptive fields. Therefore, CNN is selected to extract features from wind fields and wave heights in this study.

Figure 5 illustrates the specific process of wind field and wave height data processed through a one-dimensional Convolutional Neural Network (1D-CNN). The input data, representing a sample at a certain time from the dataset, is preprocessed and fed into the convolutional layer of the CNN in sequence form. In the convolutional layer, multiple filters (also known as convolutional kernels) slide over the input sequence, extracting local temporal features through local connectivity and weight sharing. After the convolution operation, the data moves to the pooling layer for downsampling. By selecting the maximum value (max pooling) or the average value (average pooling) within a window, the dimensionality of the feature map is reduced. This not only decreases the computational complexity of the model but also effectively prevents overfitting. After processing by the pooling layer, the dimensionality of the feature map is significantly reduced, preserving key features while lowering computation costs. Finally, these processed feature maps are flattened into a one-dimensional vector, which serves as input for subsequent fully connected layers or other models (such as LSTM or BiLSTM) for the final prediction task.

3.3 BiLSTM model

Additionally, since the current wave height is not only related to the current wind field conditions but also influenced by historical wind field and wave height changes, traditional neural networks struggle to capture this long-term dependency. LSTM, with its special architecture, can effectively retain and utilize information from long-term time series, allowing it to capture complex temporal patterns in the data. Moreover, LSTM can solve the vanishing gradient problem found in conventional Recurrent Neural Networks (RNN), making it more stable and accurate in predicting long sequences. Therefore, in wave height forecasting tasks, LSTM becomes a natural choice to better model the temporal dependency and dynamic changes in the data.

A typical LSTM unit contains three types of gates: the input gate i_t, forget gate f_t, and output gate o_t, as shown in Figure 6. In each gate, the state of the memory cell is controlled through element-wise multiplication and the Sigmoid function. The inputs to the LSTM model are the input data at the current state x_t and the output of the hidden state from the previous layer $h_{t\text{-}1}$.

The input data first passes through the forget gate, which determines which information should be discarded or retained. The equation for the forget gate is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{10}$$

Here, σ represents the Sigmoid activation function, and W_f and b_f represent the weights and biases of the forget gate, respectively. The current input x_t and the previous hidden state $h_{t\text{-}1}$ are fed into the Sigmoid function. By transforming values between 0 and 1, the forget gate determines which information needs to be updated,

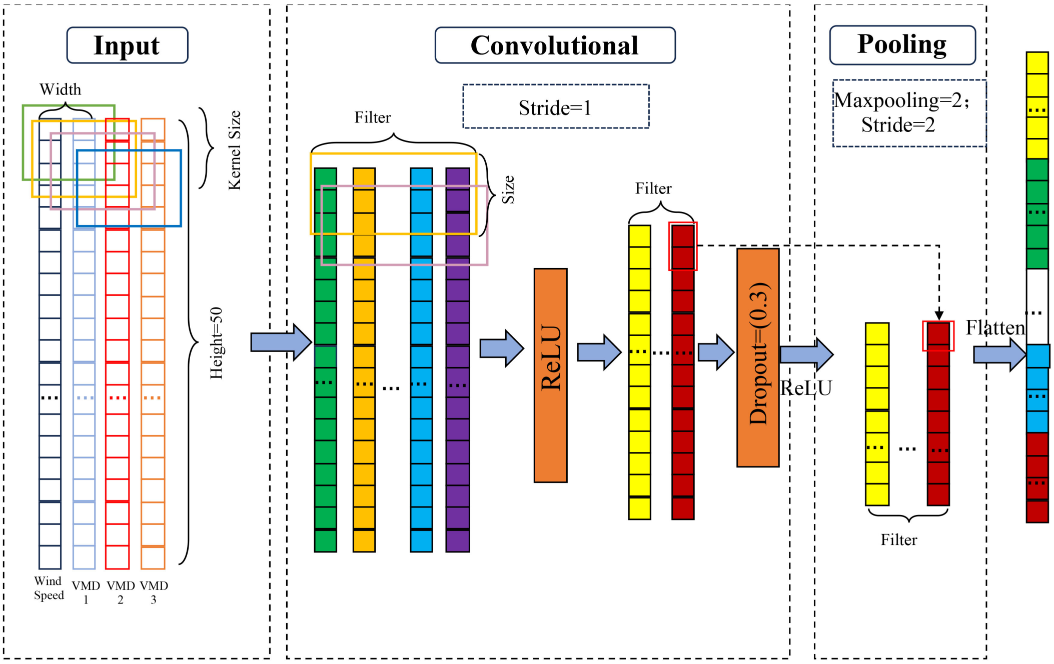


FIGURE 5
Flow chart of the one-dimensional CNN model.

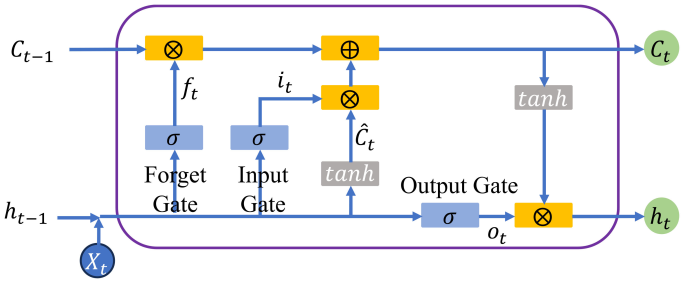


FIGURE 6
LSTM structure diagram.

where 0 represents unimportant information and 1 represents important information.

Next, the data passes through the input gate, with the calculation formula as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{11}$$

Then, the current input x_t and the hidden state $h_{t\text{-}1}$ are fed into the hyperbolic tangent function (tanh). At this point, the cell state is calculated and updated to the new cell state. The formula is as follows:

$$\begin{cases} \hat{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\ C_t = f_t \odot C_{t-1} + i_t \odot \hat{C}_t \end{cases} \tag{12}$$

Here, *tanh* is the hyperbolic tangent activation function, and ⊙ denotes the element-wise multiplication operation, with C_t being the new cell state.

Finally, the output gate selects the next hidden state. The new cell state C_t and the new hidden state h_t are passed to the next time step. The formula for the output gate is as follows:

$$\begin{cases} o_t = \sigma(W_o \cdot [h_{t-1}, C_t] + b_o) \\ h_t = o_t \odot tanh(C_t) \end{cases} \tag{13}$$

A unidirectional LSTM can only process information flow in one direction, whereas a bidirectional LSTM (BiLSTM) enhances the model's ability to understand wave height and wind field temporal evolution by analyzing both forward and backward information in parallel. BiLSTM consists of two LSTM layers operating in opposite directions, as illustrated in Figure 7. The horizontal dashed line represents the time axis flow of the time series data, while the vertical slanted lines depict the information transmission paths between network layers.

3.4 VMD-CNN-BiLSTM model

The VMD-CNN-BiLSTM model is shown in Figure 8. VMD decomposes the wave height data into several Intrinsic Mode Functions (IMFs), breaking down the non-stationary wave height time series into relatively stationary subcomponents. The CNN network extracts local features from wind speed and IMFs, while the BiLSTM network models the wave time series data to accurately predict future wave heights. The detailed process is as follows:

1. Data collection and preprocessing: Gather datasets that include wind field and wave height data, and perform preprocessing steps like data cleaning and normalization to ensure the data is suitable for model training.
2. Dataset splitting: Divide the dataset into training and testing sets to ensure that the training set has enough data for model learning, while the testing set is used to evaluate the model's performance.
3. VMD decomposition: Apply VMD to decompose the wave height data. The original wave height data is decomposed

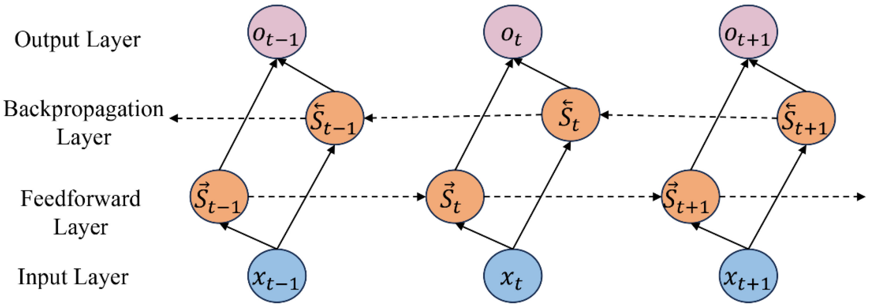


FIGURE 7
Bi-LSTM structure diagram.

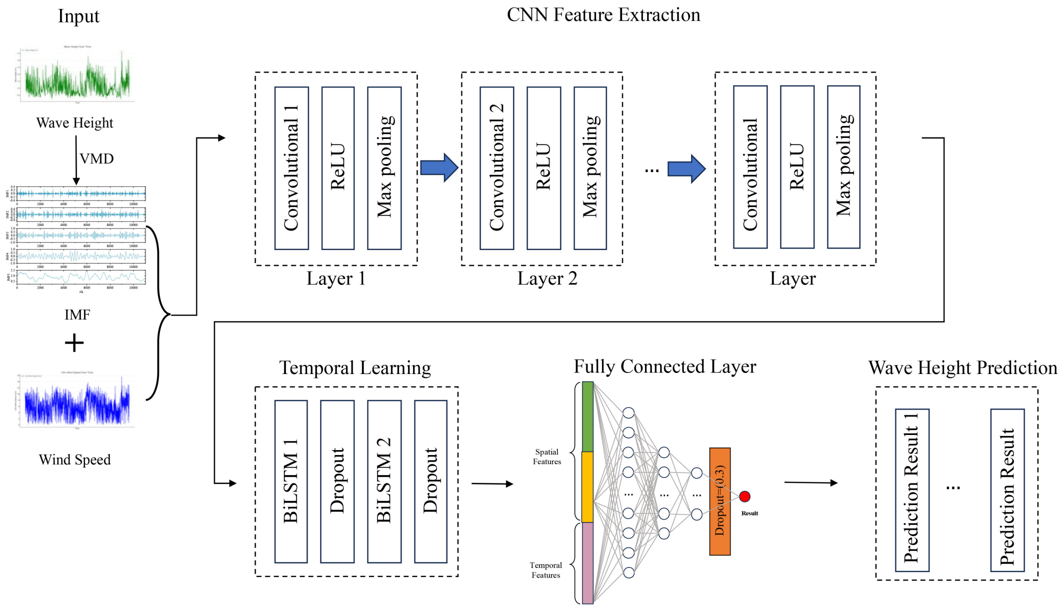


FIGURE 8
CNN-BiLSTM Flowchart.

into several IMFs, each representing different frequency components of the data. This decomposition helps CNN better extract multi-scale features from the wave data.

4. Feature extraction via CNN: Apply a multi-layer Convolutional Neural Network (CNN) to process the input data, extracting spatial features from the wind field and IMFs. The CNN layers help identify patterns and relationships between spatial data points that affect wave heights.
5. Temporal feature extraction via Bi-LSTM: Pass the spatial features extracted by CNN into a bidirectional Long Short-Term Memory network (Bi-LSTM), which extracts the temporal features from the data. The Bi-LSTM layer captures time dependencies, allowing the model to account for how event sequences and timings affect wave height variations.
6. Feature merging and fully connected layers: Merge the features extracted by CNN and Bi-LSTM and pass the merged features through fully connected layers for learning.
7. Output layer: After the last fully connected layer, a single neuron output layer is used to produce the final wave height prediction.

4 Wave height prediction

The southeastern seas of China are influenced by the monsoon climate, with prevailing northerly winds in winter and predominantly southerly winds in summer. Waves, influenced by these wind fields, exhibit a seasonal distribution characterized by lower effective wave heights in spring and summer, and higher effective wave heights in autumn and winter (Qiu et al., 2019). As shown in Figure 9, during spring and summer, the effective wave heights in the southeastern sea area range mainly from 0.2 to 1.2 meters. In autumn, the effective wave heights significantly increase, with the mean value ranging from 0.6 to 1.6 meters. In winter, the mean wave height increases further, with the maximum average reaching approximately 2.3 meters. Therefore, when using the VMD-CNN-BiLSTM model to predict effective wave heights, it is necessary to predict the wave heights for each season separately.

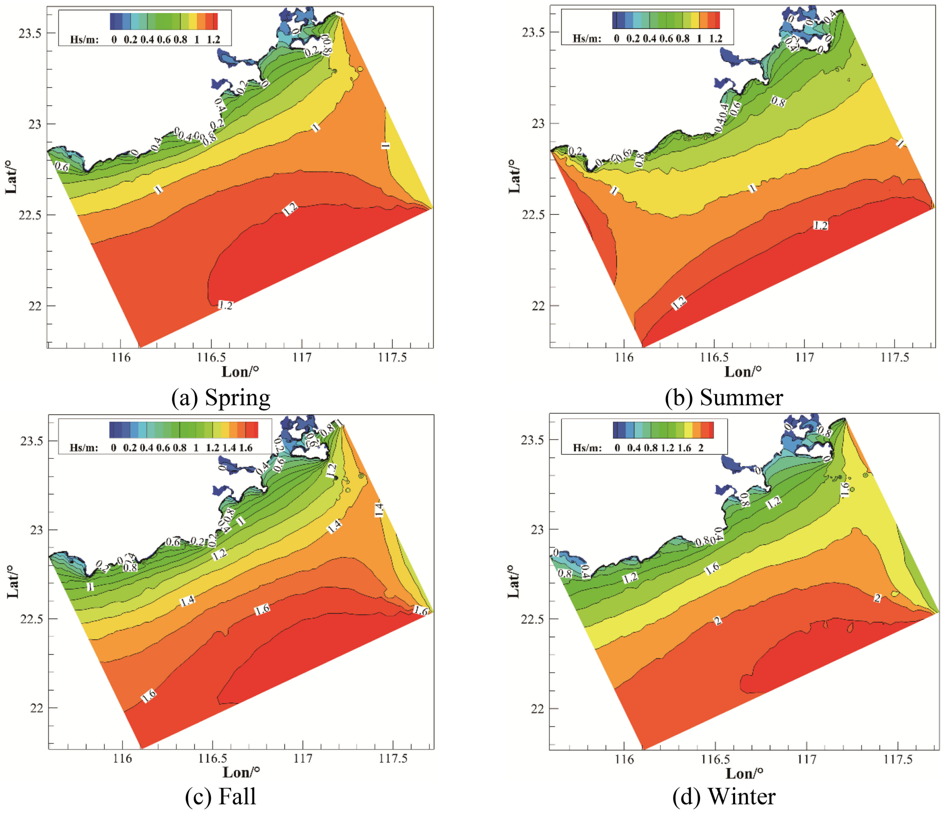
4.1 VMD decomposition

Before being input into the prediction model, the wave height dataset was normalized to a range between 0 and 1, which accelerates the model's convergence and improves prediction accuracy.

Due to the influence of the monsoon climate and typhoons in this sea area, the effective wave height sequence fluctuates greatly, requiring data processing. This paper uses Variational Mode Decomposition (VMD) to decompose the original sequences of wind fields and effective wave heights into several relatively smooth components. Taking the spring period from 2017 to 2021 as an example, with a data time interval of 1 hour, the VMD decomposition results are shown in Figure 10.

From the decomposition, we can observe that the wind field (Figure 10A) and the effective wave height (Figure 10B) sequences are decomposed into five components (IMF1 to IMF5), transitioning from high-frequency to low-frequency components.

To compare the impact of VMD decomposition of wind fields and effective wave heights on wave height prediction, two cases were designed: Case 1 includes seven vectors, namely the wind field, IMF1 to IMF5 of the effective wave height, and the original effective


Hs/m: 0 0.2 0.4 0.6 0.8 1 1.2
Hs/m: 0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6
Hs/m: 0 0.4 0.8 1.2 1.6 2
Lat/°
Lon/°
23.5
23
22.5
22
116
116.5
117
117.5

(a) Spring

(b) Summer

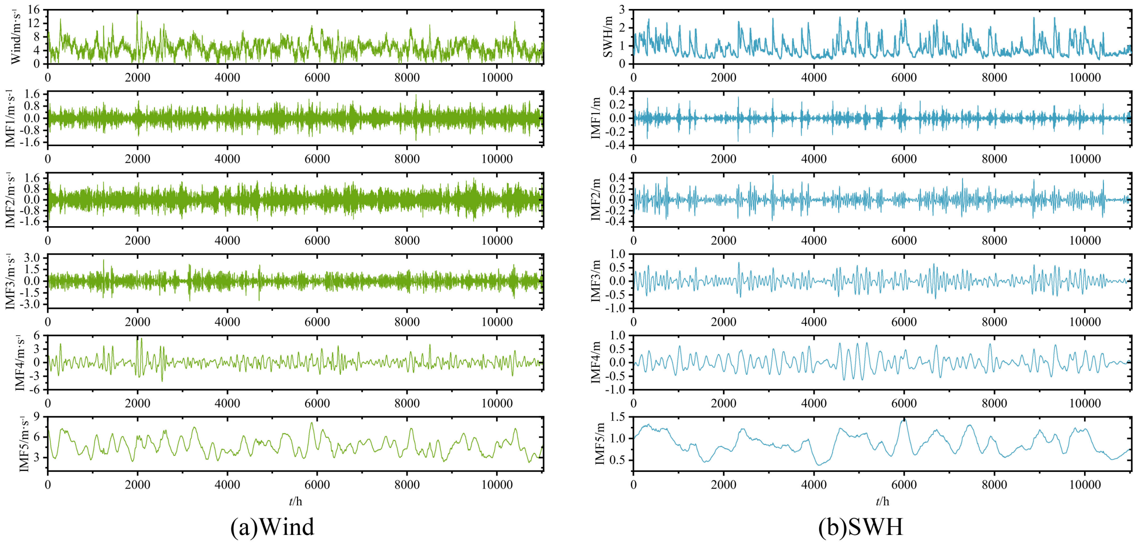
(c) Fall

(d) Winter

FIGURE 9
Seasonal distribution of significant wave heights over five years. **(A)** Spring, **(B)** Summer, **(C)** Fall, **(D)** Winter.

wave height; Case 2 includes 12 vectors, specifically IMF1 to IMF5 of the wind field, the wind field, IMF1 to IMF5 of the effective wave height, and the original effective wave height.

According to the data in Table 2, the prediction results of Case 1 and Case 2 show significant differences across different seasons. In all seasons, the errors in Case 1 are generally smaller than those in Case 2, indicating that decomposing only the effective wave height better captures its intrinsic features, while introducing the IMF components of the wind field increases the model's complexity, leading to greater errors. Notably, the computation time for Case 1



(a)Wind

(b)SWH

FIGURE 10
VMD decomposition of significant wave height sequence. **(A)** Wind, **(B)** SWH.

TABLE 2 VMD decomposition signal impact.

Season	Evaluation	Case 1	Case 2
Spring	MAE/m	0.0147	0.0214
	RMSE/m	0.0202	0.0285
	R^2	0.9981	0.9962
	CPU time/s	308	3544
Summer	MAE/m	0.0112	0.0197
	RMSE/m	0.0147	0.0258
	R^2	0.9980	0.9938
	CPU time/s	332	3345
Fall	MAE/m	0.0228	0.0340
	RMSE/m	0.0306	0.0491
	R^2	0.9971	0.9925
	CPU time/s	311	3597
Winter	MAE/m	0.0197	0.0273
	RMSE/m	0.0268	0.0363
	R^2	0.9977	0.9958
	CPU time/s	254	5098

is significantly shorter than for Case 2, especially in winter, where the CPU time for Case 2 is more than 10 times that of Case 1. This further suggests that introducing the IMF components of the wind field not only increases the model's computational complexity but also significantly prolongs the computation time. Therefore, in the subsequent predictions, to simplify the computation and improve model efficiency, only the effective wave height data, which has a more significant impact on the predictions, will be decomposed.

4.2 Univariate prediction

This experiment used data from spring, summer, autumn, and winter between 2017 and 2021 as model driving data, with data from 2017 to 2020 used as the training set and data from 2021 as the test set. In the univariate model, only significant wave height is used as the input parameter for the BiLSTM, CNN-BiLSTM, and VMD-CNN-BiLSTM models.

Figure 11 compares the univariate predictions of BiLSTM, CNN-BiLSTM, and VMD-CNN-BiLSTM models with the WW3-SWAN simulation values. As shown in Figure 11, the bidirectional LSTM (BiLSTM) is capable of considering both past and future information and performs well in predicting the overall trend, especially in periods with smaller fluctuations. However, in regions of sharp changes in wave peaks and troughs (as indicated by the black boxes in Figure 11), BiLSTM shows significant errors compared to the WW3-SWAN values. This may be due to BiLSTM's tendency to over-smooth the predictions during periods of sharp fluctuations. In contrast, the CNN-BiLSTM model is more effective at capturing the short-term fluctuations of wave peaks, particularly in areas of peak changes, outperforming BiLSTM. However, CNN-BiLSTM is less effective at capturing troughs, possibly due to limitations in its ability to extract local features. By decomposing the significant wave height data using VMD, the model can effectively extract important frequency components, and combined with CNN's ability to extract local features, it significantly improves prediction accuracy in areas of sharp changes in wave peaks and troughs. Overall, the VMD-CNN-BiLSTM model performs best in capturing changes in wave peaks and troughs.

Figure 12 and Table 3 compare the error metrics of the three models (BiLSTM, CNN-BiLSTM, and VMD-CNN-BiLSTM) in univariate significant wave height prediction, including mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). The left y-axis of Figure 11 represents the specific values of MAE, RMSE, and R^2 for each model, while the right y-axis shows the relative values of each model compared to BiLSTM. Negative values of MAE and RMSE indicate that the model performs better than BiLSTM, while positive values indicate poorer performance; for R^2, larger positive values indicate better prediction accuracy. The results show that the VMD-CNN-BiLSTM model's error is significantly lower than the other two models, especially in regions of sharp changes in wave peaks and troughs. Across all seasons, the VMD-CNN-BiLSTM model demonstrates the best prediction performance, particularly in the autumn and winter seasons, where complex wave height changes caused by typhoons and strong monsoons are present. For example, in the spring season, the MAE of the VMD-CNN-BiLSTM is 0.0159 meters, a 51.23% reduction compared to BiLSTM; across the entire year, the RMSE of the VMD-CNN-BiLSTM is 0.0256 meters, a 62.30% reduction compared to BiLSTM. Furthermore, the R^2 of the VMD-CNN-BiLSTM is the highest across all seasons and in annual statistics, reaching 0.9979 in the spring, a 1.04% improvement compared to BiLSTM. This indicates that the VMD-CNN-BiLSTM model has a stronger correlation between the predicted results and the actual observations, reflecting the actual wave height changes more accurately. Therefore, the MAE and RMSE of the VMD-CNN-BiLSTM model are significantly lower than those of other models across different seasons, indicating superior performance in capturing wave peaks and troughs. The higher R^2 value further demonstrates the model's advantage in trend prediction, particularly in handling complex fluctuations.

4.3 Multivariate forecasting

Due to the influence of the monsoon climate in the region, this study conducted a multivariate prediction research to further improve prediction accuracy. In the experiment, both wind speed and significant wave height were used as parameters for the prediction model, considering the impact of the wind field. Similarly, data from spring, summer, autumn, and winter between 2017 and 2021 were used as model driving data, with data from 2017 to 2020 as the training set and data from 2021 as the test set.

Figure 13 presents the comparison curves of multivariate predictions from BiLSTM, CNN-BiLSTM, and VMD-CNN-BiLSTM models with WW3-SWAN simulation values. The figure shows that

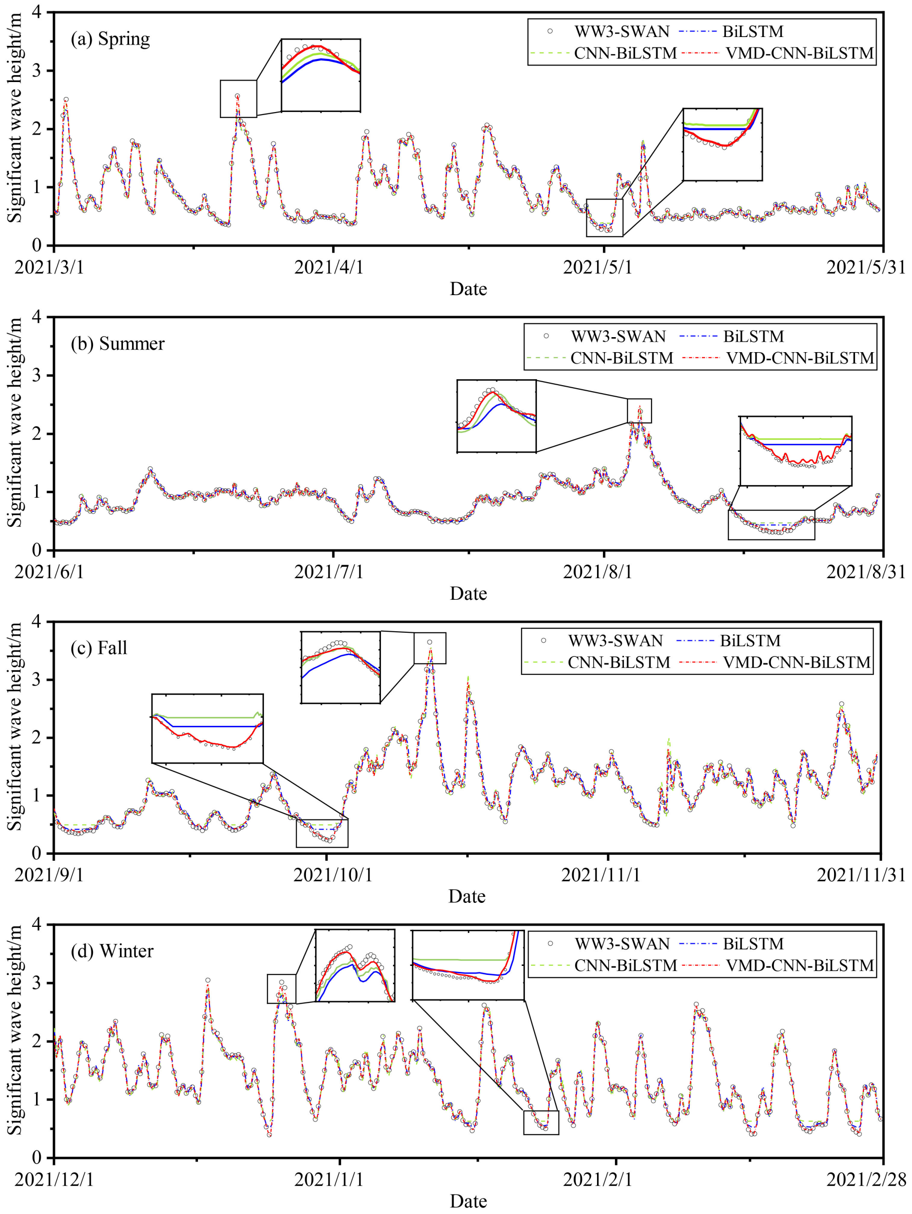
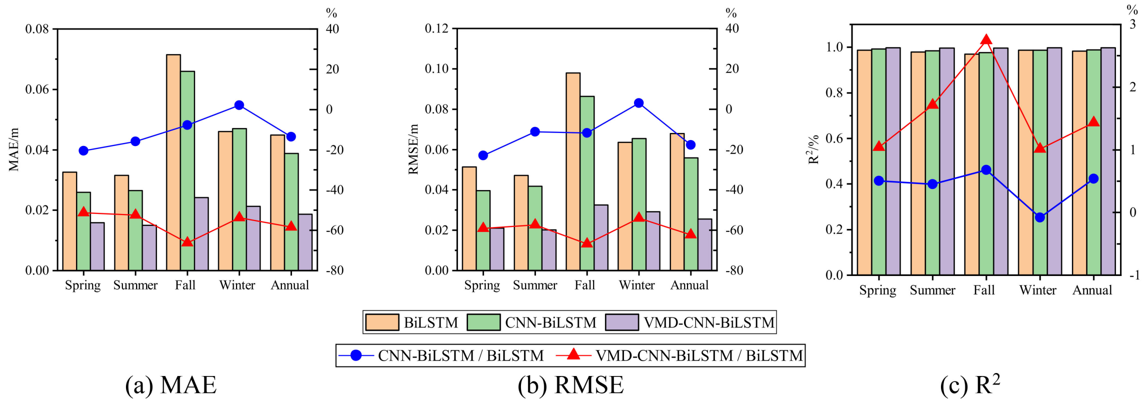


FIGURE 11
The WW3-SWAN simulated and predicted values of SWH of univariate. **(A)** Spring, **(B)** Summer, **(C)** Fall, **(D)** Winter.

the monsoon climate significantly affects the wind field and wave height variations in the Southeast China Sea, especially in summer and autumn, where the frequent occurrence of typhoons exacerbates the complexity of wave height changes. Therefore, considering multivariate factors such as the wind field is crucial for improving the accuracy of wave height predictions. Spring is a transitional period from the winter to summer wind directions, with complex wind field changes, especially during the impact of typhoon "Shuriki," where wind speed and wave height fluctuations significantly increase. Figure 13A shows that, compared to univariate predictions, multivariate predictions more accurately capture the overall trend of spring wave heights. Notably, the VMD-CNN-BiLSTM model, by effectively integrating instantaneous changes in the wind field, can accurately predict wave peak and trough changes, with its prediction curve highly aligning with WW3-SWAN simulation values, demonstrating high prediction accuracy. In summer, the prevailing southeast monsoon leads to a relatively stable wind field. Figure 13B indicates that, compared to univariate predictions, the VMD-CNN-BiLSTM model performs particularly well when considering wind field factors, with its prediction curve closely matching the WW3-SWAN simulation values, especially during August 17 to August 20, when VMD-CNN-BiLSTM accurately captures the characteristics of wave troughs. Autumn is a transitional period from summer to winter winds, with frequent typhoons and significant wave height changes. Figure 13C shows that the VMD-CNN-BiLSTM model better utilizes the intense changes in wind field data to accurately capture extreme wave peak values. The model performs excellently under extreme weather conditions such as typhoons, with its prediction curve closest to the WW3-SWAN



(a) MAE (b) RMSE (c) R^2

FIGURE 12
The WW3-SWAN simulated and predicted values of SWH of multivariate. **(A)** Spring, **(B)** Summer, **(C)** Fall, **(D)** Winter.

simulation values. In winter, the Northeast monsoon prevails in the Southeast China Sea, with strong winds and long durations, resulting in higher overall wave height levels and frequent fluctuations. Figure 13D shows that in winter, the VMD-CNN-BiLSTM model effectively captures the overall trend and local fluctuations of wave heights, with its prediction curve highly consistent with the WW3-SWAN simulation values, demonstrating the best prediction performance.

Figure 14 and Table 4 show the error performance of multivariate significant wave height prediction models (BiLSTM, CNN-BiLSTM, and VMD-CNN-BiLSTM) in different seasons and annual statistics, including mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2), as well as the ratios of each model relative to BiLSTM. The influence of the Southeast China Sea monsoon climate and typhoons was considered to evaluate each model's performance under complex meteorological conditions. The data in Figure 14 and Table 3 indicate that, within the same season, the VMD-CNN-BiLSTM model has significantly lower errors than the other two models. Particularly, after considering multivariate factors such as the wind field, the prediction performance of VMD-CNN-BiLSTM has significantly improved. Seasonal differences in prediction performance indicate that VMD-CNN-BiLSTM performs exceptionally well in autumn and winter, accurately capturing the drastic wave height changes brought by typhoons and strong monsoons. VMD-CNN-BiLSTM shows optimal performance in MAE and RMSE across all seasons, indicating that this model

TABLE 3 Statistics of univariate SWH prediction error.

Season	Error	BiLSTM	CNN-BiLSTM	VMD-CNN-BiLSTM
Spring	MAE/m	0.0326	0.0259 (-20.55)	0.0159 (-51.23)
	RMSE/m	0.0514	0.0396 (-22.96)	0.0210 (-59.14)
	R^2/%	0.9876	0.9926 (0.51)	0.9979 (1.04)
Summer	MAE/m	0.0315	0.0265 (-15.87)	0.0150 (-52.38)
	RMSE/m	0.0471	0.0418 (-11.25)	0.0201 (-57.32)
	R^2/%	0.9794	0.9838 (0.45)	0.9962 (1.72)
Fall	MAE/m	0.0715	0.0660 (-7.69)	0.0242 (-66.15)
	RMSE/m	0.0979	0.0863 (-11.85)	0.0325 (-66.80)
	R^2/%	0.9701	0.9767 (0.68)	0.9967 (2.74)
Winter	MAE/m	0.0460	0.0470 (2.17)	0.0213 (-53.70)
	RMSE/m	0.0635	0.0654 (2.99)	0.0292 (-54.02)
	R^2/%	0.9873	0.9865 (-0.08)	0.9973 (1.01)
Annual	MAE/m	0.0449	0.0388 (-13.58)	0.0187 (-58.35)
	RMSE/m	0.0679	0.0558 (-17.82)	0.0256 (-62.30)
	R^2/%	0.9836	0.9889 (0.54)	0.9977 (1.43)

The values in parentheses represent the percentage improvement of each model's performance indicators compared to BiLSTM.

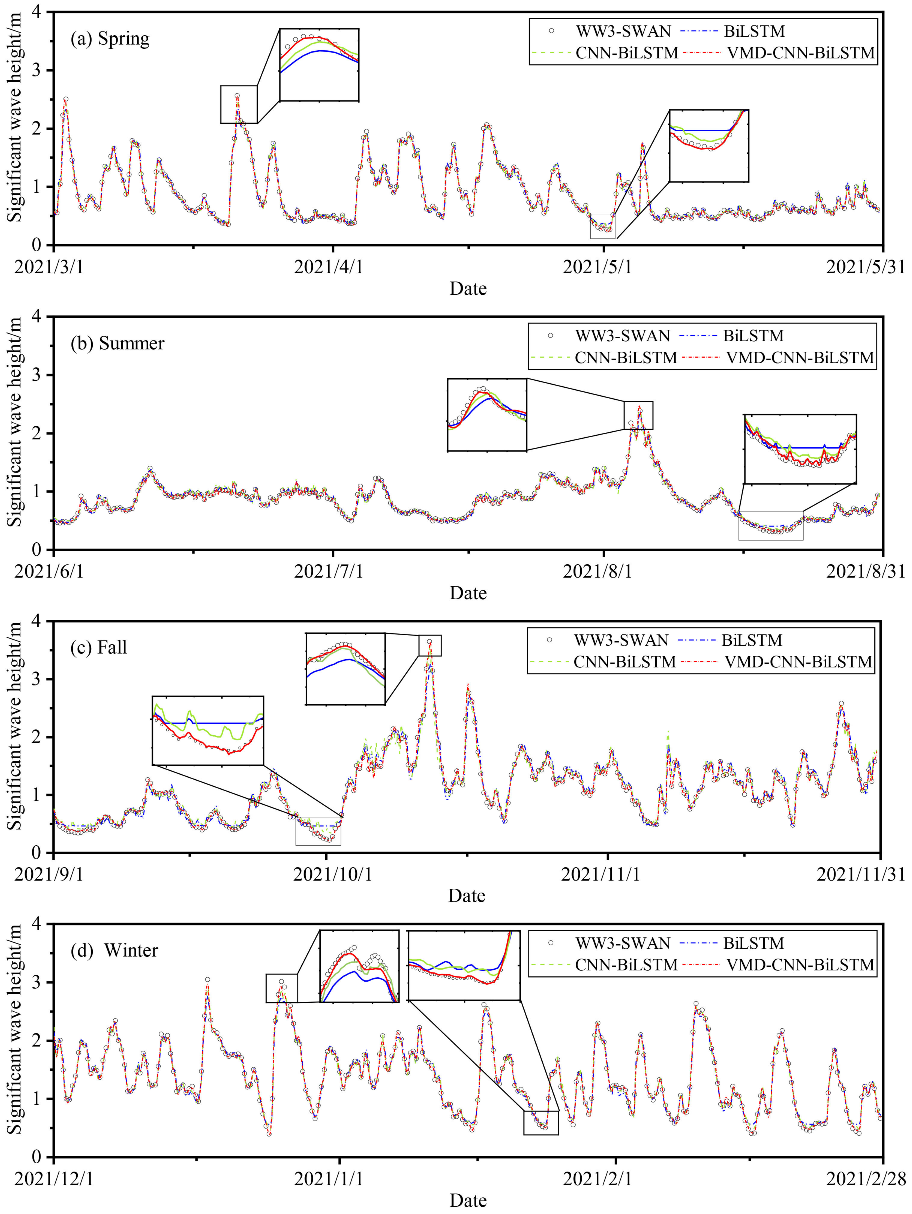


FIGURE 13
The WW3-SWAN simulated and predicted values of SWH of multivariate.

significantly outperforms BiLSTM and CNN-BiLSTM in multivariate prediction accuracy. For example, in spring, the MAE of VMD-CNN-BiLSTM is 0.0147 meters, a 51.48% reduction compared to BiLSTM; in annual statistics, the RMSE of VMD-CNN-BiLSTM is 0.0244 meters, a 61.81% reduction compared to BiLSTM. Additionally, the R^2 value of VMD-CNN-BiLSTM is the highest across all seasons and annual statistics, reaching 0.9981 in spring, a 0.79% improvement compared to BiLSTM, indicating stronger correlation and consistency in multivariate predictions. The inclusion of wind speed significantly improved the predictive performance of the VMD-CNN-BiLSTM model, particularly under complex meteorological conditions in autumn and winter, resulting in lower MAE and RMSE, as well as higher R^2. This indicates that the model is more effective at capturing the complex relationship between wind fields and wave heights, thereby enhancing the accuracy and stability of wave height predictions.

5 Conclusion

This study employs Variational Mode Decomposition (VMD) to extract significant features of significant wave height as intrinsic mode functions, combines Convolutional Neural Networks (CNN) to capture complex internal mappings of wind and waves, and integrates with Bidirectional Long Short-Term Memory (BiLSTM) networks to establish the VMD-CNN-BiLSTM model. The research focuses on the Southeast China Sea, with datasets provided by

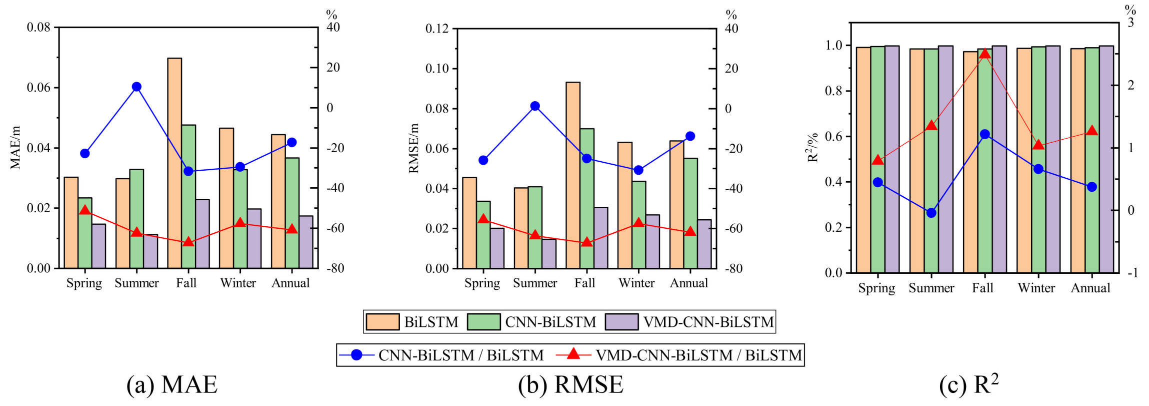


FIGURE 14
Multivariate significant wave height prediction error, **(A)** MAE, **(B)** RMSE, **(C)** R^2.

ECMWF and WW3-SWAN simulations. The case study and prediction results lead to the following conclusions:

1. Compared to models like BiLSTM and CNN-BiLSTM, the VMD-CNN-BiLSTM model is able to more accurately capture the peaks and smooth trends of wave height, resulting in higher prediction accuracy.
2. After incorporating wind field data, the MAE and RMSE of each prediction model decrease. Specifically, the VMD-CNN-BiLSTM model's MAE and RMSE are reduced to 0.0174 meters and 0.0244 meters respectively for annual statistics, with the coefficient of determination (R^2) increasing to 0.9979, outperforming other prediction models.
3. The VMD-CNN-BiLSTM model exhibits optimal prediction performance across all four seasons, particularly in winter under the influence of strong northeast monsoons and during summer and autumn when typhoons and extreme weather events occur. Its prediction performance significantly surpasses that of BiLSTM and CNN-BiLSTM models, demonstrating the model's excellent adaptability to complex sea conditions.

TABLE 4 Statistics of multivariate SWH prediction error.

Season	Evaluation	BiLSTM	CNN-BiLSTM	VMD-CNN-BiLSTM
Spring	MAE/m	0.0303	0.0234 (-22.77)	0.0147 (-51.48)
	RMSE/m	0.0455	0.0337 (-25.93)	0.0202 (-55.60)
	R^2/%	0.9903	0.9947 (0.44)	0.9981 (0.79)
Summer	MAE/m	0.0298	0.0329 (10.40)	0.0112 (-62.42)
	RMSE/m	0.0404	0.0409 (1.24)	0.0147 (-63.61)
	R^2/%	0.9848	0.9844 (-0.04)	0.9980 (1.34)
Fall	MAE/m	0.0697	0.0476 (-31.71)	0.0228 (-67.29)
	RMSE/m	0.0932	0.0699 (-25.00)	0.0306 (-67.17)
	R^2/%	0.9729	0.9847 (1.21)	0.9971 (2.49)
Winter	MAE/m	0.0465	0.0328 (-29.46)	0.0197 (-57.63)
	RMSE/m	0.0631	0.0437 (-30.74)	0.0268 (-57.53)
	R^2/%	0.9875	0.9940 (0.66)	0.9977 (1.03)
Annual	MAE/m	0.0444	0.0367 (-17.34)	0.0174 (-60.81)
	RMSE/m	0.0639	0.0551 (-13.77)	0.0244 (-61.81)
	R^2/%	0.9855	0.9892 (0.38)	0.9979 (1.26)

The values in parentheses represent the percentage improvement of each model's performance indicators compared to BiLSTM.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

WS: Data curation, Methodology, Visualization, Writing – original draft. ZY: Conceptualization, Methodology, Writing – review & editing. YZ: Formal analysis, Validation, Writing – review & editing. XW: Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We are very grateful to ECMWF for providing the wind field data, GEBCO for providing the bathymetry data, and CCCC-Fourth Harbor Engineering Institute Co., Ltd. for providing the wave buoy data.

Conflict of interest

Authors WS and YZ were employed by company CCCC Fourth Harbor Engineering Institute Co., Ltd. Authors ZY and XW were employed by companies CCCC Fourth Harbor Engineering Institute Co., Ltd. and CCCC.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abed-Elmdoust, A., and Kerachian, R. (2012). Wave height prediction using the rough set theory. *Ocean Eng.* 54, 244–250. doi: 10.1016/j.oceaneng.2012.07.020

Ahmed, A. A. M., Jui, S. J. J., Al-Musaylh, M. S., Raj, N., Saha, R., Deo, R. C., et al. (2024). Hybrid deep learning model for wave height prediction in Australia's wave energy region. *Appl. Soft Comput.* 150, 111003. doi: 10.1016/j.asoc.2023.111003

Berbíc, J., Ocvirk, E., Carevíc, D., and Loňcar, G. (2017). Application of neural networks and support vector machine for significant wave height prediction. *Oceanologia* 59.3, 331–349. doi: 10.1016/j.oceano.2017.03.007

Booij, N., Holthuijsen, L., and Ris, R. (1996). The "Swan" Wave model for shallow water. *Coast. Eng.* 1996, 668–676. doi: 10.1061/9780784402429.0

Bottcher, A. B., Whiteley, B. J., James, A. I., and Hiscock, J. G. (2012). Watershed assessment model (WAM): model use, calibration, and validation. *Trans. ASABE* 55.4, 1367–1383. doi: 10.13031/2013.42248

Deo, M. C., Jha, A., Chaphekar, A. S., and Ravikant, K. (2001). Neural networks for wave forecasting. *Ocean Eng.* 28.7, 889–898. doi: 10.1016/S0029-8018(00)00027-5

Fan, C., Wang, X., Zhang, X., and Gao, D. (2019). A newly developed ocean significant wave height retrieval method from Envisat AS-AR wave mode imagery. *Acta Oceanologica Sin.* 38, 120–127. doi: 10.1007/s13131-019-1480-2

Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* 3, 115–143.

Hao, P., Li, S., and Gao, Y. (2023). Significant wave height prediction based on deep learning in the South China Sea. *Front. Mar. Sci.* 9, 1113788. doi: 10.3389/fmars.2022.1113788

Kaloop, M. R., Kumar, D., Zarzoura, F., Roy, B., and Hu, J. W. (2020). A wavelet—Particle swarm optimization—Extreme learning machine hybrid modeling for significant wave height prediction. *Ocean Eng.* 213, 107777. doi: 10.1016/j.oceaneng.2020.107777

Li, J., and Hu, Y. (2021). Assessment of typhoons in ERA-Interim and ERA-5 reanalysis datasets. *Hydro-Science Eng.* (2021), 62–69. doi: 10.12170/20200222001

Liang, B., Gao, H., and Shao, Z. (2019). Characteristics of global waves based on the third-generation wave model SWAN. *Mar. Structures* 64, 35–53. doi: 10.1016/j.marstruc.2018.10.011

Liu, Q., Rogers, W. E., Babanin, A. V., Young, I. R., Romero, L., Zieger, S., et al. (2019). Observation-based source terms in the third-generation wave model WAVEWATCH III: Updates and verification. *J. Phys. Oceanogr.* 49, 489–517. doi: 10.1175/JPO-D-18-0137.1

Lixiang, S., Yunyue, C., Zonghui, M., and Songgui, C. (2023). Research on wave height prediction in Longkou sea area based on Attention-LSTM network. *J. Waterway Harbor* 44, 196–201. doi: 10.3969/j.issn.1005-8443.2023.02.006

Mentaschi, L., Besio, G., Cassola, F., and Mazzino, A. (2015). Performance evaluation of wavewatch III in the Mediterranean sea. *Ocean Model.* 90, 82–94. doi: 10.1016/j.ocemod.2015.04.003

Mikolov, T., Kombrink, S., Burget, L., Cernocky, J., and Khudanpur, S. (2021). "Extensions of recurrent neural network language model," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 5528–5531 (Piscataway, NJ, USA: IEEE).

Minghao, H., Lingling, X., Mingming, L., and Peng, L. (2024). The application of the Rayleigh parameter in machine learning prediction of wave height. *Oceanologia Limnologia Sin.* 55, 318–331. doi: 10.11693/hyhz20230900180

Pradnya, D., and Londhe, S. (2016). Prediction of extreme wave heights using neuro wavelet technique. *Applied Ocean Research.* 58, 241–252. doi: 10.1016/j.apor.2016.04.011

Pushpam, P. M. M., and Enigo V.S., F. (2020). "Forecasting significant wave height using RNN-LSTM models," in *Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 1141–1146 (New York: IEEE).

Rogers, W. E., Hwang, P. A., and Wang, D. ,. W. (2003). Investigation of wave growth and decay in the SWAN model: three regional-scale applications. *Oceanography* 33, 366–389. doi: 10.1175/1520-0485(2003)033<0366:IOWGAD>2.0.CO;2

Simmons, H. L., Jayne, S. R., Laurent, L. C. S., and Weaver, A. J. (2004). Tidally driven mixing in a numerical model of the ocean general circulation. *Ocean Model.* 6.3-4, 245–263. doi: 10.1016/S1463-5003(03)00011-8

Swain, J., Umesh, P. A., and Balchand, A. N. (2019). WAM and WAVEWATCH-III intercomparison studies in the North Indian Ocean using Oceansat-2 Scatterometer winds. *J. Ocean Climate* 9, 1–24. doi: 10.1177/2516019219866569

Tolman, H. L. (2009). User manual and system documentation of WAVEWATCH III TM version 3.14. *Tech. Note MMAB Contrib.* 276.

Yifan, Q., Feng, L., and Jie, Z. (2024). Research on deep learning models for predicting significant wave height. *Mar. Sci. Bull.* 43, 382–390. doi: 10.11840/j.issn.1001-6392.2024.03.00 9

Yin, J., Wang, H., Wang, N., and Wang, X. (2023). An adaptive real-time modular tidal level prediction mechanism based on EMD and Lipschitz quotients method. *Ocean Eng.* 289, 116297. doi: 10.1016/j.oceaneng.2023.116297

Yu, T., and Wang, J. (2021). A spatiotemporal convolutional gated recurrent unit network for mean wave period field forecasting. *J. Mar. Sci. Eng.* 9.4, 383. doi: 10.3390/jmse9040383

Zhang, X., Li, Y., and Gao, S. (2021). Ocean wave height series prediction with numerical long short-term memory. *J. Mar. Sci. Eng.* 9.5, 514. doi: 10.3390/jmse9050514

TYPE Original Research
PUBLISHED 11 December 2024
DOI 10.3389/fmars.2024.1499231

Check for updates

OPEN ACCESS

EDITED BY
Phoebe Koundouri,
Athens University of Economics and Business, Greece

REVIEWED BY
Guanqiu Qi,
Chung-Ang University, Republic of Korea
Weiliang Qiao,
Dalian Maritime University, China

*CORRESPONDENCE
Shiguan Liao
sgliao0113@szpu.edu.cn

RECEIVED 20 September 2024
ACCEPTED 26 November 2024
PUBLISHED 11 December 2024

CITATION
Li Z, Zhu X, Liao S, Gao K and Hu S (2024) Risk performance analysis model of escort operation in Arctic waters via an integrated FRAM and Bayesian network.
Front. Mar. Sci. 11:1499231.
doi: 10.3389/fmars.2024.1499231

Risk performance analysis model of escort operation in Arctic waters via an integrated FRAM and Bayesian network

Zhuang Li[1], Xiaoming Zhu[2], Shiguan Liao[3]*, Kaixian Gao[1] and Shenping Hu[2]

[1]Naval Architecture and Shipping College, Guangdong Ocean University, Zhanjiang, China, [2]Merchant Marine College, Shanghai Maritime University, Shanghai, China, [3]School of Management, Shenzhen Polytechnic University, Shenzhen, China

Escort operation is an effective mean to ensure the safety of ship navigation in the Arctic ice area and expand the window period for ship navigation. At the same time, the operation mode between icebreaker and escorted ship may also causes collision accident. In order to scientifically reflect the complex coupling relationship in the escort operation system in Arctic waters and effectively manage the navigation risks. This study proposes to use the functional resonance analysis method (FRAM) to identify the risk factors of ship escort operation in Arctic waters, and uses the Bayesian network (BN) method to establish a risk assessment model for escort operation collision accident. The cloud model is used to process the uncertain data information. The proposed method is applied during the actual escort operation of a commercial ship on the Arctic Northeast Passage. According to the model simulation results, the risk performance of ship escort operation in Arctic waters is quantitatively analyzed, and the key risk causes are further analyzed. This study has positive significance for better understanding the risk evolution mechanism of ship escort operation in Arctic ice area and helping relevant management departments to take risk control measures.

KEYWORDS
risk performance, functional resonance accident model, Bayesian network, cloud model, escort operation, Arctic waters

1 Introduction

The shipping industry undertakes the transportation of most commodities in the world's foreign trade and has made great contributions to the industrial development of countries around the world (Otheitis and Kunc, 2015; Lenzen et al., 2023). In recent years, people have been seeking more economical and convenient ways of maritime transportation. The opening of the Arctic route has brought new opportunities for the development of the world's shipping industry (Ryan et al., 2021). The Arctic route has greatly shortened the transportation distance from Asia to Europe, more and more commercial ships have begun to try to use the Arctic route for ocean-going cargo transportation. Due to the influence of extreme natural phenomena such as sea ice and strong winds in Arctic waters, there are great safety issues in conducting commercial navigation in Arctic waters. In order to ensure the safety of ships in Arctic waters, most ships will choose icebreaking escort services during navigation in the Arctic ice area. Under the leadership of icebreakers, ordinary commercial ships pass through the ice area by following, which greatly improves the navigation safety level of ships in the Arctic ice area (Zhang et al., 2017). However, in this operation mode, how to effectively avoid collisions between icebreaker and escorted ship becomes the key to escort operation. Therefore, it is necessary to introduce new theories and methods to scientifically analyze and manage the collision accident risk during escort operation in Arctic waters.

For the escort operation in Arctic waters, it consists of a series of operational tasks, and there exists a complex spatial and temporal correlation between different tasks, which is a typical complex system. Process safety is a commonly used safety analysis method in the industrial field. It emphasizes that the occurrence of a dangerous event will not only affect the current situation, but also the safety of the next link and subsequent processes. It is often used to analyze systemic risks (Amin et al., 2019; Behari, 2019). When quantitatively analyzing the risk of collision accident in ship escort operation in Arctic waters, it is necessary to combine the analytical ideas of process safety, introduce relevant complex system theoretical methods, characterize the complex coupling relationship in the ship escort operation system in Arctic waters, and clarify the law of changes in the navigation risks of ships during escort operations.

In order to scientifically analyze the complex coupling mechanism and risk quantification characteristics of ship escort operation in Arctic waters, this study proposes to combine the functional resonance analysis method (FRAM) with the Bayesian network (BN) to establish a quantitative analysis model for the collision accident risk of ship escort operation in Arctic waters, and adopts the cloud model (CM) for uncertain information processing to quantitatively analyze the risk of escort operation in specific scenarios.

This study is organized as follows. Section 2 analyzes and summarizes the current research status. Section 3 describes in detail the proposed method for quantitative analysis of collision accident risks of ship escort operation in Arctic waters. Section 4 applies the proposed method in combination with the specific scenario of ship escort operation in Arctic waters. Section 5 discusses the methods and results proposed in this study. Section 6 is the conclusion of this study.

2 Literature review

2.1 Ship escort operation risk in Arctic waters

As seasonal navigation in Arctic waters becomes a reality, the issue of ship navigation risks has received a lot of attention (Khan et al., 2020; Yao et al., 2024; Kandel and Baroud, 2024). In order to effectively expand the navigation window in Arctic waters, icebreaker escort operation has become a common mean (Moe and Brigham, 2017; Zhang et al., 2018). Relevant studies on the ship escort operation risks in Arctic waters has continued to increase. Zhang et al. (2019a) analyzed the safety risks of ship formation operations from the perspective of safe distance and speed in the multi-ship following mode of escort operations in Arctic ice area. Fu et al. (2022) combined failure mode and effects analysis (FMEA) and FRAM to simulate the evolution of ship traffic accident scenarios in Arctic waters, and further evaluated the changes in the probability of navigation risks of nuclear-powered icebreakers in Arctic waters under independent navigation and escort operation modes. Xu and Kim (2023) established a hybrid causal logic model for ship–ship and ship–ice collision accidents in Arctic waters and conducted a quantitative evaluation of the risks of collision accidents. Xu et al. (2023) used ship networking technology to develop an intelligent micro-model to analyze the stability of ship formations based on the movement trend and sea ice conditions of ship formations in Arctic waters to ensure the safety of ship formation operations in Arctic waters. Zhu et al. (2024) identified the risk factors affecting ship escort operations in Arctic waters from the perspective of complex system theory, and combined the BN analysis method to predict and analyze the risk characterization of escort operations. According to the current research status, compared with the study of navigation risk under the state of independent navigation of ships, escort operations need to consider traffic accidents between ship formations under the coordinated navigation of multiple ships. Due to the complexity of the task of escort operations in Arctic waters, in the research related to the risk of escort operations in Arctic waters, relevant methods suitable for complex system analysis are often used to analyze the causes of accidents (Fu et al., 2022; Zhu et al., 2024).

2.2 FRAM method for risk analysis

When analyzing the risk of ship navigation, a combination of qualitative and quantitative analysis is often used. FRAM can effectively explain the deep logic of complex system accidents by analyzing system functions and focusing on the changes and coupling relationships of system functions when accidents occur, and is widely used in risk factor identification and correlation analysis (Tian et al., 2016; Li et al., 2019; Yousefi et al., 2019; Ma et al., 2023). França and Hollnagel (2023) used FRAM to model and analyze human factors in accidents when analyzing production accidents in the industrial field, and analyzed the main human factors affecting process safety. Liu et al. (2024a) combined FRAM and reinforcement learning to construct an emergency plan for blowout accidents, in which FRAM was used to simulate the emergency response process. Yu et al. (2024) analyzed the functions related to maritime accidents based on the maritime accident investigation report for grounding accidents in Arctic waters, and constructed a FRAM model to analyze the functional resonance of accidents. Zheng et al. (2024) designed a FRAM model consisting of four stages: understanding, designing, analyzing, and enhancing the response process, to provide a decision-making basis for emergency response to fuel storage accidents.

Ship navigation operations are a complex social system. Water traffic accidents have the characteristics of low probability and serious consequences, and the causes of accidents are often difficult to clarify. FRAM is suitable for analyzing the causes of complex system accidents and has been widely used in the study of water traffic accident risk analysis (Lee et al., 2020; Salihoglu and Besikci, 2021). The Arctic waters are well known for their harsh navigation environment (Abbassi et al., 2017; Yao et al., 2022; Li et al., 2023b). Ship escort operations increase the complexity of operational tasks in such harsh environments. FRAM can help analyze the occurrence process of escort operations in Arctic waters and clarify the complex interaction relationship between system modules when accidents occur.

2.3 BN risk quantification method

The BN analysis method can effectively integrate various risk factors and carry out BN reasoning by combining various subjective and objective data information. It is widely used in the quantitative reasoning of ship navigation risks (Baksh et al., 2018; Zhang et al., 2019; Zhang et al., 2019b; Xu and Kim, 2023; Li et al., 2024). BN analysis methods have also been widely used in the quantitative analysis of ship navigation risks in Arctic waters. Vanhatalo et al. (2021) combined AIS data, satellite data, and accident data to predict the probability risk of ship ice entrapment accidents in the Arctic ice zone using the BN modeling method. Wang et al. (2022) collected relevant data sets and conducted a quantitative analysis of environmental risks on the Arctic Northwest Passage based on the establishment of a dynamic BN risk assessment model. Fu et al. (2023) used an object-oriented Bayesian network (OOBN) analysis method to quantitatively analyze the risks of multi-type ship traffic accidents in the Arctic ice area. Afenyo et al. (2023) proposed a hybrid method based on the Bayesian loss function to evaluate the losses caused by oil spill accidents in Arctic waters. Liu et al. (2024b) quantitatively analyzed the propagation mechanism of the Arctic navigation network under uncertain interference based on a data-driven BN model and proposed corresponding resilience enhancement strategies.

In the quantitative analysis of ship navigation risks, the uncertainty information band has a great impact on the quantification of risks (Nguyen et al., 2021). Due to the particularity of the navigation environment in Arctic waters, these uncertainties are more intense. The advantage of the BN analysis method is that it can integrate multi-source and multi-category information, and can perform network reasoning on a few basis. In addition, the Bayesian analysis method mostly adopts the form of network reasoning, which can effectively reflect the correlation between risk factors of complex systems. Therefore, it is appropriate to use the BN analysis method for various uncertainties in the risk quantitative analysis of ship escort operations in Arctic waters.

2.4 Cloud model for uncertain information

Since the risk factors affecting the occurrence of ship traffic accidents are diverse, many of which often have no direct data source, information uncertainty is the main obstacle to the quantification of ship navigation risks. In the process of quantifying uncertain information, cloud model is an uncertain artificial intelligence method that can realize the conversion between qualitative judgment and quantification, and has received extensive attention from relevant scholars in recent years (Peng et al., 2017; Ma et al., 2022). Wang et al. (2015) proposed a decision-making method based on cloud model for multi-criteria decision-making problems with unknown decision-makers' weights. Guo et al. (2020) proposed a comprehensive evaluation method based on cloud model for the randomness and fuzziness of subjective information, which realized the simultaneous consideration of randomness and fuzziness in the process of quantification of subjective information. Cloud model has also been widely used in the study of water traffic accident risks. Wu et al. (2019) combined Markov chain and cloud model to quantitatively reason about the process risk of bauxite ships during maritime transportation. Liu et al. (2020) developed a ship collision accident risk reasoning system based on cloud model for the risk of ship collision accidents. Li et al. (2023a) used cloud model to quantify the uncertainty information in the evolution characteristics of the Maritime Autonomous Surface Ship navigation risk. Xi et al. (2024) integrated evidence theory and cloud model to fuse multi-source subjective information when analyzing human operational errors during ship icebreaking escort operations in Arctic waters. It can be seen that since cloud model can handle uncertain information with fuzziness and randomness, it is suitable for quantitative analysis of the uncertainty in ship navigation risks.

2.5 Contributions of this study

Ship escort operation in Arctic waters is a special operation mode. During the escort operation, the risk factors affecting different operation links are different, and there are complex coupling relationships between risk factors. These unique characteristics bring a series of new challenges to understanding the random phenomenon of system risks caused by the dynamic changes of risk factors and to quantitatively analyze the risks of escort operation processes. This study uses a system analysis method to analyze the dynamic behavior and nonlinear coupling effects in the process risks of complex technical systems in escort operations in Arctic waters. By introducing FRAM into escort operation in Arctic waters, key events or factors affecting the risks of escort operations are identified. Based on the network topological relationship between the risk factors of escort operations in Arctic waters, the cloud model is used to quantify the uncertain information, and the BN analysis method based on conditional probability reasoning is used to achieve quantitative analysis of escort operation process risks.

3 Methodology

3.1 A hybrid approach for escort operation risk assessment

The main framework of the risk assessment method for escort operation in Arctic waters proposed by this study is shown in Figure 1. In this method, the complexity analysis method FRAM is used to identify risk factors, and the BN analysis method is used to establish a risk assessment model. For the uncertain information in the risk factors, the cloud model is proposed to achieve quantitative processing of subjective information. The main contents of the proposed method are divided into the following three steps.

1. By analyzing the main behavioral characteristics of the escort operation system composed of icebreaker and escorted ship, the escort operation system is converted into multiple functional modules. The characteristics of each functional module are described from six aspects: I, O, C, P, R, and T. By analyzing the main factors leading to functional failure, the failure links between functional modules are determined, and the functional failure network diagram is completed by combining chain connections. In-depth analysis of each functional module is carried out to find out the potential key risk factors that lead to accident.
2. According to the risk factors identified by the FRAM model, with the help of expert knowledge or reference to other literature, the influence relationship between risk factors is judged. With risk factors as network nodes, the influence relationship between factors forms directed edges, and the risk assessment BN structure is preliminarily constructed. On this basis, according to the type of each node in the BN, the model input required for each risk factor is judged as *a priori* probability table or conditional probability table.
3. For the model input required for each risk factor in the risk assessment BN model, it is judged whether it has an objective data source. For risk factors with direct data sources, their prior probability tables are obtained through probability statistics. For risk factors without direct data sources, the cloud model is used to quantify expert knowledge, thereby further obtaining its prior probability table. On this basis, the BN reasoning is used to quantitatively assess the risk of escort operation collision accident in Arctic waters.

3.2 Escort operation risk in Arctic waters

The risk of ship navigation is often understood as a combination of the possibility of a ship traffic accident and the severity of the accident consequences. In the research related to the risk assessment of ship navigation in polar waters, the probability of a certain type of accident is often quantitatively inferred (Afenyo et al., 2021). This probabilistic risk assessment method has been widely used. Escort operation in Arctic waters is a formation consisting of an icebreaker and a following escorted ship. During the escort operation, collision accident between ship formations is the main threat faced during navigation. The risk of collision accident is often caused by the complex coupling of multiple risk factors (Khan et al., 2020), which can be expressed by Equation 1. How to

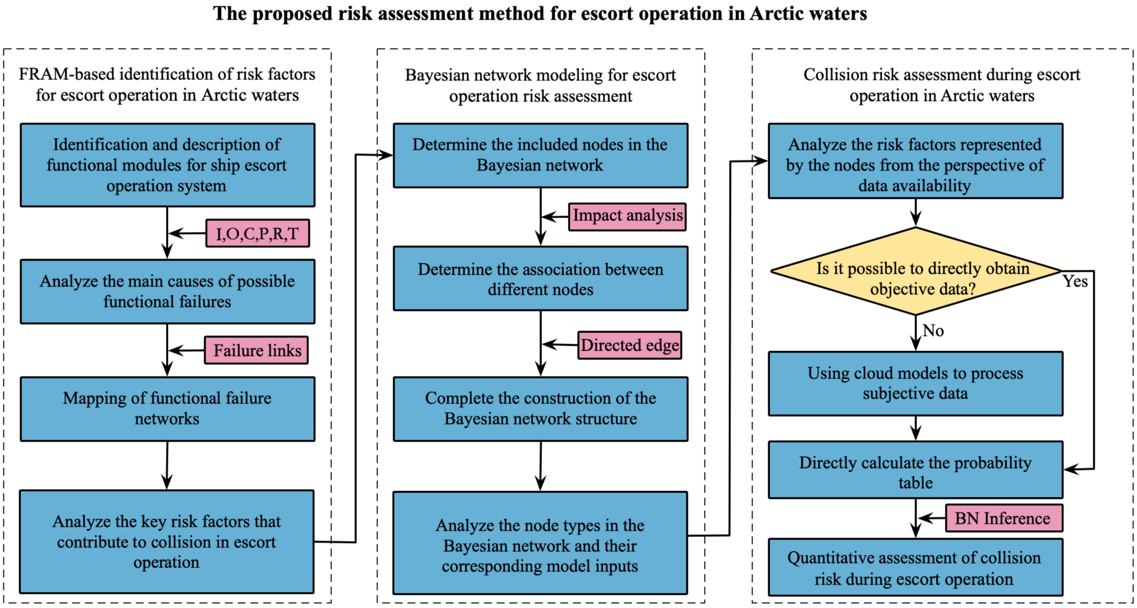

The proposed risk assessment method for escort operation in Arctic waters
FRAM-based identification of risk factors for escort operation in Arctic waters
Identification and description of functional modules for ship escort operation system
I,O,C,P,R,T
Analyze the main causes of possible functional failures
Failure links
Mapping of functional failure networks
Analyze the key risk factors that contribute to collision in escort operation
Bayesian network modeling for escort operation risk assessment
Determine the included nodes in the Bayesian network
Impact analysis
Determine the association between different nodes
Directed edge
Complete the construction of the Bayesian network structure
Analyze the node types in the Bayesian network and their corresponding model inputs
Collision risk assessment during escort operation in Arctic waters
Analyze the risk factors represented by the nodes from the perspective of data availability
Is it possible to directly obtain objective data?
Yes
No
Using cloud models to process subjective data
Directly calculate the probability table
BN Inference
Quantitative assessment of collision risk during escort operation

FIGURE 1
Methodology framework.

objectively reflect the complex correlation between risk factors and use quantitative analysis methods to evaluate the risk of collision accidents between ship formations is the key issue to be solved in this study.

$$R = R_1 \otimes R_2 \otimes \cdots \otimes R_n \tag{1}$$

3.3 FRAM method

FRAM is a method for describing and analyzing the functions and activities of socio-technical systems based on the perspective of functional resonance (Patriarca et al., 2017; Kim and Yoon, 2021). In the process of using FRAM for accident investigation and risk assessment, the focus is mainly on the connection between the subcomponents of the system. The system is often decomposed into multiple functional modules for identification and analysis, which is conducive to the positive control of system safety. The functional modules of FRAM are generally represented in the form of functional hexagons, representing six aspects of information: input (I), output (O), control (C), premise (P), resource (R) and time (T). These aspects of information are connected using chains according to the specific situation of the system to complete the construction of the FRAM model.

Safety-II focuses on the safety of the system and emphasizes how to make the system safer through various measures. As a system method from the Safety-II perspective, FRAM not only analyzes the complex nonlinear coupling between system functional modules, but also emphasizes the internal changes of the system based on human, technical and management factors, and can achieve good phenomenon expression. When using FRAM to analyze the collision accident risk of escort operations in Arctic waters, the construction of the FRAM model mainly follows the following steps.

1. Identify and describe the basic functions of the system. For the ship escort operation system in Arctic waters, it is converted into multiple functional modules through analysis. From the functional characteristics of I, O, C, P, R, and T, the characteristics of each functional module and the relationship between the functions are further described.
2. Analyze the potential changes of each functional module of the system. Analyze the main factors that may cause functional failure from the two aspects of the system inside and outside.
3. Draw a functional failure network diagram. Analyze the coupling relationship between the functional modules of the ship escort operation system in Arctic waters, determine the failure links, and complete the drawing of the functional failure network diagram by combining the chain connection.
4. Analyze the cause of the accident. Combined with the functional failure network diagram, deeply analyze each functional module to find out the key risk factors leading to collision accidents in Arctic waters escort operation.

3.4 BN analysis method

A BN structure is a directed acyclic graph consisting of variable nodes and directed edges. It can intuitively reflect the causal relationship between factors, and can perform network learning and reasoning under limited information. It is often used to assess and predict the risk of ship traffic accidents (Sakar et al., 2021; Basnet et al., 2023). In the process of constructing a BN model, once the nodes, directed edges, and probability tables are determined, the BN model is established.

When constructing the BN model of escort operation collision accident risk in Arctic waters, the risk factors that affect the occurrence of accident represent nodes, and the coupling relationship between risk factors is represented by directed edges. The relationship strength between nodes is represented by a conditional probability table. Nodes without parent nodes need to use a bright probability table to express information. In this study, the nodes and directed edges in the BN can be obtained through the constructed FRAM model. The conditional probability table is generally obtained through logical judgment. The prior probability table often needs to be obtained through statistical calculations based on the objective discrete data set of risk factors. However, in practice, some risk factors often have missing data. Expert knowledge is the main means to make up for the missing data, but it is often necessary to use certain methods to convert subjective information into objective data information (Chen et al., 2022).

3.5 Cloud model

The cloud model is often used for the quantitative processing of uncertain information. It can complete the conversion between qualitative concepts and quantitative data. The cloud model uses three characteristic quantities: expectation (*Ex*), entropy (*En*) and hyperentropy (*He*) to express the randomness and fuzziness of qualitative concepts, and overall reflects the quantitative expression of qualitative concepts (Li et al., 2022). Among them, the mathematical expression of *Ex* is shown in Equation 2, which is similar to the concept of mean and represents its concentration. X_i represents the subjective evaluation value of the expert on the uncertain information. n represents the number of evaluation values. The mathematical expression of *En* is shown in Equation 3, which represents the uncertainty of the target of interest. The larger its value, the wider the cloud droplet. The mathematical expression of *He* is shown in Equation 4, which represents the discreteness of *En*. The larger its value, the wider the thickness of the cloud droplet. In the field of ship navigation risk management research, cloud models are often used to quantify the uncertainty information of risk factors (Li et al., 2023a; Xi et al., 2024).

$$E_x = \sum_{i=1}^{n} X_i / n \tag{2}$$

$$E_n = \sqrt{\frac{\pi}{2}} \sum_{i=1}^{n} (X_i - E_x) / n \tag{3}$$

$$H_e = \sqrt{\left|\sum_{i=1}^{n}[(X_i - E_x)^2 - {E_n}^2]/(n-1)\right|} \quad (4)$$

In order to complete the conversion from qualitative description to quantitative expression based on the three characteristic quantities of *Ex*, *En* and *He*, a forward cloud generator is often used to implement this function. For a random realization of a risk factor, its membership function needs to satisfy Equation 5. Through multiple random accumulations of cloud droplets, a large number of cloud droplets can be generated to form a cloud model.

$$\mu(x) = exp[-(x - E_x)^2/2E_n^2] \quad (5)$$

4 Case study

4.1 Scenario description

The Northeast Passage in the Arctic is the busiest waters for commercial ships to conduct navigation operations in Arctic waters. The seasonal navigation characteristics of the Northeast Passage in the Arctic are obvious. It can be freely navigated in the summer months with the highest temperature. In the seasons close to the summer months, the ice conditions are good, and commercial ships reinforced with ordinary ice class can conduct commercial navigation under the escort of icebreaker. During the escort operation, the icebreaker in front has a higher icebreaking level, which can directly break the sea ice that hinders navigation and form a navigable channel. The escorted ship behind follows the navigation track of the icebreaker in front to complete the ice zone navigation operation. This method greatly improves the navigation efficiency and safety of ordinary ice class reinforced ships in the Arctic ice zone. However, the escort operation itself is still a high-risk water transportation activity. First of all, the width of the icebreaker and the escorted ship will affect the level of risk of the escort operation. If the width of the icebreaker is smaller than that of the escorted ship, the width of the channel formed by the icebreaker will not meet the navigation needs of the escorted ship, greatly increasing the navigation risk. Secondly, the closing speed of the channel formed by the icebreaker has a great impact on the risk of escort operations. If the ice channel closes quickly, the escorted ship will maintain a high speed, and the icebreaker and the escorted ship will also maintain a close distance, which will directly increase the risk of collision accidents. In addition, the size, strength and thickness of the remaining floating ice in the ice channel will also have a great impact on the risk of escort operations. Therefore, for high-risk water transportation activities such as escort operations, it is necessary to take scientific means to quantify its risk level.

In this study, a formation mode consisting of an icebreaker and an escorted ship is used to quantitatively analyze the safety risks of its operation process. The escort operation risk of the TIAN HUI ship in 2018 is taken as the research object. The ship departed from the Barents Sea on July 20 and arrived at the Bering Strait on August 3. During the Arctic voyage, the voyage from July 29 to August 2 was escorted by the icebreaker VAYGACH. In Figure 2, relying on AIS data, the route information of the ship from the departure point to the Bering Strait section and the changes in the ship's position every day are shown.

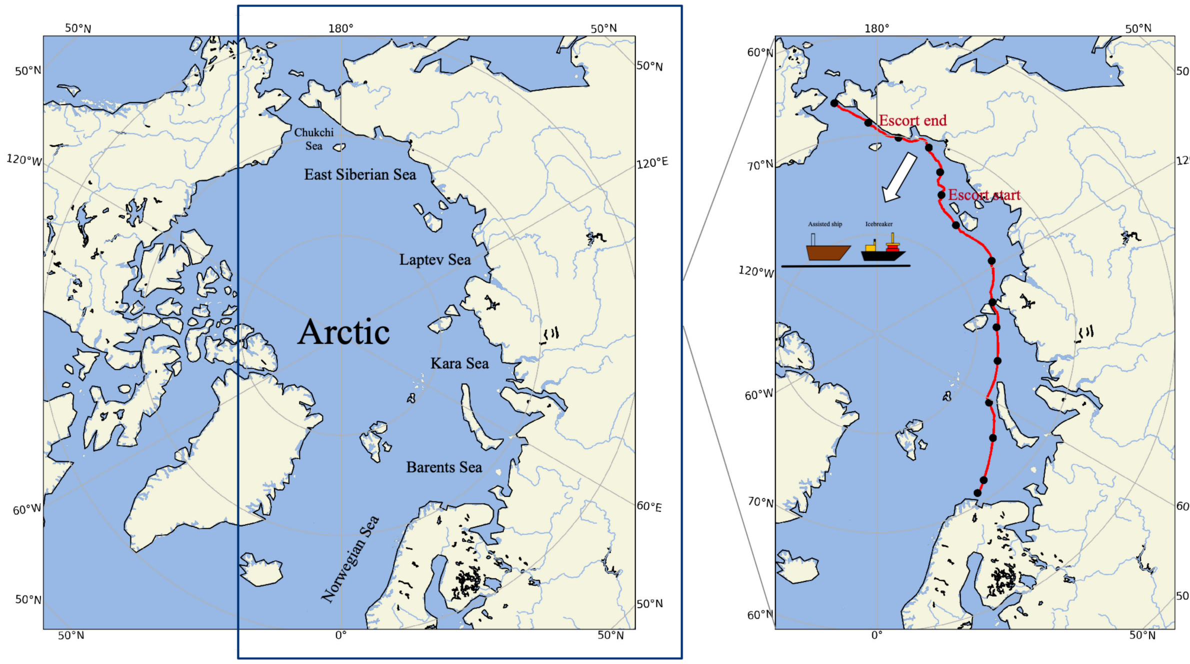


FIGURE 2
Study area and escort operation information.

4.2 Model establishment

According to the timing characteristics of ship escort operation in Arctic waters, it can be mainly divided into the preparatory stage before the operation, the icebreaking escort operation stage, and the end of the escort operation. In the preparatory stage before the operation, it is necessary to combine the weather forecast information to formulate a comprehensive navigation plan. Before the escort operation officially begins, it is also necessary to select a suitable icebreaker and conduct a comprehensive safety inspection. After the escort operation mission begins, the icebreaker must choose a suitable forward strategy in front to ensure that it can effectively break the sea ice and form a reliable navigable waterway. By maintaining timely and effective communication between the two ships, ensure that the ship speed remains within a safe and efficient range. At the same time, ensure that the distance between the two ships is not too long or too short. During the icebreaking escort operation stage, due to the dynamic adjustment of the ship movement between the two ships at any time, this stage is also a high-incidence period for collision accidents. It is necessary to analyze and judge the collision risk in a timely manner and make effective risk prevention and control strategies.

By analyzing the main operational tasks of each stage and further subdividing the key subtasks of each step, the escort operation system in Arctic waters can be mainly divided into nine functional modules: "Make a suitable sailing plan", "Choosing the right icebreaker", "Effectiveness of communication", "Comprehensive safety inspection", "Identify effective ice-breaking strategies", "Maintain a suitable speed", "Keep a safe distance", "Judge the collision risk", and "Risk control measures". Combined with the main steps of FRAM model construction, the functional characteristics of each functional module are analyzed from six aspects: I, O, C, P, R, and T. Taking the functional module "Make a suitable sailing plan" as an example, its functional description is shown in Table 1.

On this basis, the relationship between different functional modules of the escort operation system in Arctic waters is further described, and the main factors that may lead to functional failure are analyzed from the two aspects of the system internal and external, as shown in Table 2. According to the correlation and potential changes between the functional modules, the failure links are analyzed to establish the functional failure network diagram of the ship icebreaking escort operation in Arctic waters, as shown in Figure 3.

TABLE 1 Functional characterization of "Make a suitable sailing plan".

Function name	Dimension	Description
Make a suitable sailing plan	I	Prepare to make a sailing plan
	O	Sailing plan for the voyage
	C	Environment, relevant laws and regulations
	P	Reasonable crewing, ship's condition, cargo allocation, fuel and supplies, etc.
	R	Nautical charts, sailing directions, etc.
	T	–

According to the main factors leading to functional failure analyzed in Table 2 and the functional failure network of ship escort operation in Arctic waters described in Figure 3. Considering that in the Arctic waters escort operation system, abnormal changes in functional output will lead to an increase in the risk of collision accidents through the coupling resonance of upper and lower related functional modules. Through the functional resonance analysis between functional modules, combined with the specific task characteristics of ship escort operation, the risk factors leading to escort operation collision accident are further analyzed in depth. The analysis results are shown in Table 3.

Through FRAM, the basic functions of the system and potential changes of modules of ship escort operation in Arctic waters are analyzed, and the risk factors affecting the safety of escort operation are preliminarily obtained. Further, through the analysis of the coupling relationship between modules, a functional failure network diagram is drawn. On this basis, the key risk factors leading to collision accidents of escort operations in polar waters are identified. These analyses can provide a direct basis for the establishment of a BN model for risk assessment of ship escort operation in Arctic waters. Combined with the risk factors of collision accident of escort operation in Arctic waters obtained from the analysis in Table 3, according to the rules of establishing

TABLE 2 Identification of function module changes.

Function name	Function type	Source of change	Relevant factors
Make a suitable sailing plan	Organization	External changes	Inadequate voyage planning
Choosing the right icebreaker	Organizations	External changes	Inappropriate choice of icebreaker
Effectiveness of communication	Human	Internal changes	Inadequate communication between ships
Comprehensive safety inspection	Organizations	External changes	Inadequate safety inspections
Identify effective ice-breaking strategies	Organizations	Internal changes	Unreasonable ice-breaking strategies
Maintain a suitable speed	Technology	Internal changes	Speed too fast/slow
Keep a safe distance	Technology	Internal changes	Ships too close/far
Judge the collision risk	Human	Internal changes	Insufficient awareness of risks
Risk control measures	Human	Internal changes	No risk control measures are taken

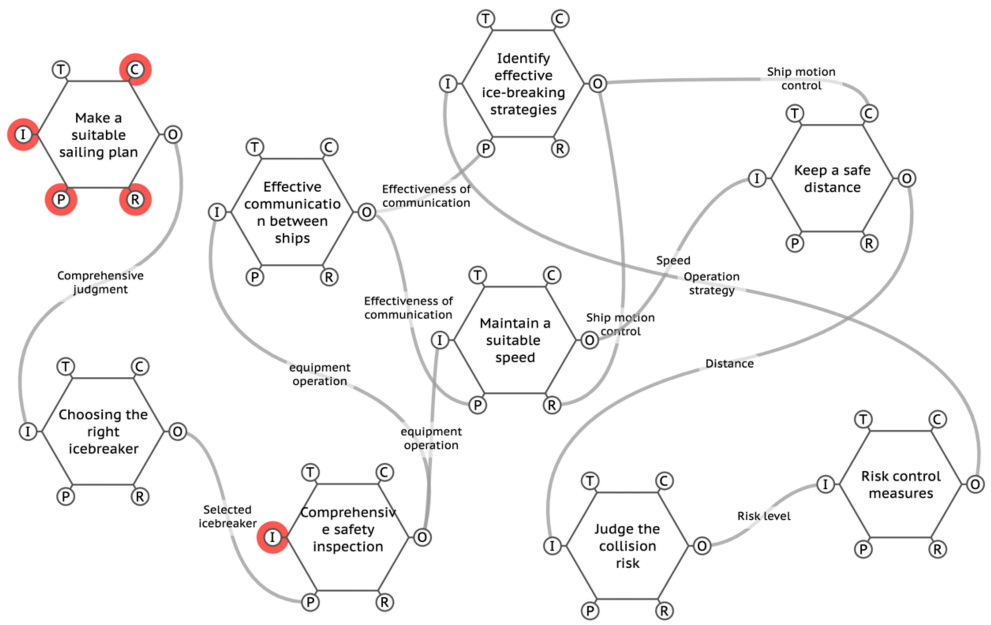

Identify effective ice-breaking strategies
Ship motion control
Make a suitable sailing plan
Keep a safe distance
Effective communication between ships
Effectiveness of communication
Speed
Operation strategy
Comprehensive judgment
Effectiveness of communication
Maintain a suitable speed
Ship motion control
Distance
equipment operation
Choosing the right icebreaker
equipment operation
Risk control measures
Selected icebreaker
Comprehensive safety inspection
Judge the collision risk
Risk level
T
C
I
O
P
R

FIGURE 3
Functional resonance analysis for escort operation in Arctic waters.

the BN model, the risk factors affecting the occurrence of collision accidents of escort operation in Arctic waters are taken as nodes, and the coupling relationship between risk factors is taken as directed edges, and the BN structure of collision accident during escort operation in Arctic waters is obtained, as shown in Figure 4.

TABLE 3 Risk factor analysis of escort operations in Arctic waters.

No	Functional failure	Root cause
1	Inadequate voyage planning	Violations; Inadequate professional skills; Insufficient reading of relevant materials
2	Error in judgement	Insufficient information; Cognitive bias; Harsh environmental conditions
3	Inappropriate choice of icebreaker	Error in judgement; Inadequate voyage planning
4	Inadequate communication	Equipment damage; Inadequate look-out; Inadequate professional skills; Harsh environmental conditions
5	Inadequate safety inspections	Insufficient attention; Relevant system is not perfect
6	Unreasonable ice-breaking strategies	Inadequate communication; Error in judgement
7	Improper speed	Unreasonable ice-breaking strategies; Equipment damage; Inadequate safety inspections; Inappropriate choice of icebreaker
8	Improper distance	Improper speed; Harsh environmental conditions
9	Level of collision risk	Error in judgement; Improper distance; Too much broken ice in the waterway

4.3 Model input

In the established BN model for risk assessment of escort operation in Arctic waters, the node states are divided into two types: “yes” and “no”. The model inputs required in the model include two types: prior probability table and conditional probability table. The conditional probability table can be obtained through logical judgment, and the prior probability table needs to be obtained through statistical calculation based on the objective discrete data set of risk factors. In the model, the node that needs prior probability input is the root node, including “harsh environmental conditions” “too much broken ice in the waterway” “insufficient information” “cognitive bias” “inadequate professional skills” “inadequate look-out” “equipment damage” “insufficient reading of relevant materials” “violations” “relevant system is not perfect” “insufficient attention”. Nowadays, a variety of marine observation and prediction methods are commonly used to facilitate the acquisition of ship navigation data (Yin et al., 2023). Among them, the environmental risk factors can be statistically calculated through objective data. The environmental data used in this study are the fifth-generation reanalysis data provided by the European Centre for Medium-Range Weather Forecasts. This dataset combines model data with observational data from all over the world to form a globally complete and consistent dataset. The dataset contains data on a variety of meteorological and oceanographic environmental parameters, including temperature, air pressure, precipitation, snowfall, wind speed, wave height, sea ice, etc. Figure 5 shows the environmental distribution of Arctic waters, among which (Figure 5A) shows the overall environmental distribution of the entire Arctic waters. Here, sea ice concentration is used as an example for display. Due to space limitations, other environmental factors are not displayed one by one. In (Figures 5B–D), the dynamic distribution of temperature, sea ice

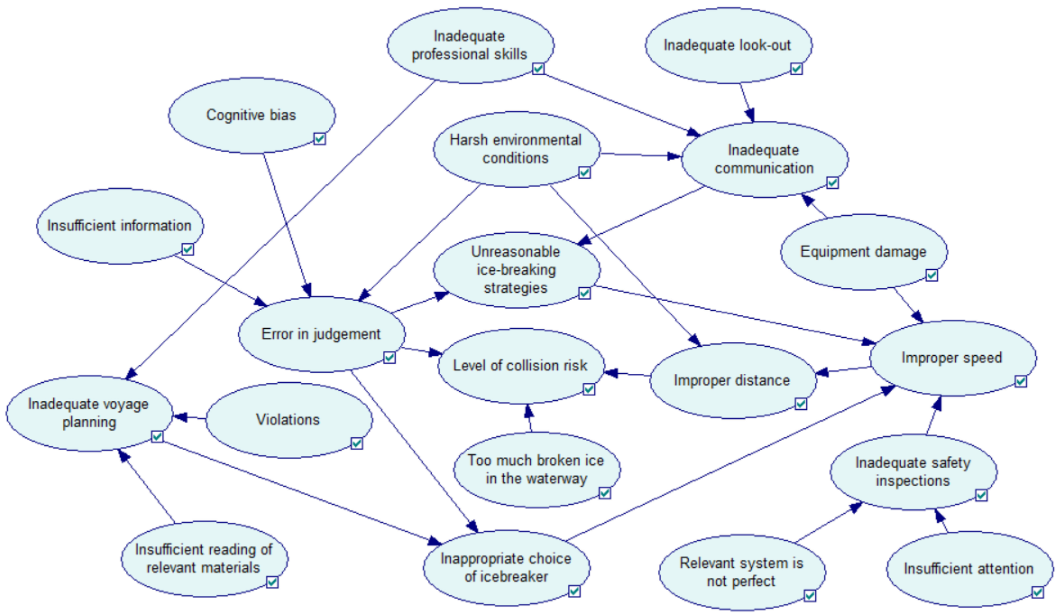

Inadequate professional skills
Inadequate look-out
Cognitive bias
Harsh environmental conditions
Inadequate communication
Insufficient information
Unreasonable ice-breaking strategies
Equipment damage
Error in judgement
Level of collision risk
Improper distance
Improper speed
Inadequate voyage planning
Violations
Too much broken ice in the waterway
Inadequate safety inspections
Insufficient reading of relevant materials
Inappropriate choice of icebreaker
Relevant system is not perfect
Insufficient attention

FIGURE 4
BN risk assessment model for escort operations in Arctic waters.

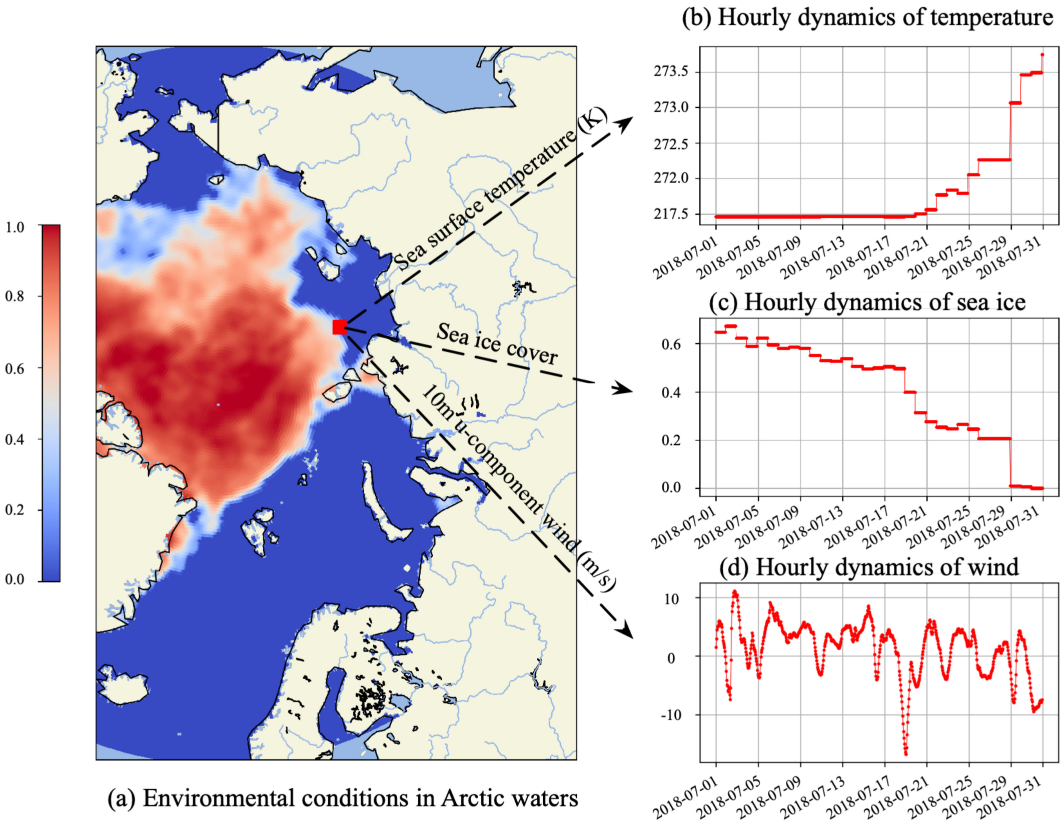


FIGURE 5
Natural environmental conditions for navigation in Arctic waters. **(A)** shows the overall environment of the Arctic waters. **(B)** shows the temperature changes at fixed coordinate points. **(C)** shows the sea ice changes at fixed coordinate points. **(D)** shows the wind speed changes at fixed coordinate points.

and wind speed at the ship station of TIAN HUI on the eighth day is shown respectively. Referring to the classification of the impact of environmental factors in Arctic waters on ship navigation safety in Li et al. (2023b), these discrete environmental data sets can provide a reference for solving the prior probability of environmental risk factors such as "Too much broken ice in the waterway" and "harsh environmental conditions". Taking the node "Too much broken ice in the waterway" as an example, when calculating the prior probability distribution of its ship station every day, the change of sea ice concentration at the location is extracted according to the grid where the longitude and latitude of its ship station are located, as shown in Figure 6. Considering the strong icebreaking performance of the icebreaker, the value of sea ice concentration of 0.55 is considered as the state threshold, and then the prior probability distribution table of the node "Too much broken ice in the waterway" at each ship station can be directly obtained by statistics, as shown in Table 4.

In the BN risk assessment model, there are still some nodes that do not have direct data sources. According to the proposed cloud model quantitative analysis method, the cloud characteristic distribution value of the risk factor is obtained by relying on expert knowledge, and the random sample distribution of the risk factor is further obtained through cloud simulation. Here, five experts from related fields are invited to provide expert knowledge for the quantification of subjective information. Among them, three experts are professors from Shanghai Maritime University, who have been engaged in the research of risk management of ship navigation in Arctic waters for a long time. Two experts are captains with rich experience in driving polar commercial ships, from COSCO SHIPPING Special Transport Co., Ltd. For each risk factor, a subjective evaluation of the quantitative value of the factor is carried out based on their experience and knowledge. Based on the evaluation, the *Ex*, *En*, and *He* of risk factor are obtained by combining Equations 2–4, and the cloud simulation results are further obtained according to Equation 5.

Take the prior probability distribution calculation of the node "inadequate professional skills" as an example. During the voyage of the TIAN HUI ship in the Arctic waters, the voyage from July 29 to August 2 was escorted by the icebreaker VAYGACH. According to its icebreaking escort operation characteristics, its icebreaking escort voyage was divided into three stages, namely the first day of icebreaking escort operation (Stage 1), the middle stage of icebreaking escort operation (Stage 2), and the last day of icebreaking escort operation (Stage 3). For these three stages of the ship's voyage in the Arctic waters, the three characteristic quantities of the risk factor "inadequate professional skills" in each stage are obtained through expert judgment, and its cloud droplet distribution is further obtained according to cloud simulation, as shown in Figure 7. Taking the quantitative value of 0.75 as the threshold for state division, the prior probability distribution of each stage of ship navigation operations in Arctic waters is calculated by counting the number of cloud droplets in each cloud droplet distribution, as shown in Table 5.

4.4 Results

According to the prior probability distribution and conditional probability distribution of each node, the quantitative assessment results of the navigation risk of the TIAN HUI ship during the icebreaker escort operation in the Arctic waters are obtained through BN reasoning, as shown in Figure 8. According to the results of the risk quantification assessment, during the icebreaker escort operation, the navigation risk of the ship showed a

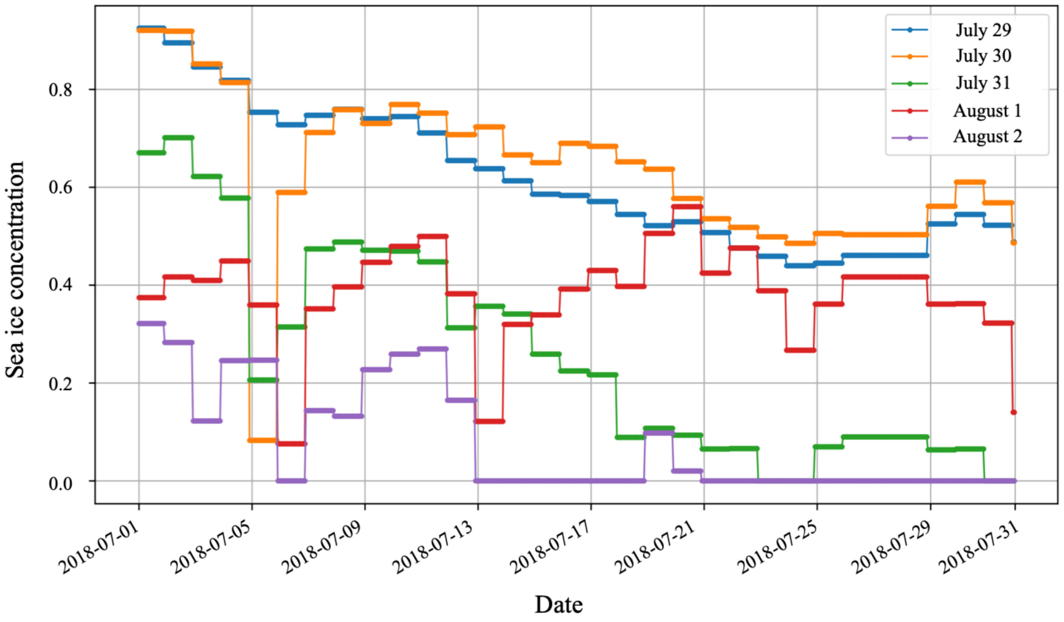


FIGURE 6
Changes in sea ice concentration at the ship's position point on a daily basis.

TABLE 4 Prior probability distribution of “too much broken ice in the waterway” during ship navigation in Arctic waters.

Date	July 29	July 30	July 31	August 1	August 2
Yes	0.55	0.71	0.13	0.32	0.01
No	0.45	0.29	0.87	0.68	0.99

fluctuating downward trend. The average risk value per day during the ship’s icebreaker escort operation is 0.087. The highest risk time occurred on the second day of the escort operation, and the lowest risk time occurred on the last day. The difference between the highest and lowest risk quantitative values is 0.176. This shows that during the icebreaker escort operation in Arctic waters, the risk level fluctuates greatly. Carrying out icebreaker escort operations is a high-risk water transportation activity, and special attention should be paid to the safety threats brought by sudden risk events in the operation. Specifically, on the first day of the icebreaking escort operation, the risk of ship navigation is at a high level, with a quantitative value of 0.157. On the second day of the icebreaking escort operation, the risk of ship navigation reached the highest value, with a quantitative result of 0.185. Subsequently, the risk of ship navigation gradually decreased, and the risk level is the lowest on the last day of the escort operation, with a quantitative value of 0.009.

In order to further analyze the key risk factors during the escort operation in Arctic waters, the node “Level of collision risk” is set as the target node in the BN model, and the BN sensitivity analysis is carried out. The main risk factors affecting the occurrence of collision accident during the escort operation in Arctic waters are obtained, as shown in Figure 9. The top ten risk factors affecting the occurrence of collision accident are obtained through analysis, namely “Harsh environmental conditions (R1)” “Too much broken ice in the waterway (R2)” “Improper distance (R3)” “Error in judgement (R4)” “Insufficient information (R5)” “Cognitive bias (R6)” “Improper speed (R7)” “Inappropriate choice of icebreaker (R8)” “Unreasonable ice-breaking strategies (R9)” “Inadequate professional skills (R10)”. It can be seen that among the main risk factors affecting the occurrence of collision accident during escort operations in Arctic waters, the most important risk factor comes from the navigation environment, followed by inappropriate ship following distance and mistakes of ship operators.

5 Discussion

5.1 Risk performance of escort operation in Arctic waters

It is a high-risk water traffic activity for ordinary commercial ships to carry out navigation in Arctic waters. In order to ensure navigation safety, escort operation is an effective operation mode. However, while this operation mode improves the navigation safety level of ships, it may cause collision accidents due to the following mode between icebreaker and escorted ship. This study quantitatively analyzes the risk level of collision accident during escort operation in Arctic waters. In order to verify the reliability of the risk assessment model established in this study and the effectiveness of escort operation on the navigation safety level in Arctic waters, it is necessary to conduct comparative verification analysis. During the five days when the TIAN HUI ship carried out escort operations, considering the impact of the icebreaking level of icebreakers on navigation risks, the risk assessment results were compared and analyzed. At a lower icebreaking level, the impact of sea ice on the navigation safety of ships will be greater. Assuming that the icebreaking performance of icebreakers is reduced, the possibility of sea ice threats increases by 20%. Therefore, in the BN risk assessment model in Figure 4, the model input of the node “too much broken ice in the waterway” is changed to analyze the navigation risk during escort operations. The results are shown in Figure 10.

From the quantitative results, it can be seen that with the decrease in icebreaking performance of icebreakers, the fluctuation trend of the

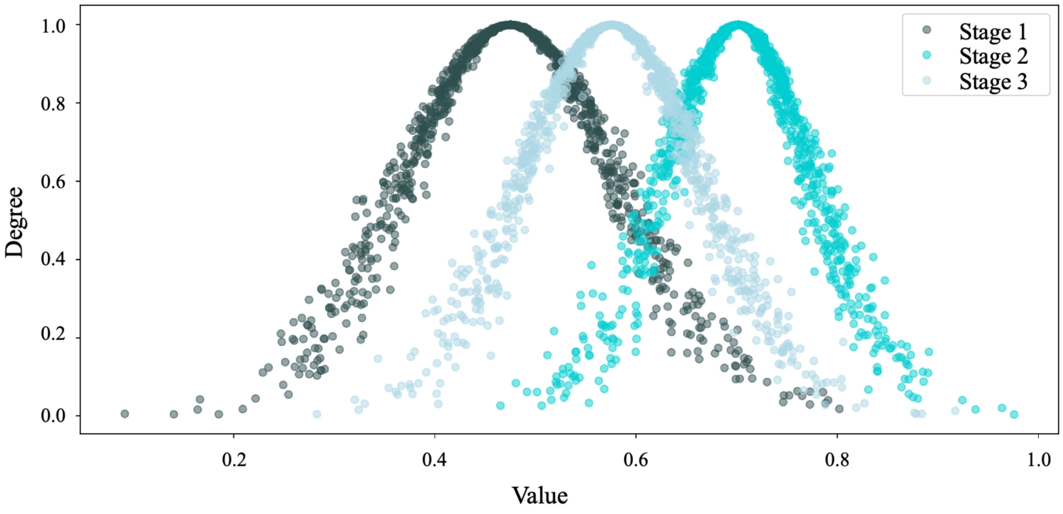


FIGURE 7
Quantitative cloud of information on risk factors for ship escort operations in Arctic waters.

TABLE 5 Prior probability distributions for each stage of ship navigation in Arctic waters.

State	Stage 1	Stage 2	Stage 3
Yes	0.13	0.45	0.24
No	0.87	0.55	0.76

escort operation risk in Arctic waters is consistent with the previous trend. However, under the strong condition of reduced icebreaking performance of icebreakers, the overall level of navigation risk is higher than that of icebreakers with high icebreaking levels. This is consistent with the actual navigation conditions of ships in Arctic waters, which shows the reliability of the model. Specifically, during the five days when TIAN HUI ship took icebreaking escort operations, the risk difference in the first two days was much higher than that in the last three days. By analyzing the sea ice conditions in Table 4, it was found that the value of ice cover in the first two days of icebreaking escort operations was much higher than that in the last three days. This also shows that when the sea ice conditions are more severe, it is very necessary to choose a reliable high-level icebreaker.

5.2 Uncertainty of risk information during escort operation in Arctic waters

Ship navigation risk management often faces the problem of missing information, and how to solve the uncertainty in it has become one of the main problems to be solved in current research. The uncertainty of risk information often comes from several aspects. On the one hand, in the stage of risk identification, in the process of identifying risk factors and determining the logical relationship between them, different scholars have different logical thinking, and the various theoretical methods used are different, so the results often have strong uncertainty. On the other hand, in the process of quantitative analysis of ship navigation risk, there is often the problem of missing data sources for some risk factors, thus bringing uncertainty in data information. Compared with ordinary waters, the navigation environment in Arctic waters is more severe, and the task of escort operation is very complex, so the uncertainty faced in the process of quantitative analysis and management of navigation risk is even stronger, and new ideas are urgently needed to solve these problems.

In this study, to address the uncertainty in the risk modeling process, based on the functional resonance perspective, the FRAM method is used to describe and analyze the functions and activities of the escort operation system in Arctic waters, to establish a functional failure network of the ice-breaking escort operation in Arctic waters, and to analyze in depth the main risk factors that lead to collision accident during escort operation. This method effectively overcomes the problem of excessive subjectivity in risk factor identification under the traditional "man-machine-environment" perspective. Moreover, the logical relationship between the risk factors can be sorted out in a scientific and logical way, which can help to establish the risk network model of collision accident of escort operation in Arctic waters. In the process of quantitative risk analysis, this study proposes the use of FRAM and BN method effectively combined to establish a quantitative model for risk assessment, which has the advantage of integrating multi-source information and carrying out quantitative analysis of risk by means of network reasoning, and it is a reliable method of quantitative risk analysis for the characteristics of the escort operation in Arctic waters. In addition, in the processing of uncertain data information, this study proposes the use of a cloud model to realize the conversion of subjective experience to objective data, and cloud simulation through the eigenvalues of risk factors, thus providing a direct data source for BN risk inference. This approach brings new ideas for dealing with the uncertainty of data information in ship navigation risk management.

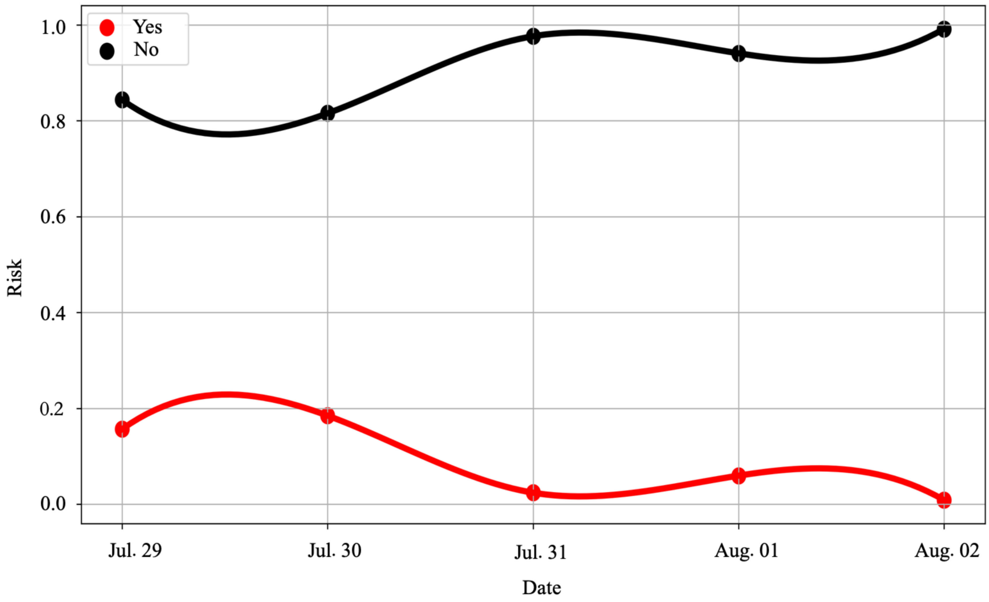


FIGURE 8
Ship escort operation risks in Arctic waters.

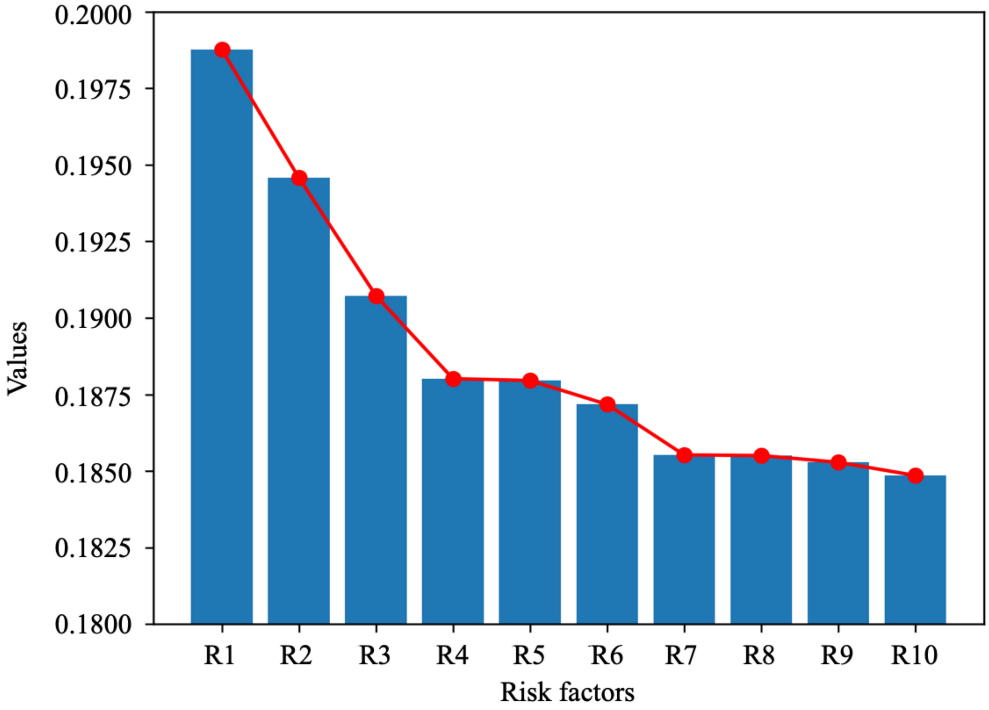


FIGURE 9
Main risk factors leading to collision accident.

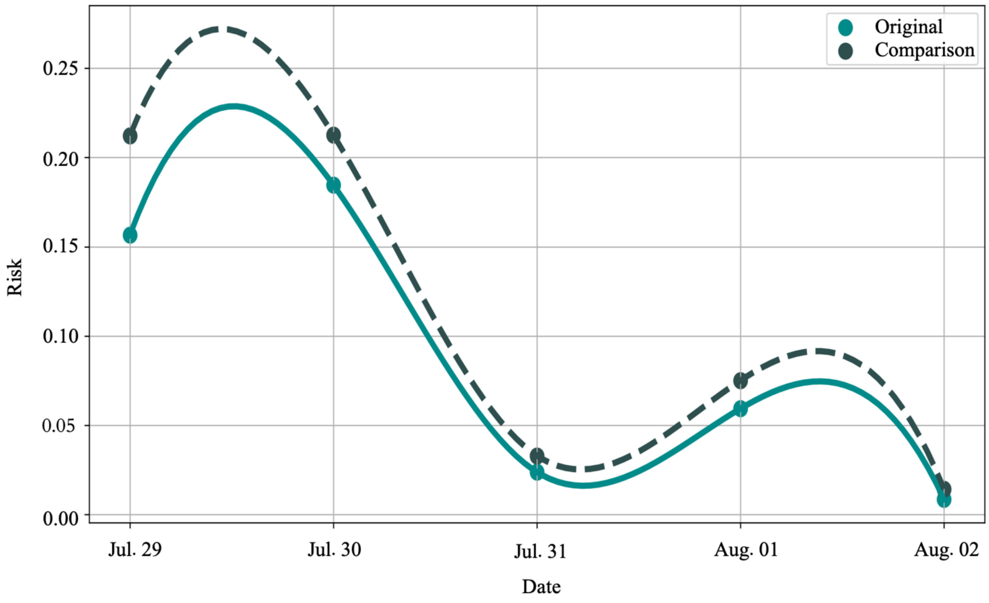


FIGURE 10
Navigation risks at different icebreaking levels.

6 Conclusion

This study proposes an approach combining FRAM and BN to quantitatively analyze the risk of ship escort operation in Arctic waters. In the proposed method, to address the complexity of the ship escort operation system in Arctic waters, FRAM is used to model and analyze the complex coupling relationship between the systems, and the key risk factors in the system are further analyzed. Based on the correlation relationship between risk factors, a BN analysis method is used to establish a collision risk assessment model for ship escort operation in Arctic waters. Relying on multi-source data, the subjective information is quantified by combining the cloud model of uncertain information processing. For the specific scenarios of ice-breaking escort operation of ships in Arctic waters, the escort operation risk is quantitatively analyzed.

The results show that during the escort operation, the average value of ship navigation risk is 0.087, which is at a high level, and the maximum difference in risk reaches 0.176, with a large level of risk fluctuation during the escort operation. Among the main risk factors affecting the occurrence of collision accident during ship escort operation in Arctic waters, the risk factor with the greatest impact is "Harsh environmental conditions", followed by "Too much broken ice in the waterway". Taken together, among the main risk factors affecting the occurrence of collision accidents during escort operation in Arctic waters, the most important risk factor comes from the navigational environment, especially the influence of sea ice conditions. Unsuitable following distance of the ship and the error of the ship operator are also the key reasons affecting the collision accident. During ship escort operation in Arctic waters, it is essential to select icebreakers of the appropriate icebreaking class according to the sea ice conditions.

The method proposed in this study is effective in understanding the level of navigational risk during ship escort operation in Arctic waters, clarifying the key risk factors involved, and improving the safety level of escort operations. In future research, the model can be further expanded to introduce decision analysis methods to assess the effectiveness of various types of risk control measures by proposing them, so as to maximize the safety level of ship escort operations in Arctic waters.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

ZL: Conceptualization, Data curation, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. XZ: Writing – review & editing, Data curation. SL: Data curation, Supervision, Writing – review & editing. KG: Writing – review & editing. SH: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work is financially supported by National Natural Science Foundation of China (NSFC) (Grant No. 52402422), by the Philosophy and Social Sciences Planning Project of Guangdong Province under Grant GD24XGL066. This work is also financially supported by the Zhanjiang Science and Technology Plan Project (Project No. 2024B01001), by the Business Administration Program Construction Project of Shenzhen Polytechnic University.

Acknowledgments

The authors gratefully acknowledge the reviewers for their valuable comments and suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbassi, R., Khan, F., Khakzad, N., Veitch, B., and Ehlers, S. (2017). Risk analysis of offshore transportation accident in Arctic waters. *Int. J. Maritime Eng.* 159, A213–A223. doi: 10.3940/rina.ijme.2017.a3.351

Afenyo, M., Jiang, C. M., and Ng, A. K. Y. (2023). A Bayesian-loss function-based method in assessing loss caused by ship-source oil spills in the arctic area. *Risk Anal.* 43, 1557–1571. doi: 10.1111/risa.14025

Afenyo, M., Ng, A. K. Y., and Jiang, C. M. (2021). A multiperiod model for assessing the socioeconomic impacts of oil spills during arctic shipping. *Risk Anal.* 9, 13773. doi: 10.1111/risa.13773

Amin, M. T., Khan, F., and Amyotte, P. (2019). A bibliometric review of process safety and risk analysis. *Process Saf. Environ. Prot.* 126, 366–381. doi: 10.1016/j.psep.2019.04.015

Baksh, A. A., Abbassi, R., Garaniya, V., and Khan, F. (2018). Marine transportation risk assessment using Bayesian Network: Application to Arctic waters. *Ocean Eng.* 159, 422–436. doi: 10.1016/j.oceaneng.2018.04.024

Basnet, S., Bahootoroody, A., Chaal, M., Lahtinen, J., Bolbot, V., Banda, O. V., et al. (2023). Risk analysis methodology using STPA-based Bayesian network- applied to remote pilotage operation. *Ocean Eng.* 270, 113569. doi: 10.1016/j.oceaneng.2022.113569

Behari, N. (2019). Assessing process safety culture maturity for specialty gas operations: A case study. *Process Saf. Environ. Prot.* 123, 1–10. doi: 10.1016/j.psep.2018.12.012

Chen, P., Zhang, Z., Huang, Y., Dai, L., and Hu, H. (2022). Risk assessment of marine accidents with Fuzzy Bayesian Networks and causal analysis. *Ocean Coast. Manage.* 228, 106323. doi: 10.1016/j.ocecoaman.2022.106323

França, J. E. M., and Hollnagel, E. (2023). Analyzing human factors and complexities of mining and O&G process accidents using FRAM: Copiapo (Chile) and FPSO CSM (Brazil) cases. *Process Saf. Prog.* 42, S9–S18. doi: 10.1002/prs.12428

Fu, S., Yu, Y., Chen, J., Han, B., and Wu, Z. (2022). Towards a probabilistic approach for risk analysis of nuclear-powered icebreakers using FMEA and FRAM. *Ocean Eng.* 2022, 260. doi: 10.1016/j.oceaneng.2022.112041

Fu, S. S., Zhang, Y., Zhang, M. Y., Han, B., and Wu, Z. D. (2023). An object-oriented Bayesian network model for the quantitative risk assessment of navigational accidents in ice-covered Arctic waters. *Reliability Eng. System Saf.* 238, 109459. doi: 10.1016/j.ress.2023.109459

Guo, P., Li, X. F., Jia, Y. L., and Zhang, X. (2020). Cloud model-based comprehensive evaluation method for entrepreneurs' Uncertainty tolerance. *Mathematics* 8, 1491. doi: 10.3390/math8091491

Kandel, R., and Baroud, H. (2024). A data-driven risk assessment of Arctic maritime incidents: Using machine learning to predict incident types and identify risk factors. *Reliability Eng. system Saf.* 243, 109779. doi: 10.1016/j.ress.2023.109779

Khan, B., Khan, F., and Veitch, B. (2020). A Dynamic Bayesian Network model for ship-ice collision risk in the Arctic waters. *Saf. Sci.* 130, 104858. doi: 10.1016/j.ssci.2020.104858

Kim, Y. C., and Yoon, W. C. (2021). Quantitative representation of the functional resonance analysis method for risk assessment. *Reliability Eng. System Saf.* 214, 107745. doi: 10.1016/j.ress.2021.107745

Lee, J., Yoon, W. C., and Chung, H. (2020). Formal or informal human collaboration approach to maritime safety using FRAM. *Cognition Technol. Work* 22, 861–875. doi: 10.1007/s10111-019-00606-y

Lenzen, M., Tzeng, M., Floerl, O., and Anastasija, Z. (2023). Application of multi-region input-output analysis to examine biosecurity risks associated with the global shipping network. *Sci. Total Environ.* 854, 158758. doi: 10.1016/j.scitotenv.2022.158758

Li, P. C., Wang, Y. H., and Yang, Z. L. (2024). Risk assessment of maritime autonomous surface ships collisions using an FTA-FBN model. *Ocean Eng.* 309, 118444. doi: 10.1016/j.oceaneng.2024.118444

Li, W., Chen, W. J., Hu, S. P., Xi, Y. T., and Guo, Y. L. (2023a). Risk evolution model of marine traffic via STPA method and MC simulation: A case of MASS along coastal setting. *Ocean Eng.* 281, 114673. doi: 10.1016/j.oceaneng.2023.114673

Li, W. C., He, M., Sun, Y. B., and Cao, Q. G. (2019). A proactive operational risk identification and analysis framework based on the integration of ACAT and FRAM. *Reliability Eng. System Saf.* 186, 101–109. doi: 10.1016/j.ress.2019.02.012

Li, Z., Hu, S. P., Zhu, X. M., Gao, G. P., Yao, C. Y., Han, B., et al. (2022). Using DBN and evidence-based reasoning to develop a risk performance model to interfere ship navigation process safety in Arctic waters. *Process Saf. Environ. Prot.* 162, 357–372. doi: 10.1016/j.psep.2022.03.089

Li, Z., Zhu, X. M., Li, R. H., Liao, S. G., and Gao, K. X. (2023b). A comprehensive method for causation analysis of ship–ice collision risk in Arctic waters. *Environ. Sci. pollut. Res.* 31, 40366–40382. doi: 10.1007/s11356-023-28249-7

Liu, Y., Ma, X. X., Qiao, W. L., Ma, L. H., and Han, B. (2024b). A novel methodology to model disruption propagation for resilient maritime transportation systems-a case study of the Arctic maritime transportation system. *Reliability Eng. System Saf.* 241, 109620. doi: 10.1016/j.ress.2023.109620

Liu, X., Meng, H. X., An, X., and Xing, J. D. (2024a). Integration of functional resonance analysis method and reinforcement learning for updating and optimizing emergency procedures in variable environments. *Reliability Eng. System Saf.* 241, 109655. doi: 10.1016/j.ress.2023.109655

Liu, H., Zhang, L., and Liu, S. (2020). Modeling of ship collision risk based on cloud model. *IEEE Access* 8, 221162–221175. doi: 10.1109/ACCESS.2020.3042811

Ma, Q., Lu, L., Li, Q., and Wang, Z. P. (2022). Green construction grade evaluation of large channels based on uncertain AHP-multidimensional cloud model. *Sustainability* 14, 6143. doi: 10.3390/su14106143

Ma, L. H., Ma, X. X., Liu, Y., Deng, W. Y., and Lan, H. (2023). Risk assessment of coupling links in hazardous chemicals maritime transportation system. *J. loss Prev. process industries* 82, 105011. doi: 10.1016/j.jlp.2023.105011

Moe, A., and Brigham, L. (2017). Organization and management challenges of Russia's icebreaker fleet. *Geographical Rev.* 107, 48–68. doi: 10.1111/j.1931-0846.2016.12209.x

Nguyen, S., Chen, P. S. L., Du, Y. Q., and Thai, V. V. (2021). An operational risk analysis model for container shipping systems considering uncertainty quantification. *Reliability Eng. System Saf.* 209, 107362. doi: 10.1016/j.ress.2020.107362

Otheitis, N., and Kunc, M. (2015). Performance measurement adoption and business performance An exploratory study in the shipping industry. *Manage. Decision* 53, 139–159. doi: 10.1108/MD-02-2014-0108

Patriarca, R., Di Gravio, G., and Costantino, F. (2017). A Monte Carlo evolution of the Functional Resonance Analysis Method (FRAM) to assess performance variability in complex systems. *Saf. Sci.* 91, 49–60. doi: 10.1016/j.ssci.2016.07.016

Peng, B., Zhou, J. M., and Peng, D. H. (2017). Cloud model based approach to group decision making with uncertain pure linguistic information. *J. Intelligent Fuzzy Syst. Appl. Eng. Technol.* 32, 1959–1968. doi: 10.3233/JIFS-161473

Ryan, C., Huang, L. F., Li, Z. Y., Ringsberg, J. W., and Thomas, G. (2021). An Arctic ship performance model for sea routes in ice-infested waters. *Appl. Ocean Res.* 117, 102950. doi: 10.1016/j.apor.2021.102950

Sakar, C., Toz, A. C., Buber, M., and Koseoglu, B. (2021). Risk analysis of grounding accidents by mapping a fault tree into a bayesian network. *Appl. Ocean Res.* 113, 1–12. doi: 10.1016/j.apor.2021.102764

Salihoglu, E., and Besikci, E. B. (2021). The use of Functional Resonance Analysis Method (FRAM) in a maritime accident: A case study of Prestige. *Ocean Eng.* 219, 108223. doi: 10.1016/j.oceaneng.2020.108223

Tian, J., Wu, J. Y., Yang, Q. B., and Zhao, T. D. (2016). FRAMA: A safety assessment approach based on Functional Resonance Analysis Method. *Saf. Sci.* 85, 41–52. doi: 10.1016/j.ssci.2016.01.002

Vanhatalo, J., Huuhtanen, J., Bergstrm, M., Helle, I., Mäkinen, J., Kujala, P., et al. (2021). Probability of a ship becoming beset in ice along the Northern Sea Route – A Bayesian analysis of real-life data. *Cold Regions Sci. Technol.* 184, 103238. doi: 10.1016/j.coldregions.2021.103238

Wang, C. Y., Ding, M. H., Yang, Y. D., Wei, T., and Dou, T. F. (2022). Risk assessment of ship navigation in the northwest passage: historical and projection. *Sustainability* 14, 5591. doi: 10.3390/su14095591

Wang, J. Q., Peng, J. J., Zhang, H. Y., Liu, T., and Chen, X. H. (2015). An uncertain linguistic multi-criteria group decision-making method based on a cloud model. *Group Decision Negotiation* 24, 171–192. doi: 10.1007/s10726-014-9385-7

Wu, J. J., Hu, S. P., Jin, Y. X., Fei, J. G., and Fu, S. S. (2019). Performance simulation of the transportation process risk of bauxite carriers based on the markov chain and cloud model. *J. Mar. Sci. Eng.* 7, 108. doi: 10.3390/jmse7040108

Xi, Y. T., Zhang, X., Han, B., Zhu, Y., Fan, C. L., Kim, E., et al. (2024). Advanced human reliability analysis approach for ship convoy operations via a model of IDAC and DBN: A case from ice-covered waters. *J. Mar. Sci. Eng.* 12, 1536. doi: 10.3390/jmse12091536

Xu, S., and Kim, E. (2023). Hybrid causal logic model for estimating the probability of an icebreaker–ship collision in an ice channel during an escort operation along the Northeast Passage. *Ocean Eng.* 284, 115264. doi: 10.1016/j.oceaneng.2023.115264

Xu, K. Y., Liu, J. G., and Meng, H. (2023). Stability and energy consumption analysis of arctic fleet: modeling and simulation based on future motion of multi-ship. *Environ. Sci. pollut. Res.* 31, 40352–40365. doi: 10.1007/s11356-023-27787-4

Yao, S. Y., Hu, H., and Yang, J. B. (2022). A probabilistic safety assessment framework for arctic marine traffic using the evidential reasoning approach. *Int. J. Shipping Transport Logistics* 14, 266–301. doi: 10.1504/IJSTL.2022.122419

Yao, S. Y., Wu, Q. H., Qi, K., Qi, K., Chen, Y. W., Lu, Y., et al. (2024). An interpretable XGBoost-based approach for Arctic navigation risk assessment. *Risk Anal.* 44, 14175. doi: 10.1111/risa.14175

Yin, J. C., Wang, H. F., Wang, N. N., and Wang, X. G. (2023). An adaptive real-time modular tidal level prediction mechanism based on EMD and Lipschitz quotients method. *Ocean Eng.* 289, 116297. doi: 10.1016/j.oceaneng.2023.116297

Yousefi, A., Hernandez, M. R., and Pea, V. L. (2019). Systemic accident analysis models: A comparison study between AcciMap, FRAM, and STAMP. *Process Saf. Prog.* 38, e12002. doi: 10.1002/prs.12002

Yu, Y. R., Liu, K. Z., Fu, S. S., and Chen, J. H. (2024). Framework for process risk analysis of maritime accidents based on resilience theory: A case study of grounding accidents in Arctic waters. *Reliability Eng. System Saf.* 249, 110202. doi: 10.1016/j.ress.2024.110202

Zhang, W. B., Goerlandt, F., Kujala, P., and Qi, Y. (2018). A coupled kinematics model for icebreaker escort operations in ice-covered waters. *Ocean Eng.* 167, 317–333. doi: 10.1016/j.oceaneng.2018.08.035

Zhang, M. Y., Zhang, D., Fu, S. S., Yan, X. P., and Goncharov, V. (2017). Safety distance modeling for ship escort operations in Arctic ice-covered waters. *Ocean Eng.* 146, 202–216. doi: 10.1016/j.oceaneng.2017.09.053

Zhang, X. X., Zhang, Q. N., Yang, J., et al. (2019b). Safety risk analysis of unmanned ships in inland rivers based on a fuzzy bayesian network. *J. advanced transportation*, 4057195. doi: 10.1155/2019/4057195

Zhang, W. B., Zou, Z. Y., Goerlandt, F., et al. (2019a). A multi-ship following model for icebreaker convoy operations in ice-covered waters. *Ocean Eng.* 180, 238–253. doi: 10.1016/j.oceaneng.2019.03.057

Zheng, Q. H., Liu, X. W., Yang, M., Wang, W. Z., and Adriaensen, A. (2024). Enhancing emergency response planning for natech accidents in process operations using functional resonance analysis method (FRAM): A case of fuel storage tank farm. *Process Saf. Environ. Prot.* 188, 514–527. doi: 10.1016/j.psep.2024.05.132

Zhu, X. M., Hu, S. P., Li, Z., Wu, J. J., Yang, X., Fu, S. S., et al. (2024). Risk performance analysis approach for convoy operations via a hybrid model of STPA and DBN: A case from ice-covered waters. *Ocean Eng.* 302, 117570. doi: 10.1016/j.oceaneng.2024.117570

frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 22 January 2025
DOI 10.3389/fmars.2024.1503482

EDITED BY
Phoebe Koundouri,
Athens University of Economics and Business, Greece

REVIEWED BY
Dongdong Mu,
Dalian Maritime University, China
Liang Zhao,
Zhejiang University, China

*CORRESPONDENCE
Xiaoguang Mou
mouxg@gdou.edu.cn

RECEIVED 29 September 2024
ACCEPTED 23 December 2024
PUBLISHED 22 January 2025

CITATION
Wang H, Mao S, Mou X, Zhang J and Li R (2025) Path planning for unmanned surface vehicles in anchorage areas based on the risk-aware path optimization algorithm. *Front. Mar. Sci.* 11:1503482. doi: 10.3389/fmars.2024.1503482

Path planning for unmanned surface vehicles in anchorage areas based on the risk-aware path optimization algorithm

Hongbo Wang[1,2,3], Shuaiwei Mao[1], Xiaoguang Mou[4]*, Jinfeng Zhang[2] and Ronghui Li[1]

[1]Naval Architecture and Shipping College, Guangdong Ocean University, Zhanjiang, China, [2]Hubei Key Laboratory of Inland Shipping Technology, Wuhan University of Technology, Wuhan, China, [3]Guangdong Provincial Key Laboratory of Intelligent Equipment for South China Sea Marine Ranching, Guangdong Ocean University, Zhanjiang, China, [4]School of Mechanical Engineering, Guangdong Ocean University, Zhanjiang, China

In dense anchorage areas, the challenge of navigation for Unmanned Surface Vehicles is particularly pronounced, especially regarding path safety and economy. A Risk-Aware Path Optimization Algorithm is proposed to enhance the safety and efficiency of Unmanned Surface Vehicle navigating in anchorage areas. The algorithm incorporates risk assessment based on the A* algorithm to generate an optimized path and employs a Dual-Phase Smoothing Strategy to ensure path smoothness. First, the anchorage area is spatially separated using a Voronoi polygon, the Risk-Aware Path Optimization Algorithm includes a grid risk function, derived from the ship domain and Gaussian influence function, in the path evaluation criteria, directing Unmanned Surface Vehicle to successfully bypass high-risk areas and as a result. Then the Dual-Phase Smoothing Strategy is used to decrease path turning points and boost path continuity, which in turn improves path economy. Simulation results demonstrate that this method significantly reduces the path length and the number of turning points, enhancing Unmanned Surface Vehicle navigation safety and economy in anchorage areas.

KEYWORDS

unmanned surface vehicles, anchorage areas, risk-aware path optimization, ship domain, Gaussian influence function, dual-phase smoothing strategy

1 Introduction

Ships need to anchor in anchorage waters for quarantine, waiting for berths, tide waiting (Yin et al., 2023), unloading at anchorage, or sheltering from typhoons. Anchorage areas are typically densely populated, with ships varying in size and type, as illustrated in Figure 1. Navigating vessels are typically needed to avoid these waters to prevent collisions.

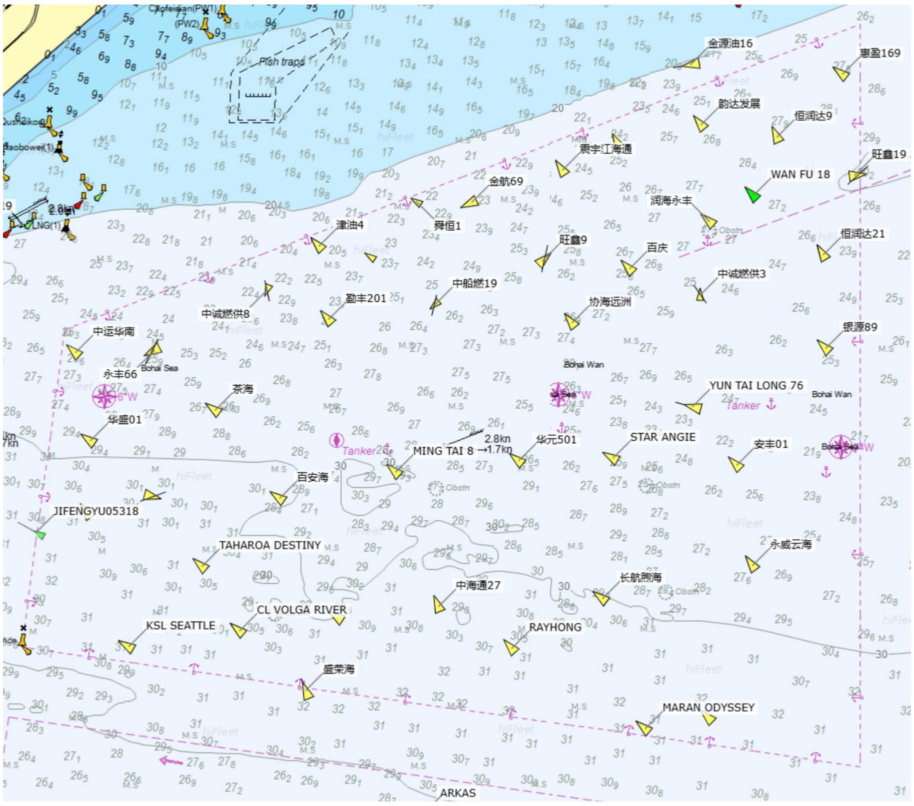


FIGURE 1
Anchorage area layout and ship distribution.

The application of intelligent ships is becoming increasingly common (Zhou et al., 2024). For example, USVs could decrease the risk of collisions for tasks such as maritime monitoring and transporting materials in complex navigation environments. USVs are autonomous surface vessels capable of navigating without onboard personnel (Specht et al., 2017). Generally, USV is smaller in size and do not require human operation, which can significantly enhance safety when navigating through anchorage areas, improve operational efficiency, and reduce labor costs. USVs can keep an eye on the marine environment and the status of anchored vessels in real time, which effectively boosts the efficiency of safety management (Wang et al., 2023). Also, USVs can effectively carry materials in a range of weather and sea conditions, which makes them especially fit for high-risk environments or those not suitable for human operations (Bae and Hong, 2023).

The core task of path planning is to design a collision-free route on a map from the starting point to the endpoint (Yin and Wang, 2021). Path planning is crucial in the navigation systems of USVs (Liu and Bucknall, 2015). It involves devising the optimal route for USVs from a starting point to a destination, primarily considering navigational safety and path efficiency. The goal of path planning is to minimize navigational risks and path costs as much as possible while ensuring mission completion by the USVs (Shu et al., 2023). Currently, various path planning algorithms can be applied in different scenarios, such as the A* algorithm, Dijkstra's algorithm, Artificial Potential Field (APF), Rapidly-Exploring Random Tree (RRT), Genetic Algorithm (GA), and Particle Swarm Optimization (PSO).

The Dijkstra algorithm is a traditional shortest path search algorithm (Dijkstra, 1959). This algorithm identifies the shortest route from an origin to a destination, and finding paths using it is quite simple (Cover and Hart, 1967). Dijkstra's algorithm, however, computes all nodes during path searches, which leads to poor efficiency. Improving computational efficiency involves the selection of the nearest nodes and the exclusion of unnecessary ones (Julius Fusic et al., 2018), which greatly reduces computational load and speeds up the path planning process. The optimal path can be found by calculating the number of turns and travel time through the introduction of a travel time calculation function and in complicated environments, the best route may still not be achievable (Qing et al., 2017).

The A* algorithm, as a heuristic search algorithm (Sang et al., 2021), finds the shortest path between two points. It evaluates the cost from the current node to the target using a heuristic function and expands the most promising nodes. A poorly designed heuristic function can adversely affect the smoothness and continuity of the path (Julius Fusic et al., 2018). Traditional A* can only generate piecewise linear paths, which often results in unsmooth trajectories (Dolgov et al., 2010). Dynamic simplification of the A* algorithm can reduce computation time (Lima et al., 2019). However, the adaptability of the algorithm is insufficient; especially in different scenarios, multiple adjustments of algorithm parameters are required to adapt to changing environments. To obtain safer paths, methods incorporating safe distance maintenance and heuristic function optimization were introduced (Singh et al., 2018). However, manual adjustment of safe distance parameters is required in different scenarios. Additionally, three path smoothing techniques were integrated into the A* algorithm (Song et al., 2019), generating smoother paths with fewer turns. However, the smoothing effect of this algorithm depends on parameter selection and lacks adaptability to different environments.

The basic idea of the APF method is to construct repulsive potential fields around obstacles and an attractive potential field at the target point. The attraction pulls the USVs toward the target, while the repulsion pushes the USVs away from obstacles. This method has a simple computational principle and fast operation speed but easily falls into local optima (Peng et al., 2024). Incorporating Genetic Algorithms into the APF method can effectively alleviate local minima and oscillation problems (Pan et al., 2022). However, the generated paths exhibit frequent turns, and parameter tuning becomes complex, with the design of the fitness function depending on the task scenario. Introducing the temperature parameters of a deterministic annealing strategy into the APF method (Wu et al., 2023) allows the system to increase the temperature when trapped in local minima to escape them. However, this method relies on the initial setting of temperature parameters and cooling rate; improper settings may lead to excessively long paths or failure in obstacle avoidance.Combining Model Predictive Control (MPC) with the APF forms the Model Predictive Artificial Potential Field (MPAPF) method (He et al., 2023). This approach considers the vessel's kinematic constraints and incorporates the International Regulations for Preventing Collisions at Sea (COLREGs), effectively solving the local optimum problem of the traditional APF. However, the path changes direction frequently, affecting the vessel's operational stability.

The RRT is a sampling-based path planning algorithm proposed by LaValle in 1998 (LaValle, 1998). This algorithm takes the starting point as the root node and performs searches in the space using random sampling, continuously adding leaf nodes to form a random tree until it reaches the endpoint. Although this algorithm is highly effective, the process of randomly generating nodes consumes a significant amount of time, and the resulting path is not smooth. By integrating AIS information and Douglas-Peucker (DP) compression to improve the traditional RRT algorithm (Gu et al., 2023), the convergence speed is increased, redundant turning points are reduced, and path smoothness is optimized. However, performance may be limited in areas with insufficient AIS data. By combining Voronoi diagrams to improve the Artificial Potential Field (APF) method (Chi et al., 2022), it guides the sampling of RRT, solves the local optimum problem, and enhances efficiency. However, in environments with fewer obstacles, the path may become longer due to detours. The improved heuristic bidirectional RRT algorithm (Zhang et al., 2022) uses a heuristic biased sampling strategy to reduce ineffective random sampling and increase convergence speed. It also reduces unnecessary turning points through path reorganization. However, in uncertain environments, inaccurate heuristic information may cause the path planning to deviate from the optimal route.

The GA is a bioinspired algorithm for optimisation that identifies the best solution to a problem by mimicking biological processes such as natural selection, inheritance, crossover, and mutation but can act as a general search technique to address path planning problems (Niu et al., 2022). The Genetic Algorithm, however, results in a high computational load, a slow convergence speed, and a tendency to fall into local optima. The addition of a new genetic mutation operator to the GA (Qu et al., 2013) can successfully stop the algorithm from reaching local optima and boost its convergence speed. The GA still raises computational complexity when dealing with extensive data and thus the combination of Voronoi diagrams with the GA (Niu et al., 2020) can markedly reduce the number of redundant nodes in the path, which helps to lower energy consumption and improve path smoothness. However, the algorithm is sensitive to parameter selection. Path planning can be considered a multi-objective optimization problem. By introducing different fitness functions for various objectives (Cheng et al., 2020), the feasibility of the path is ensured, and optimization is performed in terms of time, smoothness, and safety. However, its generality in different environments requires further verification. Using a heuristic median insertion method to generate a high-quality initial population (Li et al., 2021) and optimizing the Genetic Algorithm through multi-objective fitness functions (path length, safety, energy consumption) improved the convergence speed and shortened the path length. However, this method did not perform detailed optimizations on path smoothness.

The PSO (Kennedy and Eberhart, 1995) is another biologically inspired algorithm. It was originally designed to simulate the movement of particles in a solution space, iteratively updating their positions and velocities to search for the optimal solution to a function. The AquaFeL-PSO algorithm (Jara Ten Kathen et al., 2024), which integrates multimodal PSO, Gaussian Processes (GP), and Federated Learning (FL), reduces the likelihood of getting trapped in local optima, while improving both convergence speed and algorithm robustness. However, the Gaussian Process modeling may lead to high computational complexity. Traditional PSO-based path planning algorithms typically assume a static environment, making them less effective in complex dynamic scenarios. To address this limitation, the OkayPlan algorithm (Xin et al., 2024) combines dynamic obstacle motion modeling, Dynamic Priority Initialization (DPI), and a relaxation strategy, significantly enhancing both the safety and real-time performance of path planning. However, the conservative planning strategy of OkayPlan may compromise the optimality of path length. ACO-based path planning, through parameter optimization and adjustment of its search strategies (Heng et al., 2024), can identify the shortest obstacle-free path while ensuring safety. However, in complex environments, ACO is prone to falling into local optima, failing to achieve a truly global optimal path.

In anchorage areas, the high density of anchored vessels complicates traditional path planning, making it difficult to guarantee both safety and efficiency. The close proximity between vessels increases the collision risk for USVs. Therefore, an algorithm that can recognize and avoid risk areas while maintaining path efficiency is proposed. In this paper, a modified A* algorithm, named RAPO, is introduced, which incorporates risk awareness and models the risk field using a Gaussian influence function. After the path is optimized, the DPSS is applied to smooth the path, ensuring its smoothness and feasibility. The main contributions of this paper are as follows:

- The RAPO is proposed, which effectively incorporates the risk characteristics of anchorage areas, thus improving both path safety and economic efficiency.
- A Gaussian influence function is used to model the risk field in the anchorage area, addressing the limitations of the traditional A* algorithm in complex environments.
- The DPSS is applied to smooth the optimized path, ensuring its navigability and smoothness, thereby enhancing its applicability in real-world scenarios.

2 Methodology

2.1 Traditional A* algorithm

The A* algorithm is one of the most widely used methods in path planning. The basic idea involves define the starting point S as the parent node, to estimate the cost to the surrounding nodes n, and selecting the node with the lowest cost as the next parent node until the target node G is identified. Commonly used search directions consist of 4-connected and 8-connected grid searches. The 4-connected mode considers only horizontal and vertical movements, whereas the 8-connected mode additionally accounts for diagonal movements. Due to the complex movement characteristics of USVs in anchorage areas, this paper uses an 8-connected grid search to support more flexible and efficient navigation and thus the node evaluation function consists of two components, as shown in Equation 1:

$$f(n) = g(n) + h(n) \tag{1}$$

where $f(n)$ is the total cost of the current node, $g(n)$ represents the minimum path cost from the starting point S to the current node n and $h(n)$ represents the estimated minimum cost from the current node n to the target node G.

The traditional A* algorithm typically uses heuristic functions such as the Euclidean distance and the Manhattan distance. This paper employs the Euclidean distance, which calculates the straight-line distance between two points to provide an accurate estimation of the path cost. The direct use of the straight-line distance between two points allows for the estimation of movement cost in path planning. Thus the Euclidean distance in the A* algorithm effectively directs the search process to favour paths that are physically nearer to the target, thereby improving search efficiency and reducing computational costs. The heuristic function $h(n)$ is shown in Equation 2, the actual cost $g(n)$ is shown in Equation 3, and the path cost is shown in Equation 4:

$$h(n) = \sqrt{(x_n - x_G)^2 + (y_n - y_G)^2} \tag{2}$$

$$g(n) = \sum_{i=1}^{n} cost(i-1, i) \tag{3}$$

$$cost(i-1, i) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{4}$$

where x_n is the x-coordinate of any node n, y_n is the y-coordinate of node n, x_G is the x-coordinate of the target node G, y_G is the y-coordinate of the target node G, and i is the index of the nodes in the path.

2.2 Risk-aware path optimisation algorithm

The RAPO algorithm was proposed to improve the safety and efficiency of USVs navigation through anchorage areas. The RAPO integrates risk assessment with a dual-phase smoothing strategy. Risk assessment guides the A* algorithm to avoid high-risk areas by evaluating each grid based on a ship domain model and Gaussian influence function. The DPSS smooths the path in two phases. First, Bresenham's algorithm is used to reduce the number of sharp turns. Second, cubic B-spline path smoothing is applied to enhance path continuity.

2.2.1 Risk assessment

The ship domain (Pietrzykowski and Uriasz, 2009) is a concept used to represent the safe area around a vessel. It is typically defined as a two-dimensional area surrounding the vessel, which other ships should avoid to prevent collisions. The size and shape of this domain can vary on the basis of the vessel's size, speed, and navigational environment. The ship domain is usually quantified by boundary radii in four directions around the vessel: forwards (bow), aft (stern), left (port side), and right (starboard side), expressed in multiples of the ship's length (L). The establishment of an unnavigable zone around a ship prevents collision accidents. A typical ship domain representation is illustrated in Figure 2, where the boundary radii in each direction are used to depict the safe zones around the vessel in different orientations.

A reasonable establishment of unnavigable zones can significantly reduce collision risk, improve navigation efficiency, and enhance overall safety (Goerlandt and Kujala, 2014). A dodecagonal forbidden zone model (Kundakçı et al., 2023), which closely approximates an elliptical shape, was proposed by Kundakçı et al. As shown in Figure 3, the dark purple area represents the forbidden zone. In this paper, an elliptical shape was directly adopted for the forbidden zone. Using an elliptical shape for the unnavigable zone around the anchored ship has significant advantages. The long axis of the elliptical unnavigable zone aligns with the longitudinal axis of the ship, providing greater fore-and-aft safety distance. The short axis provides the lateral safety distance, preventing other ships from approaching the sides of the anchored ship and reducing the collision risk. In this paper, elliptical unnavigable zones were set up on the basis of the captain's navigational experience. If the ship's length is L, then the semimajor axis would be $1.2L$; if the ship's width is W, then the semiminor axis would be $2W$. The risk value for grids within the unnavigable zones is set to infinity.

When a USV navigates through an anchorage area, the anchored ships pose a certain risk to the USV. This risk can be characterised by the Gaussian function (Liu and Ma, 2023). The Gaussian function was introduced by the German mathematician Carl Friedrich Gauss. It was first introduced in his work in the early 19th century and has been widely applied in probability theory and

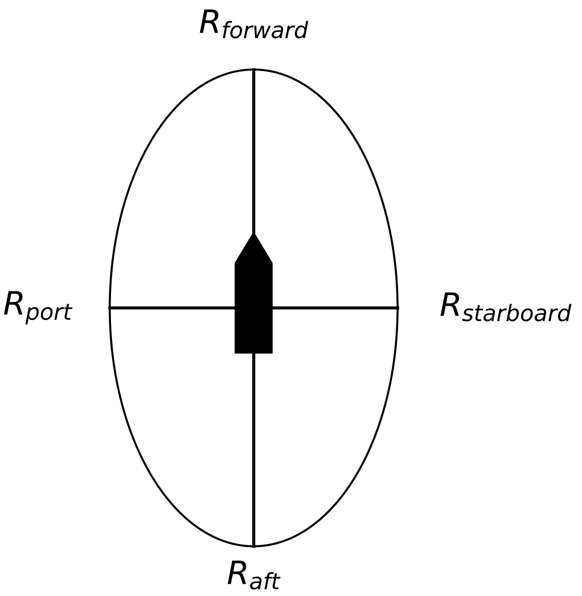


FIGURE 2
Ship domain.

statistics, especially in normal distributions. The normal distribution is one of the most important distributions in statistics and describes the distributions of many natural phenomena and experimental data. The standard form of the Gaussian function is shown in Equation 5, and the graph of the Gaussian function is shown in Figure 4:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

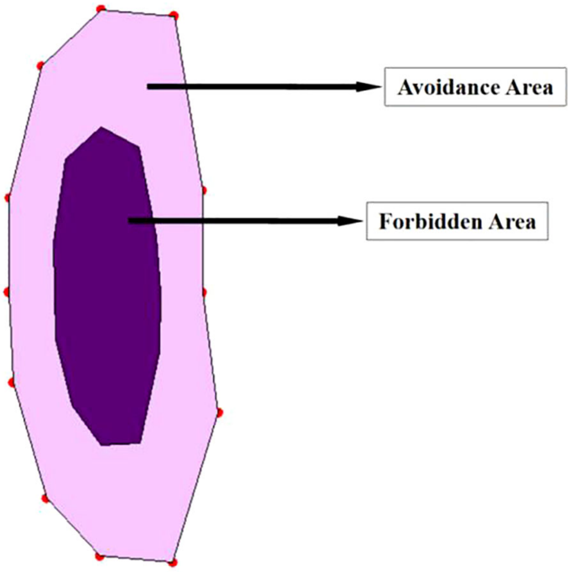


FIGURE 3
Distribution of ship domain with forbidden and avoidance areas.

In Equation 5, μ is the mean, indicating the central position of the Gaussian distribution. It is the symmetric centre of the Gaussian curve, determining its position and controlling the peak position of the curve, which reaches its maximum at $x = \mu$. The σ is the standard deviation, representing the width of the Gaussian distribution, which determines the degree of data dispersion: the larger the standard deviation is, the wider and flatter the curve; the smaller the standard deviation is, the narrower and steeper the curve. In statistics, the normal distribution has an important property known as the three-sigma rule (68-95-99.7 rule), which states that in a normal distribution, approximately 68.27% of the data lie within one standard deviation of the mean $[\mu - \sigma, \mu + \sigma]$, approximately 95.45% of the data lies within two standard deviations $[\mu - 2\sigma, \mu + 2\sigma]$, and approximately 99.73% of the data lies within three standard deviations $[\mu - 3\sigma, \mu + 3\sigma]$.

The Gaussian influence function is a variant of the Gaussian function and is used mainly to describe the exponential influence of a quantity with distance or time. Its form is shown in Equation 6. The three-sigma rule of the Gaussian function also applies to the Gaussian influence function. In the Gaussian influence function, the values range from (0, 1), which aligns with the typical range of risk values.

$$f(x) = e^{-\frac{x^2}{2\sigma^2}} \tag{6}$$

The Gaussian influence function is used to describe the ship domain and assess risks (Im and Luong, 2019), this method is highly reliable and effective. In risk assessment, the Gaussian influence function represents the attenuation of risk with distance, providing an intuitive and computationally simple model for path planning and obstacle avoidance. Its smoothness and symmetry ensure continuity and uniformity in risk distribution, making it especially effective for representing the high risk near anchored ships, where risk diminishes gradually with increasing distance.

In this paper, the map is divided into Voronoi polygons. The distance from each ship to the Voronoi polygon boundary is half of the ship spacing, the risk posed by each anchored ship is confined to the area within its assigned Voronoi polygon. For example, a Gaussian influence function with a parameter of $\sigma = 80$ can be used to depict the risk values, as shown in Figure 5.

The Gaussian influence function ensures that the risk gradually decreases with distance, naturally simulating the risk posed by the anchored ship to its surroundings. By calculating the distance d from a point to the boundary of the unnavigable zone and applying the Gaussian influence function, precise risk assessments can be provided for path planning, thus enhancing navigation safety and the effectiveness of path selection. The map is converted to grids, with d being the distance from the centre of the grid to the boundary of the unnavigable zone, which is calculated as shown in Equation 7. Every grid outside the unnavigable zone has a risk value with the range set to [1,2), and the grid risk function derived from the modified Gaussian influence function is shown in Equation 8.

$$d = \sqrt{(x - x_{edge})^2 + (y - y_{edge})^2} \tag{7}$$

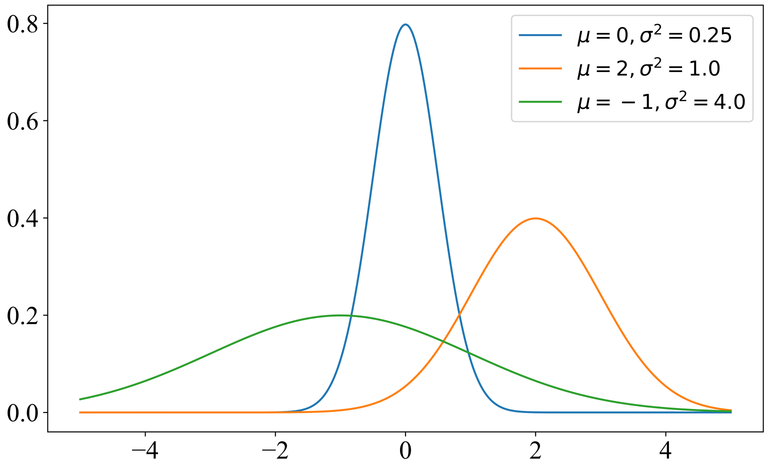


FIGURE 4
Gaussian function plot.

In Equation 7, x and y are the 2-dimensional coordinates of the grid centre, whereas x_{edge} and y_{edge} are the 2-dimensional coordinates of the corresponding point on the ellipse boundary.

$$D(n) = \begin{cases} 1 + e^{-\frac{d^2}{2\sigma^2}}, & d > 0 \\ \infty \ , & d \le 0 \end{cases} \tag{8}$$

In Equation 8, n represents the index or identifier of the current grid point, which is used to indicate its position within the overall risk matrix. $D(n)$ is the grid risk degree function, and d represents the distance from a point to the boundary of the unnavigable zone. When $d > 0$, the point is outside the unnavigable zone, and the risk decreases as the distance increases. When $d = 0$, the point lies on the boundary, and the risk is set to infinity (∞). When $d < 0$, the point is inside the unnavigable zone, and the risk is also set to infinity (∞).

The risk caused by a single ship to its surroundings is displayed on the grid map, with white representing the forbidden zone, and yellow to purple indicating gradually decreasing risk levels, as shown in Figure 6.

The traditional A* algorithm uses only path length as its heuristic function, causing planned paths to often approach obstacles and fail to guide USVs to navigate safely and smoothly. To address this issue, the RAPO incorporates the ship domain and Gaussian influence function to determine the risk zones formed by anchored ships for other vessels. The risk degree function is included as part of the RAPO evaluation function for path planning.

The evaluation function is shown in Equation 9:

$$f(n) = p(n) + h(n) \tag{9}$$

$$p(n) = \sum_{i=1}^{n} cost(i-1, i) \times D(i) \tag{10}$$

$$h(n) = \sqrt{(x_G - x_n)^2 + (y_G - y_n)^2} \tag{11}$$

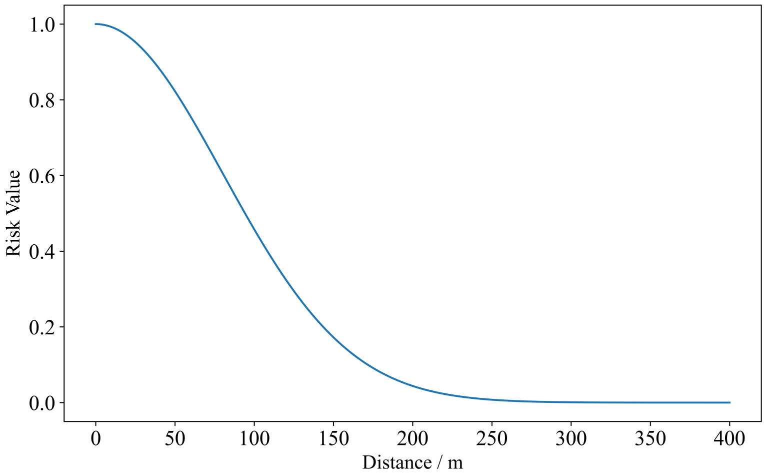


FIGURE 5
Gaussian influence function plot (σ = 80).

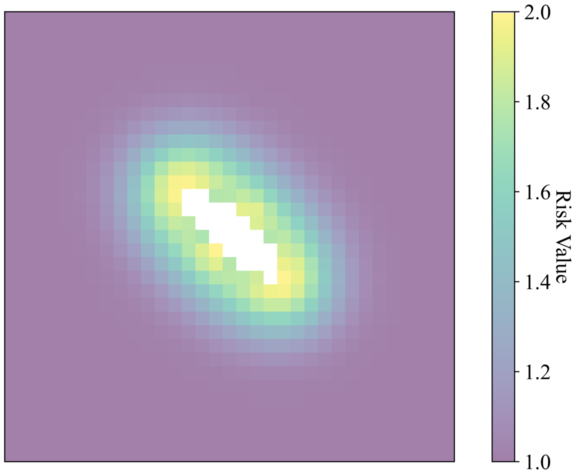


FIGURE 6
Risk distribution around an anchored ship.

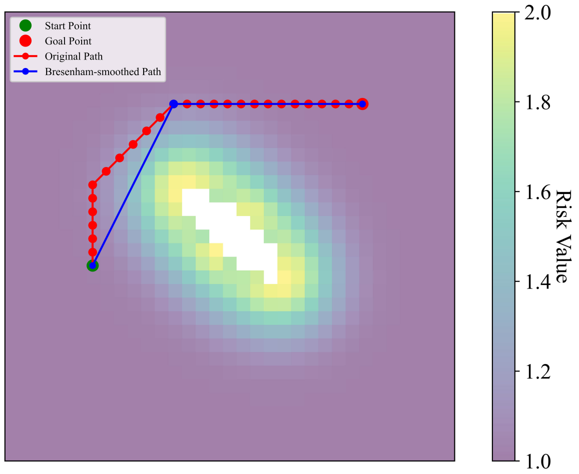


FIGURE 7
Schematic diagram of Bresenham-based path smoothing.

where $f(n)$ is the total cost of the current node $D(n)$ is the grid risk degree function, $p(n)$ represents the path cost from the start point S to the current point n after including the risk, $h(n)$ represents the estimated minimum cost from the current point n to the goal node G, x_G is the x-coordinate of the target node G, and y_G is the y-coordinate of the target node G.

The superiority of Equation 9 over Equation 3 lies in its better consideration of potential collision risks. By introducing the grid risk degree function $D(n)$, USVs can effectively avoid entering unnavigable zones.

2.2.2 Dual-phase smoothing strategy

2.2.2.1 Bresenham-based path smoothing

The RAPO, which incorporates risk assessment, is limited by the heuristic search principle, which does not allow cross-grid search, resulting in many redundant turning points in the planned path. Path smoothing aims to improve the continuity and feasibility of the USV path and lower the energy consumption. In practical applications, path smoothing can significantly enhance the navigation performance and task execution efficiency of USVs. By introducing a path smoothing strategy, the path length can be optimised, removing redundant nodes and unnecessary turns.

The initial path, generated by the RAPO, which incorporates risk assessment, may contain many redundant nodes and turns. To optimise this path, the Bresenham line algorithm (Wang et al., 2024) is used to check the connections between every pair of adjacent nodes, and a schematic of Bresenham line path smoothing is shown in Figure 7. If the risk values of all intermediate nodes between the current node and a distant node are within an acceptable range (below the set threshold), these nodes can be directly connected. By doing so, intermediate redundant nodes are skipped. The pseudocode for the initial smoothing is shown in Algorithm 1.

2.2.2.2 Cubic B-spline-based path smoothing

After the initial smoothing by the Bresenham algorithm, although redundant nodes and sharp turns have been partially reduced, significant angular changes may persist. These changes can lead to large turning angles, increasing energy consumption and operational difficulty for USVs during actual operation. To further optimise the smoothness and continuity of the path, a path smoothing method based on cubic B-splines (Muñoz, 2008) was introduced in the second stage of the DPSS. The mathematical definition of the B-spline curve is shown in Equation 12:

$$\mathrm{C}(u)=\sum_{i=0}^{n}N_{i,k}(u)P_i \tag{12}$$

In Equation 12, $\mathrm{C}(u)$ represents the point on the curve at parameter u, P_i is the i_{th} control point, and $N_{i,k}(u)$ is the B-spline basis function, with k=3 indicating a cubic B-spline.

The recursive definition of the cubic B-spline basis function $N_{i,3}(u)$ is as follows:

For the zeroth-degree B-spline basis function, as shown in Equation 13:

$$N_{i,0}(u)=\begin{cases}1 \ \text{if} \quad u_i\le u<u_{i+1}\\ 0 \ \text{otherwise}\end{cases} \tag{13}$$

For higher-degree B-spline basis functions, as shown in Equation 14:

$$N_{i,k}(u)=\frac{u-u_i}{u_{i+k}-u_i}N_{i,k-1}(u)+\frac{u_{i+k+1}-u}{u_{i+k+1}-u_{i+1}}N_{i+1,k-1}(u) \tag{14}$$

To generate smooth B-spline curves, uniformly distributed knot vectors were adopted. If there are n+1 control points, the knot vectors are typically defined as:

$$u=\{u_0,u_1,\dots,u_{k-1},u_k,\dots,u_n,u_{n+1},\dots,u_{n+k}\} \tag{15}$$

```
Algorithm: BresenhamLineSmoothPath
Input: path - a list of points forming the initial path
       risk_matrix - a 2D grid representing risk values of the area
       threshold - maximum acceptable risk value for a path to be
       considered safe
Output: smoothed_path - a list of points forming the smoothed path
Initialise smoothed_path with the first point of path
Set skip to 0
For each point i from 1 to the second last point of path
    If skip is not zero then
        Decrement skip
        Continue to the next iteration of the loop
    End If
    For each point j from end of path down to i + 1
        Generate all points on the line from the last point of smoothed_path to path[j] using the BresenhamLine function
        If all points on the line have a risk value<= threshold Then
            Add path[j] to smoothed_path
            Set skip to j - i - 1
            Break the inner loop
        EndIf
    EndFor
    If no suitable connection point was identified Then
        Add path[i] to smoothed_path
    End If
End For
Return smoothed_path
```

ALGORITHM 1 Bresenham-based path smoothing pseudocode..

These uniformly distributed knot vectors ensure a smooth transition between control points in the B-spline curve.

In accordance with the standards set forth in the U.S. Navy's "Navy USV Master Plan", USVs with lengths ranging from 3 to 11 metres are widely employed in various mission scenarios. In this paper, a typical 10-metre USV with a turning radius of approximately 30 metres was selected. The 30-meter insertion interval not only meets the requirements for path smoothing but also aligns with the maneuvering characteristics of the USV, ensuring that the generated path is both operationally stable and feasible. Consequently, control points were inserted every 30 metres. A schematic of the second smoothing is shown in Figure 8. The pseudocode for the second path smoothing is shown in Algorithm 2.

The cubic B-spline method significantly improved path smoothness, reduced the number of sharp turns, enhanced the navigational stability of the USV, and optimised the path's continuity and length. As a result, the economic efficiency and safety of the generated path in complex environments were effectively improved.

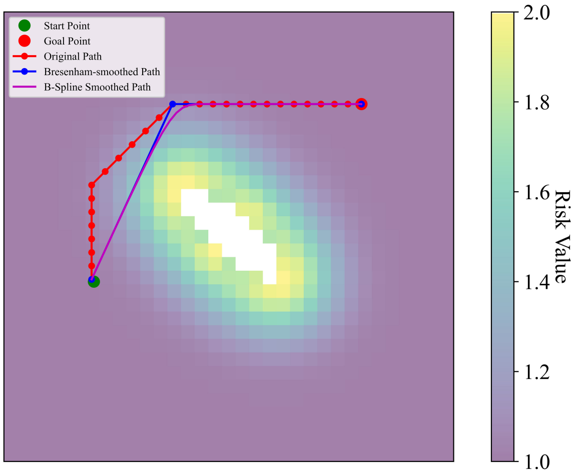


FIGURE 8
Schematic diagram of the path smoothing process using cubic B-Splines.

3 Simulation experiments

3.1 Experimental environment setup

All simulations were conducted on a computer with Microsoft Windows 11 as the operating system, an Intel i5 3.10 GHz twelve-core CPU, and 16 GB of RAM. To validate the rationality and efficiency of the RAPO algorithm proposed in this paper, simulations were carried out on a 2D static grid map with PyCharm as the development environment.

3.2 Anchorage area model construction

3.2.1 Ship positioning and Voronoi polygon partitioning

In this paper, anchorages and anchored ships in Beibu Gulf waters were referred to. The simulated anchorage size was set to 5.5 km × 4.8 km. Sixty anchored ships, each with lengths ranging from 90 to 150 m, were included. The distance between ships was set to 500 to 750 metres. The heading of each ship was uniformly distributed within the range of 135° to 165°. The coordinates of the anchored ships were set to determine their positions. Thirteen ships with lengths of 90 to 110 m are represented by green dots. Twenty-four ships with lengths of 110 to 130 m are represented by blue dots. Twenty-three ships with lengths of 130 to 150 m are represented by red dots. The anchored ships were used as points P_i to partition the anchorage area via the Voronoi polygon. This process prepares for the introduction of risk from the anchored ships. With Voronoi polygon partitioning, the distribution of the simulated ships in the defined anchorage area is shown in Figure 9.

3.2.2 Grid-based processing and risk evaluation

When processing environmental maps, grid-based maps are the most commonly used form of representation and processing, as they effectively convey spatial information and support the application of various algorithms. The grid size was set to 30 m × 30 m, considering that the normal length of a USV is approximately 10 m. The resulting grid map used in the risk assessment is shown in Figure 10.

Before the simulation experiments, risk values were assigned to each grid on the basis of the Gaussian influence function. Each anchored ship formed a risk area. The anchored ships were used as seed points for the Voronoi polygons. Each polygon was a risk assessment unit. It was assumed that each anchored ship affected only the navigable waters within its corresponding Voronoi polygon. The distance between anchored ships ranges from 500 to 750 m, and the shortest distance from an anchored ship to the boundary of its Voronoi polygon is approximately 250 m. Considering that the main influence range of the Gaussian distribution is concentrated within $[-3\sigma,+3\sigma]$, corresponding to an actual risk range of 250 m, $3\sigma=250$ is set, yielding $\sigma \approx 80$. So, the parameter σ in the Gaussian influence function was set to 80.

```
Algorithm: B-SplineSmoothPath
Input: smoothed_path: A list of points forming the smoothed path after the first smoothing. interval: The distance interval for inserting control points along the smoothed path (set to 30 metres). degree: The degree of the B-Spline (set to 3).
Output: b_spline_path: A list of points forming the final smoothed B-Spline path.
Initialise control_points as an empty list.
For each pair of consecutive points (start_point, end_point) in smoothed_path:
    Calculate the segment_length between start_point and end_point.
    If segment_length > 0:
        Calculate the number of control points to insert (num_points = segment_length//interval).
        For each j from 0 to num_points:
            Calculate the interpolated point between start_point and end_point using linear interpolation.
            Add the interpolated point to control_points.
    Else:
        Skip the segment (if start_point and end_point are identical).
Add the last point of smoothed_path to control_points.
Generate a uniform knot vector based on the number of control_points and the degree of the B-Spline.
Create a B-Spline curve using the control_points and the knot vector.
Generate a dense set of points along the B-Spline curve to represent the final smoothed path (b_spline_path).
Return b_spline_path.
```

ALGORITHM 2 B-Spline-based path smoothing pseudocode..

After the map was converted to grids, each grid was assigned a risk value. The risk value for unnavigable zones was set to infinity. Grids in this area are displayed in white. The risk values for risk zones ranged from 1 to 2, with colours representing the risk value from purple (low risk) to yellow (high risk), transitioning through cyan and green. A risk distribution map of anchored ships is shown in Figure 11, where "Start" is the starting point and "Goal" is the ending point.

3.3 Path planning and smoothing

The RAPO algorithm was used to plan safe and efficient paths for USVs in anchorage areas. First, a modified Gaussian influence function was used to conduct risk assessments to minimize potential risks. Then, the algorithm optimises the path through two stages of DPSS path smoothing. In the first phase, a Bresenham-

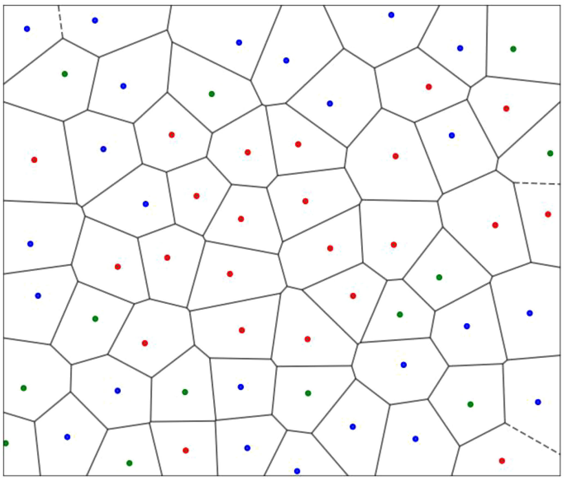


FIGURE 9
Simulated ship positions and Voronoi polygon partitioning in the experiment.

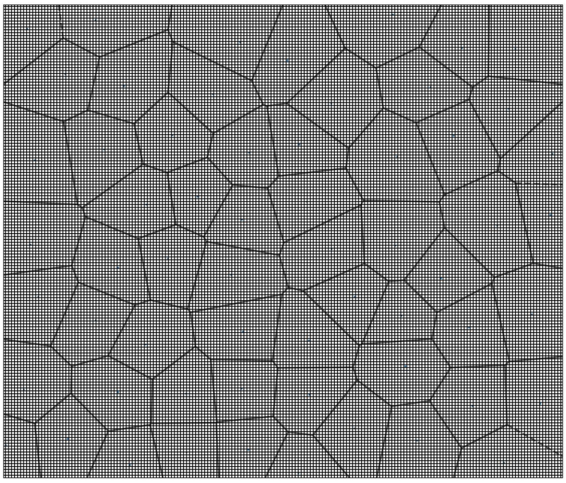


FIGURE 10
Simulated ship positions and Voronoi polygon partitioning in the experiment.

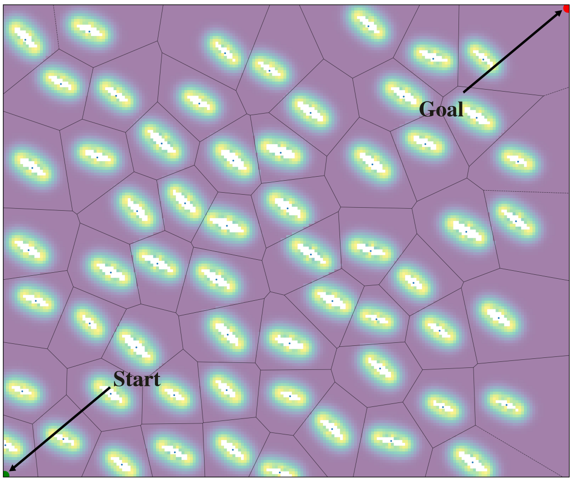


FIGURE 11
Risk distribution map of anchored ships.

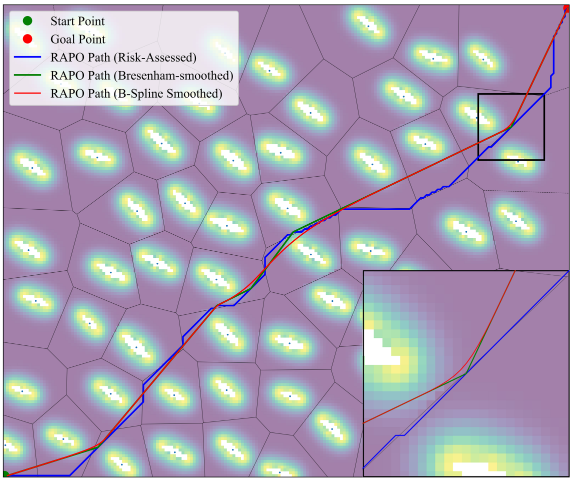


FIGURE 12
Path planning outcomes under a risk tolerance of 1.5 using the RAPO algorithm.

based path smoothing method is employed to eliminate unnecessary turns and redundant nodes in the initial path. In the second phase, a cubic B-spline-based path smoothing method is used to further smooth the path obtained from the first phase, inserting a control point every 30 meters on the path obtained from the first-phase smoothing, and then applying a cubic B-spline curve to smooth the path. After the DPSS, the number of turns is significantly reduced, and the smoothness of the path is improved.

3.4 Simulation results

The RAPO algorithm integrates risk assessment and the DPSS, to verify that the RAPO algorithm can be applied to path planning in anchorage areas, simulation experiments were conducted. The path planning results of the RAPO algorithm at a path risk value of 1.5 are shown in Figure 12, where the blue solid line represents the initial path from the risk-improved A* algorithm within RAPO, the black solid line indicates the path after the first-phase smoothing based on Bresenham's algorithm, and the red solid line shows the final path after the second-phase smoothing using a cubic B-spline.

The path planning results indicate that the RAPO algorithm, which includes DPSS, significantly improves both path length and the number of turns across different path risk tolerances. Table 1 presents the path lengths, number of turns, and maximum turning angles for the RAPO algorithm under path risk tolerances of 1.2, 1.5, and 1.8. Compared with the original paths generated through risk assessment within the RAPO algorithm, the lengths of the smoothed paths were reduced by 7.13%, 7.60%, and 7.70%, respectively. The number of turns decreased by 81.13%, 90.57%, and 94.34%, respectively, while the maximum turning angle was reduced by 17.78%, 13.33%, and 11.11%, respectively. When comparing the smoothed paths at different risk tolerances, the path length with a risk tolerance of 1.5 was reduced by 4.9%, and the number of turns decreased by 57.14% compared with the smoothed path with a risk tolerance of 1.2. The path length at a risk tolerance of 1.8 is reduced by 0.51% compared to that at a risk tolerance of 1.5, and the number of turns decreases by 50%. The path length at a risk tolerance of 1.8 is reduced by 0.61% compared to that at a risk tolerance of 1.2, and the number of turns decreases by 70%. As the risk tolerance increases, the resulting path length continuously shortens, and the number of turns decreases, thereby reducing the operational difficulty and energy consumption of the USV, thus ensuring the economy of the path.

4 Discussions

The RAPO algorithm proposed in this paper first assesses the risk of anchored ships and then plans a route, while also smoothing the route to ensure a safe and economical path for USVs in anchorage areas. The results from simulation experiments demonstrate that the RAPO algorithm outperforms the A* algorithm (Hart et al., 1968), the Voronoi-based A* algorithm

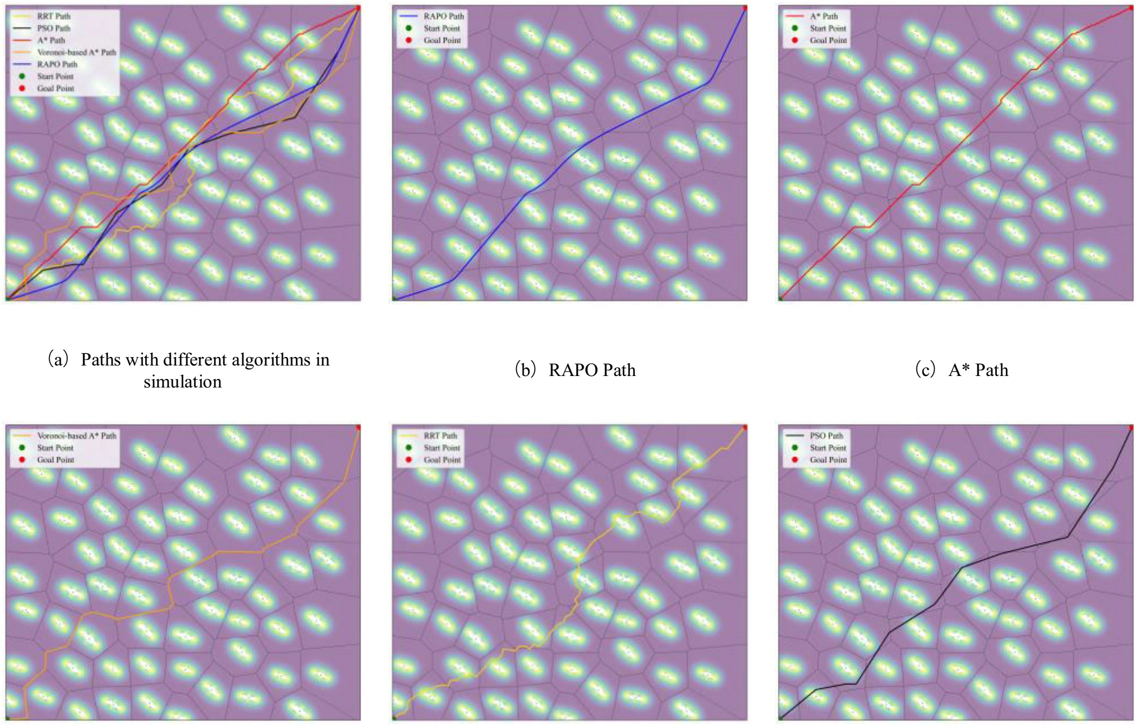
TABLE 1 Comparison of path planning of the RAPO algorithm under different risk tolerances.

Risk Tolerance	Original Path Length (Risk-Assessed) (m)	DPSS Smoothed Path Length (m)	Original Number of Turns	Smoothed Number of Turns	Original Max Turning Angle (°)	Smoothed Max Turning Angle (°)
1.2	8.641	8.025	53	10	45°	37°
1.5	8.641	7.984	53	5	45°	39°
1.8	8.641	7.976	53	3	45°	40°

(Fedorenko and Gurenko, 2016), RRT algorithm (LaValle, 1998) and PSO algorithm (Kennedy and Eberhart, 1995) in terms of path length, the number of turns as well as a path smoothness.

Figure 13 illustrates the paths obtained by the five algorithms. The blue solid line represents the path generated by the RAPO algorithm, the red solid line represents the path produced by the traditional A* algorithm, the orange solid line shows the path from the Voronoi-based A* algorithm, the golden yellow solid line represents the path obtained by the RRT algorithm, and the black solid line represents the path from the PSO algorithm.

When the risk tolerance is 1.5, path planning was conducted using five different algorithms, and the simulation results are shown in Table 2. In terms of path length, the RAPO algorithm resulted in a path length of 7.984 km, which is significantly shorter than the paths obtained by the other four algorithms. It can be seen that while considering risk factors to ensure path safety, its path length is also the shortest, and the overall path length was further optimized after applying the DPSS. Regarding the number of turns, the path obtained by the RAPO algorithm has only 5 turns, which is significantly fewer than the 39 turns of the traditional A* algorithm, the 20 turns of the Voronoi-based A* algorithm, the 68 turns of the RRT algorithm, and the 7 turns of the PSO algorithm. In terms of the maximum turning angle, the path generated by the RAPO algorithm has a maximum turn of only 40°, which is significantly lower than the 45° of the traditional A* algorithm, the 90° of the Voronoi-based A* algorithm, the 128° of the RRT algorithm, and the 57° of the PSO algorithm. It can be seen that the path smoothing phase in the RAPO algorithm effectively reduces unnecessary turns, enhances path smoothness, and decreases the operational difficulty and energy consumption of USVs. Additionally, the maximum risk value of the path obtained by the RAPO algorithm is 1.484, which, although higher than that of the path obtained by the Voronoi-based A* algorithm, is still within the



(a) Paths with different algorithms in simulation

(b) RAPO Path

(c) A* Path

(d) Voronoi-based A* Path

(e) RRT Path

(f) PSO Path

FIGURE 13
Path planning outcomes by different algorithms.

TABLE 2 Comparison of different path planning algorithms with risk tolerance of 1.5.

Algorithm Type	Path Length (km)	Number of Turns	Maximum Turning Angle (°)	Maximum Risk Value
RAPO	7.984	5	40°	1.484
Traditional A*	8.013	39	45°	2
Voronoi-based A*	9.257	20	90°	1.214
RRT	9.299	68	128°	2
PSO	8.991	7	57°	1.97

set range. Therefore, the RAPO algorithm can plan a safe path for USVs in anchorage areas.

The traditional A* algorithm focuses solely on finding the shortest path, without considering path safety, resulting in poor overall path safety. Additionally, the traditional A* algorithm generates paths with numerous redundant turns, which increases operational complexity and energy consumption. Although the Voronoi-based A* algorithm considers path safety, it does not optimize path length, resulting in longer paths. Furthermore, the paths are constrained by the boundaries of Voronoi polygons, leading to more sharp turns, which further increases operational difficulty and energy consumption. The RRT algorithm lacks path smoothness in path planning, generating longer paths with excessive sharp turns and limited overall optimization capability. Although the PSO algorithm demonstrates certain global optimization capabilities, its generated paths perform poorly in risk avoidance, making it difficult to ensure path safety.

The RAPO algorithm mainly combines risk assessment with the DPSS. The addition of risk assessment to path planning allows the path to successfully bypass high-risk areas. The DPSS process eliminates a large quantity of unneeded turns and improves the flow of the path. The RAPO algorithm is capable of designing routes for USVs in challenging environments, ensuring both the safety and economy of the path, and also making USVs operations less difficult and less energy intensive.

5 Conclusions

This paper proposes the RAPO algorithm to enhance the safety and efficiency of USVs in anchorage areas. The algorithm integrates a grid-based risk function derived from the ship domain model, a Gaussian influence function, and the DPSS. By defining prohibited zones using the ship domain and conducting risk assessments on waters outside these zones with the Gaussian influence function, the algorithm effectively avoids high-risk areas, improving the safety of path planning. Furthermore, the DPSS reduces the number of turns, resulting in a smoother and more efficient planned path.

Still, the algorithm has some inherent limitations. Initially, the algorithm's computational burden is quite high, which leads to an increase in time needed and the smoothing results are contingent upon the parameter settings. Then, the algorithm is currently mostly focused on static environments, which may influence its use in real complex marine settings.

In future research, the role of ocean currents in the navigation environment will be examined to better understand USVs navigation in anchorage areas. In addition, the exploration of path planning for USVs in dynamic environments with both static and dynamic obstacles will be undertaken to further improve the practicality of the RAPO algorithm, allowing it to perform well in static environments and to provide safe and effective path planning in complex dynamic settings.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HW: Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. SM: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. XM: Writing – original draft. JZ: Writing – original draft. RL: Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (51479157, 52171346), China Scholarship Council Program (202208440272), Hubei Key Laboratory of Inland Shipping Technology (NHHY2020002), the Special Projects of Key Fields of Universities in Guangdong Province (2023ZDZX3003) and Teaching Quality and Teaching Reform Project of Guangdong Undergraduate Colleges and Universities (010201132401).

Acknowledgments

Special thanks go to the funding support provided for this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bae, I., and Hong, J. (2023). Survey on the developments of unmanned marine vehicles: intelligence and cooperation. *Sensors* 23, 4643. doi: 10.3390/s23104643

Cheng, K. P., Mohan, R. E., Khanh Nhan, N. H., and Le, A. V. (2020). Multi-objective genetic algorithm-based autonomous path planning for hinged-tetro reconfigurable tiling robot. *IEEE Access* 8, 121267–121284. doi: 10.1109/ACCESS.2020.3006579

Chi, W., Ding, Z., Wang, J., Chen, G., and Sun, L. (2022). A generalized voronoi diagram-based efficient heuristic path planning method for RRTs in mobile robots. *IEEE Trans. Ind. Electron.* 69, 4926–4937. doi: 10.1109/TIE.2021.3078390

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numer. Math.* 1, 269–271. doi: 10.1007/BF01386390

Dolgov, D., Thrun, S., Montemerlo, M., and Diebel, J. (2010). Path planning for autonomous vehicles in unknown semi-structured environments. *Int. J. Robotics Res.* 29, 485–501. doi: 10.1177/0278364909359210

Fedorenko, R., and Gurenko, B. (2016). Local and global motion planning for unmanned surface vehicle. *MATEC Web Conferences* 42, 1005. doi: 10.1051/matecconf/20164201005

Goerlandt, F., and Kujala, P. (2014). On the reliability and validity of ship–ship collision risk analysis in light of different perspectives on risk. *Saf. Sci.* 62, 348–365. doi: 10.1016/j.ssci.2013.09.010

Gu, Q., Zhen, R., Liu, J., and Li, C. (2023). An improved RRT algorithm based on prior AIS information and DP compression for ship path planning. *Ocean Eng.* 279, 114595. doi: 10.1016/j.oceaneng.2023.114595

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybernetics* 4, 100–107. doi: 10.1109/TSSC.1968.300136

He, Z., Chu, X., Liu, C., and Wu, W. (2023). A novel model predictive artificial potential field based ship motion planning method considering COLREGs for complex encounter scenarios. *ISA Trans.* 134, 58–73. doi: 10.1016/j.isatra.2022.09.007

Heng, H., Ghazali, M. H. M., and Rahiman, W. (2024). Exploring the application of ant colony optimization in path planning for Unmanned Surface Vehicles. *Ocean Eng.* 311, 118738. doi: 10.1016/j.oceaneng.2024.118738

Im, N., and Luong, T. N. (2019). Potential risk ship domain as a danger criterion for real-time ship collision risk evaluation. *Ocean Eng.* 194, 106610. doi: 10.1016/j.oceaneng.2019.106610

Jara Ten Kathen, M., Peralta, F., Johnson, P., Jurado Flores, I., and Gutiérrez Reina, D. (2024). AquaFeL-PSO: An informative path planning for water resources monitoring using autonomous surface vehicles based on multi-modal PSO and federated learning. *Ocean Eng.* 311, 118787. doi: 10.1016/j.oceaneng.2024.118787

Julius Fusic, S., Ramkumar, P., and Hariharan, K. (2018). "Path planning of robot using modified dijkstra Algorithm," in *2018 National Power Engineering Conference (NPEC)* (IEEE, Madurai), 1–5. doi: 10.1109/NPEC.2018.8476787

Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). IEEE. doi: 10.1109/ICNN.1995.488968

Kundakçı, B., Nas, S., and Gucma, L. (2023). Prediction of ship domain on coastal waters by using AIS data. *Ocean Eng.* 273, 113921. doi: 10.1016/j.oceaneng.2023.113921

LaValle, S. (1998).Rapidly-exploring random trees : a new tool for path planning. In: *Research Report 9811*. Available online at: https://cir.nii.ac.jp/crid/1573950399665672960 (Accessed July 23, 2024).

Li, K., Hu, Q., and Liu, J. (2021). Path planning of mobile robot based on improved multiobjective genetic algorithm. *Wireless Commun. Mobile Computing* 2021, 8836615. doi: 10.1155/2021/8836615

Lima, J., Costa, P., Costa, P., Eckert, L., Piardi, L., Moreira, A. P., et al. (2019). A* search algorithm optimization path planning in mobile robots scenarios. *AIP Conf. Proc.* 2116, 220005. doi: 10.1063/1.5114223

Liu, Y., and Bucknall, R. (2015). Path planning algorithm for unmanned surface vehicle formations in a practical maritime environment. *Ocean Engineering.* 97, 126–144. doi: 10.1016/j.oceaneng.2015.01.008

Liu, Y., and Ma, Y. (2023). A field theory-based novel algorithm for navigational hazard index. *J. Mar. Sci. Eng.* 11, 178. doi: 10.3390/jmse11010178

Muñoz, J. J. (2008). Modelling unilateral frictionless contact using the null-space method and cubic *B*-Spline interpolation. *Comput. Methods Appl. Mechanics Eng.* 197, 979–993. doi: 10.1016/j.cma.2007.09.022

Niu, H., Ji, Z., Savvaris, A., and Tsourdos, A. (2020). Energy efficient path planning for Unmanned Surface Vehicle in spatially-temporally variant environment. *Ocean Eng.* 196, 106766. doi: 10.1016/j.oceaneng.2019.106766

Niu, Y., Zhang, J., Wang, Y., Yang, H., and Mu, Y. (2022). "A Review of Path Planning Algorithms for USV," in *Proceedings of 2021 International Conference on Autonomous Unmanned Systems (ICAUS 2021).* Eds. M. Wu, Y. Niu, M. Gu and J. Cheng (Springer, Singapore), 263–273. doi: 10.1007/978-981-16-9492-9_27

Pan, Z., Zhang, C., Xia, Y., Xiong, H., and Shao, X. (2022). An improved artificial potential field method for path planning and formation control of the multi-UAV systems. *IEEE Trans. Circuits Syst. II: Express Briefs* 69, 1129–1133. doi: 10.1109/TCSII.2021.3112787

Peng, B., Zhang, L., and Xiong, R. (2024). Smooth path planning with subharmonic artificial potential field. doi: 10.48550/arXiv.2402.11601

Pietrzykowski, Z., and Uriasz, J. (2009). The ship domain – A criterion of navigational safety assessment in an open sea area. *J. Navigation* 62, 93–108. doi: 10.1017/S0373463308005018

Qing, G., Zheng, Z., and Yue, X. (2017). "Path-planning of automated guided vehicle based on improved Dijkstra algorithm," in *2017 29th Chinese Control And Decision Conference (CCDC)* (IEEE, Chongqing, China), 7138–7143. doi: 10.1109/CCDC.2017.7978471

Qu, H., Xing, K., and Alexander, T. (2013). An improved genetic algorithm with co-evolutionary strategy for global path planning of multiple mobile robots. *Neurocomputing* 120, 509–517. doi: 10.1016/j.neucom.2013.04.020

Sang, H., You, Y., Sun, X., Zhou, Y., and Liu, F. (2021). The hybrid path planning algorithm based on improved A* and artificial potential field for unmanned surface vehicle formations. *Ocean Eng.* 223, 108709. doi: 10.1016/j.oceaneng.2021.108709

Shu, Y., Zhu, Y., Xu, F., Gan, L., Lee, P. T.-W., Yin, J., et al. (2023). Path planning for ships assisted by the icebreaker in ice-covered waters in the Northern Sea Route based on optimal control. *Ocean Eng.* 267, 113182. doi: 10.1016/j.oceaneng.2022.113182

Singh, Y., Sharma, S., Sutton, R., Hatton, D., and Khan, A. (2018). A constrained A* approach towards optimal path planning for an unmanned surface vehicle in a maritime environment containing dynamic obstacles and ocean currents. *Ocean Eng.* 169, 187–201. doi: 10.1016/j.oceaneng.2018.09.016

Song, R., Liu, Y., and Bucknall, R. (2019). Smoothed A* algorithm for practical unmanned surface vehicle path planning. *Appl. Ocean Res.* 83, 9–20. doi: 10.1016/j.apor.2018.12.001

Specht, C., Świtalski, E., and Specht, M. (2017). Application of an autonomous/unmanned survey vessel (ASV/USV) in bathymetric measurements. *Polish Maritime Res.* 24, 36–44. doi: 10.1515/pomr-2017-0088

Wang, J., Zhou, K., Xing, W., Li, H., and Yang, Z. (2023). Applications, evolutions, and challenges of drones in maritime transport. *JMSE* 11, 2056. doi: 10.3390/jmse11112056

Wang, S., Li, J., He, Y., and Gen, N. (2024). "Path Planning of Mobile Robot Based on Bresen-Ham Line Algorithm Improved A* Algorithm," in *Advances in Mechanical Design*. Eds. J. Tan, Y. Liu, H.-Z. Huang, J. Yu and Z. Wang (Springer Nature, Singapore), 2055–2066. doi: 10.1007/978-981-97-0922-9_130

Wu, Z., Dai, J., Jiang, B., and Karimi, H. R. (2023). Robot path planning based on artificial potential field with deterministic annealing. *ISA Trans.* 138, 74–87. doi: 10.1016/j.isatra.2023.02.018

Xin, J., Kim, J., Chu, S., and Li, N. (2024). OkayPlan: Obstacle Kinematics Augmented Dynamic real-time path Planning via particle swarm optimization. *Ocean Eng.* 303, 117841. doi: 10.1016/j.oceaneng.2024.117841

Yin, J., and Wang, N. (2021). Predictive trajectory tracking control of autonomous underwater vehicles based on variable fuzzy predictor. *Int. J. Fuzzy Syst.* 23, 1809–1822. doi: 10.1007/s40815-020-00898-7

Yin, J., Wang, H., Wang, N., and Wang, X. (2023). An adaptive real-time modular tidal level prediction mechanism based on EMD and Lipschitz quotients method. *Ocean Eng.* 289, 116297. doi: 10.1016/j.oceaneng.2023.116297

Zhang, X., Zhu, T., Du, L., Hu, Y., and Liu, H. (2022). Local path planning of autonomous vehicle based on an improved heuristic bi-RRT algorithm in dynamic obstacle avoidance environment. *Sensors* 22, 7968. doi: 10.3390/s22207968

Zhou, M., Cao, L., Liu, J., Zhang, Z., Liang, Z., Cui, Z., et al. (2024). Research on intelligent three-dimensional anchor position detection method for ships utilizing Traversal and Monte Carlo algorithms. *Front. Mar. Sci.* 11. doi: 10.3389/fmars.2024.1471328

TYPE Original Research
PUBLISHED 14 April 2025
DOI 10.3389/fmars.2025.1495822

Improved deep learning method and high-resolution reanalysis model-based intelligent marine navigation

Zeguo Zhang[1,2,3†], Liang Cao[1,2,3†] and Jianchuan Yin[1,2,3*]

[1]Naval Architecture and Shipping College, Guangdong Ocean University, Zhanjiang, China, [2]Guangdong Provincial Key Laboratory of Intelligent Equipment for South China Sea Marine Ranching, Zhanjiang, China, [3]Guangdong Provincial Engineering Research Center for Ship Intelligence and Safety, Zhanjiang, China

OPEN ACCESS

EDITED BY
Jin Liu,
Shanghai Maritime University, China

REVIEWED BY
Xinjian Wang,
Dalian Maritime University, China
Bing Wu,
Wuhan University of Technology, China

*CORRESPONDENCE
Jianchuan Yin
✉ jianchuanyin_gdou@163.com

†These authors have contributed equally to this work

RECEIVED 13 September 2024
ACCEPTED 11 March 2025
PUBLISHED 14 April 2025

CITATION
Zhang Z, Cao L and Yin J (2025) Improved deep learning method and high-resolution reanalysis model-based intelligent marine navigation.
Front. Mar. Sci. 12:1495822.
doi: 10.3389/fmars.2025.1495822

Large-scale weather forecasting is critical for ensuring maritime safety and optimizing transoceanic voyages. However, sparse meteorological data, incomplete forecasts, and unreliable communication hinder accurate, high-resolution wind system predictions. This study addresses these challenges to enhance dynamic voyage planning and intelligent ship navigation. We propose IPCA-MHA-DSRU-Net, a novel deep learning model integrating incremental principal component analysis (IPCA) with a spatial-temporal depthwise separable U-Net. Key components include: (1) IPCA preprocessing to reduce dimensionality and noise in 2D wind field data; (2) depthwise-separable convolution (DSC) blocks to minimize parameters and computational costs; (3) multi-head attention (MHA) and residual mechanisms to improve spatial-temporal feature extraction and prediction accuracy. The framework is optimized for real-time onboard deployment under communication constraints. The model achieves high accuracy in high-resolution wind predictions, validated through reanalysis datasets. Experiments demonstrated enhanced path planning efficiency and robustness in dynamic oceanic conditions. The IPCA-MHA-DSRU-Net balances computational efficiency and accuracy, making it viable for resource-limited ships. This novel IPCA application provides a promising alternative for preprocessing large-scale meteorological data.

KEYWORDS

extreme wind forecast, machine learning, marine navigation, incremental principal component analysis, depthwise-separable convolution

1 Introduction

Marine transportation has been recognized as one of the indispensable transport models for developing a global logistics network. In recent years, with the rapid development of global trade and the vast expansion of the supply chain network, the demand for reliable and efficient marine transport has increased sharply (Koukaki and Tei, 2020). Yet, potential challenges and risks arise for sea-going vessels when it comes to long-distance path planning due to the instability and unpredictability of the meteorological environment resulting in too much uncertainty (Lau et al., 2024). This is especially so when encountering adverse sea conditions, such as extreme wind and wave scenarios, that can significantly impede ship navigation, thus, requiring timely speed reduction and route deviation so as to ensure safety (Rawson et al., 2021). Ocean state conditions can significantly impact the safety and decision-making of marine vehicles. Although shipping route recommendations could be obtained from weather routing companies (Szlapczynski et al., 2023), real-time access to weather forecasts is becoming more crucial for underway ships. Accurate and timely weather forecasting can support the captain in designing and determining the ship's path in advance and further ensure the safety of mariners and ships. More importantly, efficient and handy onboard weather predictions can provide invaluable marine environment references for intelligent navigation (He et al., 2022). Accurate and fine-grid weather predictions are essential for the seaworthiness and safety of sea-going ships, especially during long transoceanic voyages, where vessels are exposed to the open ocean's full range of meteorological and oceanographic phenomena. These voyages can last days or weeks, during which weather and sea states can change rapidly and drastically, impacting both the physical safety of the vessel and the efficiency of its journey. Fine-scale weather predictions play a critical role in enhancing situational awareness for shipping operations, enabling them to anticipate and mitigate risks associated with severe sea conditions, such as strong winds, and intense storms. For instance, accurate, high-resolution weather forecasts enable route planning to avoid severe weather, which reduces fuel consumption, lowers operational costs, and minimizes emissions. Given the substantial size and fuel requirements of ocean-going vessels, even minor deviations from optimal weather conditions can result in significant additional fuel consumption, which contributes to both increased costs and environmental impact. Fine-grid predictions allow for precise navigational adjustments that align with favorable weather patterns, helping ships follow safer and more efficient routes. Moreover, a precise forecast of extreme wind on a fine grid can give shipping operators and crew sufficient warning to take preventive measures, such as adjusting speed, changing course, or securing loose cargo. For crews, these predictions mean better preparation and safety measures, reducing the likelihood of accidents or fatalities. As a consequence, providing accurate and efficient meteorological prediction is crucial for achieving intelligent, safe, and green ship path planning (Zis et al., 2020).

Classical ocean and meteorology forecasting relies on the numerical weather prediction model (NWP). It uses the collected meteorological parameters, geographical boundaries, and initial conditions to predict weather variability based on a physical conservation equation (Cheng et al., 2013; Hur, 2021). Nevertheless, the inherent instability and stochasticity characteristics of earth system evolution make it challenging to forecast global weather using deterministic weather forecast models. In addition, with the increasing complexity, higher uncertainty, and variability of earth systems due to global climate changes, traditional numerical forecasting models tend to fail to capture abrupt and intricate spatial-temporal disturbances and dependencies inherited in earth-evolving systems (Ouyang et al., 2017; Wu et al., 2023). The computational cost of a physical model-based numerical method is very high. These intricate numerical models pose significant challenges in development and maintenance, yet, are quite rigid for real-time applications (Cai et al., 2020; Yan et al., 2023). Moreover, the spatial-temporal resolution of a numerical model would have a significant impact on prediction accuracy, such as the grid and temporal interval resolution. Improving the grid resolution will achieve longer processing times and higher computational requirements. Most weather forecast and weather observation systems mainly provide sparse low-resolution data samples. For instance, as illustrated in Figure 1, there is missing wind forecasting or observational data in different large regions, and as marine meteorology is vast and complex, the observational and monitoring costs of the marine environment are much higher than those of the continents. Only certain parts of the ocean region where data samples are available can be validated.

Tremendous efforts have been implemented to explore ship path planning and optimization based on ocean forecasts, such as dynamic programming, A-star algorithm, and genetic algorithm (Chen et al., 2021b; Khan et al., 2022). For example, a new stability-related, dynamic route constraint was proposed for path optimization (Krata and Szlapczynska, 2018). Du (2022) developed an improved 3D dynamic programming algorithm for ship path planning, which takes the meteorological conditions, constraints of engine power, and safety into consideration. Yet, many previous ship path planning approaches primarily focused on realizing the shortest navigation time. Those optimization methods usually neglected the comprehensive energy consumption and motion response factors, especially when encountering severe sea states. Currently, the marine industry is paying more and more attention to shipping energy efficiency, thus, more comprehensive factors, including fuel consumption, the safety of mariners and vessels, reduction of greenhouse gas emissions, and so on, have to be taken into account to achieve greener route planning (Moradi et al., 2022; Chen and Mao, 2024). For example, a multi-objective route optimization methodology was proposed (Vettor and Soares, 2016) by employing the genetic evolution algorithm while realizing route and speed optimization simultaneously. Ma et al. (2021) established a ship routing and speed multi-objective optimization framework for minimizing greenhouse gas emissions by selecting appropriate plans. A genetic algorithm is employed to derive the optimal route based on a ship heading or on both heading and propulsion power information. Yet, a low-resolution sea state dataset was integrated into this study and their main focus was to achieve fuel savings (Kytariolou and Themelis, 2022). Important weather and sea state information is often absent for the ship sensors, thus, a hybrid data fusion and machine learning model was proposed to evaluate the relationship between fuel consumption rate and the voyage's weather situation. This study attempted to aggregate meteorological data and sensor information for the purpose of enhancing the accuracy of machine learning (ML) models, and they focused on quantifying ship fuel consumption based on weather conditions, sailing speed, and sea conditions (Du et al., 2022). A novel study established a hybrid genetic algorithm to optimize ship path planning for safe transoceanic

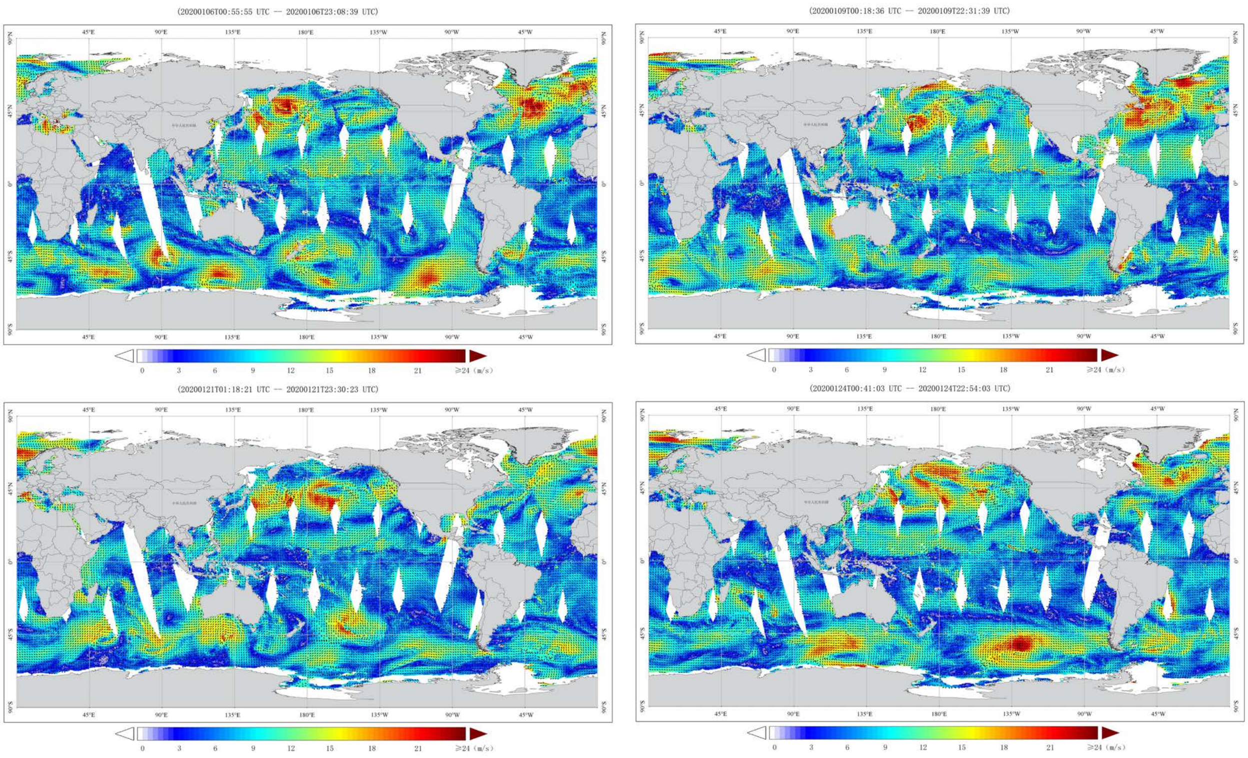

(20200106T00:55:55 UTC — 20200106T23:08:39 UTC)
(20200109T00:18:36 UTC — 20200109T22:31:39 UTC)
(20200121T01:18:21 UTC — 20200121T23:30:23 UTC)
(20200124T00:41:03 UTC — 20200124T22:54:03 UTC)
45°E
90°E
135°E
180°E
135°W
90°W
45°W
90°N
45°N
0°
45°S
90°S
0
3
6
9
12
15
18
21
≥24 (m/s)

FIGURE 1
Global sea surface wind from National Satellite Ocean Application Service (http://www.nsoas.org.cn/eng/column/141.html).

navigation with complicated sea conditions. They mainly focused on the voyage time and fuel consumption as the optimization criteria, yet overlooked the issue of the ship's own structure's resistance to wind and waves and the workload of personnel during high wind and wave weather (Zhou et al., 2023). An improved A* algorithm was proposed for ship collision avoidance path planning by integrating the multi-target point artificial potential field method (MPAPF). They analyzed the static environment and ship navigating situation, thus, the dynamic weather information may be lost (Huang et al., 2024). The Non-Dominated Sorting Genetic Algorithm III (NSGA-III) model was employed to realize ship weather routing tasks by integrating ship heading angle and speed. The main aim was to optimize operational costs and CO_2 emissions (Ma et al., 2024). A constrained policy optimization (CPO) perspective was proposed for a multi-objective path planning model to investigate Pareto-optimal paths, and the results demonstrated that adapting the potential policy factors into the ship path planning model could achieve an advantageous result in complex environments (Zhu et al., 2025a). In order to reduce fuel consumption during a ship voyage, a route planning model that is able to identify energy-efficient routes in complicated sea conditions was proposed by combining ocean currents into the traditional level set method. They proved that ocean environmental factors, such as ocean currents, were very useful for energy-efficient ship voyage planning (Zhu et al., 2025b).

The above studies focused on ship path planning from different perspectives. Nevertheless, most of these approaches employed meteorological forecasts with very low spatial-temporal resolution. It has been emphasized that low spatial and temporal resolution weather forecasting data usually result in inaccuracy in shipping path optimization (Wu et al., 2023). In addition, high-resolution ocean weather prediction plays a major role in ensuring the safe navigation of intelligent autonomous marine vehicles (Chen et al., 2021a; Qiao et al., 2023).

Deep learning methods have been demonstrated to show promise in mitigating the gaps in numerical weather forecasting models and marine environment monitoring systems (Kochkov et al., 2024; Zhao et al., 2024). A deep learning-based weather prediction model has exhibited great potential in uncovering underlying climatic patterns from historical records, enabling the acquisition of high-resolution forecasting data, which provides a new perspective for improving the reliability of highly efficient and intelligent ship path planning. Many researchers have been attempting to explore different kinds of ML methods for obtaining accurate natural wind estimations (Wang et al., 2021). However, the intricate non-linear spatiotemporal properties of large-scale spatial-temporal weather systems represent great challenges for traditional machine learning which attempts to extract sequential evolutionary trends from past records (Khodayar and Wang, 2018).

For the purpose of alleviating the above-mentioned limitations and research gaps, an incremental principal component analysis (IPCA) based on a spatial-temporal depthwise separable U-Net model by aggregating an attention and residual learning scheme, the IPCA-MHA-DSRU-Net, was developed for fine-grid large-scale extreme wind speed field system predictions. Specifically, the depthwise-separable convolution (DSC) blocks first introduced into this proposed method can provide an effective way to improve forecasting efficiency and performance while reducing their computational and memory requirements. The depthwise separable blocks greatly reduce the number of parameters and computation requirements compared to traditional convolutions. They can allow for better feature extraction and aggregation by separating the spatial and channel-wise information in the input data (Zhou et al., 2024; Xu et al., 2024). Incremental principal component analysis (IPCA) is also employed for 2D wind field preprocessing, which can effectively filter the feature space of data samples by reducing dimensionality and redundant noise effects. IPCA is an adaptive version of principal component analysis (PCA) designed for large or streaming datasets. Instead of processing the entire dataset at once, IPCA updates the principal components incrementally as new data arrives, making it memory-efficient and suitable for real-time or large-scale applications. Moreover, a sequential sliding-data window scheme (Yin et al., 2023), obeying a strictly chronological order, was mixed into the tensor-preparation phase, which would enable the accurate preservation of wind temporal-dependent variabilities within consecutive spatial patterns. The framework of the developed wind system forecast model is displayed in Figure 2.

As can be seen in Figure 2, the wind system forecast model demonstrates the structure and workflow of the IPCA-based

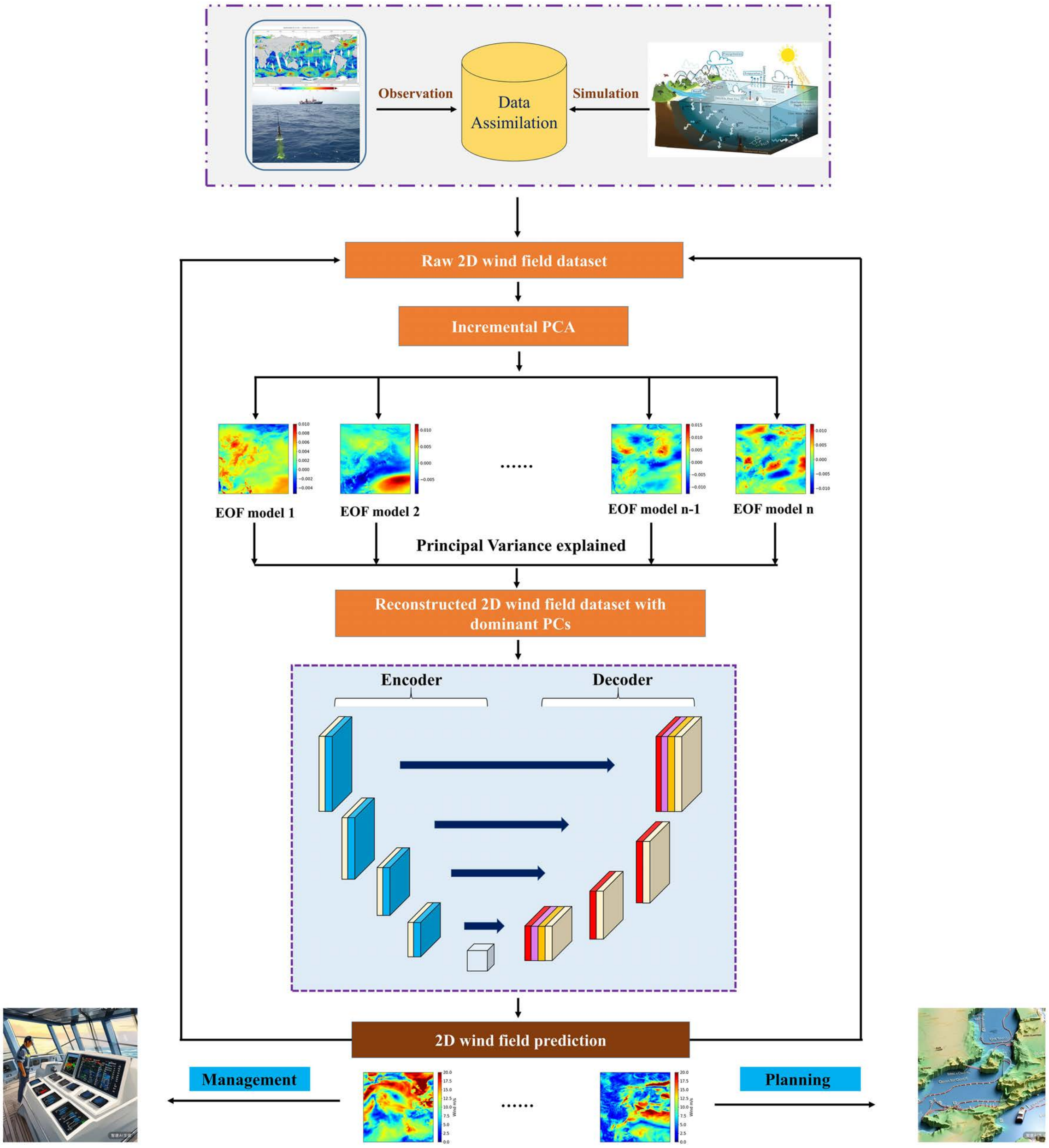


FIGURE 2
The diagram of the proposed wind forecast model.

spatial-temporal depthwise separable U-Net model. U-Net is a convolutional neural network (CNN) architecture designed for computer vision tasks. It features a symmetric U-shaped structure with an encoder-decoder pathway: the encoder captures contextual information by downsampling the input, while the decoder reconstructs precise localization by upsampling. Skip connections between corresponding encoder and decoder layers help preserve spatial details, making U-Net highly effective for tasks like 2D image processing and object detection. This developed forecasting model utilizes modular IPCA preprocessing to effectively reduce the dimensionality of the input data while preserving key spatial-temporal patterns, which are essential for forecasting wind systems. By incorporating depthwise separable convolutions, the model achieves computational efficiency, allowing the processing of large-scale spatial-temporal datasets with reduced complexity. The attention mechanism selectively focuses on the most critical regions in the input data, enhancing the model's ability to capture significant features that influence wind predictions. Meanwhile, the residual learning scheme aids in preserving finer details and mitigates the vanishing gradient problem, allowing deeper layers to learn more nuanced patterns in the data.

First, the reanalysis dataset, which assimilates real observations with numerical simulation, is employed as the input, and then the input data sample is preprocessed by the employed IPCA method to filter noise and retain principle components of wind variability. Next, the processed wind dataset is fed into the proposed forecasting hybrid U-Net model. The last step is to aggregate the forecasting output from the hybrid U-Net model and analyze the forecasting performance. The figure provides a step-by-step visual representation of the data flow, making it easier to understand the contributions of each component in achieving accurate and efficient wind forecasting. This comprehensive architecture, with its innovative use of IPCA, depthwise separable convolutions, attention, and residual connections, demonstrates a balanced approach to handling complex spatial-temporal wind data for forecasting applications.

Our study developed a novel deep learning model for onboard weather prediction during large-scale ocean voyages. It provided us with a fully complete large-scale sea surface wind field forecasting with very high resolution and accuracy, which is very important and valuable for voyage scheduling to avoid severe sea states and ensure the safety of seafarers and ship transoceanic navigation. In addition, the transferability of the proposed model is also verified by utilizing two different geospatial regions with various weather characteristics. By mapping weather observational gaps into a fine-grid and complete spatial perspective, the proposed approach, implemented on a single laptop, aims to enhance the timeliness and accuracy of onboard ship routing, thereby enhancing ship navigation safety. The main aim and focus of this study is to provide ships undertaking transoceanic ship voyages with a highly accurate and high-resolution sea state forecasting model onboard while taking the factors of ship structure safety and seafarer workload into consideration. Finally, the model provides instantaneous extreme wind system pattern mapping, helping achieve adaptive and intelligent path planning for marine vehicles, especially for sea-going navigations in large-scale oceans.

The main contributions of this study can be summarized as follows:

1. A novel intelligent neural-learning model was developed by aggregating a depthwise-separable convolution-based U-Net framework with attention and residual learning blocks.
2. Incremental principal component analysis was first introduced to preprocess a fine-grid wind dataset, filter Empirical Orthogonal Function (EOF) models, and retain principal wind evolution information.
3. The DSC-based methodology was developed to achieve fine-grid spatial-temporal extreme wind field forecasting on a large scale.
4. The fine-grid wind prediction model can enhance the navigation safety of sea-going vessels.
5. A sequential sliding-data window is adopted for the aggregation of input-target tensor pairs to better preserve the temporal wind evolution information.
6. A sensitivity trial was implemented to explore wind forecasting model parameter adjustment and optimization.
7. The transferability of the intelligent neural learning model was validated by employing two geographic regions with different wind patterns.

The remainder of this article is arranged as follows. Section 2 introduces the developed spatial-temporal wind prediction approach. The targeted experimental case is presented in Section 3 with the quantitative forecasting analysis, and Section 4 validates the model transferability. Finally, Section 5 summarizes the work and outlines future directions.

2 Methodology

The novel hybrid wind systems forecasting model, IPCA-MHA-DSRU-Net, integrates IPCA with a spatial-temporal depthwise separable U-Net architecture, enhanced by attention and residual learning mechanisms. This model aims to achieve fine-grid, large-scale wind system predictions, improving voyage planning and navigation safety. The use of DSC blocks significantly reduces model parameters and computational complexity. By leveraging the strengths of modular IPCA preprocessing, residual learning, multi-head attention, and the depthwise separable CNN-based U-Net architecture, this hybrid framework is optimized to predict complex, spatial-temporal variations in extreme wind signals. Detailed explanations for each component of the proposed model are as follows.

2.1 Incremental principal component analysis

The basic theory of PCA is to generate a set of independent composite indicators by recombining the raw variables, thereby reducing the dimensionality of the original data samples while retaining most of the original/principal information features.

Specifically, PCA performs data transformation on the original data and projects it onto a new coordinate system, resulting in the projected data having the largest variance. The main merits of PCA include reducing data dimensionality, decreasing computational complexity and model complexity, reducing the impact of noise, improving the signal-to-noise ratio of data, identifying the most important features in data samples through dimensionality reduction, and removing some redundant features, thereby reducing the risk of overfitting and improving the model's generalization ability (Xu et al., 2023a; Xiao et al., 2023; Zhang et al., 2024b).

Provided that the targeted data-sample size is m x n, the data sample matrix is represented in the Equation 1 as follows:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \tag{1}$$

subtract the average value of each column in the Equation 2:

$$P = \begin{bmatrix} p_{11} - b_1 & p_{12} - b_1 & \cdots & p_{1n} - b_1 \\ p_{21} - b_1 & p_{22} - b_1 & \cdots & p_{2n} - b_1 \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} - b_1 & p_{m2} - b_1 & \cdots & p_{mn} - b_1 \end{bmatrix} \tag{2}$$

where bi is the average of each column in the Equation 3:

$$b_i = \frac{1}{m}\sum_{i=1}^{m} p_{ji} \tag{3}$$

The covariance CM is an $m \times m$ matrix, and the CM_{ij} of the covariance matrix indicates the covariance value of the targeted variables p_i and p_j. Next, the eigenvalues of the covariance CM are derived and the computed eigenvalues are filtered in descending order. The eigenvectors related to the first k eigenvalues are employed to aggregate a new feature matrix. Finally, after the dimensionality reduction operation, the projection of P on the new eigenvector matrix is computed to represent the eigenvectors.

IPCA decomposes a large-scale sample into multiple small-batch datasets through gradual iterations and performs principal component analysis on each small-batch dataset. This avoids the memory and computing resource consumption caused by processing the entire dataset at once. After conducting principal component analysis on each small-batch dataset, the obtained principal components are merged so as to obtain the principal components of the entire dataset. Compared to traditional PCA algorithms, IPCA has lower computational complexity and can obtain principal components with greater efficiency. It can also perform incremental updates when new data arrives without recalculating the principal components of the entire dataset, thus achieving real-time data processing. IPCA employs singular value decomposition to perform linear dimensionality reduction on target data samples, retaining only the most important singular vectors, and then processing/projecting the data samples into a lower dimensional feature space. It finds principal components by calculating singular value decomposition, processing only one batch of samples in one iteration to reduce memory consumption (Greenacre et al., 2022; Weng et al., 2003). The principal component is calculated by the Equations 4 and 5:

$$\widetilde{PC}_i(n) = PC_i(n-1) + \alpha_i(n)u(n)u^T(n)PC_i(n-1) \tag{4}$$

$$\begin{aligned} &PC_i(n) = orthonormalize\widetilde{PC}_i(n) \;\; with \;\; respect \;\; to \;\; PC_i(n), \\ &j = 1, 2, \ldots, i-1 \end{aligned} \tag{5}$$

where the $PC_i(n)$ denotes the projection of the ith dominant eigenvector for the derived sample covariance matrix $CM = E\{u(n)$ $u^T(n)\}$. The a_i indicates a stochastic approximation gain. The u_n is a m-dimensional vector.

The full wind speed field can be reconstructed by the linear combination of the leading principal components (PCs) and their corresponding EOFs after filtering redundant features and noise signals. The EOF analysis is a statistical technique used to identify dominant patterns or structures in spatial-temporal datasets, such as climate or geophysical data. It decomposes the data into EOFs that capture the maximum variance, with associated time coefficients describing their temporal evolution. A given wind field $Wind_t$, at time step t can be calculated as follows in the Equation 6:

$$Wind_{m,t} = \sum_{n=1}^{k} PC_{n,t} EOF_{m,n} \tag{6}$$

where m denotes the grid index of the wind field, t indicates the time index, and k is the total number of retained PCs.

IPCA is an adaptation of PCA that allows for processing data in an incremental manner, rather than requiring the entire dataset to be available in memory at once. Thus, instead of computing the covariance matrix from the entire dataset at once, the algorithm updates the principal components incrementally as new data arrives. The key idea is that there is no requirement to store the whole dataset, but data is processed in small batches (minibatches) and the principal components are updated as new data is fed into the model.

The application of IPCA in 2D extreme wind field preprocessing offers a strategic approach to handle data dimensionality and mitigate noise interference. It is very crucial for improving prediction accuracy in wind field forecasting models. Here is a detailed explanation of IPCA's principles, and how they rationalize its application in this context. Since wind field data are typically represented as large 2D grids, with each cell corresponding to specific wind metrics (e.g., speed and direction) at that spatial point. Processing such high-dimensional input directly in deep learning models would lead to high computational costs and increase the risk of overfitting, especially with limited training data. The IPCA reduces the spatial dimensions, retaining only essential components that reflect the primary spatial patterns in wind fields, making the data manageable without significant information loss. By focusing on principal components, the IPCA naturally discards lower-variance components, which are likely to be noise. This selective filtering of information means that the data entering the forecasting model is "cleaner," which supports better model training and more accurate predictions. In addition, as a ship navigates through different ocean regions, wind patterns will vary significantly. The IPCA's incremental nature allows it to adapt to these changes by updating principal components with incoming data. This ongoing adaptation ensures that the data fed into the forecasting model always reflects current environmental conditions, enhancing prediction accuracy, which allows forecasting models to focus on essential features without the burden of excessive, redundant information. This burden reduction lowers computational demands, allowing models to train faster and reducing the risk of overfitting. Additionally, because the model is working with a cleaner, lower-dimensional dataset, prediction accuracy tends to improve.

In summary, the choice of IPCA in 2D wind field preprocessing is rational due to its ability to reduce dimensionality, handle real-time data, and filter out noise, all while requiring limited resources. This pre-processing step enhances the predictive model's accuracy and efficiency by supplying a refined, lower-dimensional input that captures the most relevant spatial patterns in the wind field data. As a result, IPCA-based preprocessing is a practical and effective solution to prepare 2D wind data for deep learning models in a constrained, dynamic environment like that on a ship.

2.2 Depthwise separable convolution

In general, the basic U-Net framework is prone to overfitting and is computationally heavy with traditional convolution operations. In this study, we introduced the DSC block to reduce the basic U-Net model size and trainable parameters (Chollet, 2017). The DSC block separates a complete convolution operation into two steps: pointwise convolution (PTC) and depthwise convolution (DC). The operation of PTC is similar to classical convolution, and its convolution kernel has a size of 1 × 1. Unlike the classical convolution computing process, a kernel of DSC is responsible for one channel. Therefore, the entire model parameters are greatly reduced. Each input channel was applied by a single convolutional kernel in the depthwise convolution and outputs the respective feature maps.

In a standard 2D convolutional operation, a kernel spans all input channels (or depth) and slides over the spatial dimensions (height and width) of the input, creating output channels by combining information from all input channels. However, in depthwise convolution, each input channel has its own independent kernel. Specifically, instead of applying a single kernel across all input channels, the depthwise convolution applies one filter per implementation independently. This process captures spatial information within each channel but does not combine information across different channels, which limits its expressive power. Thus, the next step, pointwise convolution, is introduced to address this issue. The pointwise convolution can adjust the number of output channels and helps to combine the channel-wise information produced by the depthwise convolution. By performing these two operations sequentially, the depthwise-separable convolution emulates the effect of a standard convolution while significantly reducing the computational cost.

Considering the input feature map $\boldsymbol{I}$ is (D_I, D_I, M), the target-output $\boldsymbol{O}$ is (D_O, D_O, N); and the standard convolutional operation kernel $\boldsymbol{K}$ indicates (D_K, D_K, M, N), of which M and N represent the number of inputs and target channels, correspondingly. D denotes the size of convolved high-dimensional feature maps. Specifically, the kernel K is divided into two convolutional modulations: the depthwise (D_K, D_K, 1, N) and pointwise convolution (1, 1, M, N). In addition, the classical 1 x 1 convolutional kernel is employed in the pointwise convolution modulation, and the channel features derived by depthwise convolution operation are then projected onto the deeper and higher channel space. The pointwise convolution was applied after the depthwise operation, using N convolutional kernels with 1 x 1 x M size for the purpose of representing the M D_K x D_K feature maps. The weighted combination operation is then performed in the depth direction in order to generate the N D_K, D_K x 1 feature maps $\boldsymbol{O}$ (D_O, D_O, N). The two convolutional modulations are illustrated in Figure 3.

The formula of standard convolution is expressed in the Equation 7:

$$O_{k,l,n} = \sum_{i,j,m} K_{i\,.\,j,m,n} \quad \cdot \quad I_{k+i-1,l+j-1,m} \tag{7}$$

and the formula of depthwise separable convolution is shown in in the Equation 8:

$$\hat{O}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \quad \cdot \quad I_{k+i-1,l+j-1,m} \tag{8}$$

2.3 Multi-head attention

The attention strategy in deep learning is widely used in image, natural language processing, speech recognition, and so on. The core task of the attention mechanism is to optimally extract critical information from mass data samples quickly and accurately. Compared with the standard convolution mechanism, the attention strategy is characterized by fewer parameters, high accuracy, and lower computational cost. The basic scaled dot-production attention block consists of multi-head attention modulation. It has been demonstrated that multi-head attention is able to better catch and preserve underlying high-dimensional features (Xu et al., 2023b, 2024). In particular, the attention mechanism has been proven to be helpful for spatial-temporal wind speed forecasting (Yu et al., 2023), and more non-linear dynamics could potentially be reproduced, especially for the dynamic fluid field (Niu et al., 2021; Che et al., 2022) based on the multi-head strategy. The attention can be understood as a key-value query, which maps queries and key values to the target output. The essence of the attention is the processing of weighted summation for values based on keys and queries, together with the weight redistribution.

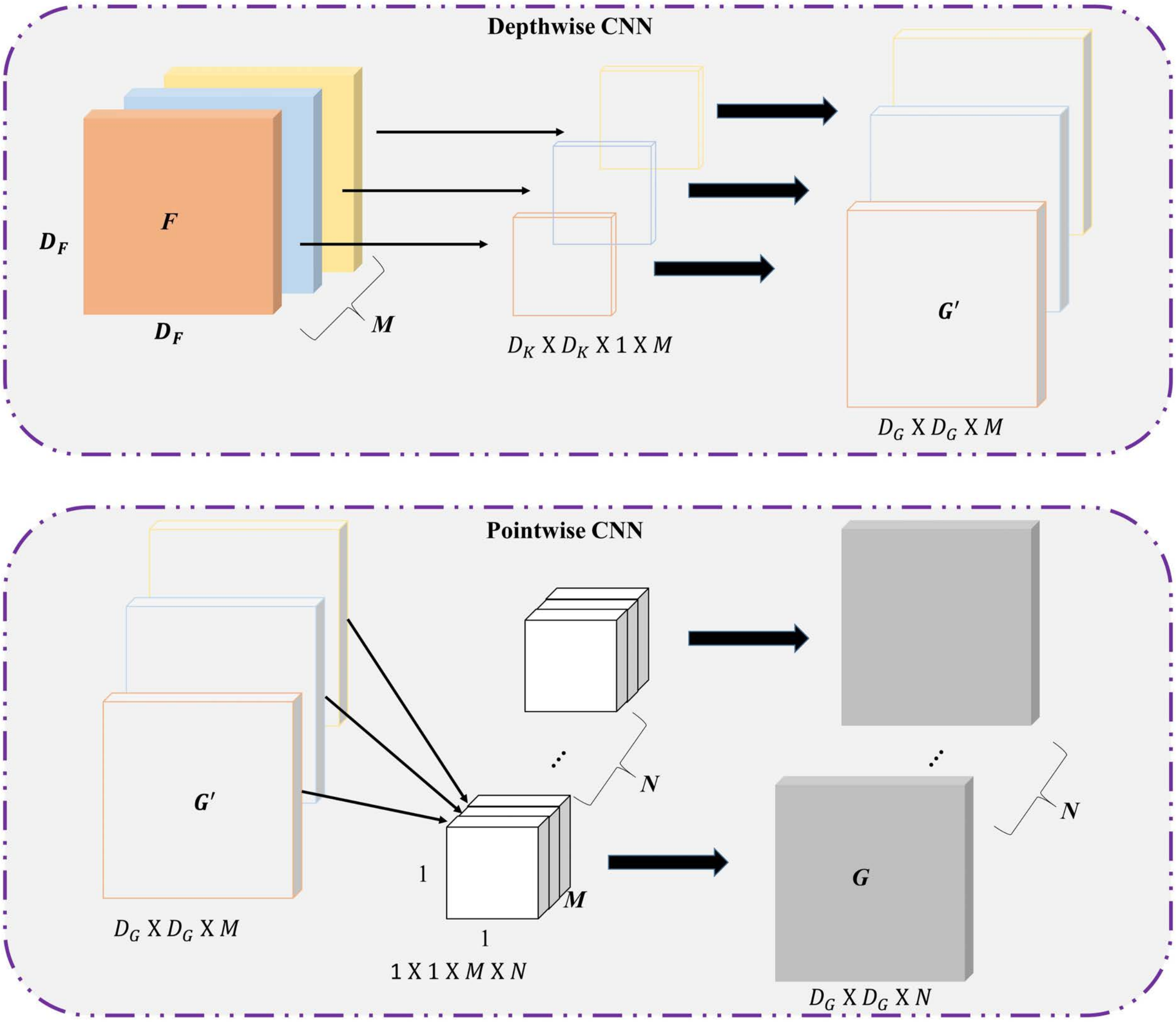


FIGURE 3
The Schematic illustration of depthwise separable CNN.

The multi-head attention strategy exhibits lower complexity compared to the scaled dot-product attention, allowing the forecasting model to deeply map different high-dimensional representations while avoiding the loss of small targets. In this study, the query matrices were linearly projected three times on the sequential wind-speed tensors. Then, the projected weight matrices were concatenated to generate the refinement forecasting outputs.

In wind field forecasting, multi-head attention enhances the model's ability to interpret the spatial and temporal relationships within the wind data. By simultaneously attending to multiple areas of the input grid, the model captures subtle, location-specific patterns (e.g., shifts in wind intensity across regions and changes over time) that a standard convolutional layer may miss. Moreover, traditional convolutional layers have a fixed receptive field and struggle with long-range dependencies, particularly in spatial-temporal data. Multi-head attention addresses this limitation by dynamically focusing on relevant areas across both spatial and temporal dimensions. In the U-Net model, this allows the encoder-decoder structure to more effectively aggregate spatial-temporal information, which is crucial for fine-grid forecasting of fluctuating wind conditions. More importantly, for real-time applications on ships, balancing latency with model accuracy is essential. Multi-head attention, while enhancing predictive accuracy through improved feature attention, may introduce latency due to the processing load. Efficient implementation techniques, such as attention approximation methods (e.g., sparse or low-rank approximations), can be considered to reduce the burden of multi-head attention.

H and W are the height and width of the 2D input matrix and the C indicates the feature number for the input-sequential tensors. Providing that the sequential series represent $\boldsymbol{Wind} = [\boldsymbol{x}_{1,\ldots}, \boldsymbol{x}_N] \in \mathbb{R}^{\mathrm{HxWxC}}$, the dot-product will aggregate and derive the $\boldsymbol{K}$ and the $\boldsymbol{Q}$

together with the $\boldsymbol{V}$ query terms using three projected matrices $\boldsymbol{W}_k \in \mathbb{R}^{D_x\text{x}D_k}$, $\boldsymbol{W}_q \in \mathbb{R}^{D_x\text{x}D_q}$, and $\boldsymbol{W}_v \in \mathbb{R}^{D_x\text{x}D_v}$ in the Equation 9:

$$\begin{aligned}\boldsymbol{K} &= \boldsymbol{WinsW_k} \in \mathbb{R}^{H\text{x}W\text{x}D_k}\\ \boldsymbol{Q} &= \boldsymbol{WindW_q} \in \mathbb{R}^{H\text{x}W\text{x}D_k}\\ \boldsymbol{V} &= \boldsymbol{WindW_v} \in \mathbb{R}^{H\text{x}W\text{x}D_k}\end{aligned} \tag{9}$$

In the attention-based data-processing stage, a specific normalization term $\xi(\boldsymbol{q}_i^T\boldsymbol{k}_j) \in \mathbb{R}^1$ will be introduced to calculate the similarity between the ith query $\boldsymbol{q}_i^T \in \mathbb{R}^{D_k}$ and the jth key $\boldsymbol{k}_j \in \mathbb{R}^{D_k}$. Then at a designated position i, the attention weight is derived by the Equation 10

$$\vartheta(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \xi(\frac{\boldsymbol{QK}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{10}$$

Larger d_k derived from the input sequential tensors with higher dimensions will then lead to the softmax-normalization trapped into local optima with extremely small gradients. The scaled term $\frac{1}{\sqrt{d_k}}$, laterally aggregated into the weighted summation, will alleviate this traditional vanishing gradient issue.

The ith row weights can be derived as the Equation 11

$$\vartheta(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^N e^{\boldsymbol{q}_i^T\boldsymbol{k}_j}\boldsymbol{v}_j}{\sqrt{d_k}\sum_{j=1}^N e^{\boldsymbol{q}_i^T\boldsymbol{k}_j}} \tag{11}$$

sequentially, it can be simplified as

$$\vartheta(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\Psi(\boldsymbol{q}_i)^T\sum_{j=1}^N \rho(\boldsymbol{k}_j)\boldsymbol{v}_j^T}{\sqrt{d_k}\Psi(\boldsymbol{q}_i)^T\sum_{j=1}^N \rho(\boldsymbol{k}_j)} \tag{12}$$

The Equation 12 can, then, be illustrated when different types of normalization functions $\phi()$ were aggregated

$$\vartheta(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^N \phi(\boldsymbol{q}_i, \boldsymbol{k}_i)\boldsymbol{v}_i}{\sum_{j=1}^N \phi(\boldsymbol{q}_i, \boldsymbol{k}_i)} \tag{13}$$

$\phi(\boldsymbol{q}_\text{i}, \boldsymbol{k}_j)$ function will calculate the correlated similarities between $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$.

A constraint term can be illustrated as the $ker(x, y)$ $\mathbb{R}^{2\text{x}F} \rightarrow \mathbb{R}_+$, which would ensure that the specific attention blocks are non-negative. The Equation 13 can be expressed as

$$\vartheta(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\sum_{j=1}^N \varsigma(\boldsymbol{q}_i)^T\varsigma(\boldsymbol{k}_i)\boldsymbol{v}_i}{\sum_{j=1}^N \varsigma(\boldsymbol{q}_i)^T\varsigma(\boldsymbol{k}_i)} \tag{14}$$

The associative property of the matrix multiplication was used to rewrite Equation 14

$$\vartheta(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = V' = \frac{\varsigma(\boldsymbol{q}_i)^T\sum_{j=1}^N \varsigma(\boldsymbol{k}_i)\boldsymbol{v}_i^T}{\varsigma(\boldsymbol{q}_i)^T\sum_{j=1}^N \varsigma(\boldsymbol{k}_i)} \tag{15}$$

Equation 15 can, subsequently, be simplified when the numerator is in the vector form as the Equation 16:

$$(\varsigma(\boldsymbol{Q})\varsigma(\boldsymbol{K})^T)\boldsymbol{V} = \varsigma(\boldsymbol{Q})(\varsigma(\boldsymbol{K})^T\boldsymbol{V}) \tag{16}$$

2.4 Spatial-temporal forecasting network

The underlying spatial-temporal features inherited in the sequential wind speed systems with low-level nonlinearities are mapped by the encoder module of the U-Net backbone, and high-level semantic representations will then be extracted into the decoder modulation (Ronneberger et al., 2015). Yet, the ordinary skip-connection operations would usually lead to insufficient exploration of potential semantic and contextual features, especially for fine-grid 2D wind speed system mapping tasks. Thus, in this study, two additional multi-head attention blocks together with deep residual learning (Manucharyan et al., 2021) are introduced together with depthwise separable convolutional modulation to mitigate these issues. The residual learning block mitigates the vanishing gradient problems that would usually occur in very deep networks. It enables the constructed wind mapping network to be deep enough. In addition, in the context of wind field forecasting, residual learning allows the model to refine spatial-temporal representations by focusing on differences in wind patterns across time and space. This focus is especially important for forecasting applications where subtle changes in wind conditions need to be captured accurately. Residual learning supports the model's ability to detect and propagate important spatial-temporal features throughout the network, improving forecasting accuracy. The IPCA-based dimensionality reduction further enhances residual learning by streamlining the data. With IPCA pre-compressing high-dimensional inputs, residual layers can focus on fine-tuning only the most critical components of the compressed data, which reduces both computation and memory usage without compromising model performance. Finally, residual learning enables the model to adapt to rapidly changing wind conditions by emphasizing residuals, or deviations, in the wind field data. This adaptability is particularly valuable in marine environments where weather and wind conditions can shift quickly. With residual learning, the model becomes better equipped to capture these subtle changes, leading to more accurate and timely forecasts.

The diagram of the wind system mapping based U-Net model combination is illustrated in Figure 4.

The residual block was only integrated into two layers of the Decoder part, which would alleviate the total computational burden. Specifically, one block was incorporated into the last layer of the decoder, and the other one was located in the first layer of decoder modulation. The attention block in between the Bottleneck layer and 2D depthwise separable CNN layers can query and reproduce more embedded spatial wind system features with refinement operations. The second one further augments original feature maps aggregated by skip-connections, deeper refinement, and feature augmentation realized by the attention operations can improve the final forecasting performance (Vaswani et al., 2017). These newly introduced modifications, including DSC modulation, attention blocks, and residual learning strategy, for the raw U-Net-backbone, can enhance the reproduction performance of underlying fine-grid 2D wind spatial variabilities. In addition,

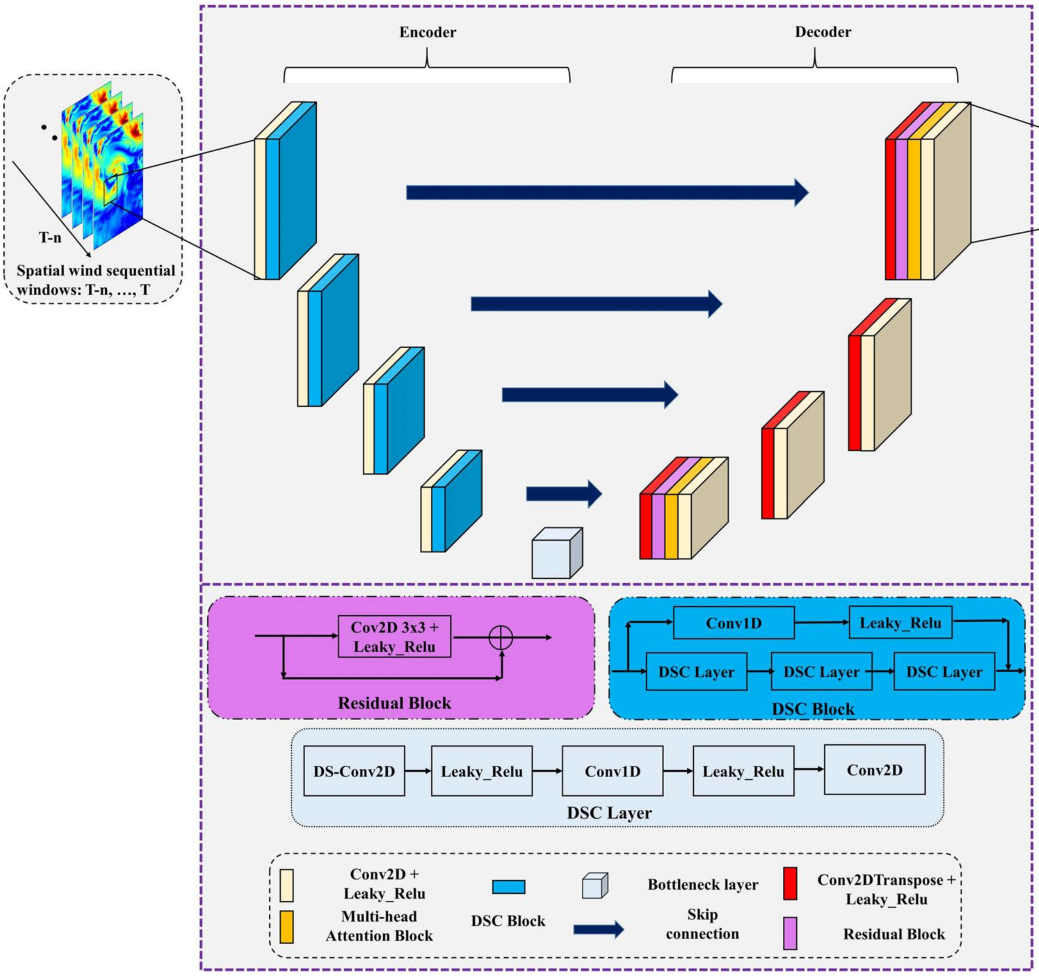


FIGURE 4
The wind system forecasting network.

dropout layers were retained in the forecasting operations due to the dropout being a potential Bayesian approximation that could mitigate the predictive uncertainty for deep learning regression tasks (Gal and Ghahramani, 2016).

The core architecture of this hybrid model is a depthwise separable CNN-based U-Net-like structure, as illustrated in Figure 4. The model adopts a U-Net-like architecture, which is characterized by an encoder-decoder structure with skip connections. This design is particularly effective for capturing multi-scale features, making it suitable for spatiotemporal data such as wind fields, and the 2D DSC layers employed in the U-Net framework can process spatial data (e.g., wind speed maps) across time steps, enabling it to learn spatial patterns and temporal dynamics simultaneously. One of the major innovation points of this proposed model is that the depthwise separable convolutions are employed as the main CNN block, as shown in Figure 3, which could reduce computational complexity and the number of trainable parameters. This convolution operation separates spatial filtering (depthwise convolution) from channel-wise feature combinations (pointwise convolution), making the model more efficient. This depthwise separable CNN block enhances the model's ability to extract localized spatial features from extreme wind data samples, which is critical for capturing fine-grained patterns in wind fields. The other innovation of this model is that multi-head attention is integrated into the proposed network to capture long-range dependencies and interactions across both spatial and temporal dimensions. This mechanism allows the model to focus on the most relevant regions of the input data at different scales. By computing attention scores across multiple heads, the model can dynamically weight the importance of different spatial and temporal features, improving its ability to model complex wind dynamics. In addition, the residual connections are also incorporated to facilitate gradient flow during training, mitigating issues such as vanishing gradients and enabling the training of deeper networks. These connections allow the model to reuse features from earlier layers, enhancing its ability to learn hierarchical representations of wind field data. As can be seen in Figures 2, 4, a reanalysis of the extreme wind field dataset, which combines the real observation and numerical model simulations using the data assimilation method, was aggregated from the ERA5 model, and we then implemented z-score normalization in the raw extreme wind dataset and transformed the dataset into a standard form with a mean of 0 and a standard deviation of 1. Then, the IPCA approach was employed for 2D wind field decomposition, which can effectively filter the feature space of data samples by reducing dimensionality and redundant noise effects. The autocorrelation analysis, as outlined in section 3.1, is employed to obtain a comprehensive perspective on the temporal dependency of the overall extreme wind speed field sequential lagging. The sequential wind field time lag is determined as 12 time steps, and the target wind field is a one-time step. We then split the dataset into 70% training and 30% testing parts. Afterward, several batch-size wind map data samples with the aggregated wind tensors were fed into the developed forecasting network for parameter training and optimization, and the rest 30% testing data sample was used to test the model performance compared to the reanalysis target.

A novel architecture was designed specifically for spatiotemporal significant extreme wind signal prediction in a large-scale perspective, which leverages the strengths of U-Net framework for precise feature extraction. The IPCA approach was employed for 2D wind field decomposition, which can effectively filter the feature space of data samples by reducing dimensionality and redundant noise effects. The depthwise separable convolution block was incorporated to reduce computational complexity and improve model efficiency without sacrificing performance. In addition, the multi-head attention mechanism was introduced to enhance the model's ability to capture complex spatiotemporal dependencies in wind data. Finally, the residual learning block was also aggregated into the new framework to address potential vanishing gradient issues in deep networks, ensuring stable training and improved feature representation.

3 Experimental results and discussion

3.1 Case study

This study utilized a Linux platform as the simulation environment based on the Tensorflow framework by employing a single NVIDIA-A100 GPU. The forecasting experiment covers the Asia-Pacific region within a longitude of 96.5-160°E and a latitude of 6-69.5°N, and 2 years of hourly wind data samples spanning from 2016 to 2017 with fine-grid 256 x 256 spatial resolution were selected. One year of hourly samples from 2016 were utilized to train the forecasting model, and the independent validation dataset covers 3 months of data samples from January to March in 2017 (UTC). The weather forecast ERA5 data was provided by the European Centre for Medium-Range Weather Forecasts (ECMWF), while the weather observation data was provided by the National Satellite Ocean Application Service (IMOS). The spatial resolution of the hourly weather forecast was 0.25° × 0.25°. The reanalysis data had global horizontal coverage. The temporal coverage was from 1940 to the present. The dataset size utilized in this research was approximately 9Gb, covering a time period from 2016 to 2017 with 256x256 spatial resolution (the pixel size is 256x256 for each hourly wind field snapshot).

3.2 Wind field decomposition

The dominant variability of spatial-temporal wind speed patterns could be decomposed into a certain number of principal EOF models, and the derived EOF time series is able to represent wind spatial variation patterns associated with its corresponding temporal PC time series. Based on the IPCA data-preprocessing approach, the reconstruction of the wind speed pattern, after cleaning redundant wind features and noise signal, is calculated by multiplying the decomposed PCs with retained EOFs models (Zhang et al., 2022):

$$Wind_{recon} = \widetilde{PC_i} EOF_i \tag{17}$$

As can be seen from Figure 5, far more than 1,000 principal wind variability components were decomposed from the original raw wind data samples in Panel (a), which explained most of the wind evolutional variance, yet, a certain portion of noise signals and irrelevant features have already been coupled and embedded within the original data samples due to the stochasticity and non-linearity of the evolved earth system. Panel (b) clearly illustrates that the first 25 PC models would be capable of explaining almost 70% of the total wind evolutional variance. In order to save computational resources, reduce time consumption, and further clean up the additional redundant noise signals with potentially irrelevant features, the first 25 EOFs (as displayed in Figure 6) were selected as the primary evolutional variability model of wind speed patterns. Finally, the cleaned input wind data samples were reconstructed by employing the 25 principal EOF models with their corresponding PC time series based on Equation 17.

3.3 The sliding-data window method

The autocorrelation analysis was employed to obtain a comprehensive perspective on the temporal dependency of the overall wind field sequential-lagging, which can usually explore the relatively optimal historical time lags, coupled with the most inter-correlated sequential information, for the aggregated wind samples by showing time-series correlation maps of both regionally averaged and randomly selected grid-cell based wind series. In Figure 7, the bounds of the derived 95% confidential interval are represented as the shadow blue band.

Given a time series for correlation analysis with its delayed values, the formula of correlation can be calculated based on the following Equations 18, 19 and 20:

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{18}$$

$$corr(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{19}$$

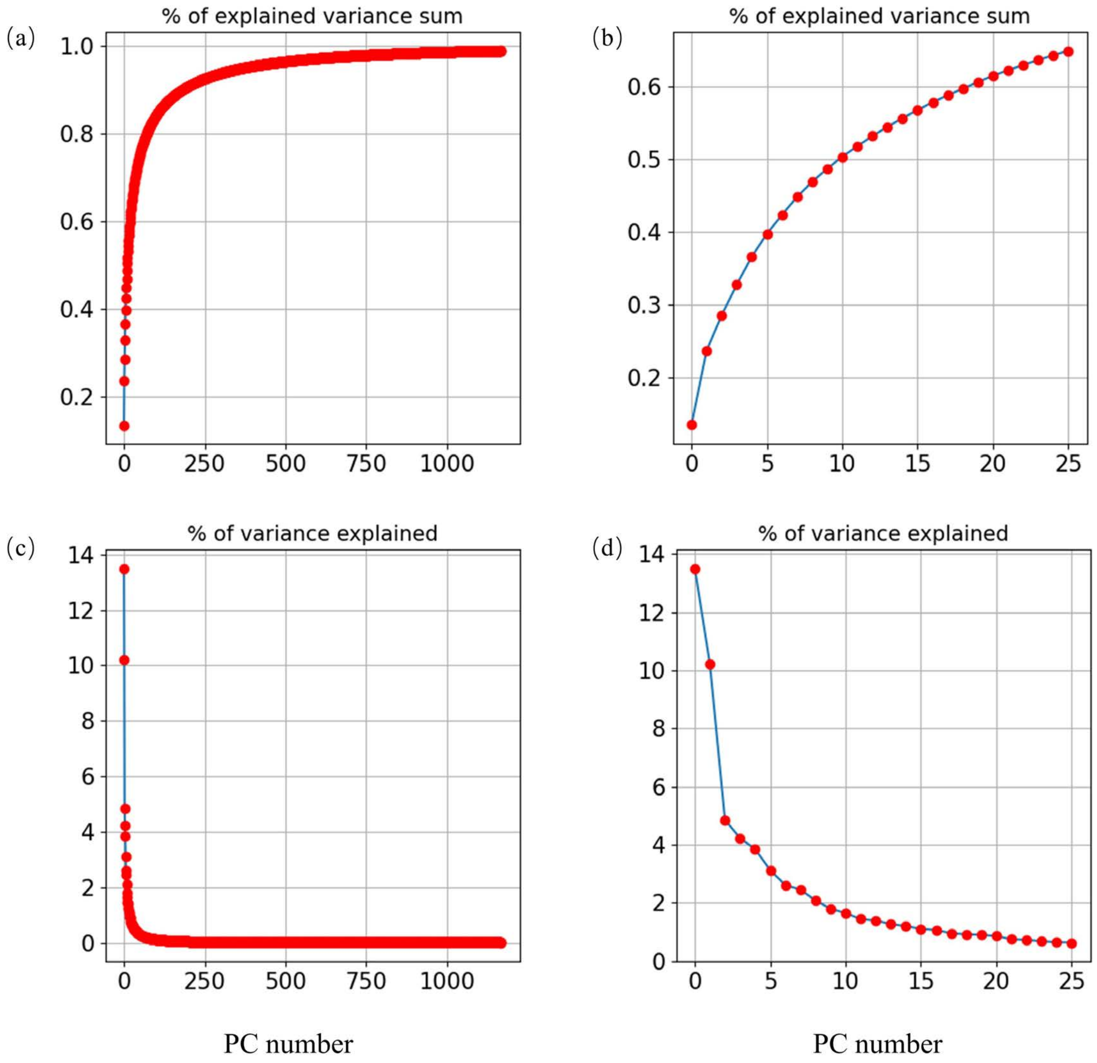


FIGURE 5
The variance explained based on the decomposed spatial wind patterns. Panels **(a, c)** indicate the complete explained wind evolutional variance, panels **(b, d)** represent 70% explained wind variance.

$$corr(X, Y) = \frac{E[X, Y] - E[X]E[Y]}{\sqrt{E[X^2] - E[X]^2}\sqrt{E[Y^2] - E[Y]^2}} \quad (20)$$

for wind time-series G in time t step, $X = G_{t+1}$ and $Y = G_t$.

The partial autocorrelation function (PACF) also employs the same correlation formula to derive the autocorrelation in between time lags, yet the PACF disregards the indirect correlations between Gt+1 and Gt. The Equation 21 is as follows given $k \geq 2$:

$$PACF(k) = corr(G_{t-k} - P_{t,k}(G_{t+k}), W_t - P_{t,k}(G_t)) \quad (21)$$

where $P_{t,k}$ (x) indicates the subjective operator of the orthogonal projection for x onto the linear subspace of Hilbert spanned by $G_{t+1},\ldots,G_{t+\mathrm{k}}$.

As shown in Figure 7, the PACF within the shadow blue band occurred at lag step 12 and lag step 7, correspondingly. Note that the correlation values distributed within the shadow blue band indicate these time lags were not significant. Thus, in this study, the wind time series ranging from historical time-lag t-1 to t-7 was finally filtered to aggregate the input-tensor depth. In this study, the wind pattern time series consists of 256 x 256 ($Width \times Height$) grids. Based on the optimal correlated time lags, the sequential sliding data window with a fixed window size of 7 was set. Each pair of the training and validation sample contains seven wind field snapshots with strict chronological order as $SSW_{\mathrm{t}} = (Wind_{\mathrm{t-10}}, \ldots, Wind_{\mathrm{t-2}}, Wind_{\mathrm{t-1}})$, combining one or more output-wind speed maps with a specific given leading time-steps. Specifically, the prepared modeling data sample was normalized into the value range [-1,1] to speed up convergence efficiency. In addition, the scale consistency will be eliminated between data samples by implementing normalization pre-processing. The learning rate of the selected Adam optimizer in the wind-forecasting model was set to 1e-4, the batch size was set as 200, and the loss function employed Huber loss, which was minimized by using the gradient descent approach. An early-stopping criterion that the training iteration will be

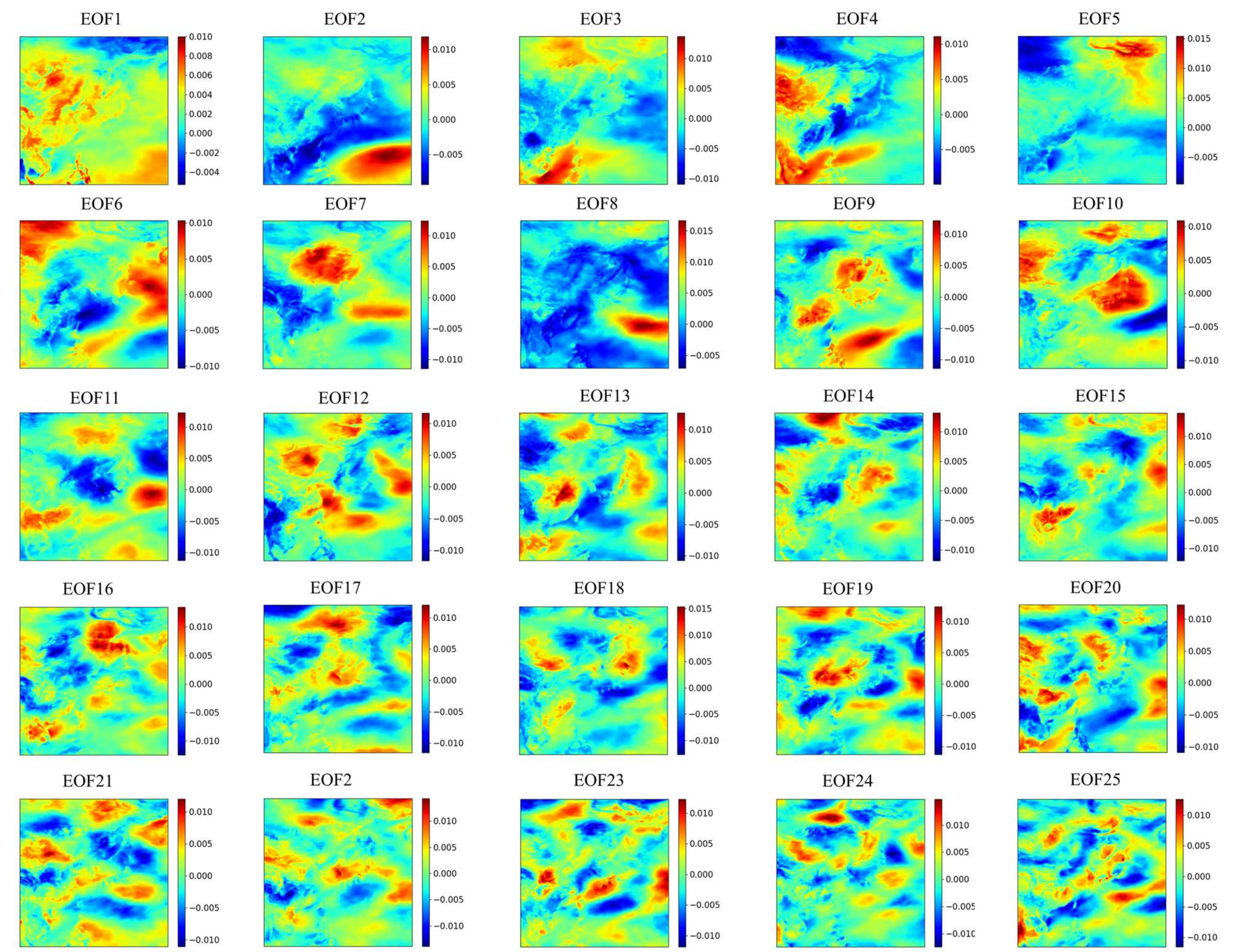


FIGURE 6
The first 25 decomposed EOF models of raw wind pattern.

terminated if the loss metric has stopped improving after consecutive 12-iterations was further employed. The Huber loss Equation 22 is as follows:

$$L_{\varsigma}(O,\boldsymbol{\vartheta}(X)) = \begin{cases} \frac{1}{2}(O-\boldsymbol{\vartheta}(X))^2 \\ \varsigma|0-\boldsymbol{\vartheta}(X)|\,\frac{1}{2}\varsigma^2 \end{cases} \tag{22}$$

where O is the reanalysis model and ϑ denotes the deep neural learning model. In this study, the ς was tested and set as 1.0. The Huber loss is usually less sensitive to outliers, since it can approach an L2 loss if the ς approximate to 0, and approaches L1 when the ς is positive infinity. The flowchart of the established wind pattern forecasting network is presented in Figure 8, so as to provide an clear model operation process.

3.4 Model sensitivity analysis

The rationale concerning how to determine the model hyperparameter settings is very important to evaluate its robustness and uncertainty. In this study, we tested a range of hyperparameters, consisting of the batch size, activation function, learning rates, and loss function, to assess the model's robustness based on forecasting performance.

The statistics forecasting skills for wind pattern prediction are illustrated in the Appendix (Supplementary Table S1–S4), employing the varied hyperparameters. Note that we implemented forecasting experiments using different parameter settings, yet, for the optimization algorithm, the reasonable parameter range settings are also determined by preliminary experiments and domain knowledge (Parri and Teeparthi, 2024). Also, it has been emphasized that optimizing hyperparameters of machine learning models is a laborious process (Zhang et al., 2024a). Moreover, one can better monitor the comprehensive model performance and robustness by applying model sensitivity experiments in which varied model parameter settings are explored, which can provide us with a deeper insight into a better understanding of which hyperparameters might have a potential impact on the predictive capability. More importantly, it has been illustrated that a sensitivity trial can provide a basis for model parameter adjustment and optimization, and further enable quantification of the potential model uncertainties (Asheghi et al., 2020). The uncertainty of the specific model settings can be quantified by exploring the underlying

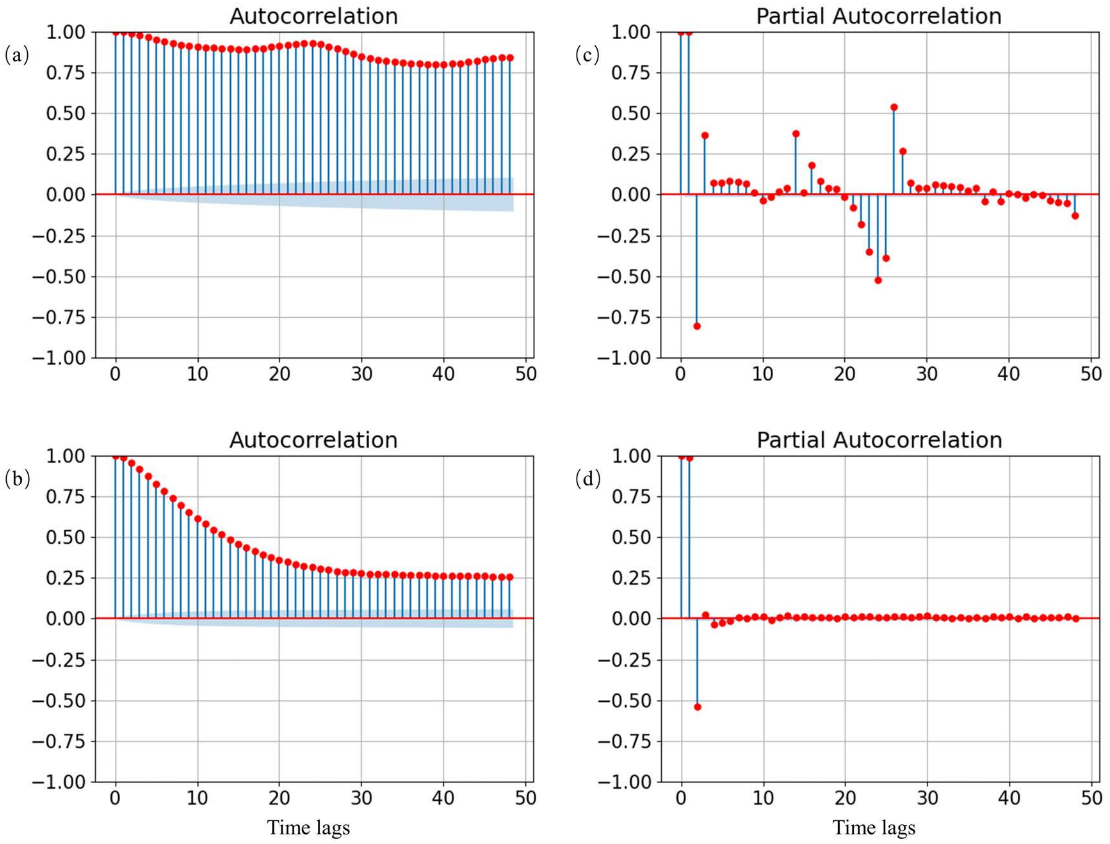


FIGURE 7
The Autocorrelation analysis of wind component pattern variability with field-mean time series panels **(a, b)**, the random selected grid-point time series panels **(c, d)**.

impact of these hyperparameters on predictive performance. Thus, potential model uncertainties together with its robustness derived from varied parameter settings could furnish us with a valuable reference concerning optimization and adjustment of the developed framework, and better show the confidence interval of the model settings (Abbaszadeh et al., 2022).

3.5 Wind system prediction

In order to evaluate the prediction errors, several methods including recurrent neural network (RNN), Long-short term memory network (LSTM), CNN-LSTM, Encoder-decoder, ResU-Net, and MHA-ResU-Net were used for a comparison with the proposed approach. For the prediction experiments, the mean absolute error (MAE) derived using Equation 23 and the root mean square error (RMSE) derived using Equation 24 were employed as model-evaluation metrics to reveal the performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_{prediction,i} - Y_{observation,i}\right| \tag{23}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{prediction,i} - Y_{observation,i}\right)^2} \tag{24}$$

where the $Y_{observation,i}$ denotes the reanalysis 2D wind map and the $X_{prediction,i}$ indicates the predicted snapshot.

The forecasting metrics are illustrated in Figure 9. The lowest forecasting errors were obtained by the proposed prediction model amongst all individual experiments, which verified that the proposed deep neural learning model outperforms the rest of the models, especially in fine-grid spatial-temporal 2D wind system mapping. The derived area-mean RMSEs for 1-hour-ahead and 12-hour-ahead predictions were less than 0.15 m/s and 0.53 m/s, respectively.

The spatial-resolved wind gust speed predictions were derived and are shown in Figure 10, to further explore the model performance in a fine-grid spatial perspective. *Pre* indicates model forecasting, *ob* represents the reanalysis samples. As displayed in Figure 10, the proposed neural-learning method can preserve the spatial-temporal sequential wind system variabilities, which shows that the spatial-temporal wind evolution patterns were well reproduced for each single wind field snapshot. In addition, extreme wind signals were also well captured continually within the sequential wind evolving trend. Longer leading-step predictions with corresponding deviation maps are shown in Figure 11.

In order to explore the effectiveness of deep-learning-based weather prediction for ship path planning, two types of weather predictions were employed to evaluate an empirical shipping route.

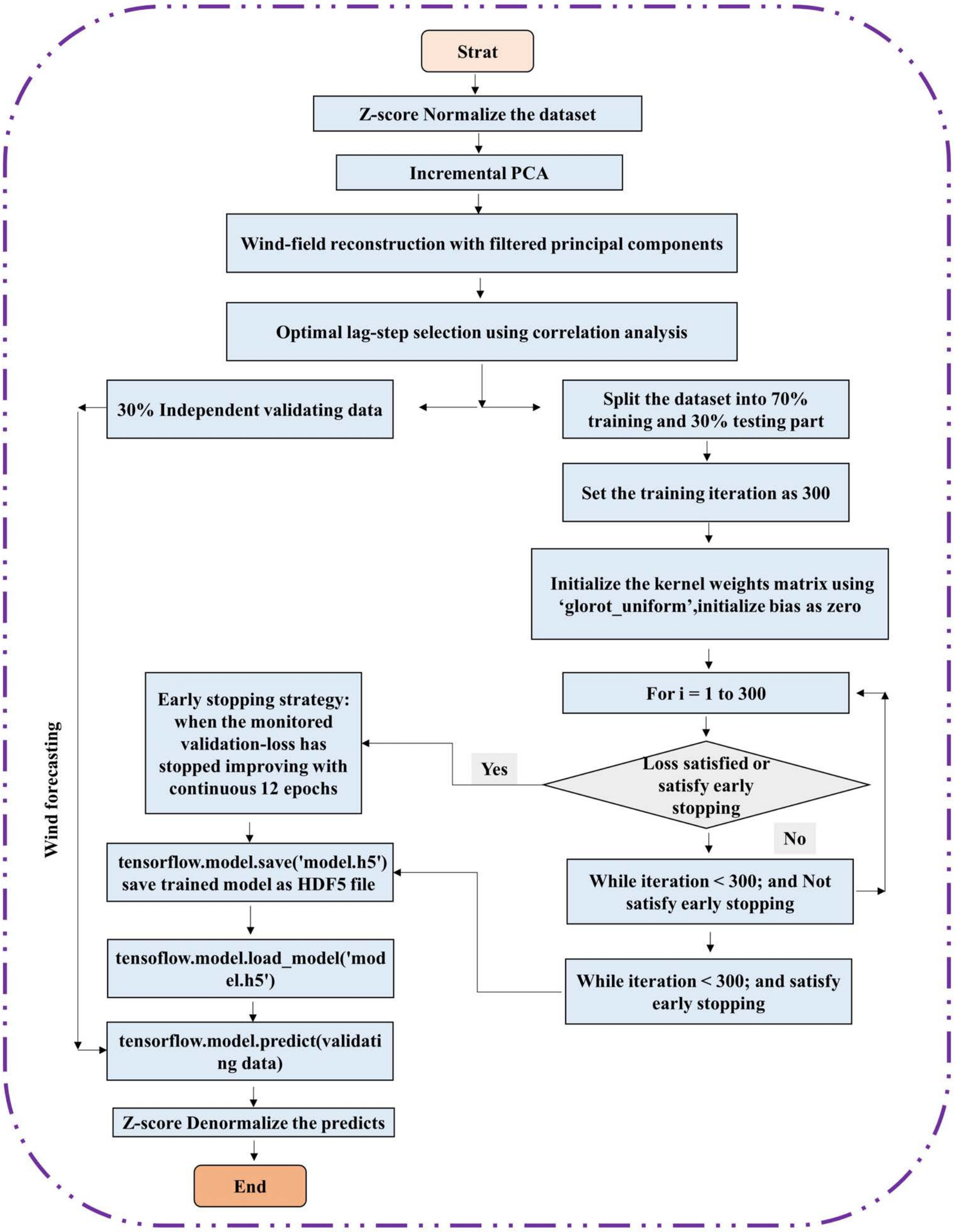

Strat
Z-score Normalize the dataset
Incremental PCA
Wind-field reconstruction with filtered principal components
Optimal lag-step selection using correlation analysis
30% Independent validating data
Split the dataset into 70% training and 30% testing part
Set the training iteration as 300
Initialize the kernel weights matrix using 'glorot_uniform',initialize bias as zero
For i = 1 to 300
Early stopping strategy: when the monitored validation-loss has stopped improving with continuous 12 epochs
Yes
Loss satisfied or satisfy early stopping
No
tensorflow.model.save('model.h5') save trained model as HDF5 file
While iteration < 300; and Not satisfy early stopping
tensoflow.model.load_model('model.h5')
While iteration < 300; and satisfy early stopping
tensorflow.model.predict(validating data)
Z-score Denormalize the predicts
End
Wind forecasting

FIGURE 8
The flowchart of the established wind pattern forecasting network.

It has been proven that if the raw numerical model forecasting data with sparse-grid resolution and with 24-h intervals are utilized to schedule the voyage path, the extreme wind field could not be identified by the ship route optimization software (Yuan et al., 2022; Wu et al., 2023). On the contrary, the developed spatial-temporal deep learning model is able to provide continuous weather forecasting with a very high spatial resolution of 0.25° × 0.25° and an hourly time scale, which will help the path optimization software to identify dangerous navigation regions with accurate area boundaries where severe sea states exist, as displayed in Figure 12. More importantly, the proposed model is able to offer continuous weather forecasting updates, even on a single laptop. This means that the proposed framework combined with reanalysis data samples is very convenient and practical for adaptive path planning of marine vehicles, especially for sea-going navigation in large-scale oceans.

Moreover, a shipping path application was evaluated based on the deep learning-based wind forecast for the sake of better illustrating its effectiveness for efficient and intelligent route planning. Generally speaking, the major part of the experimental shipping route would directly pass through the high sea-state region, if the sparse weather forecasting and weather observation system could not recognize the severe sea state. However, the adaptive ship route based on the proposed continuous fine-grid

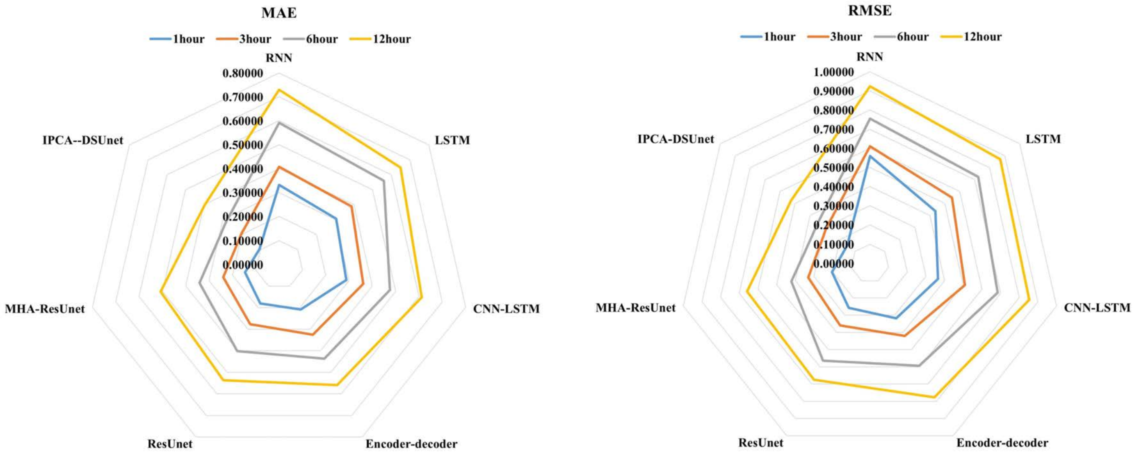


FIGURE 9
Wind pattern forecasting error analysis.

wind forecasting model would accurately avoid adverse weather conditions as much as possible, since the variability of the sea state would be perceived based on weather routing software (Vettor and Soares, 2016; Wu et al., 2023). In addition, it can be seen in Figure 12 that the fine-grid sea-state region detection can help to adjust the experimental path planning accurately using 1-day weather forecasting, and from a sea-going navigation practical perspective, a longer prediction time-span that exceeds 1 day would provide a timely reference for future voyage adjustment. Moreover, with the efficient and intelligent identification of severe weather conditions, autonomous marine vehicles would be able to achieve active obstacle avoidance and intelligent route adjustment, which will lay a solid foundation for intelligent ocean environment perception and the development of smart ships.

4 Model transferability

Deep learning model transferability is a strategy that involves transferring knowledge obtained from the source domain to solve the tasks in a related target domain (Pan and Yang, 2009; Hu et al., 2016). This study provides a machine learning approach that can be employed to transfer the weather forecasting model knowledge gained for available trained jobs from one specific geospatial region to another region's field and time span. It provides an opportunity to transfer information between different datasets and different geospatial regions. Model transferability, including the model hyperparameters and model weights relocation, demonstrates whether a newly developed machine learning method can be transferred directly to an unknown region to realize specific

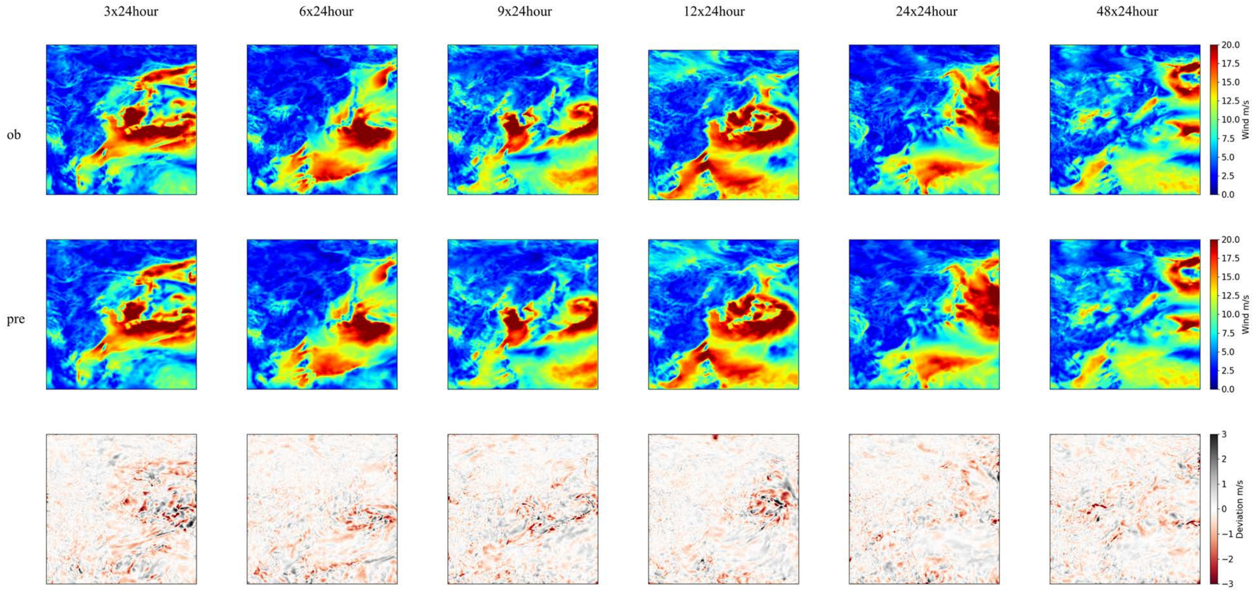


FIGURE 10
Snapshots of spatial-resolved wind speed patterns forecasting.

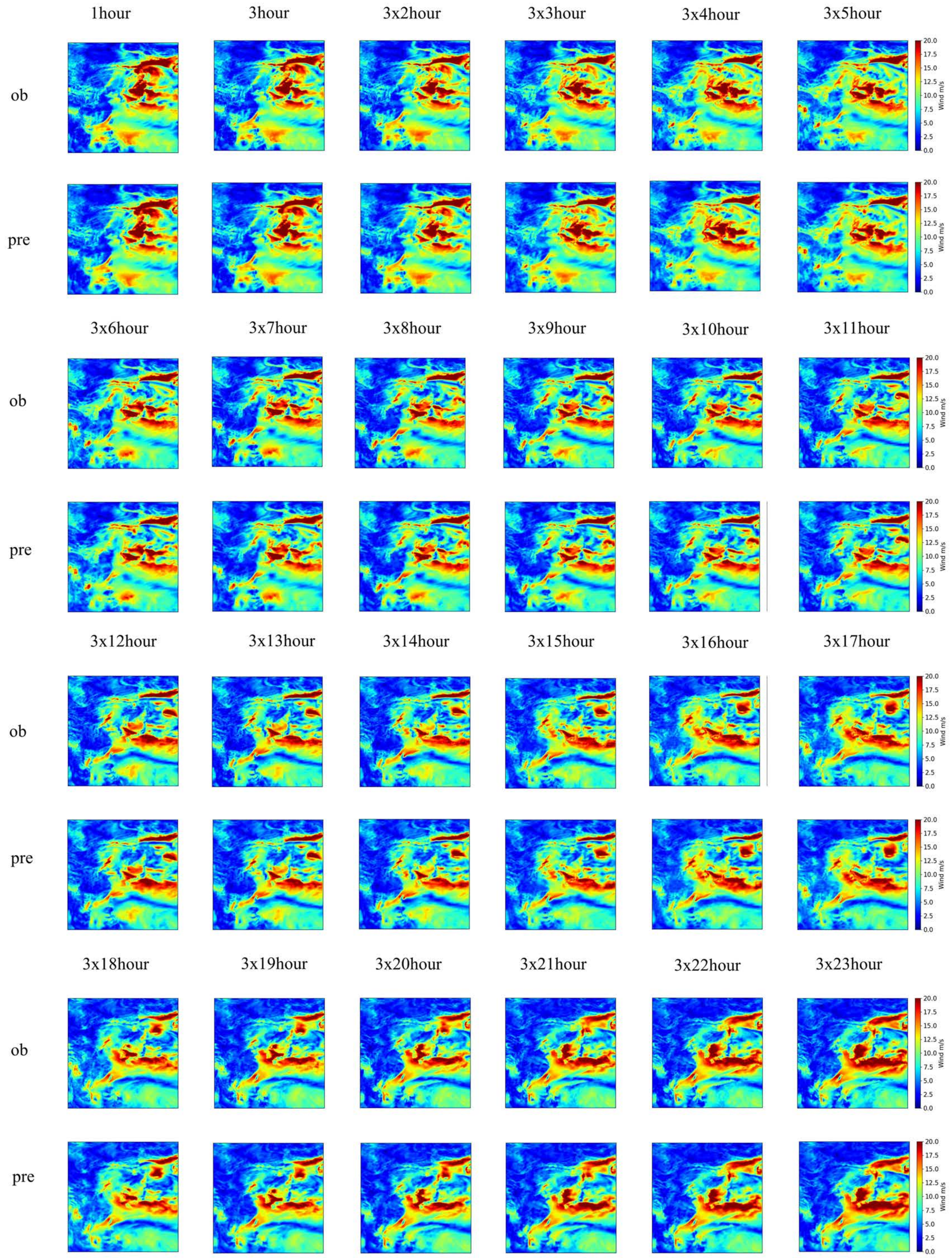


FIGURE 11
Longer surface extreme wind forecasting.

weather forecasting-based ship path planning tasks. A square area covering the North Atlantic Ocean within 6.25 - 70°N, -53.75 - 10°E was selected as the modeling region to realize the same model-hyperparameter transferability-based wind field forecasting directly.

It is illustrated in Figure 13 that the developed neural-learning approach reproduced the spatial-temporal sequential wind system variabilities again. This indicates that the spatial-temporal wind distribution patterns located in different geospatial regions were well preserved for each single field snapshot. The extreme wind signals were also continually captured within the sequential wind-evolving trend. The corresponding longer leading-step forecasting with its deviation fields is displayed in Figure 14.

A new shipping path was evaluated using the deep learning-based wind forecast for the sake of better illustrating its effectiveness on the

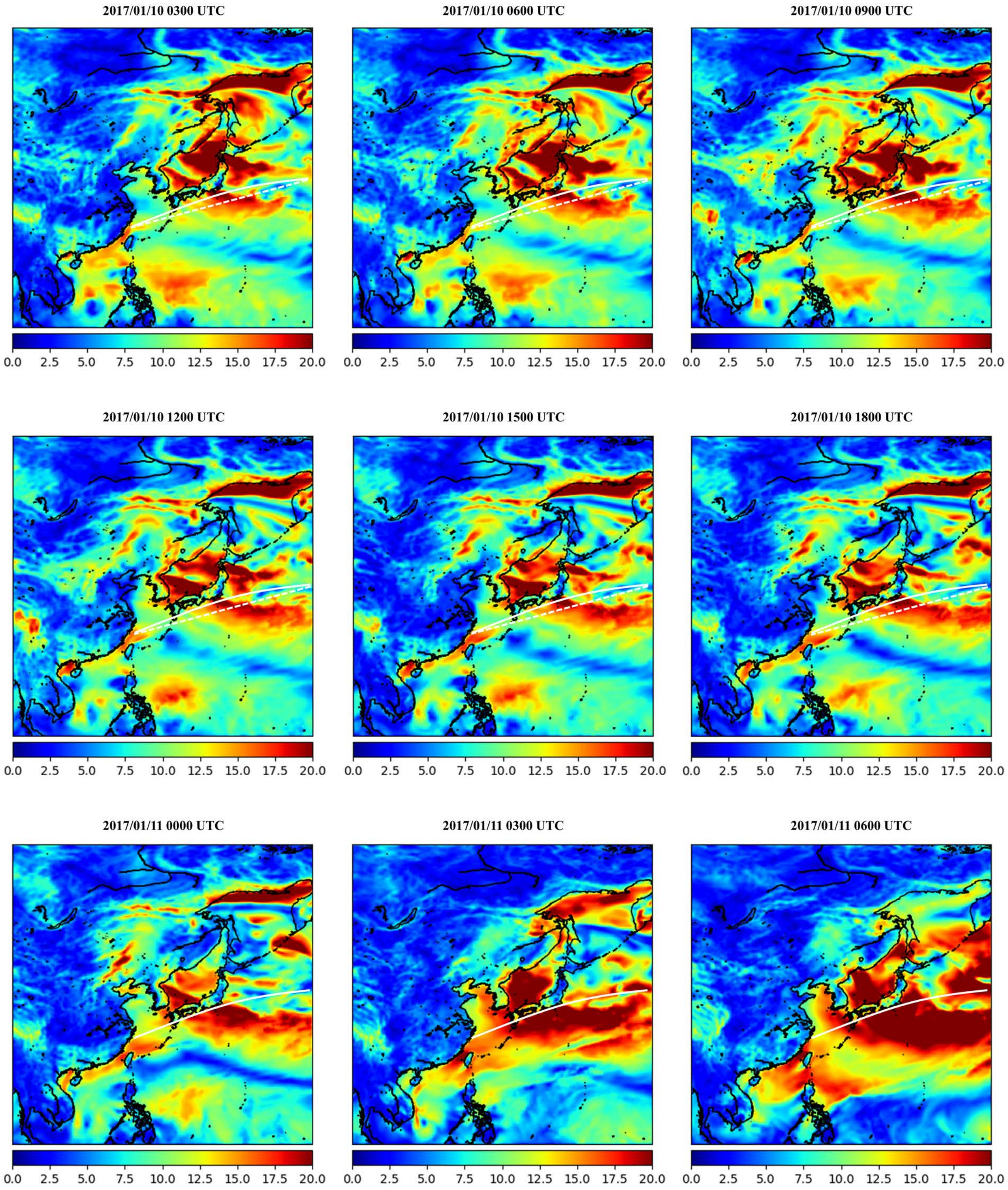


FIGURE 12
Ship path planning based on the wind field forecasting, the white dash-line indicates adaptive path and white line is experimental route.

route plan in the North Atlantic Ocean. It can be seen that the major part of the experimental shipping route directly passes through the high sea state region in Figure 15. However, the adaptive ship route based on the proposed continuous fine-grid wind forecasting model was able to avoid adverse weather conditions as accurately as possible. This can not only ensure the safety of marine vehicles and navigators but also provide voyage planning with timely or real-time path adjustment. The smart shipping industry will greatly benefit from the efficient and intelligent detection of severe large-scale sea states using the proposed wind forecasting model.

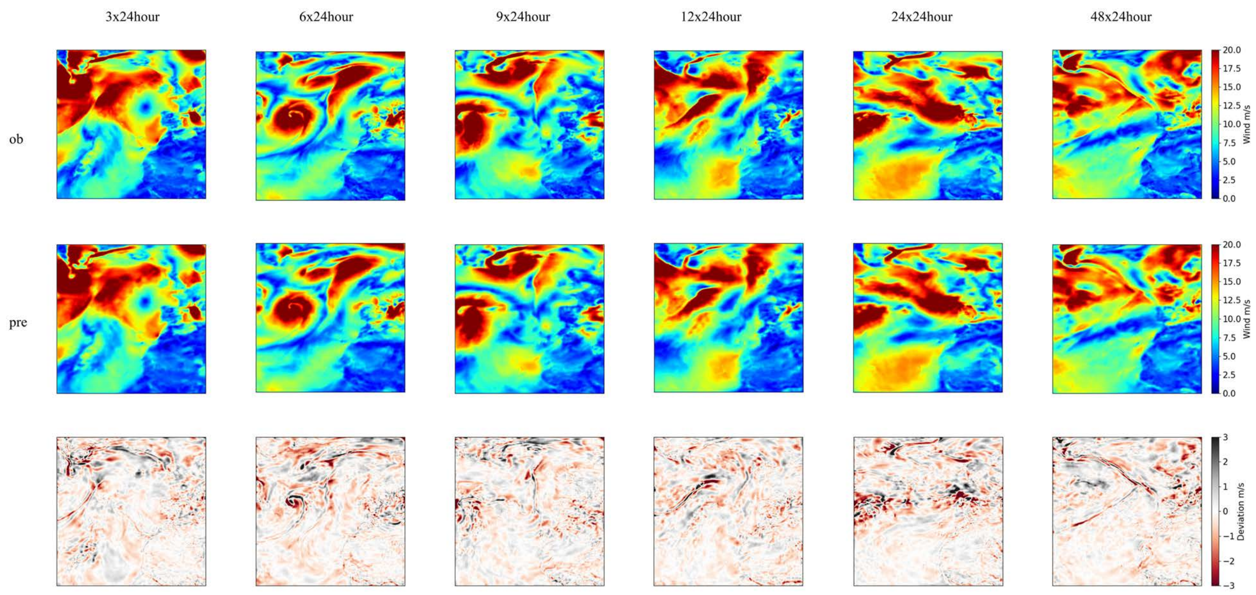


FIGURE 13
Snapshots of spatial-resolved wind speed patterns forecasting as Figure 10, but for the North Atlantic Ocean region.

5 Discussion and conclusion

5.1 Discussion of the model's potential applications and limitations

A depthwise separable U-Net with spatial-temporal attention layers typically has a lower parameter count than standard convolutional U-Nets. Yet, the model is still complex and requires significant computing power for real-time inference. Wind field forecasting involves large volumes of spatial-temporal data, often requiring high-resolution inputs over a continuous time frame. Real-time processing is necessary for effective forecasting, meaning the model must handle frequent data updates without lag. IPCA facilitates dimensionality reduction, which helps manage data size, but there is still a need for fast data preprocessing pipelines to feed into the model without creating bottlenecks. The IPCA-based model necessitates sufficient memory to handle large input matrices (spatial-temporal wind data), intermediate activations, and model weights. The memory requirement can be reduced by applying IPCA to pre-process and compress the input data, but this is still contingent on having enough capacity to maintain intermediate data during real-time inference.

Most ships are limited in terms of the onboard processing power available, typically having less powerful central processing units (CPUs) and possibly limited or no GPUs. While some larger vessels may have limited GPU capacity, deploying such a GPU-based model requires specialized hardware, such as embedded systems with tensor-processing units (TPUs) or compact GPUs. Alternatively, high-performance CPUs capable of supporting multithreading and parallel processing may also be viable, though potentially slower. In addition, other constraints are critical on ships where energy resources are shared among navigation, communication, and other systems. Depthwise separable U-Nets help in reducing computation costs by focusing only on the most relevant filters in the spatial-temporal data. Additionally, IPCA can reduce the data dimensions, resulting in lower power consumption. Nevertheless, the system should be designed to operate within the ship's power constraints, often requiring energy-efficient processors. IPCA provides an advantage by enabling incremental updates, essential for real-time processing on ships, where data is generated continuously and model re-training is impractical. IPCA reduces data dimensions iteratively, which is efficient, but still requires sufficient processing power to perform real-time updates. A balance is necessary between the model's forecasting accuracy and the latency in delivering these forecasts. The depthwise separable U-Net offers computational efficiency, but the real-time application might still necessitate simplifying the model further or accepting coarser forecasting to ensure timely output.

In summary, implementing an IPCA-based spatial-temporal depthwise separable U-Net model on ships requires hardware capable of efficient parallel processing, compact design, and low power consumption. Compact GPUs or embedded TPUs are ideal but may not always be feasible, especially on smaller vessels. CPU-based implementations are possible but might face latency issues. Reducing model complexity and utilizing IPCA for dimensionality reduction can mitigate some hardware limitations, but ongoing trade-offs between computational power, accuracy, and latency will be required to make this model operational on actual ships. For stakeholders, understanding these constraints is crucial for planning resource allocation, assessing deployment feasibility, and selecting suitable hardware for maritime forecasting applications.

Concerning the model's limitations, in marine environments, wind patterns are highly variable and can be influenced by various factors such as ship movements and surrounding weather systems. The IPCA's

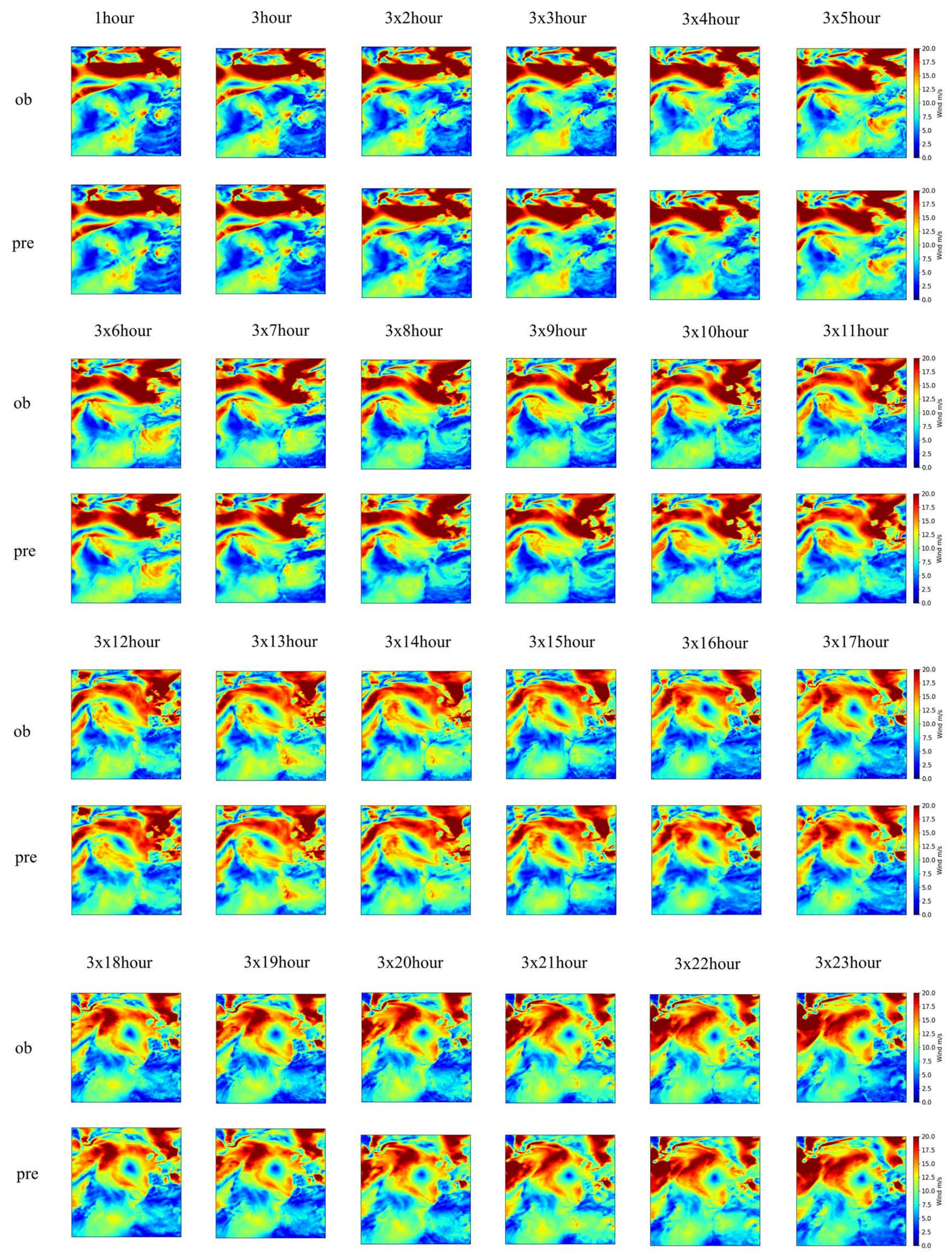


FIGURE 14
The same as Figure 11, but for the North Atlantic Ocean.

incremental learning approach may not fully capture this complexity, as it assumes incremental changes to the learned principal components, which may not adapt quickly enough to abrupt shifts or highly dynamic wind fields. Furthermore, incremental updates in IPCA rely on frequent model retraining with new data. This approach risks underperforming if updates are too infrequent or if older components fail to capture emerging patterns. This can lead to model drift, where the U-Net model's depthwise separable convolutions become misaligned with the shifting data distributions. Moreover, onboard computing systems may be limited in memory and processing power, restricting the model's ability to perform complex IPCA transformations alongside the spatial-temporal depthwise separable U-Net operations. This constraint could necessitate simplifying the model at the cost of predictive accuracy.

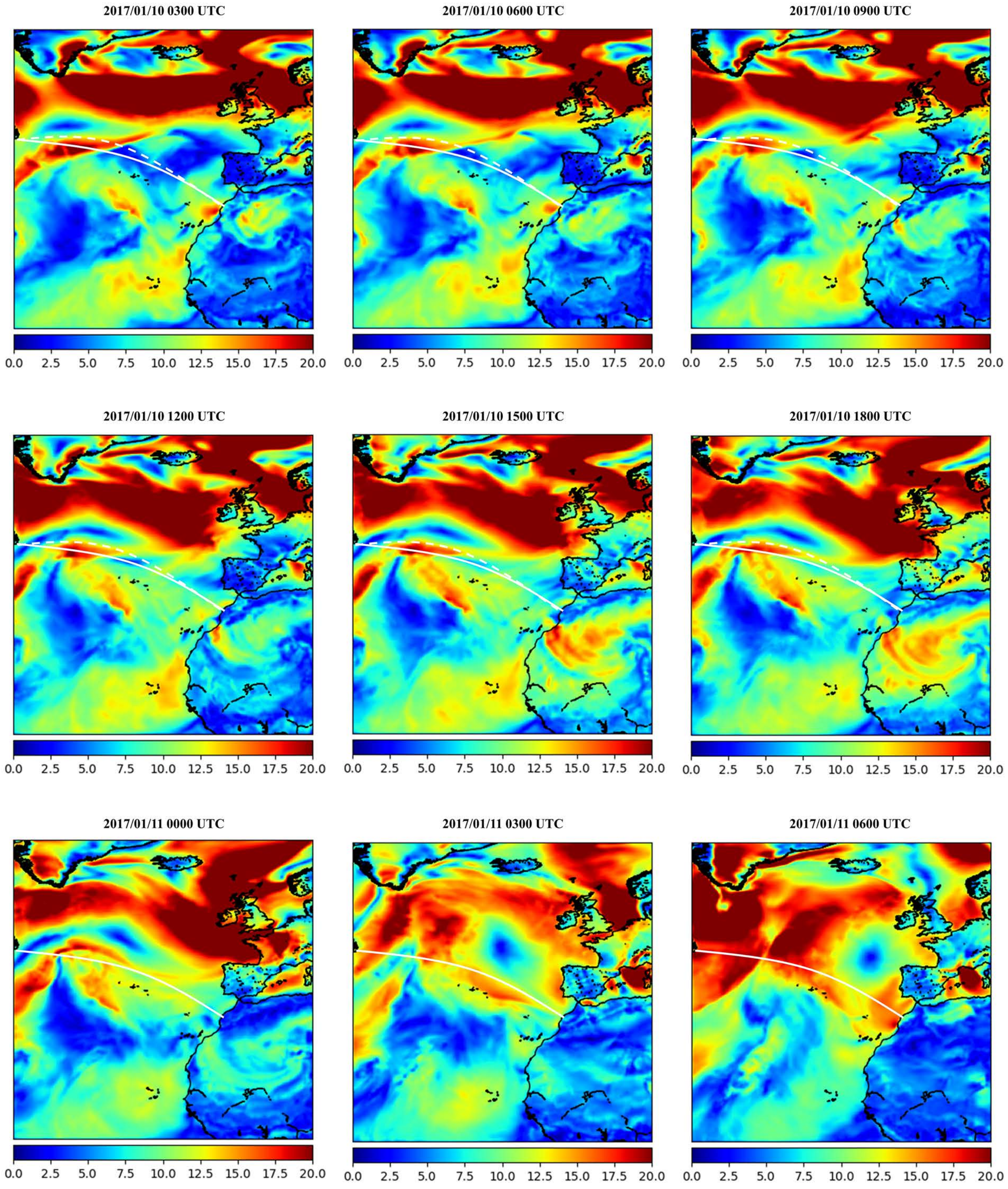


FIGURE 15
The same as Figure 12, but for ship path planning at the North Atlantic Ocean.

While depthwise separable convolutions reduce computation by splitting spatial and channel-wise filtering, combining them with IPCA can lead to a loss in detail, particularly in fine-grid scenarios where capturing spatial intricacies is critical. Depthwise operations, while efficient, may not fully exploit the principal components' spatial relationships, leading to potential oversimplification. Depthwise separable convolutions, when paired with IPCA, might over-rely on a limited number of components, as selecting too many can offset efficiency gains. Choosing an appropriate number of components becomes crucial but challenging in achieving a balance between spatial detail and computational feasibility. In addition, ships' routes, speeds, and maneuvers might introduce unique challenges in wind field predictions. These unpredictable movements can make it difficult for an IPCA-based U-Net model to maintain consistent predictive accuracy, as rapid course or speed changes could invalidate previously learned components or spatial patterns. Addressing these limitations would involve strategies such as incorporating more adaptive or hierarchical components within the IPCA process, leveraging advanced real-time data filtering, or incorporating more sophisticated recurrent mechanisms within the U-Net architecture in the future steps to handle temporal dynamics better.

5.2 Conclusions

In order to provide instantaneous extreme wind system pattern mapping tasks, and provide adaptive and intelligent path planning for marine vehicles, especially for sea-going navigations in large-scale oceans, a spatial-temporal 2D depthwise separable convolutional based neural-learning model was developed by integrating the multi-head feature-concentrated attention scheme. Specifically, incremental principal component analysis was first employed to filter the feature space of 2D wind data samples by reducing dimensionality and redundant features. The proposed wind forecasting network was employed to capture and preserve the intermittence and non-linearity of spatial-temporal wind system evolutions between the future wind pattern distributions and the historical wind time-series snapshots. The historical wind time lags with a strict chronological order were determined by further introducing a sequential sliding-data window approach and the established spatial-temporal feature mapping methodology was then able to capture the underlying temporal dependencies and variabilities from the consecutive wind maps. In addition, the transferability of the proposed model was verified by employing two geospatial regions with different weather characteristics. By mapping weather observational gaps into a fine-grid and complete spatial format, the proposed approach, implemented in a single laptop, aimed to improve the timeliness and accuracy of onboard ship routing, thereby enhancing ship navigation safety. Based on the efficient and intelligent identification of severe weather conditions, autonomous marine vehicles will be able to achieve active obstacle avoidance and intelligent route adjustment, which will lay a solid foundation for intelligent ocean environment perception for the development of smart shipping.

The experimental findings in this study demonstrate that the developed deep learning-based methodology can accurately and effectively detect severe wind fields. Yet, some limitations remain. For example, other meteorological factors such as atmosphere pressure and wave height conditions were not fully taken into account. Furthermore, issues such as fuel consumption were not considered, which could impact intelligent weather routing-based predictions and ship navigation safety, thus, future research is required to better consider ship navigation performance and its efficiency index and realize a more reliable smart ship path planning task.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.

Author contributions

ZZ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. LC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. JY: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the National Natural Science Foundation of China under Grants 52271361 and 52231014, the Special Projects of Key Areas for Colleges and Universities in Guangdong Province under Grant 2021ZDZX1008, the Natural Science Foundation of Guangdong Province of China under Grant 2023A1515010684.

Acknowledgments

We would like to acknowledge the organizations that provided the sources of the data used in this work, namely, the European Centre for Medium-Range Weather Forecasts.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2025.1495822/full#supplementary-material

References

Abbaszadeh, S., Shan, C., and Larsson, S. (2022). A novel approach to uncertainty quantification in groundwater table modeling by automated predictive deep learning. *Nat. Resour.* 31, 1351–1373. doi: 10.1007/s11053-022-10051-w

Asheghi, R., Hosseini, S. A., Saneie, M., and Shahri, A. A. (2020). Updating the neural network sediment load models using different sensitivity analysis methods: a regional application. *J. HYDROINFORM.* 22, 562–577. doi: 10.2166/hydro.2020.098

Cai, H., Jia, X., Feng, J., Li, W., Hsu, Y. M., and Lee, J. (2020). Gaussian process regression for numerical wind speed prediction enhancement. *Renew Energ.* 146, 2112–2123. doi: 10.1016/j.renene.2019.08.018

Che, H., Niu, D., Zang, Z., Cao, Y., and Chen, X. (2022). ED-DRAP: Encoder–decoder deep residual attention prediction network for radar echoes. *IEEE GEOSCI Remote* S, 1–5. doi: 10.1109/LGRS.2022.3141498

Chen, C., Sasa, K., Prpić-Oršić, J., and Mizojiri, T. (2021a). Statistical analysis of waves' effects on ship navigation using high-resolution numerical wave simulation and shipboard measurements. *Ocean eng.* 229, 108757. doi: 10.1016/j.oceaneng.2021.108757

Chen, G., Wu, T., and Zhou, Z. (2021b). Research on ship meteorological route based on A-star algorithm. *MATH PROBL ENG.* 2021, 1–8. doi: 10.1155/2021/9989731

Chen, Y., and Mao, W. (2024). An isochrone-based predictive optimization for efficient ship voyage planning and execution. *IEEE Trans. Intell Transp. Syst.* 25, 18078–18092 doi: 10.1109/TITS.2024.3416349

Cheng, W. Y., Liu, Y., Liu, Y., Zhang, Y., Mahoney, W. P., and Warner, T. T. (2013). The impact of model physics on numerical wind forecasts. *Renew. Energ.* 55, 347–356. doi: 10.1016/j.renene.2012.12.041

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR, Honolulu, HI, USA), 1251–1258.

Du, Y., Chen, Y., Li, X., Schönborn, A., and Sun, Z. (2022). Data fusion and machine learning for ship fuel efficiency modeling: Part III–Sensor data and meteorological data. *Commun. Transport Res.* 2, 100072. doi: 10.1016/j.commtr.2022.100072

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML'16: proceedings of the 33rd international conference on international conference on machine learning* (New York, NY, USA: ICML), 1050–1059.

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., and Tuzhilina, E. (2022). Principal component analysis. *Nat. Rev. Methods Primers.* 2, 100. doi: 10.1038/s43586-022-00184-w

He, Y., Liu, X., Zhang, K., Mou, J., Liang, Y., Zhao, X., et al. (2022). Dynamic adaptive intelligent navigation decision making method for multi-object situation in open water. *Ocean Eng.* 253, 111238. doi: 10.1016/j.oceaneng.2022.111238

Hu, Q., Zhang, R., and Zhou, Y. (2016). Transfer learning for short-term wind speed prediction with deep neural networks. *Renew Energ.* 85, 83–95. doi: 10.1016/j.renene.2015.06.034

Huang, Y., Zhao, S., and Zhao, S. (2024). Ship trajectory planning and optimization via ensemble hybrid A* and multi-target point artificial potential field model. *J. Mar. Sci. Eng.* 12, 1372. doi: 10.3390/jmse12081372

Hur, S. H. (2021). Short-term wind speed prediction using Extended Kalman filter and machine learning. *Energy Rep.* 7, 1046–1054. doi: 10.1016/j.egyr.2020.12.020

Khan, S., Grudniewski, P., Muhammad, Y. S., and Sobey, A. J. (2022). The benefits of co-evolutionary Genetic Algorithms in voyage optimisation. *Ocean Eng.* 245, 110261. doi: 10.1016/j.oceaneng.2021.110261

Khodayar, M., and Wang, J. (2018). Spatio-temporal graph deep neural network for short-term wind speed forecasting. *IEEE T SUSTAIN ENERG* 10, 670–681. doi: 10.1109/TSTE.2018.2844102

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature* 632, 1060–1066. doi: 10.1038/s41586-024-07744-y

Koukaki, T., and Tei, A. (2020). Innovation and maritime transport: A systematic review. *Case Stud. Transp. Policy.* 8, 700–710. doi: 10.1016/j.cstp.2020.07.009

Krata, P., and Szlapczynska, J. (2018). Ship weather routing optimization with dynamic constraints based on reliable synchronous roll prediction. *Ocean Eng.* 150, 124–137. doi: 10.1016/j.oceaneng.2017.12.049

Kytariolou, A., and Themelis, N. (2022). Ship routing optimisation based on forecasted weather data and considering safety criteria. *J. Navig.* 75, .1310–.1331. doi: 10.1017/S0373463322000613

Lau, Y., Chen, Q., Poo, M. C. P., Ng, A. K., and Ying, C. (2024). Maritime transport resilience: A systematic literature review on the current state of the art, research agenda and future research directions. *Ocean Coast. Manage.* 251, 107086. doi: 10.1016/j.ocecoaman.2024.107086

Ma, W., Ma, D., Ma, Y., Zhang, J., and Wang, D. (2021). Green maritime: A routing and speed multi-objective optimization strategy. *J. Clean Prod.* 305, 127179. doi: 10.1016/j.jclepro.2021.127179

Ma, D., Zhou, S., Han, Y., Ma, W., and Huang, H. (2024). Multi-objective ship weather routing method based on the improved NSGA-III algorithm. *J. Ind. Inf. Integr.* 38, 100570. doi: 10.1016/j.jii.2024.100570

Manucharyan, G. E., Siegelman, L., and Klein, P. (2021). A deep learning approach to spatiotemporal sea surface height interpolation and estimation of deep currents in geostrophic ocean turbulence. *J. Adv. Model. Earth Sy* 13, e2019MS001965. doi: 10.1029/2019MS001965

Moradi, M. H., Brutsche, M., Wenig, M., Wagner, U., and Koch, T. (2022). Marine route optimization using reinforcement learning approach to reduce fuel consumption and consequently minimize CO2 emissions. *Ocean Eng.* 259, 111882. doi: 10.1016/j.oceaneng.2022.111882

Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing* 452, 48–62. doi: 10.1016/j.neucom.2021.03.091

Ouyang, T., Zha, X., and Qin, L. (2017). A combined multivariate model for wind power prediction. *Energ Convers Manage.* 144, 361–373. doi: 10.1016/j.enconman.2017.04.077

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Parri, S., and Teeparthi, K. (2024). SVMD-TF-QS: An efficient and novel hybrid methodology for the wind speed prediction. *Expert Syst. Appl.* 249, 123516. doi: 10.1016/j.eswa.2024.123516

Qiao, Y., Yin, J., Wang, W., Duarte, F., Yang, J., and Ratti, C. (2023). Survey of deep learning for autonomous surface vehicles in marine environments. *IEEE Trans. Intell. Transp. Syst.* 24, 3678–3701. doi: 10.1109/TITS.2023.3235911

Rawson, A., Brito, M., Sabeur, Z., and Tran-Thanh, L. (2021). A machine learning approach for monitoring ship safety in extreme weather events. *Saf. Sci.* 141, 105336. doi: 10.1016/j.ssci.2021.105336

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference* (Springer International Publishing, Munich, Germany), 234–241.

Szlapczynski, R., Szlapczynska, J., and Vettor, R. (2023). Ship weather routing featuring w-MOEA/D and uncertainty handling. *Appl. Soft Comput.* 138, 110142. doi: 10.1016/j.asoc.2023.110142

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 5998–6008. doi: 10.48550/arXiv.1706.03762

Vettor, R., and Soares, C. G. (2016). Development of a ship weather routing system. *Ocean Eng.* 123, 1–14. doi: 10.1016/j.oceaneng.2016.06.035

Wang, Y., Zou, R., Liu, F., Zhang, L., and Liu, Q. (2021). A review of wind speed and wind power forecasting with deep neural networks. *Appl. Energy.* 304, 117766. doi: 10.1016/j.apenergy.2021.117766

Weng, J., Zhang, Y., and Hwang, W. S. (2003). Candid covariance-free incremental principal component analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 1034–1040. doi: 10.1109/TPAMI.2003.1217609

Wu, Z., Wang, S., Yuan, Q., Lou, N., Qiu, S., Bo, L., et al. (2023). Application of a deep learning-based discrete weather data continuousization model in ship route optimization. *Ocean Eng.* 285, 115435. doi: 10.1016/j.oceaneng.2023.115435

Xiao, Y., Zou, C., Chi, H., and Fang, R. (2023). Boosted GRU model for short-term forecasting of wind power with feature-weighted principal component analysis. *Energy.* 267, 126503. doi: 10.1016/j.energy.2022.126503

Xu, H., Hu, F., Liang, X., Zhao, G., and Abugunmi, M. (2024). A framework for electricity load forecasting based on attention mechanism time series depthwise separable convolutional neural network. *Energy* 299, 131258. doi: 10.1016/j.energy.2024.131258

Xu, X., Hu, S., Shi, P., Shao, H., Li, R., and Li, Z. (2023a). Natural phase space reconstruction-based broad learning system for short-term wind speed prediction: Case studies of an offshore wind farm. *Energy* 262, 125342. doi: 10.1016/j.energy.2022.125342

Xu, L., Ou, Y., Cai, J., Wang, J., Fu, Y., and Bian, X. (2023b). Offshore wind speed assessment with statistical and attention-based neural network methods based on STL decomposition. *Renew Energ.* 216, 119097. doi: 10.1016/j.renene.2023.119097

Yan, B., Shen, R., Li, K., Wang, Z., Yang, Q., Zhou, X., et al. (2023). Spatio-temporal correlation for simultaneous ultra-short-term wind speed prediction at multiple locations. *Energy* 284, 128418. doi: 10.1016/j.energy.2023.128418

Yin, J., Wang, H., Wang, N., and Wang, X. (2023). An adaptive real-time modular tidal level prediction mechanism based on EMD and Lipschitz quotients method. *Ocean Eng.* 289, 116297. doi: 10.1016/j.oceaneng.2023.116297

Yu, C., Yan, G., Yu, C., and Mi, X. (2023). Attention mechanism is useful in spatio-temporal wind speed prediction: Evidence from China. *Appl. Soft Comput.* 148, 110864. doi: 10.1016/j.asoc.2023.110864

Yuan, Q., Wang, S., Zhao, J., Hsieh, T. H., Sun, Z., and Liu, B. (2022). Uncertainty-informed ship voyage optimization approach for exploiting safety, energy saving and low carbon routes. *Ocean Eng.* 266, 112887. doi: 10.1016/j.oceaneng.2022.112887

Zhang, Z., Lin, L., Gao, S., Wang, J., and Zhao, H. (2024a). Wind speed prediction in China with fully-convolutional deep neural network. *RENEW SUST ENERG Rev.* 201, 114623. doi: 10.1016/j.rser.2024.114623

Zhang, C., Tao, Z., Xiong, J., Qian, S., Fu, Y., Ji, J., et al. (2024b). Research and application of a novel weight-based evolutionary ensemble model using principal component analysis for wind power prediction. *Renew Energ.* 232, 121085. doi: 10.1016/j.renene.2024.121085

Zhang, Z., Wagner, S., Klockmann, M., and Zorita, E. (2022). Evaluation of statistical climate reconstruction methods based on pseudoproxy experiments using linear and machine-learning methods. *Clim Past.* 18, 2643–2668. doi: 10.5194/cp-18-2643-2022

Zhao, Q., Peng, S., Wang, J., Li, S., Hou, Z., and Zhong, G. (2024). Applications of deep learning in physical oceanography: a comprehensive review. *Front. Mar. Sci.* 11. doi: 10.3389/fmars.2024.1396322

Zhou, Y., Kang, X., Ren, F., Lu, H., Nakagawa, S., and Shan, X. (2024). A multi-attention and depthwise separable convolution network for medical image segmentation. *Neurocomputing* 564, 126970. doi: 10.1016/j.neucom.2023.126970

Zhou, P., Zhou, Z., Wang, Y., and Wang, H. (2023). Ship weather routing based on hybrid genetic algorithm under complicated sea conditions. *J. Ocean Univ. China* 22, 28–42. doi: 10.1007/s11802-023-5002-1

Zhu, M., Kong, M., Wen, Y., Gu, S., Xue, B., and Huang, T. (2025a). A multi-objective path planning method for ships based on constrained policy optimization. *Ocean Eng.* 319, 120165. doi: 10.1016/j.oceaneng.2024.120165

Zhu, J., Shen, H., Tang, Q., Qin, Z., and Yu, Y. (2025b). Energy-efficient route planning method for ships based on level set. *Sensors.* 25, 381. doi: 10.3390/s25020381

Zis, T. P., Psaraftis, H. N., and Ding, L. (2020). Ship weather routing: A taxonomy and survey. *Ocean Eng.* 213, 107697. doi: 10.1016/j.oceaneng.2020.107697

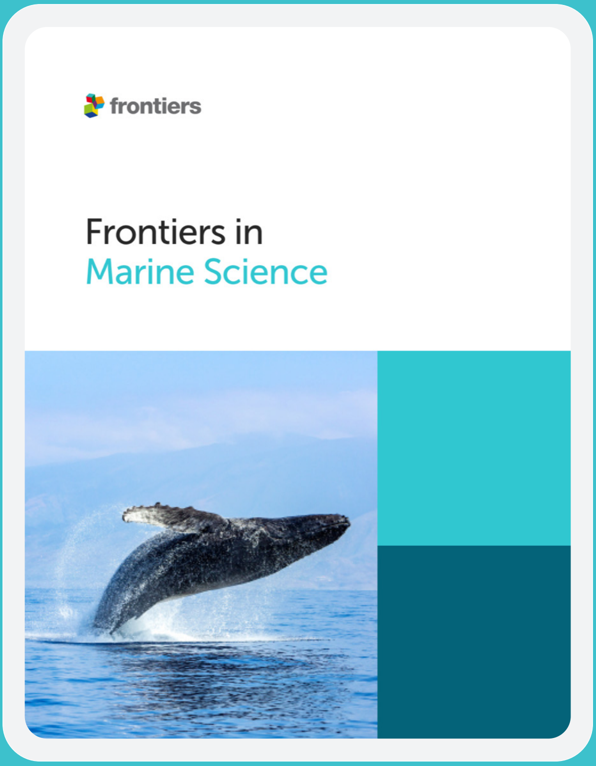
Frontiers in Marine Science

Explores ocean-based solutions for emerging global challenges

The third most-cited marine and freshwater biology journal, advancing our understanding of marine systems and addressing global challenges including overfishing, pollution, and climate change.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Marine Science

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact