PHYSIOCEAN: A SPATIO-TEMPORAL OPERATOR WITH PHYSICS-GUIDED LEARNING FOR COASTAL DISSOLVED OXYGEN FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting dissolved oxygen (DO) across monitoring locations and depth layers is challenging because coastal DO dynamics involve multiscale environmental forcing, strong spatial heterogeneity, and sparse observations. Existing data-driven approaches typically model individual stations independently or rely on fixed spatial decompositions, limiting their ability to jointly represent temporal dynamics and continuous spatial variation. We introduce **PhysiOcean**, a spatio-temporal operator that maps historical environmental forcing and spatial coordinates to DO concentrations. PhysiOcean combines a multi-scale causal temporal encoder, a learnable Fourier spatial decoder, and a multi-head nonlinear fusion mechanism to learn regional DO fields end-to-end. EOF-guided initialization and physics-guided regularization further incorporate dominant spatial structure and domain knowledge during training. We analyze the expressivity of the proposed fusion mechanism and the complexity of an idealized physics-constrained hypothesis class. Experiments on a 32-year monitoring record from a large, environmentally heterogeneous estuary containing 40 monitoring stations and 800 station-depth locations show that PhysiOcean outperforms sequence, decomposition-based, and operator-learning baselines. Component analyses demonstrate that the temporal encoder, continuous spatial representation, fusion mechanism, EOF initialization, and physics-guided losses provide complementary improvements. PhysiOcean also maintains stronger performance at monitoring locations and depth layers excluded from training, demonstrating within-estuary generalization to unobserved spatial coordinates. Our code is available at `https://anonymous.4open.science/r/physiocean/`.

1 INTRODUCTION

Forecasting spatiotemporal variability of aquatic elements is central to environmental sustainability. Coastal hypoxia is growing global threat to marine ecosystems (Diaz & Rosenberg, 2008; Breitburg et al., 2018), causing mass mortality of benthic organisms. The number and extent of coastal hypoxic sites have increased rapidly in recent decades (Breitburg et al., 2018), driven by anthropogenic nutrient loading, altered wind patterns, and strengthened stratification (Carstensen et al., 2014; Du et al., 2018). Therefore, accurate DO forecasting is urgent and essential for ecosystem management. However, DO dynamics are governed by nonlinear and multiscale interactions among physical transport, stratification, and biogeochemical processes (Fennel & Testa, 2019; Kemp et al., 2005), with substantial delayed responses to external forcings and sparse and irregular observations. These features make DO forecasting a challenging spatiotemporal learning problem.

Existing approaches to DO prediction can be broadly grouped into three categories, including **(i) Mechanistic numerical models** couple three-dimensional hydrodynamic and biogeochemical equations to represent DO dynamics (Cerco & Noel, 2013; Testa et al., 2014; Irby et al., 2016). However, they are computationally expensive, require extensive calibration, and accumulate errors across coupled model components. **(ii) Individual-spatial data-driven models** predict DO at specific locations using locally collected physicochemical parameters (Wang et al., 2024; Liang et al., 2023; Li et al., 2021). Although often accurate at individual sites, they treat each station independently and therefore cannot generalize to unobserved locations or effectively leverage regional-scale external

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

forcings, such as river discharge and air temperature. **(iii) Decomposition-based regional models**, exemplified by the Empirical Orthogonal Functions+Neural Network (EOF+NN) approach (Yu et al., 2020), decompose the DO field into a set of fixed spatial modes using EOF and then train separate neural networks for each temporal component. While this approach reduces dimensionality, the spatial modes are predetermined during training data and cannot adapt to unseen locations. More broadly, across all three categories, existing machine learning approaches neglect or fail to incorporate physical constraints, such as saturation bounds and spatial smoothness, that govern DO variability.

These limitations motivate a coordinate-conditioned forecasting framework that jointly learns temporal forcing-response relationships and continuous spatial structure, while allowing domain knowledge to guide training. We introduce **PhysiOcean**, a spatio-temporal operator built on a multi-head Deep Operator Network (DeepONet) (Lu et al., 2021). PhysiOcean maps a history of environmental forcing and a spatial coordinate to the corresponding DO concentration. Our main contributions are as follows:

1. **Spatio-temporal operator architecture.** We develop a multi-scale causal temporal encoder, a learnable Fourier spatial decoder, and a multi-head fusion mechanism with nonlinear correction. The resulting operator represents temporal forcing-response relationships and continuous spatial structure jointly, and is more expressive than the standard bilinear DeepONet (Theorem 1).
2. **EOF- and physics-guided learning.** We use EOF modes to initialize the spatial trunk and formulate four domain-specific physical priors (bounds, smoothness, stratification, and temporal continuity) as soft differentiable regularizers with a three-phase schedule. We analyze the expressivity of the multi-head operator, the complexity of an idealized constrained hypothesis class (Theorem 2), and local optimization from EOF initialization under stated assumptions (Proposition 1).
3. **Evaluation of forecasting and within-estuary spatial generalization.** We evaluate PhysiOcean on a 32-year record from a large and environmentally heterogeneous estuary, using standard temporal testing together with held-out stations, held-out depth layers, and sparse vertical training. These experiments assess generalization to unobserved coordinates within the study estuary rather than transfer across geographically distinct coastal systems.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

We consider the problem of forecasting dissolved oxygen concentrations across space and time from external environmental forcings. We use $\mathcal{S} \subset \mathbb{R}^{d_s}$ to denote the spatial domain of a water body, where each point $\boldsymbol{x} \in \mathcal{S}$ encodes geographic coordinates and bathymetric features. We use $\boldsymbol{f}_t \in \mathbb{R}^{d_f \times W}$ to denote a window of W months of d_f-dimensional forcing variables (e.g., river discharge, nutrient loading, air temperature, wind, and solar radiation) ending at month t. The goal is to learn a mapping

$$\mathcal{G} : (\boldsymbol{f}_t, \boldsymbol{x}) \mapsto \mathrm{DO}(\boldsymbol{x}, t) \in \mathbb{R}, \tag{1}$$

where $\mathcal{G}$ accurately predicts DO concentrations at any specific location $\boldsymbol{x}$ and time t given the forcing context $\boldsymbol{f}_t$, including at locations not observed during training.

In this setting, $\boldsymbol{f}_t$ represents regionally aggregated external drivers (e.g., river discharge and meteorology conditions) that are spatially uniform, that is, they are shared across all query locations $\boldsymbol{x}$. Given a training set $\mathcal{D} = \{(\boldsymbol{f}_{t_i}, \boldsymbol{x}_j, y_{ij})\}$ where y_{ij} is the observed DO at time t_i and location $\boldsymbol{x}_j$, we aim to minimize the expected prediction error while satisfying known physical constraints (detailed in §2.2).

2.2 PHYSICS OF DISSOLVED OXYGEN IN ESTUARIES

The dynamics of DO in estuarine waters are governed by well-established physical and biogeochemical principles that provide strong inductive biases for learning. We identify four key constraints that hold universally across coastal systems:

(P1) Non-negativity and saturation bounds. DO concentration is physically non-negative and generally bounded above by the oxygen saturation concentration $C_{\text{sat}}(T, S)$, which depends on

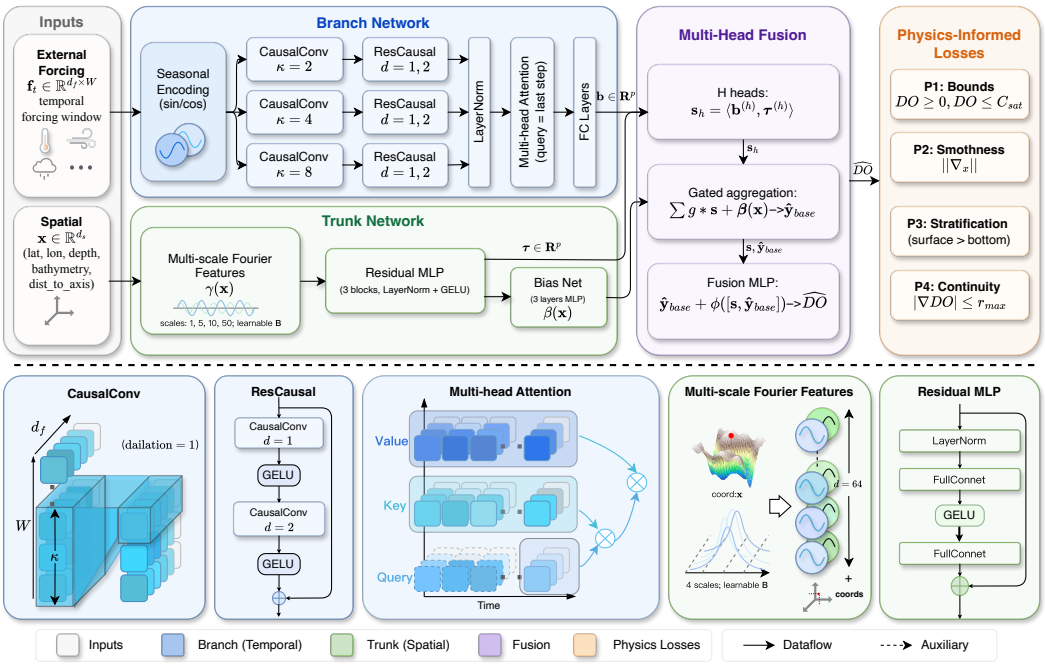


Figure 1: Overall framework of PhysiOcean. The Branch Network encodes multi-scale temporal forcing via parallel causal convolutions and self-attention into basis coefficients $\boldsymbol{b}$. The Trunk Network maps spatial coordinates through learnable Fourier features and a residual MLP to basis functions $\boldsymbol{\tau}$. The multi-head fusion mechanism computes per-head inner products, then applies gated aggregation and a correction MLP. Physics-informed losses (P1–P4) are progressively introduced during training.

temperature T and salinity S (Weiss, 1970):

$$0 \leq \mathrm{DO}(\boldsymbol{x}, t) \leq C_{\mathrm{sat}}(T, S). \tag{2}$$

(P2) Spatial smoothness. DO fields in large water bodies are governed by advection-diffusion processes, which ensure spatial smoothness. The governing equation implies bounded spatial gradients $\|\nabla_{\boldsymbol{x}}\mathrm{DO}\|$ in the absence of sharp fronts (Fennel & Testa, 2019).

(P3) Summer stratification. During summer months (May–September), thermal stratification in estuaries restricts vertical mixing, leading to oxygen depletion in bottom waters. This produces a robust vertical gradient, with surface DO exceeding bottom DO (Kemp et al., 2005; Murphy et al., 2011).

(P4) Temporal continuity. Monthly-averaged DO changes are bounded by the net rates of biological production, consumption, and physical transport. Abrupt month-to-month changes exceeding $\sim$3 mg/L are physically implausible under normal conditions.

These constraints, while individually simple, collectively define a physically meaningful hypothesis class that is substantially smaller than the unconstrained function space. We formalize this intuition in §3.4 and §C.

3 PHYSIOCEAN

PhysiOcean (Figure 1) is a coordinate-conditioned spatio-temporal operator with three architectural components: (1) a *Branch Network* that encodes temporal forcing into basis coefficients, (2) a *Trunk Network* that maps spatial coordinates to continuous basis functions, and (3) a *multi-head fusion mechanism* that combines temporal and spatial representations beyond the standard bilinear DeepONet formulation. EOF-guided initialization and physics-guided regularization are training enhancements that provide additional spatial structure and domain knowledge.

3.1 BRANCH NETWORK: MULTI-SCALE CAUSAL TEMPORAL ENCODER

The Branch Network transforms a forcing window $\boldsymbol{f}_t \in \mathbb{R}^{d_f \times W}$ into a vector of basis coefficients $\boldsymbol{b} \in \mathbb{R}^p$. Coastal DO dynamics are driven by forcing processes operating across disparate timescales, ranging from synoptic weather events ($\sim$days) to seasonal accumulation ($\sim$months). To capture this multi-scale structure, we employ parallel causal convolutional pathways with different kernel sizes.

Multi-scale causal convolutions with dilation. We employ S parallel branches with kernel sizes κ, each consisting of a causal convolution followed by two residual causal blocks with increasing dilation:

$$\boldsymbol{h}^{(s)} = \text{ResCausal}_{d=2}^{(s)} \circ \text{ResCausal}_{d=1}^{(s)} \circ \text{CausalConv}_{\kappa_s}(\tilde{\boldsymbol{f}}_t), \quad s = 1, \ldots, S, \tag{3}$$

where each ResCausal_d block applies two dilated causal convolutions with a residual connection, and $\tilde{\boldsymbol{f}}_t$ is the forcing input augmented with seasonal encoding (below). Each ResCausal block with dilation d adds $2(\kappa_s{-}1)d$ to the receptive field (two convolutions), so the total receptive field is $\kappa_s + 2(\kappa_s{-}1)(1{+}2) = 7\kappa_s - 6$ time steps per scale, reaching up to 50 months for $\kappa_s = 8$.

Seasonal encoding. We append explicit seasonal features $[\sin(2\pi(m{-}1)/12),\, \cos(2\pi(m{-}1)/12)]$ to the forcing channels, where $m \in \{1, \ldots, 12\}$ is the target month. This provides a strong inductive bias for the dominant seasonal cycle of DO.

Temporal attention and projection. The multi-scale features are concatenated along the channel dimension to form $\boldsymbol{h} \in \mathbb{R}^{(S \cdot d_h) \times W}$, normalized via LayerNorm, and processed by multi-head self-attention (Vaswani et al., 2017) with the final time step as query:

$$\boldsymbol{b} = \text{FC}\big(\text{Attn}(\boldsymbol{h}_W, \boldsymbol{h}_{1:W}, \boldsymbol{h}_{1:W})\big) \in \mathbb{R}^p, \tag{4}$$

where FC is a three-layer MLP with GELU activations. The attention mechanism enables the network to adaptively weight different historical time steps, learning which past forcings are most relevant for the current prediction. This is analogous to the time-lag selection in the ASFT module of Yu et al. (2020), but learned end-to-end rather than via greedy forward selection.

3.2 TRUNK NETWORK: MULTI-SCALE FOURIER SPATIAL DECODER

The Trunk Network maps spatial coordinates $\boldsymbol{x} \in \mathbb{R}^{d_s}$ to basis function values $\boldsymbol{\tau} \in \mathbb{R}^p$ and a spatial bias $\beta \in \mathbb{R}$. Standard MLPs are known to suffer from spectral bias (Rahaman et al., 2019): they learn low-frequency components rapidly but struggle to capture fine-grained spatial variations. This is problematic for DO fields, which exhibit both large-scale latitudinal gradients and fine-scale depth structure.

Multi-scale learnable Fourier features. We employ random Fourier features (Rahimi & Recht, 2007; Tancik et al., 2020) at multiple frequency scales. The frequency matrix $\boldsymbol{B} \in \mathbb{R}^{n_\omega \times d_s}$ is initialized from a mixture of Gaussians at scales $\{\sigma_1, \ldots, \sigma_L\}$ and made *learnable*:

$$\gamma(\boldsymbol{x}) = \big[\sin(2\pi \boldsymbol{B}\boldsymbol{x}),\ \cos(2\pi \boldsymbol{B}\boldsymbol{x}),\ \boldsymbol{x}\big] \in \mathbb{R}^{2n_\omega + d_s}, \tag{5}$$

where each scale contributes n_ω/L frequencies. The low scales capture global spatial trends while high scales resolve fine bathymetric and depth variations. Making $\boldsymbol{B}$ learnable allows the network to adapt its frequency allocation during training.

Residual MLP with LayerNorm. The Fourier features are processed by a residual MLP:

$$\boldsymbol{\tau} = W_{\text{out}} \cdot \text{ResMLP}(\gamma(\boldsymbol{x})), \quad \text{ResMLP}(\boldsymbol{h}) := \boldsymbol{h} + W_2\, \text{GELU}(W_1\, \text{LN}(\boldsymbol{h})), \tag{6}$$

This module employs $D = 3$ blocks. A separate bias network produces $\beta(\boldsymbol{x}) \in \mathbb{R}$ representing the spatial mean structure.

EOF-guided pre-training. Rather than learning the spatial decomposition from scratch, we initialize the Trunk Network by pre-training it to approximate the leading EOF spatial modes extracted from the training data. Given the top-K EOF spatial modes $\boldsymbol{U}_K \in \mathbb{R}^{N_s \times K}$ and the spatial mean $\bar{\boldsymbol{f}} \in \mathbb{R}^{N_s}$ from Eq. (13), we minimize:

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{N_s} \sum_{j=1}^{N_s} \|\boldsymbol{\tau}(\boldsymbol{x}_j)_{1:K} - \boldsymbol{U}_K[j,:]\|^2 + \frac{\lambda_b}{N_s} \sum_{j=1}^{N_s} (\beta(\boldsymbol{x}_j) - \bar{f}_j)^2 \tag{7}$$

for a fixed number of steps before main training begins. This provides a warm start that is empirically close to the optimal spatial decomposition, while the subsequent end-to-end training allows the model to refine and go beyond the linear EOF structure. Proposition 1 (§C) formalizes why this initialization aids optimization: aligning the trunk with the leading-variance EOF subspace reduces the observable trunk mismatch. Under idealized local strong convexity and local output identifiability near an optimum, any initialization with a smaller trunk mismatch yields a smaller local parameter distance and hence requires fewer gradient steps to reach a target accuracy.

3.3 Multi-Head Fusion Mechanism

Standard DeepONet computes a single inner product $\langle \boldsymbol{b}, \boldsymbol{\tau} \rangle$ that treats the p-dimensional basis as a monolithic representation. We propose a *multi-head* factorization with nonlinear post-fusion that enhances expressivity (Theorem 1).

Multi-head inner product. We partition the basis coefficients and functions into H heads of dimension $m = p/H$:

$$s_h(t, \boldsymbol{x}) = \langle \boldsymbol{b}_t^{(h)}, \boldsymbol{\tau}_{\boldsymbol{x}}^{(h)} \rangle = \sum_{k=1}^{m} b_{(h-1)m+k}(\boldsymbol{f}_t) \cdot \tau_{(h-1)m+k}(\boldsymbol{x}), \quad h = 1, \ldots, H. \tag{8}$$

Each head captures a distinct "mode" of spatiotemporal interaction.

Gated aggregation. The per-head signals are combined via learnable softmax gates $\boldsymbol{g} = \text{softmax}(\boldsymbol{w})$ where $\boldsymbol{w} \in \mathbb{R}^H$ is a free parameter:

$$\hat{y}_{\text{base}}(t, \boldsymbol{x}) = \sum_{h=1}^{H} g_h \cdot s_h(t, \boldsymbol{x}) + \beta(\boldsymbol{x}). \tag{9}$$

Post-fusion MLP correction. To move beyond the bilinear regime, a small MLP ϕ produces a nonlinear correction from the per-head signals:

$$\widehat{\text{DO}}(t, \boldsymbol{x}) = \hat{y}_{\text{base}}(t, \boldsymbol{x}) + \phi\big([s_1(t, \boldsymbol{x}), \ldots, s_H(t, \boldsymbol{x}),\ \hat{y}_{\text{base}}(t, \boldsymbol{x})]\big), \tag{10}$$

where $\phi : \mathbb{R}^{H+1} \to \mathbb{R}$. This allows different spatial locations to exhibit qualitatively different temporal response patterns. For instance, surface stations may respond primarily to temperature (captured by one head) while deep stations respond to stratification-mediated transport (captured by another).

3.4 Physics-Guided Regularization

We train PhysiOcean with a composite loss that combines data fidelity with four soft regularization terms derived from common properties of coastal DO fields and summarized in §2.2:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda_b(e)\mathcal{L}_{\text{bound}} + \lambda_s(e)\mathcal{L}_{\text{smooth}} + \lambda_r(e)\mathcal{L}_{\text{strat}} + \lambda_c(e)\mathcal{L}_{\text{temporal}}, \tag{11}$$

where e denotes the training epoch and $\lambda_\cdot(e)$ are epoch-dependent weights following a progressive schedule.

Data loss. Mean squared error on observed DO: $\mathcal{L}_{\text{data}} = \frac{1}{|\mathcal{D}_{\text{batch}}|} \sum_{(t,\boldsymbol{x}) \in \mathcal{D}_{\text{batch}}} (\widehat{\text{DO}}(t, \boldsymbol{x}) - \text{DO}(t, \boldsymbol{x}))^2$.

Boundary loss (P1). Penalizes violations of physical bounds: $\mathcal{L}_{\text{bound}} = \mathbb{E}\big[\text{ReLU}(-\widehat{\text{DO}})^2 + \text{ReLU}(\widehat{\text{DO}} - C_{\text{sat}})^2\big]$.

Smoothness loss (P2). Enforces spatial regularity at collocation points $\{\boldsymbol{x}_c\}$ sampled uniformly in $\mathcal{S}$: $\mathcal{L}_{\text{smooth}} = \mathbb{E}_{\boldsymbol{x}_c}\big[\|\nabla_{\boldsymbol{x}}\widehat{\text{DO}}(t, \boldsymbol{x}_c)\|^2\big]$, computed via automatic differentiation through the trunk network.

Stratification loss (P3). Penalizes bottom-DO exceeding surface-DO during summer: $\mathcal{L}_{\text{strat}} = \mathbb{E}\big[\text{ReLU}(\widehat{\text{DO}}_{\text{bot}} - \widehat{\text{DO}}_{\text{surf}})^2 \cdot \mathbf{1}_{\text{summer}}(m)\big]$.

Temporal continuity loss (P4). Bounds the monthly rate of change: $\mathcal{L}_{\text{temporal}} = \mathbb{E}\big[\text{ReLU}(|\widehat{\text{DO}}(t+1, \boldsymbol{x}) - \widehat{\text{DO}}(t, \boldsymbol{x})| - r_{\max})^2\big]$, where $r_{\max}$ is the maximum plausible monthly DO change.

Progressive scheduling. Inspired by curriculum learning (Bengio et al., 2009), we introduce the physics-guided losses gradually to avoid destabilizing early training:

- **Phase 1** ($e < e_1$): $\mathcal{L} = \mathcal{L}_{\text{data}}$ only, establishing a reasonable data fit.
- **Phase 2** ($e_1 \leq e < e_2$): Linearly ramp up $\mathcal{L}_{\text{bound}}$ and $\mathcal{L}_{\text{smooth}}$.
- **Phase 3** ($e \geq e_2$): Additionally ramp up $\mathcal{L}_{\text{strat}}$ and $\mathcal{L}_{\text{temporal}}$.

This phased approach prevents the physics terms from dominating before the network has learned basic data patterns, and ensures that more complex constraints (stratification, temporal continuity) are imposed only after simpler ones (bounds, smoothness) have been absorbed.

Connection to generalization theory. Theorem 2 provides a PAC-style bound on the squared error whose key complexity term is the Rademacher complexity $\mathfrak{R}_n(\mathcal{F}_{\mathcal{C}})$. The theorem analyzes an idealized class whose members satisfy the relevant bounds and spatial regularity constraints. Because our penalties are soft empirical regularizers, this result provides theoretical intuition for complexity control rather than a guarantee that every trained model strictly satisfies all four constraints.

4 EXPERIMENTS

In this section, we evaluate dissolved oxygen forecasting on a published 32-year monitoring record (Yu et al., 2020). We organize the empirical study around three research questions:

RQ1: What accuracy gains does PhysiOcean achieve over regional decomposition, sequence, recent forecasting, and operator-learning baselines? (Section 4.1)

RQ2: How well does PhysiOcean predict at stations and depth layers excluded from training within the study estuary? (Section 4.2)

RQ3: How much does each architectural and training component contribute to predictive performance? (Section 4.3)

Dataset. We use a 32-year dissolved oxygen monitoring record from Chesapeake Bay, a large and environmentally heterogeneous estuarine system with strong along-channel, vertical, and seasonal gradients (Yu et al., 2020). The dataset spans 1985–2016 across 40 long-term monitoring stations, each with DO profiles interpolated to 20 depth layers, yielding $N_s = 800$ station-depth locations and $N_t = 384$ monthly time steps. Following Yu et al. (2020), we use 1985–2008 ($N_{\text{train}} = 288$ months) for training and 2009–2016 ($N_{\text{test}} = 96$ months) for testing. The model ingests $d_f = 39$ globally aggregated forcing variables including air temperature, river discharge, total nitrogen loads, wind, precipitation, and shortwave radiation. A context window of $W = 24$ months is used. Throughout this paper, spatial generalization refers to prediction at unobserved coordinates within this monitored estuarine system; it does not imply transfer to a geographically distinct coastal system.

Baselines. We compare against the EOF+NN decompose-predict-reconstruct pipeline of Yu et al. (2020), recurrent and attention-based sequence models (RNN, LSTM, GRU, Transformer), four recently reproduced forecasting baselines (Timer-XL (Liu et al., 2025), TimeMixer++ (Wang et al., 2025), OLinear (Yue et al., 2025), and FACT (Wang et al., 2026)), the decomposition-ensemble IVMD-MAGRU of Wang et al. (2024), and an architecture-matched DeepONet. The latter uses the same operator backbone as PhysiOcean but omits physics-guided losses and EOF pre-training. Full configuration details for all methods are given in Appendix D.

Table 1: Test-period performance (2009–2016) on the study-estuary DO prediction task. All metrics are computed over the full spatiotemporal test grid (96×800) and reported as mean $\pm$ std over 5 random seeds. General-purpose time-series forecasting baselines are listed above the horizontal rule; dissolved-oxygen and architecture-specific baselines are listed below it. Best results are in **bold**; second best are underlined. DeepONet refers to the PhysiOcean architecture trained without physics-guided losses or EOF pre-training.

Method	$R^2 \uparrow$	**RMSE (mg/L)** $\downarrow$	**Skill** $\uparrow$
Timer-XL	$0.748_{\pm.004}$	$1.473_{\pm.011}$	$0.924_{\pm.002}$
TimeMixer++	$0.862_{\pm.006}$	$1.090_{\pm.024}$	$0.964_{\pm.002}$
OLinear	$\underline{0.873}_{\pm.001}$	$\underline{1.045}_{\pm.004}$	$\underline{0.965}_{\pm.001}$
FACT	$0.865_{\pm.001}$	$1.078_{\pm.006}$	$0.964_{\pm.001}$
RNN	$0.804_{\pm.007}$	$1.297_{\pm.024}$	$0.945_{\pm.001}$
LSTM	$0.801_{\pm.010}$	$1.306_{\pm.034}$	$0.944_{\pm.002}$
GRU	$0.821_{\pm.009}$	$1.240_{\pm.031}$	$0.950_{\pm.002}$
Transformer	$0.828_{\pm.011}$	$1.216_{\pm.031}$	$0.954_{\pm.003}$
IVMD-MAGRU	$0.090_{\pm.066}$	$2.893_{\pm.162}$	$0.636_{\pm.023}$
EOF+NN	$0.848_{\pm.005}$	$1.146_{\pm.018}$	$0.958_{\pm.001}$
DeepONet	$0.861_{\pm.005}$	$1.093_{\pm.022}$	$0.962_{\pm.002}$
PhysiOcean (ours)	$\mathbf{0.881}_{\pm.004}$	$\mathbf{1.011}_{\pm.015}$	$\mathbf{0.967}_{\pm.001}$

Spatial coordinates. Each station-depth location is described by 5 features: normalized latitude, longitude, depth (0 = deepest, 1 = surface), bathymetry (maximum station depth), and distance to the estuary's main axis, all scaled to $[0, 1]$. The main axis direction is set to longitude.

Metrics. We report three standard metrics on the held-out test period: (1) $R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$, coefficient of determination, (2) RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, root mean squared error, measured in mg/L, and (3) Willmott Skill (Willmott, 1981), defined as Skill $= 1 - \frac{\sum(\hat{y}-y)^2}{\sum(|\hat{y}-\bar{y}|+|y-\bar{y}|)^2}$, a robust measure combining correlation and bias.

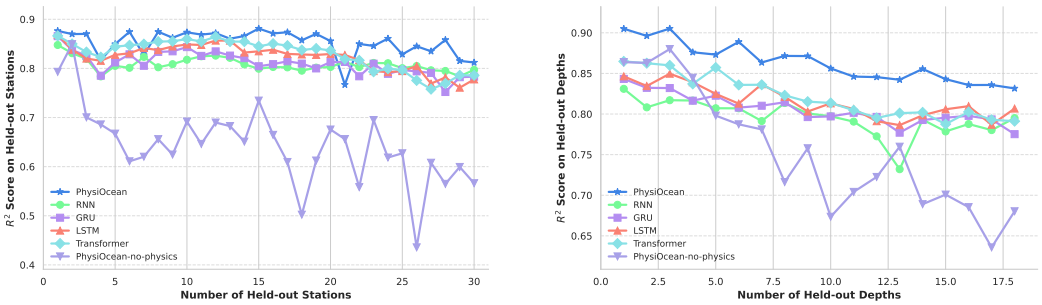
Input information. PhysiOcean and the original forcing-based baselines use global environmental forcing and spatial coordinates. Timer-XL, TimeMixer++, OLinear, and FACT additionally use the observed DO history at each target location, following their autoregressive forecasting interfaces. Their results are included as practically relevant forecasting comparisons, while spatial generalization claims are based on the coordinate-conditioned experiments using the prescribed spatial hold-out protocol.

4.1 Main Results

Table 1 summarizes test-period performance on the study-estuary task. PhysiOcean obtains $R^2 = 0.881$, RMSE $= 1.011$ mg/L, and Skill $= 0.967$. It outperforms the architecture-matched DeepONet, which omits physics-guided losses and EOF pre-training, by $+0.020$ in R^2 and 7.5% in RMSE; Table 2 examines these training enhancements separately. OLinear is the strongest newly reproduced forecaster and ranks second overall, with $R^2 = 0.873 \pm 0.001$ and RMSE $= 1.045 \pm 0.004$ mg/L. PhysiOcean improves over OLinear by 0.008 in R^2 and reduces RMSE by 3.3%, although OLinear additionally observes the target-location DO history. All four reproduced results are averaged over the same five random seeds.

Compared to the EOF+NN baseline, which similarly employs a decomposition approach, PhysiOcean achieves an R^2 improvement of 0.033. This gap is notable because EOF+NN already exploits the dominant low-rank spatial structure through explicit EOF decomposition; PhysiOcean recovers and surpasses this structure via end-to-end learning.

Among the original sequence-model baselines, RNN, LSTM, GRU, and Transformer cluster in the range $R^2 \in [0.801, 0.828]$, all below EOF+NN. The reproduced modern forecasters are stronger: OLinear, FACT, and TimeMixer++ reach R^2 values of 0.873 ± 0.001, 0.865 ± 0.001, and $0.862 \pm$



(a) Station hold-out (horizontal).

(b) Depth hold-out (vertical).

Figure 2: **(a)** R^2 on held-out stations as the number of withheld stations increases from 1 to 30. **(b)** R^2 on held-out depth layers as the number of withheld layers increases from 1 to 18, following the inside-out removal order.

0.006, respectively, whereas Timer-XL reaches 0.748 ± 0.004. FACT and OLinear also show low variation across seeds, while the mean R^2 of TimeMixer++ differs from DeepONet by only 0.001. Because these methods use additional observed-DO history, they provide strong forecasting comparisons but do not isolate continuous coordinate-conditioned spatial modeling. The IVMD-MAGRU method of Wang et al. (2024), originally designed for single-station prediction via per-station VMD decomposition, fails when extended to the full 800-point spatial grid ($R^2 = 0.090$): the global VMD optimization required to handle all stations simultaneously does not preserve the per-station mode structure that the method relies on.

4.2 Generalization to Unobserved Locations within the Study Estuary

We evaluate coordinate-level generalization within the study estuary along two axes. **Horizontal (station hold-out):** we randomly withhold $N_{s_h} \in [1, 30]$ of the 40 stations from training and test on the held-out set. **Vertical (depth hold-out):** we adopt an "inside-out" strategy, progressively removing depth layers starting from the median depth and expanding toward the surface and bottom, until only boundary layers remain. These experiments measure within-estuary transfer to unobserved monitoring locations and depth layers rather than cross-estuary transfer. All experiments use $N_t - W = 360$ evaluation time steps; IVMD-MAGRU is omitted due to its poor base performance.

Figure 2(a) shows that all methods degrade as more stations are withheld, but PhysiOcean degrades more slowly than the baselines, and the gap widens with increasing hold-out severity. Figure 2(b) reveals a more pronounced advantage in the vertical direction: PhysiOcean retains high R^2 even when most intermediate depth layers are removed, whereas the no-physics variant deteriorates rapidly. This behavior is consistent with the spatial regularity encouraged by the physics-guided objective, although the comparison also includes the removal of EOF pre-training.

Figure 3 examines the more extreme setting of training on only the surface, middle, or bottom depth category. A single intermediate-depth observation is more representative of the full water column than either boundary layer, and PhysiOcean exhibits lower cross-station variance than the baselines across all three settings.

4.3 Ablation Studies

Among the architectural components, reducing the branch network to a single convolution scale causes the steepest decline ($\Delta R^2 = -0.089$), confirming the importance of multi-scale temporal encoding for capturing forcing effects across disparate timescales (§3.1). In the trunk network, removing Fourier features ($\Delta R^2 = -0.047$) and EOF pre-training ($\Delta R^2 = -0.040$) each cause meaningful degradation: the former confirms that plain MLPs struggle with fine-grained depth structure due to spectral bias (Rahaman et al., 2019), while the latter validates the convergence advantage of EOF-guided initialization (Proposition 1). Reducing the basis dimensionality from $p = 64$ to $p = 16$ or $p = 32$ costs $\Delta R^2 = -0.075$ and -0.086 respectively, indicating that the spatiotemporal field requires sufficient representational capacity.

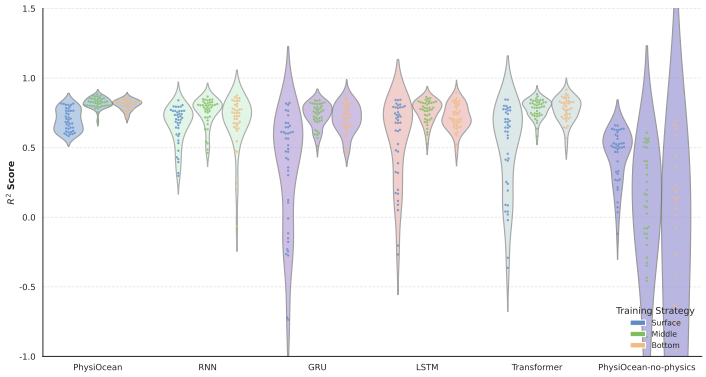


Figure 3: Per-station R^2 when training uses only surface, middle, or bottom depth layers. Violin width indicates density across stations.

Table 2: Ablation studies on PhysiOcean components. Each row removes or modifies one component from the full model.

Configuration	$R^2 \uparrow$	**RMSE** $\downarrow$	**Skill** $\uparrow$
PhysiOcean (full)	**0.8814**	**1.0107**	**0.9672**
No Fourier encoding	0.8349	1.1927	0.9552
Single-scale conv ($\kappa = 4$)	0.7921	1.3385	0.9445
$p = 16$	0.8066	1.2908	0.9464
$p = 32$	0.7956	1.3272	0.9433
No EOF pre-training	0.8412	1.1694	0.9562
Single head ($H = 1$, no fusion MLP)	0.8523	1.1284	0.9589
$H = 4$, no fusion MLP	0.8631	1.0872	0.9621
No physics loss	0.8674	1.0688	0.9639

The fusion mechanism contributes incrementally: a single-head model without fusion MLP yields $R^2 = 0.852$; adding multi-head decomposition ($H = 4$, no MLP) raises this to 0.863; including the nonlinear correction MLP reaches the full-model $R^2 = 0.881$. Both the multi-head factorization and the post-fusion nonlinearity contribute, consistent with the expressivity result of Theorem 1.

Although the individual contributions of physics-guided losses and EOF pre-training are relatively modest when training on the full spatial grid, Figure 2 illustrates that PhysiOcean maintains markedly superior within-estuary performance compared to the DeepONet baseline (which lacks both enhancements). This result suggests that the two training enhancements are particularly useful when spatial training coverage is reduced, while the ablation does not isolate the contribution of each individual physical loss.

Additional depth- and season-specific results and physical-consistency diagnostics are reported in Appendix E. Appendix F further analyzes temporal attention and learned spatial bases.

5 Conclusion

PhysiOcean combines a multi-scale temporal branch, continuous Fourier spatial trunk, and multi-head fusion mechanism to forecast coastal dissolved oxygen over continuous spatial coordinates. On a 32-year monitoring record, it achieves $R^2 = 0.881$ and RMSE $= 1.011$ mg/L, outperforming the architecture-matched DeepONet and all reproduced forecasting baselines, including OLinear, the strongest autoregressive comparison. Ablations and spatial hold-out experiments support the contributions of the architecture, EOF initialization, and physics-guided training. Evaluation across geographically distinct coastal systems remains an important direction.

AI USE STATEMENT

Generative AI tools were used to assist with software implementation and adaptation of baseline methods, manuscript restructuring and language editing, reference formatting, and plotting utilities. The authors reviewed all AI-assisted code and text, executed the experiments, and checked the reported numerical results and citations. The authors take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The model definition and training objective are provided in Sections 3–3.4; dataset splits, baselines, implementation details, architecture settings, and training hyperparameters are documented in Section 4 and Appendix D. Complete proofs are included in Appendix C, and anonymous code is linked in the abstract.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *ArXiv*, abs/1803.01271, 2018. URL https://api.semanticscholar.org/CorpusID:4747877.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Denise Breitburg, Lisa A Levin, Andreas Oschlies, Marilaure Grégoire, Francisco P Chavez, Daniel J Conley, Véronique Garçon, Denis Gilbert, Dimitri Gutiérrez, Kirsten Isensee, et al. Declining oxygen in the global ocean and coastal waters. *Science*, 359(6371):eaam7240, 2018.

Jacob Carstensen, Jesper H Andersen, Bo G Gustafsson, and Daniel J Conley. Deoxygenation of the baltic sea during the last century. *Proceedings of the National Academy of Sciences*, 111(15): 5628–5633, 2014.

Carl F Cerco and Mark R Noel. Twenty-one-year simulation of chesapeake bay water quality using the ce-qual-icm eutrophication model. *JAWRA Journal of the American Water Resources Association*, 49(5):1119–1133, 2013.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE transactions on neural networks*, 6(4):911–917, 1995.

Robert J Diaz and Rutger Rosenberg. Spreading dead zones and consequences for marine ecosystems. *science*, 321(5891):926–929, 2008.

Jiabi Du, Jian Shen, Kyeong Park, Ya Ping Wang, and Xin Yu. Worsened physical condition due to climate change contributes to the increasing hypoxia in chesapeake bay. *Science of the Total Environment*, 630:707–717, 2018.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Katja Fennel and Jeremy M Testa. Biogeochemical controls on coastal hypoxia. *Annual review of marine science*, 11(1):105–130, 2019.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Isaac D Irby, Marjorie AM Friedrichs, Carl T Friedrichs, Aaron J Bever, Raleigh R Hood, Lyon WJ Lanerolle, Ming Li, Lewis Linker, Malcolm E Scully, Kevin Sellner, et al. Challenges associated with modeling low-oxygen waters in chesapeake bay: a multiple model comparison. *Biogeosciences*, 13(7):2011–2028, 2016.

W Michael Kemp, Walter R Boynton, Jason E Adolf, Donald F Boesch, William C Boicourt, Grace Brush, Jeffrey C Cornwell, Thomas R Fisher, Patricia M Glibert, Jim D Hagy, et al. Eutrophication of chesapeake bay: historical trends and ecological interactions. *Marine ecology progress series*, 303:1–29, 2005.

Wuyan Li, Hao Wu, Nanyang Zhu, Yongnian Jiang, Jinglu Tan, and Ya Guo. Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). *Information Processing in Agriculture*, 8(1):185–193, March 2021. ISSN 2214-3173. doi: 10.1016/j.inpa.2020.02.002.

Wenzhao Liang, Tongcun Liu, Yuntao Wang, Jiu Jimmy Jiao, Jianping Gan, and Ding He. Spatiotemporal-aware machine learning approaches for dissolved oxygen prediction in coastal waters. *Science of The Total Environment*, 905:167138, December 2023. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2023.167138.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KMCJXjlDDr.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Rebecca R Murphy, W Michael Kemp, and William P Ball. Long-term trends in chesapeake bay seasonal hypoxia, stratification, and nutrient loading. *Estuaries and Coasts*, 34(6):1293–1309, 2011.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Jeremy M Testa, Yun Li, Younjoo J Lee, Ming Li, Damian C Brady, Dominic M Di Toro, W Michael Kemp, and James J Fitzpatrick. Quantifying the effects of nutrient loading on dissolved o2 cycling and hypoxia in chesapeake bay using a coupled hydrodynamic–biogeochemical model. *Journal of Marine Systems*, 139:139–158, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Huiqiang Wang, Jieming Shi, and Qing Li. FACT: Fine-grained across-variable convolution for multivariate time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j3gNYqrHtl.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1CLzLXSFNn.

Zhaocai Wang, Qingyu Wang, Zhixiang Liu, and Tunhua Wu. A deep learning interpretable model for river dissolved oxygen multi-step and interval prediction based on multi-source data fusion. *Journal of Hydrology*, 629:130637, February 2024. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2024.130637.

Ray F Weiss. The solubility of nitrogen, oxygen and argon in water and seawater. In *Deep sea research and oceanographic abstracts*, volume 17, pp. 721–735. Elsevier, 1970.

Cort J Willmott. On the validation of models. *Physical geography*, 2(2):184–194, 1981.

Xin Yu, Jian Shen, and Jiabi Du. A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in chesapeake bay as an example. *Water Resources Research*, 56(9):e2020WR027227, 2020.

Wenzhen Yue, Yong Liu, Hao Wang, Haoxuan Li, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. Olinear: A linear model for time series forecasting in orthogonally transformed domain. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 60573–60632. Curran Associates, Inc., 2025. doi: 10.52202/085713-2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/571c7e164fb1ffbcf2f84a63784451ec-Paper-Conference.pdf.

APPENDIX

A RELATED WORK

Physics-informed neural networks. Physics-informed neural networks (PINNs) (Raissi et al., 2019) embed PDE residuals directly into the training loss, enabling data-efficient learning of physical systems. While PINNs have achieved success in fluid dynamics and heat transfer, they require explicit knowledge of the governing PDE. For estuarine DO, this involves coupled hydrodynamic, biogeochemical, and sediment processes that are prohibitively complex to specify analytically. PhysiOcean instead incorporates *soft* physical constraints (bounds, smoothness, stratification priors) that hold universally without requiring the full PDE specification.

Operator learning. DeepONet (Lu et al., 2021), grounded in the universal approximation theorem for operators (Chen & Chen, 1995), learns mappings between infinite-dimensional function spaces via a branch-trunk factorization. Recent extensions include multi-output DeepONet and physics-informed variants. PhysiOcean extends the standard DeepONet with multi-scale domain-specific encoders for both branch and trunk, a multi-head fusion mechanism that exceeds the expressivity of the bilinear form, and domain-specific physics-guided training with progressive scheduling.

Data-driven ocean and water quality modeling. Machine learning approaches to ocean prediction range from station-level models to spatial decomposition frameworks. Yu et al. (2020) introduced an EOF+NN strategy that combines Empirical Orthogonal Functions with per-mode neural networks, while Wang et al. (2024) further enhanced this paradigm via the IVMD-MAGRU framework, which utilizes Variational Mode Decomposition (VMD) and an improved Sparrow Search Algorithm (SSA) for parameter optimization. However, these decomposition-based methods remain inherently constrained: EOF+NN fixes spatial modes at training time, and the computational complexity of IVMD-MAGRU limits its applicability to individual-spatial (single-spatial) prediction. Liang et al. (2023) incorporated historical data and neighboring observation data as model inputs. Yet, this approach relies heavily on local observations and requires training a separate model for each spatial coordinate, hindering its scalability. In contrast, *PhysiOcean* is, to our knowledge, the first operator learning approach for coastal DO prediction that learns spatiotemporal representations end-to-end, enabling robust generalization to unobserved locations without the need for per-site retraining.

Multi-scale temporal modeling. Temporal convolutional networks (TCNs) with dilated causal convolutions (Bai et al., 2018) have proven effective for sequence modeling, offering advantages in parallelization and stable gradients over recurrent architectures. Our Branch Network combines multi-scale dilated causal convolutions with self-attention (Vaswani et al., 2017), tailored to capture the disparate timescales, from synoptic weather events to seasonal nutrient accumulation, that characterize estuarine forcing–response dynamics.

B PRELIMINARIES

B.1 DEEP OPERATOR NETWORKS

The Deep Operator Network (DeepONet) (Lu et al., 2021) provides a principled framework for learning nonlinear operators between infinite-dimensional function spaces. Building on the universal approximation theorem for operators (Chen & Chen, 1995), DeepONet approximates a continuous operator $\mathcal{G} : \mathcal{U} \to \mathcal{V}$ via a finite-rank decomposition:

$$\mathcal{G}(\boldsymbol{u})(\boldsymbol{y}) \approx \sum_{k=1}^{p} b_k(\boldsymbol{u}) \cdot \tau_k(\boldsymbol{y}) + \beta(\boldsymbol{y}), \tag{12}$$

where the *branch network* $\{b_k\}_{k=1}^{p}$ encodes the input function $\boldsymbol{u}$ into p basis coefficients, the *trunk network* $\{\tau_k\}_{k=1}^{p}$ maps evaluation coordinates $\boldsymbol{y}$ to p basis function values, and $\beta(\boldsymbol{y})$ is a spatial bias. This factored structure is natural for our setting: the branch encodes temporal forcing dynamics while the trunk encodes spatial structure, and their inner product produces spatiotemporal predictions.

Remark 1 *The standard DeepONet in Eq. (12) is a rank-p bilinear form between the temporal and spatial representations. While the universal approximation theorem guarantees that arbitrarily accurate approximation is possible as $p \to \infty$, in practice the rank p must remain moderate for computational tractability. This motivates our multi-head extension in §3.3 that goes beyond the bilinear form.*

B.2 Empirical Orthogonal Functions

Empirical Orthogonal Functions (EOF), also known as Principal Component Analysis in the geosciences, decompose a spatiotemporal data matrix $\boldsymbol{F} \in \mathbb{R}^{N_s \times N_t}$ (with N_s spatial points and N_t time steps) via the singular value decomposition:

$$\boldsymbol{F} - \bar{\boldsymbol{f}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^\top, \tag{13}$$

where $\bar{\boldsymbol{f}}$ is the temporal mean, $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]$ contains the spatial modes (maps), $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]$ contains the temporal modes, and $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are the singular values. For Chesapeake Bay DO, the first singular value alone accounts for $\sim$87% of the total variance (Yu et al., 2020).

The EOF+NN approach of Yu et al. (2020) exploits this low-rank structure by: (1) extracting the top-K spatial modes and principal components ($PC_k = \sigma_k \boldsymbol{v}_k$) from training data, (2) independently training a neural network to predict each PC from transformed forcing variables, and (3) reconstructing the full DO field as $\widehat{\mathrm{DO}}(\boldsymbol{x}, t) = \bar{f}(\boldsymbol{x}) + \sum_{k=1}^{K} \widehat{\mathrm{PC}}_k(t) \cdot \mathrm{Map}_k(\boldsymbol{x})$. While effective, this three-stage pipeline fixes the spatial modes at training time and models each mode in isolation. PhysiOcean instead learns the spatial decomposition end-to-end, uses EOF modes only as initialization, and captures cross-modal interactions through a unified architecture.

C Theoretical Analysis

C.1 Expressivity of Multi-Head DeepONet

The standard DeepONet in Eq. (12) produces predictions via a rank-p bilinear form. We show that the multi-head extension with fusion MLP (Eq. (10)) is strictly more expressive for any basis rank $p > 4$ satisfying the head split in Theorem 1.

Theorem 1 (Multi-Head Expressivity) *Let $\mathcal{G}_p^{std}$ be the class of operators representable by a standard DeepONet with p basis functions. Let $\mathcal{G}_{p,H,m}^{MH}$ be the class representable by a multi-head DeepONet with H heads (p/H dimensions each) and a fusion MLP of width m and depth 2. Then for any $H \geq 2$, any $p > 4$ with $p/H \geq 2$, and any $m \geq 3$:*

$$\mathcal{G}_p^{std} \subsetneq \mathcal{G}_{p,H,m}^{MH}. \tag{14}$$

Furthermore, there exist operators in $\mathcal{G}_{p,H,m}^{MH}$ that require rank strictly greater than p to be represented by a standard DeepONet. In particular, for $H = 2$, the rank gap is at least $(p/2)^2 - p > 0$. (When $p = 4$, the rank-counting construction has effective rank p; see Appendix C.5.)

Proof (Proof sketch) When the fusion MLP outputs zero, the multi-head DeepONet reduces to a gated sum of rank-(p/H) bilinear forms, which spans $\mathcal{G}_p^{\mathrm{std}}$. For strict inclusion when $p > 4$ and $H = 2$, the product of two head bilinear forms expands to rank $(p/2)^2 > p$. The fusion MLP can moreover approximate continuous maps of the per-head signals (two-layer universal approximation), which need not lie in any fixed rank-p bilinear algebraic span. See Appendix C.5 for details.

C.2 Theoretical Intuition for Physics-Guided Regularization

We formalize the intuition that an idealized constrained hypothesis class can have lower complexity than the corresponding unconstrained class. This analysis motivates the regularizers but does not replace the empirical ablations.

Theorem 2 (Physics-Constrained Generalization) *Let $\mathcal{F}$ be the hypothesis class of a DeepONet with parameters bounded as $\|\theta\| \le B$, and let $\mathcal{F}_\mathcal{C} \subseteq \mathcal{F}$ be the subset satisfying physics constraints $\mathcal{C} = \{\text{(P1)–(P4)}\}$. Suppose (P1) bounds outputs as $|f| \le C_{sat}$ and (P2) imposes $\|\nabla_{\boldsymbol{x}} f(\boldsymbol{f}_t, \boldsymbol{x})\| \le L$ for all $f \in \mathcal{F}_\mathcal{C}$. Then for any $\delta > 0$, with probability at least $1-\delta$ over a training set of size n, every $f \in \mathcal{F}_\mathcal{C}$ satisfies:*

$$\mathbb{E}[(\hat{y} - y)^2] \le \hat{\mathcal{L}}_{data}(f) + 4C_{sat} \cdot \mathfrak{R}_n(\mathcal{F}_\mathcal{C}) + C_{sat}^2 \sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{15}$$

where $\hat{\mathcal{L}}_{data}$ is the empirical loss. Since $\mathcal{F}_\mathcal{C} \subseteq \mathcal{F}$, its Rademacher complexity satisfies

$$\mathfrak{R}_n(\mathcal{F}_\mathcal{C}) \le \mathfrak{R}_n(\mathcal{F}). \tag{16}$$

Moreover, when $d_s > 2$ (as in our Chesapeake Bay setting, where $d_s = 5$), for any fixed temporal input $\bar{\boldsymbol{f}}$, define the spatial slice class $\mathcal{F}_\mathcal{C}|_{\bar{\boldsymbol{f}}} = \{\boldsymbol{x} \mapsto f(\bar{\boldsymbol{f}}, \boldsymbol{x}) : f \in \mathcal{F}_\mathcal{C}\}$. Its empirical Rademacher complexity over n spatial samples satisfies

$$\mathfrak{R}_n(\mathcal{F}_\mathcal{C}|_{\bar{\boldsymbol{f}}}) \le C_{d_s} \left(C_{sat} + L \cdot diam(\mathcal{S})\right) n^{-1/d_s} \sqrt{\log n}, \tag{17}$$

for a constant C_{d_s} depending only on d_s and all $n \ge 2$.

The full spatiotemporal bound is controlled by $\mathfrak{R}_n(\mathcal{F}_\mathcal{C})$ and hence by the unconstrained class through Eq. (16). The Lipschitz rate in Eq. (17) applies only to spatial slices with fixed temporal forcing; it formalizes how (P1)–(P2) reduce the nonparametric spatial complexity, but it is not by itself an upper bound on the full class over both $(\boldsymbol{f}_t, \boldsymbol{x})$ without additional regularity in $\boldsymbol{f}_t$.

C.3 Convergence Advantage of EOF Initialization

Proposition 1 (EOF Warm-Start Convergence) *Let $\boldsymbol{U}_K$ be the top-K EOF spatial modes used for pre-training, and let $\boldsymbol{\tau}^*$ be a locally optimal trunk representation on the training spatial grid. After EOF pre-training (Eq. 7), the initialized trunk satisfies*

$$\frac{1}{N_s} \sum_{j=1}^{N_s} \left\| \boldsymbol{\tau}(\boldsymbol{x}_j; \theta_T^{(0)})_{1:K} - \boldsymbol{U}_K[j,:] \right\|^2 \le \epsilon_{opt}, \tag{18}$$

where ϵ_{opt} is the pre-training optimization error. Define the full-grid trunk mismatch

$$\Delta_T^2 := \frac{1}{N_s} \sum_{j=1}^{N_s} \left\| \boldsymbol{\tau}(\boldsymbol{x}_j; \theta_T^{(0)}) - \boldsymbol{\tau}^*(\boldsymbol{x}_j) \right\|^2. \tag{19}$$

Under Assumptions 1 and 2 below, if $\theta_T^{(0)}$ lies in the local basin $\mathcal{B}(\theta^*, r)$, then gradient descent with condition number $\kappa = \ell/\mu$ satisfies

$$\|\theta_T^{(t)} - \theta^*\| \le \left(1 - \tfrac{1}{\kappa}\right)^t \|\theta_T^{(0)} - \theta^*\| \le \left(1 - \tfrac{1}{\kappa}\right)^t \sigma_{\min}^{-1} \Delta_T, \tag{20}$$

where parameter distances are taken after quotienting redundant trunk directions as in Assumption 2. Consequently, reaching $\|\theta_T^{(t)} - \theta^*\| \le \epsilon$ requires at most $O\big(\kappa \log(\sigma_{\min}^{-1} \Delta_T / \epsilon)\big)$ iterations. Thus, any EOF warm-start that yields a smaller Δ_T than random initialization gives the corresponding logarithmic reduction in local convergence time.

Assumption 1 (Local strong convexity) *There exists a neighborhood $\mathcal{B}(\theta^*, r)$ of the optimum such that the loss $\mathcal{L}(\theta)$ is μ-strongly convex and ℓ-smooth in $\mathcal{B}(\theta^*, r)$.*

Assumption 2 (Local output identifiability) *Let $\mathcal{T}(\theta_T) = \{\boldsymbol{\tau}(\boldsymbol{x}_j; \theta_T)\}_{j=1}^{N_s}$ denote trunk outputs on the training spatial grid. After quotienting grid-invariant directions, there is a local inverse-Lipschitz constant $\sigma_{\min} > 0$ such that every $\theta_T \in \mathcal{B}(\theta^*, r)$ satisfies*

$$\|\theta_T - \theta^*\| \le \sigma_{\min}^{-1} \left(\frac{1}{N_s} \sum_{j=1}^{N_s} \|\boldsymbol{\tau}(\boldsymbol{x}_j; \theta_T) - \boldsymbol{\tau}^*(\boldsymbol{x}_j)\|^2 \right)^{1/2}. \tag{21}$$

The proposition formalizes a conditional local optimization statement. EOF pre-training directly controls the first K trunk components, which are the leading-variance spatial modes; the full mismatch Δ_T also depends on the remaining trunk components and the chosen local optimum. If this mismatch is smaller than that from a random initialization and the iterate is already in a locally well-conditioned basin, the standard linear-convergence bound gives fewer iterations. We note that Assumptions 1 and 2 are strong local assumptions and do not claim global convergence; our empirical results in §4.3 confirm the practical benefit of EOF initialization.

C.4 Formal Definitions and Assumptions

We first collect the notation and formal assumptions used throughout the paper.

Notation. Let $\mathcal{S} \subset \mathbb{R}^{d_s}$ denote the spatial domain and $\mathcal{T} = \{1, \dots, N_t\}$ the set of discrete time indices. We write $\boldsymbol{f}_t \in \mathbb{R}^{d_f \times W}$ for the forcing window at time t, $\boldsymbol{x} \in \mathcal{S}$ for a spatial coordinate, and $y_{t,\boldsymbol{x}} = \mathrm{DO}(\boldsymbol{x}, t)$ for the observed dissolved oxygen. The model prediction is $\hat{y}_{t,\boldsymbol{x}} = \mathcal{G}_\theta(\boldsymbol{f}_t, \boldsymbol{x})$ parameterized by θ. Branch output: $\boldsymbol{b} = B_\theta(\boldsymbol{f}_t) \in \mathbb{R}^p$. Trunk output: $\boldsymbol{\tau} = T_\theta(\boldsymbol{x}) \in \mathbb{R}^p$, $\beta = \beta_\theta(\boldsymbol{x}) \in \mathbb{R}$.

Assumption 3 (Bounded parameters) *The model parameters satisfy $\|\theta\| \le B_\theta$ for some $B_\theta > 0$. The branch and trunk outputs are bounded: $\|\boldsymbol{b}\| \le B_b$, $\|\boldsymbol{\tau}\| \le B_\tau$, $|\beta| \le B_\beta$.*

Assumption 4 (Spatial regularity) *The spatial domain $\mathcal{S}$ is a compact, connected subset of $\mathbb{R}^{d_s}$ with Lipschitz boundary. The trunk network T_θ is continuously differentiable with respect to $\boldsymbol{x}$.*

Assumption 5 (Physical constraints) *The hypothesis class $\mathcal{F}_\mathcal{C}$ satisfies the following constraints for all $f \in \mathcal{F}_\mathcal{C}$:*

- *(P1)* $0 \le f(\boldsymbol{f}_t, \boldsymbol{x}) \le C_{sat}$ *for all $(t, \boldsymbol{x})$, where $C_{sat} > 0$ is the saturation concentration.*
- *(P2)* $\|\nabla_{\boldsymbol{x}} f(\boldsymbol{f}_t, \boldsymbol{x})\| \le L$ *for all $(t, \boldsymbol{x})$ and some Lipschitz constant $L > 0$.*
- *(P3) For summer months $m \in \{5, \dots, 9\}$: $f(\boldsymbol{f}_t, \boldsymbol{x}_{surf}) \ge f(\boldsymbol{f}_t, \boldsymbol{x}_{bot})$ for paired surface/bottom coordinates.*
- *(P4)* $|f(\boldsymbol{f}_{t+1}, \boldsymbol{x}) - f(\boldsymbol{f}_t, \boldsymbol{x})| \le r_{max}$ *for consecutive months.*

Note that Assumptions 1 (local strong convexity) and 2 (local output identifiability) are stated in the main text alongside Proposition 1. These assumptions are used only in the convergence analysis of EOF-guided initialization and are not required for the expressivity or generalization results.

C.5 Proof of Theorem 1 (Multi-Head Expressivity)

Proof We proceed in two parts: first showing inclusion ($\mathcal{G}_p^{\text{std}} \subseteq \mathcal{G}_{p,H,m}^{\text{MH}}$), then showing strict inclusion.

Part 1: $\mathcal{G}_p^{\text{std}} \subseteq \mathcal{G}_{p,H,m}^{\text{MH}}$. Consider any operator $G \in \mathcal{G}_p^{\text{std}}$ with representation $G(\boldsymbol{u})(\boldsymbol{y}) = \sum_{k=1}^p b_k(\boldsymbol{u})\tau_k(\boldsymbol{y}) + \beta(\boldsymbol{y})$. In the multi-head DeepONet, set the gate weights $g_h = 1/H$ for all h (achievable by $\boldsymbol{w} = \boldsymbol{0}$), and set the fusion MLP to output $\phi \equiv 0$ (achievable since the MLP has bias terms that can be set to zero). Then:

$$\hat{y} = \sum_{h=1}^{H} \frac{1}{H} \sum_{k=1}^{p/H} b_{(h-1)p/H+k} \tau_{(h-1)p/H+k} + \beta = \frac{1}{H} \sum_{k=1}^{p} b_k \tau_k + \beta.$$

The factor $1/H$ can be absorbed into the branch outputs by scaling $b_k \to H \cdot b_k$, recovering the standard DeepONet exactly.

Part 2: Strict inclusion (rank argument for $p > 4$). Let $H = 2$ and $d = p/2$. Consider per-head bilinear scores

$$s_1(\boldsymbol{u}, \boldsymbol{y}) = \sum_{k=1}^{d} b_k^{(1)}(\boldsymbol{u})\, \tau_k^{(1)}(\boldsymbol{y}), \qquad s_2(\boldsymbol{u}, \boldsymbol{y}) = \sum_{k=1}^{d} b_k^{(2)}(\boldsymbol{u})\, \tau_k^{(2)}(\boldsymbol{y}),$$

and the operator $G^*(\boldsymbol{u})(\boldsymbol{y}) = s_1(\boldsymbol{u}, \boldsymbol{y})\, s_2(\boldsymbol{u}, \boldsymbol{y})$. This is realized in $\mathcal{G}^{\mathrm{MH}}_{p,2,m}$ by taking $\hat{y}_{\text{base}} \equiv 0$ and letting ϕ implement multiplication on the compact range of $[s_1, s_2, \hat{y}_{\text{base}}]$: two-layer nets with sufficiently wide nonlinear hidden layers uniformly approximate continuous maps on compact sets (Hornik et al., 1989). Expanding yields

$$G^*(\boldsymbol{u})(\boldsymbol{y}) = \sum_{j=1}^{d} \sum_{k=1}^{d} b_j^{(1)}(\boldsymbol{u})\, b_k^{(2)}(\boldsymbol{u}) \cdot \tau_j^{(1)}(\boldsymbol{y})\, \tau_k^{(2)}(\boldsymbol{y}),$$

a canonical rank-d^2 separable decomposition in the CP/tensor sense whenever the d^2 rank-one tensors are linearly independent, whereas a rank-p DeepONet allows at most p such terms simultaneously. Hence, when $d^2 > p$, i.e. $p > 4$ under $H = 2$, generic choices yield $G^* \notin \mathcal{G}^{\text{std}}_p$. For $p = 4$, $d^2 = p$ and counting rank alone does not separate the classes, consistent with Theorem 1.

C.6 Proof of Theorem 2 (Physics-Constrained Generalization)

Proof We establish the full generalization bound and then the separate spatial-slice complexity bound.

Step 1: Standard generalization bound. Consider the squared loss $\ell(f, y) = (f - y)^2$, which is bounded in $[0, C_{\text{sat}}^2]$ for $f \in \mathcal{F}_{\mathcal{C}}$ (since constraint (P1) gives $|f| \le C_{\text{sat}}$ and we assume $|y| \le C_{\text{sat}}$). By the standard Rademacher generalization bound (Bartlett & Mendelson, 2002), for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample of size n:

$$\sup_{f \in \mathcal{F}_{\mathcal{C}}} \left| R(f) - \hat{R}(f) \right| \le 2\mathfrak{R}_n(\ell \circ \mathcal{F}_{\mathcal{C}}) + C_{\text{sat}}^2 \sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{22}$$

where $R(f) = \mathbb{E}[(f(\boldsymbol{f}_t, \boldsymbol{x}) - y)^2]$ and $\hat{R}(f) = \hat{\mathcal{L}}_{\text{data}}(f)$ is the empirical risk. The squared loss $z \mapsto z^2$ restricted to $[-C_{\text{sat}}, C_{\text{sat}}]$ is $2C_{\text{sat}}$-Lipschitz, so by the contraction principle (Ledoux–Talagrand), $\mathfrak{R}_n(\ell \circ \mathcal{F}_{\mathcal{C}}) \le 2C_{\text{sat}} \cdot \mathfrak{R}_n(\mathcal{F}_{\mathcal{C}})$. Substituting gives the stated bound.

Step 2: Full-class inclusion. Since $\mathcal{F}_{\mathcal{C}} \subseteq \mathcal{F}$ by construction, monotonicity of Rademacher complexity gives $\mathfrak{R}_n(\mathcal{F}_{\mathcal{C}}) \le \mathfrak{R}_n(\mathcal{F})$.

Step 3: Spatial-slice metric entropy. Fix a temporal input $\bar{\boldsymbol{f}}$ and consider the slice $\mathcal{F}_{\mathcal{C}}|_{\bar{\boldsymbol{f}}}$. By (P2), every function in this slice is L-Lipschitz in $\boldsymbol{x}$ on the compact domain $\mathcal{S} \subset \mathbb{R}^{d_s}$, and by (P1) it takes values in $[-C_{\text{sat}}, C_{\text{sat}}]$. Classical metric-entropy bounds for bounded Lipschitz functions give, for some constants depending only on d_s (Dudley, 1967),

$$\log \mathcal{N}\big(\epsilon,\, \mathcal{F}_{\mathcal{C}}|_{\bar{\boldsymbol{f}}},\, \|\cdot\|_\infty\big) \le c_1 \left(\frac{C_{\text{sat}} + L \cdot \operatorname{diam}(\mathcal{S})}{\epsilon} \right)^{d_s}.$$

For $d_s > 2$, Dudley's entropy integral with truncation at $\epsilon_0 \asymp (C_{\text{sat}} + L \cdot \operatorname{diam}(\mathcal{S}))n^{-1/d_s}$ yields

$$\mathfrak{R}_n(\mathcal{F}_{\mathcal{C}}|_{\bar{\boldsymbol{f}}}) \le C_{d_s}\, (C_{\text{sat}} + L \cdot \operatorname{diam}(\mathcal{S}))\, n^{-1/d_s} \sqrt{\log n},$$

for all $n \ge 2$. This proves Eq. (17). This step is only a bound for fixed-forcing spatial slices; because (P2) imposes no regularity in the forcing variable $\boldsymbol{f}_t$, it cannot be unioned or marginalized into a bound for the full spatiotemporal class without additional assumptions.

C.7 Proof of Proposition 1 (EOF Warm-Start Convergence)

Proof The proof separates what EOF pre-training directly guarantees from the additional local assumptions needed for parameter-space convergence.

Part 1: EOF pre-training error. By Eq. (7), EOF pre-training minimizes the empirical squared error between the first K trunk coordinates and the top-K EOF modes. Let $\theta_T^{(0)}$ be the output of this optimization. Defining ϵ_{opt} as the achieved optimization error gives

$$\frac{1}{N_s} \sum_{j=1}^{N_s} \|\boldsymbol{\tau}(\boldsymbol{x}_j; \theta_T^{(0)})_{1:K} - \boldsymbol{U}_K[j, :]\|^2 \le \epsilon_{\text{opt}}.$$

This statement concerns only the pre-trained EOF coordinates. The remaining coordinates $K+1,\dots,p$ are not constrained by Eq. (7); their contribution is therefore kept explicitly in the full-grid mismatch Δ_T.

Part 2: From trunk mismatch to identifiable parameter distance. On the finite spatial grid, many trunk parameter directions may leave all outputs unchanged. Assumption 2 removes these null directions by working in quotient coordinates and postulates a local inverse-Lipschitz relation between grid outputs and identifiable parameters. Applying it at $\theta_T^{(0)}$ gives

$$\|\theta_T^{(0)}-\theta^*\| \le \sigma_{\min}^{-1}\sqrt{\frac{1}{N_s}\sum_{j=1}^{N_s}\|\boldsymbol{\tau}(\boldsymbol{x}_j;\theta_T^{(0)})-\boldsymbol{\tau}^*(\boldsymbol{x}_j)\|^2} = \sigma_{\min}^{-1}\Delta_T,$$

where parameter distances are understood in the quotient coordinates. This assumption is precisely the step that converts function-space closeness into parameter-space closeness; without it, finite-grid output errors control only equivalence classes of parameters.

Part 3: Local convergence. Under Assumption 1, gradient descent with step size $\eta = 1/\ell$ in the local basin satisfies the standard strongly convex and smooth convergence bound

$$\|\theta_T^{(t)}-\theta^*\| \le \left(1-\frac{\mu}{\ell}\right)^t\|\theta_T^{(0)}-\theta^*\| = \left(1-\frac{1}{\kappa}\right)^t\|\theta_T^{(0)}-\theta^*\|.$$

Combining with Part 2 yields the bound in Proposition 1. Solving $(1-1/\kappa)^t\,\sigma_{\min}^{-1}\Delta_T \le \epsilon$ gives $t = O\big(\kappa\log(\sigma_{\min}^{-1}\Delta_T/\epsilon)\big)$. Therefore, if EOF pre-training produces $\Delta_T^{\text{EOF}} < \Delta_T^{\text{rand}}$ relative to a random initialization in the same local basin, the iteration count is reduced by $O\big(\kappa\log(\Delta_T^{\text{rand}}/\Delta_T^{\text{EOF}})\big)$. This is a conditional local result; it does not assert that EOF pre-training alone controls the untrained trunk coordinates or guarantees entry into the basin.

D Experimental Setup

D.1 Baseline

EOF+NN (Yu et al., 2020): The decompose-predict-reconstruct approach using 5 EOF modes, ASFT forcing selection, and per-mode neural networks with 100-member ensembles. As the ASFT algorithm was absent from the original open-source repository, the results reported here are based on our own re-implementation. We follow the original methodology: 5 EOF modes, ASFT forcing selection with greedy forward selection (minimum $\Delta R^2 = 0.005$), per-mode 2-layer neural networks with $\tanh$ activation trained via L-BFGS, 100-member ensemble per mode.

IVMD-MAGRU (Wang et al., 2024): A decomposition-ensemble framework originally developed for single-spatial tasks and here adapted for multi-spatial. The model employs Variational Mode Decomposition (VMD) to resolve complex DO signals into K intrinsic modes. To optimize VMD hyperparameters (α, tol), Wang et al. (2024) implement a Circle Chaotic Sparrow Search Algorithm (CCSSA) with Cauchy mutation to prevent local optima. Given the computationally prohibitive cost of per-location optimization across 800 spatial points, we adopt a global strategy by minimizing the median Sample Entropy across all spatial locations. The optimized configuration ($K = 20, \alpha = 5000, tol = 1.09\times10^{-6}$) feeds into an Attention-based GRU (64 hidden units, 24-month window). For each mode, regional forcings are used as inputs if their correlation exceeds 0.3; otherwise, the mode is predicted autoregressively.

RNN / LSTM / GRU: Following Li et al. (2021), we extend each recurrent architecture from individual-spatial to multi-spatial prediction by providing external forcings and spatial coordinates as unified input. All three use 2 hidden layers with 128 units and a window length of 24. RNN uses Adam (lr $= 0.015$); LSTM and GRU use RMSprop (lr $= 0.01$).

Transformer: A 3-layer Transformer encoder ($d_{\text{model}} = 128$, 4 heads) with sinusoidal positional encoding, combined with the same spatial embedding as LSTM.

DeepONet: Same architecture as PhysiOcean but trained without physics-guided losses or EOF pre-training. This comparison isolates the combined contribution of the two training enhancements.

Timer-XL (Liu et al., 2025): A long-context Transformer forecaster adapted to one-step DO prediction. The reproduction uses the 24-month context, global forcing channels, and observed DO history, with a shared model applied across spatial points.

TimeMixer++ (Wang et al., 2025): A multiscale forecasting backbone adapted from PyPOTS. Each spatial point is treated as a sample with the global forcing channels and its observed DO history as input; the model is shared across points and predicts the next monthly DO value.

OLinear (Yue et al., 2025): The official shared-temporal-basis OLinear architecture reimplemented for one-step prediction. It uses the global forcing channels together with the observed DO history at each spatial point and shares parameters across points.

FACT (Wang et al., 2026): A dual-domain forecasting model combining time-domain and frequency-domain convolutional branches. We adapt it to the same one-step task with global forcing and observed per-point DO history, using a shared model across spatial points.

The four newly reproduced baselines use observed DO history at the target point, whereas PhysiOcean, the original RNN/LSTM/GRU/Transformer comparisons, and the EOF+NN pipeline are evaluated with their respective forcing-based or prescribed inputs. Their results are therefore reported as strong forecasting baselines under a different information setting and are not interpreted as direct tests of no-history spatial generalization. Each reported result is the mean and standard deviation over five seeds (1, 2, 3, 4, and 42).

D.2 Implementation

PhysiOcean uses $S = 3$ parallel branches and kernel sizes $\kappa \in \{2, 4, 8\}$, mixture of Gaussians at scales ($\{1, 5, 10, 50\}$), $D = 3$ and hidden dimension 256 for Residual MLP blocks, $p = 64$ basis functions, $H = 4$ heads, fusion MLP width 64, Branch hidden dimension 64 per scale, and Trunk hidden dimensions $[256, 256, 256]$. Training uses AdamW (lr $= 5 \times 10^{-4}$, weight decay 10^{-4}) with 15-epoch linear warmup followed by cosine annealing (CosineAnnealingWarmRestarts, $T_0 = 50$, $T_{\text{mult}} = 2$), gradient clipping at $\|\nabla\| \leq 1.0$, batch sizes of 32 time steps $\times$ 400 spatial points, and early stopping with patience 50 (activated after Phase 2). EOF pre-training runs for 2000 steps with the top 10 modes. To determine the C_{sat} for $\mathcal{L}_{\text{bound}}$, if the forcing variables include global temperature and salinity data, we utilize the formulation of Weiss (1970) to calculate the oxygen saturation concentration C_{sat} ; otherwise, a fixed upper bound of 14.0 mg/L is adopted. When calculating the smoothness loss $\mathcal{L}_{\text{smooth}}$, we randomly sample collocation points $\{x_c\}$ uniformly within the spatial domain $\mathcal{S}$ to evaluate $\mathcal{L}_{\text{smooth}}$ at these locations. Additionally, the gradient computation is restricted to the three-dimensional spatial dimensions (latitude, longitude, and depth). We adopted the current hyperparameter configuration following a grid search. Full details are given in Tables 3 and 4.

D.3 Model Architecture

Table 3 provides the complete architecture specification.

D.4 Training Procedure

Table 4 summarizes the training configuration. EOF pre-training uses the top 10 modes to warm-start the Trunk Network before main training begins. Early stopping on validation MSE is activated only after Phase 2 to avoid premature termination during physics loss ramp-up.

E Additional Diagnostic Results

Depth and seasonal breakdown. PhysiOcean achieves $R^2 = 0.912$ for the upper 5 depth layers and $R^2 = 0.843$ for the lower 5, reflecting the greater difficulty of predicting bottom-water DO where stratification and sediment oxygen demand dominate. Seasonally, winter months (Nov–Feb) reach $R^2 = 0.935$ while summer months (May–Sep) yield $R^2 = 0.854$. The summer–bottom combination represents the most challenging regime, with PhysiOcean achieving an R^2 of 0.843 compared to 0.712 for the no-physics variant; this result suggests that physics-guided training is particularly beneficial in hypoxia-prone conditions, without isolating the effect of the stratification term alone.

Table 3: Full architecture details of PhysiOcean.

Component	Parameter	Value
Branch Net	Input dimension	$d_f + 2 = 41$
	Kernel sizes	$\{2, 4, 8\}$ (3 scales)
	Hidden dim per scale	64
	Residual blocks per scale	2 (dilation 1, 2)
	Attention	4 heads, 192-dim
	FC layers	$192 \to 256 \to 128 \to 64$
	Dropout	0.1
Trunk Net	Input dimension	$d_s = 5$
	Fourier frequencies	64 total, scales $\{1, 5, 10, 50\}$
	MLP hidden dims	$[256, 256, 256]$ with LayerNorm + residual
	Bias network	$133 \to 128 \to 64 \to 1$
	Output dimension	$p = 64$
Fusion	Heads H	4 (head dim = 16)
	Fusion MLP	$(H+1) \to 64 \to 32 \to 1$
	Activation	GELU
Total parameters		$\sim$650K

Table 4: Training hyperparameters and configuration for PhysiOcean.

Category	Parameter	Value
Optimization	Optimizer	AdamW (β_1=0.9, β_2=0.999)
	Learning rate	5×10^{-4}, weight decay 10^{-4}
	LR schedule	15-epoch warmup $\to$ CosineWarmRestarts (T_0=50, T_{mult}=2)
	Gradient clipping	Max norm 1.0
Batching	Batch size	32 time steps $\times$ 400 spatial points
	Collocation points	1000 per batch (for $\mathcal{L}_{\text{smooth}}$)
	Validation split	20% of training time steps
EOF pre-training	Steps / optimizer	2000 / Adam (lr=10^{-3}, cosine)
	Modes	Top 10 EOF modes
Physics schedule	Phase transitions	Phase 1 $\to$ 2 at epoch 30; Phase 2 $\to$ 3 at epoch 100
	Loss weights	λ_{bound}=1.0, λ_{smooth}=0.01, λ_{strat}=0.2, $\lambda_{\text{temporal}}$=0.05
Stopping & misc.	Early stopping	Patience 50 (active after epoch 30); max 400 epochs
	Hardware	Single NVIDIA GPU ($\sim$1–2 h per model)
	Seeds	5 seeds (1, 2, 3, 4, and 42) for Table 1; seed 42 for diagnostics

Physical consistency. With physics-guided losses active, only 0.3% of test predictions violate non-negativity (P1) and 0.1% exceed saturation, versus 2.1% and 1.8% without. Summer stratification violations (bottom DO $>$ surface DO, P3) drop from 8.7% to 1.2%. These results show that the regularizers improve physical consistency in addition to their modest average-error contribution.

F INTERPRETABILITY

PhysiOcean admits two natural avenues for post-hoc interpretation, both visualized in Figure 4.

F.1 Temporal Attention Weights

The Branch Network's attention mechanism (Figure 4a) reveals which historical months receive the most weight for a given prediction target. Summer DO predictions (June–August) display a bimodal pattern: a short-range peak at lags of 4–6 months, corresponding to the winter–spring nutrient pulse that drives subsequent hypoxia (Murphy et al., 2011; Yu et al., 2020), and a secondary peak at lags of 15–18 months, indicating that the model captures interannual memory, i.e., prior-year conditions that modulate baseline nutrient loading and residual organic matter. For July predictions, approximately 31% of cumulative attention falls on the nutrient-pulse window (lags -6 to -2) and 28% on the interannual component (lags -18 to -15), so intra-annual and interannual forcing contribute nearly equally. This bimodal structure is most pronounced in summer and winter; spring and autumn predictions distribute attention more uniformly, consistent with the transitional mix of physical drivers in those seasons.

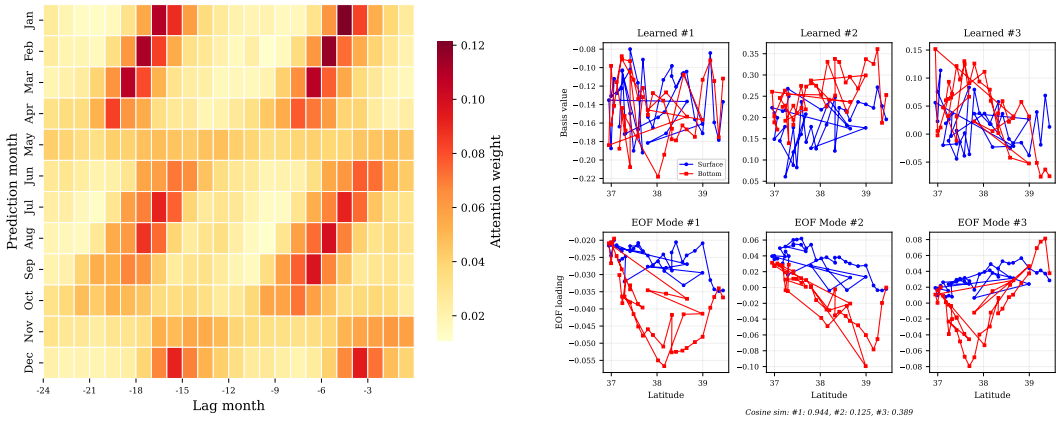
F.2 Spatial Basis Functions

The learned trunk basis functions $\tau_k(\boldsymbol{x})$ encode spatial patterns of DO variability (Figure 4b). For reference, the three leading EOF modes of the training anomaly field are: *Mode 1* ($\sim$87% variance), a uniform-sign pattern peaking in the deep channel (38–39°N) representing the bay-wide seasonal hypoxia signal; *Mode 2*, a surface–bottom contrast strongest in the mid-bay, capturing vertical stratification where surface re-aeration and bottom oxygen consumption vary out of phase; and *Mode 3*, a north–south dipole in the bottom layer reflecting the competing influence of oceanic and riverine end-members along the estuary axis.

The leading learned basis retains high similarity to EOF mode 1 (cosine similarity $= 0.94$), recovering the dominant hypoxia pattern. Higher-order bases diverge from their EOF counterparts (cosine similarity $=$ 0.13 and 0.39 for modes 2 and 3). This is expected: EOF modes are orthogonal linear projections maximizing explained variance, while the end-to-end trained bases can learn nonlinear features oriented toward the prediction objective. Specifically, learned basis 2 replaces the surface–bottom contrast of EOF mode 2 with a monotonic depth-weighted pattern whose loadings are uniformly larger in the bottom layer, suggesting a representation more directly aligned with the nonlinear sediment oxygen demand signal. Learned basis 3 retains a latitudinal dipole in the bottom layer but decorrelates it from the surface component (surface–bottom $r \approx 0$), disentangling depth and along-channel dimensions more flexibly than the orthogonality-constrained EOF decomposition. The learned functions also exhibit finer spatial detail near bathymetric transitions and tributary mouths, where the continuous Fourier representation resolves structure that the discrete EOF modes cannot.

G Sensitivity to Physics Loss Weights

We vary each physics loss weight over a factor of five from its default value (Table 4). Performance is stable within this range for λ_{strat}, $\lambda_{\text{temporal}}$, and λ_{bound}. The smoothness weight λ_{smooth} is more sensitive: values above 0.1 degrade data fidelity (RMSE rises by $\sim$5%), while values below 0.001 provide negligible regularization (vertical generalization degrades). Applying all physics losses from epoch 0 (bypassing progressive scheduling) leads to training instability and worse final R^2 ($\sim$0.85 vs. 0.881), confirming the importance of the phased introduction described in §3.4.



(a) Temporal Attention Weights

(b) Spatial Basis Functions vs EOF Modes

Figure 4: **(a)** Temporal attention weights from the Branch Network for different prediction months (y-axis) over historical lag months (x-axis). Summer predictions attend strongly to the preceding winter–spring nutrient pulse. **(b)** Comparison of learned spatial basis functions (top row) with corresponding EOF modes (bottom row) for modes 1–3. The learned functions capture finer spatial structure near bathymetric transitions.