PIER: Physics-Informed Environmental Retrieval for Time-Series Modeling

Shiyuan Luo[1], Runlong Yu[2], Chonghao Qiu[1], Yue Qin[1], Rahul Ghosh[3], Robert Ladwig[4], Paul C. Hanson[5], Yiqun Xie[6], Xiaowei Jia[1]

[1]University of Pittsburgh [2]University of Alabama [3]University of Minnesota
[4]Aarhus University [5]University of Wisconsin–Madison [6]University of Maryland

[1]{shl298,chq29,yuq9,xiaowei}@pitt.edu [2]ryu5@ua.edu
[3]ghosh128@umn.edu [4]rladwig@ecos.au.dk [5]pchanson@wisc.edu [6]xie@umd.edu

arXiv:2607.20230v1 [cs.LG] 22 Jul 2026

Abstract. Accurate modeling of environmental systems is fundamental to scientific understanding and decision-making, yet remains challenging because observations are limited and physical dynamics vary across systems. Retrieval-augmented approaches offer a natural path to transfer knowledge across systems, but standard embedding-based retrieval does not guarantee consistency of underlying physical processes, since scenarios with similar embeddings may arise from different underlying mechanisms. We propose Physics-Informed Environmental Retrieval (PIER), a model-agnostic framework that augments embedding-based retrieval with a physics-aware stream that scores candidates by flux-response consistency with the target, using local verifiers trained on physics-derived flux features. A weight adjustment mechanism then learns per-scenario weights that adaptively balance the two retrieval streams based on diagnostic features summarizing physics-stream reliability. Experiments on 356 lakes across the Midwestern United States spanning 41 years show that PIER consistently outperforms baselines for water temperature and dissolved oxygen prediction, and serves as a general augmentation strategy across diverse backbones.

1 Introduction. Accurate modeling of environmental systems is fundamental to advancing scientific understanding of the natural world. Environmental processes span a wide range of spatial and temporal scales and involve nonlinear interactions among physical, chemical, and biological components, making direct observation and experimentation alone insufficient for comprehensive analysis and understanding. Computational models bridge this gap by integrating observational data with theoretical knowledge to simulate system behavior under varying scenarios and support decision-making. In particular, the scientific community has long relied on process-based physical models that encode fundamental physical principles, such as conservation of mass, energy, and momentum, to simulate system behavior. Such models are widely used across aquatic science [6, 12], agriculture [9], and geology [21]. A key advantage of these models is that they produce physically meaningful intermediate states and flux quantities that reflect the governing processes of each system, providing interpretable and mechanistically grounded representations of system dynamics. However, their predictive accuracy is often limited by imperfect and simplified parameterization, which can introduce systematic biases and constrain model fidelity [10].

Machine learning (ML) offers a complementary path. By learning directly from data, ML models can capture nonlinear relationships and complex temporal patterns, and have shown strong results across a range of environmental applications [35, 33, 25, 15]. However, in real-world environmental problems, observational data are often limited, since consistent environmental monitoring requires expensive field studies, infrastructure, and ongoing maintenance, making comprehensive observation networks economically infeasible [26, 38]. A common strategy is to train a single global model using data from a large number of observed sites and over long time periods. Such a global model works well in general, but real-world environmental problems exhibit complex and nonstationary patterns whose periodicities and dynamics vary across spatiotemporal contexts [4, 30, 11]. For example, the temporal dynamics of water quality vary substantially across lakes owing to differences in geometric characteristics (e.g., depth, and area), trophic state, and local meteorological forcing; likewise, hydrological responses in watersheds are shaped by heterogeneous geological substrates, land cover distributions, and climate regimes [37, 34]. More importantly, some of these conditions may not be observable and thus cannot be included in data. While a global model benefits from broad training coverage, it tends to converge toward an averaged representation that can obscure the

localized dynamics unique to each individual system. On the other hand, separate modeling of each system is often constrained by sparse observation data. These limitations highlight the need for models that capture both the shared structure that enables knowledge transfer across systems and the distinctions in the underlying processes governing each individual system.

Retrieval-augmented approaches offer a natural path toward this goal by identifying relevant systems and leveraging their data to augment local predictions with information drawn from the broader system population. These approaches have proven effective in language modeling [8, 3, 18] and, more recently, in general time-series forecasting [4, 13] and environmental modeling [16, 32]. The typical approach is to create data embeddings that summarize each scenario's characteristics and then retrieve other scenarios with similar embeddings. However, such conventional embedding-based retrieval can be limited by the quality of embeddings learned from sparse data. Additionally, embedding similarity does not guarantee consistency of underlying physical processes, since scenarios with a similar response may arise from different mechanisms. For example, a shallow lake may warm rapidly because of its low thermal mass, while a deeper lake in a warmer climate may reach similar temperatures through a different radiative balance. Although their embeddings may be similar, the underlying flux dynamics are not. Retrieving such a scenario can introduce dynamics that conflict with the target system and degrade prediction performance. This raises a question: How can we leverage physical knowledge to enhance embedding-based retrieval, ensuring that retrieved samples remain physically consistent with the target scenario's dynamics?

To address this challenge, we propose **P**hysics-**I**nformed **E**nvironmental **R**etrieval (**PIER**), a retrieval framework for environmental time-series modeling. In addition to the standard embedding-based retrieval, PIER introduces a physics-aware retrieval stream that scores candidates by how similarly they respond to physical fluxes relative to the target, using local verifiers trained on physics-derived flux features. The two retrieval streams are complementary: embedding similarity uses embeddings extracted by a global model and captures shared temporal patterns of environmental drivers, while local physical consistency identifies candidates whose underlying dynamics match those of the target. However, their relative reliability varies across scenarios, e.g., systems with sparse local observations may depend more on embedding similarity, whereas well-observed systems can build reliable local verifiers and place greater weight on physical consistency. PIER addresses this through a weight adjustment mechanism that learns per-scenario weights to adaptively balance the two streams.

We evaluate the proposed method for modeling two important water quality variables, water temperature and dissolved oxygen (DO) dynamics using a comprehensive dataset of 356 lakes across the Midwestern United States spanning 41 years. These lakes differ in size, depth, morphometry, and ecological state, while observations vary dramatically in coverage. These two variables are key indicators of water quality and ecosystem health [28, 22, 19]. Water temperature regulates metabolic rates, chemical reactions, and oxygen solubility, shaping overall ecosystem dynamics [23]. DO concentrations are crucial for aquatic biodiversity and water safety since oxygen fluctuations in lakes often reflect their ecological balance and overall health [5, 1]. Our contributions are:

- We propose PIER, a model-agnostic retrieval framework that augments embedding-based retrieval with a physics-aware stream using local verifiers, ensuring retrieved scenarios are both broadly relevant and physically grounded.
- We develop a weight adjustment mechanism that adaptively balances the two retrieval streams per scenario based on the reliability of local verifiers.
- We validate PIER across 356 real-world freshwater lakes, demonstrating consistent improvements over strong baselines.

2 Problem Definition and Background

2.1 Problem Definition. We consider a collection of environmental systems indexed by $\ell \in \{1, \ldots, L\}$, each observed over a particular time period. For each target system, we have access to its environmental features $\mathbf{x}_t$ at each time t, which may include geographic and morphometric attributes, meteorological forcing, ecological indicators, and land-use characteristics. In addition to these inputs, the prediction target y_t is observed on a subset of days, where the number and temporal distribution of observations vary substantially across systems.

In general, a scenario is a fixed-length environmental context drawn from a single system, capturing its local conditions and observations. In our lake dataset, we define a scenario as a yearly data segment from a specific lake. Multiple scenarios may originate from the same system (e.g., different years of the same lake) or different systems. Retrieval is performed over all scenarios in the training set. Our objective is to predict the target variable at all time steps for scenarios in the testing period. Because observations are typically insufficient to capture the dynamics of each system, we aim

to retrieve the most relevant scenarios from all available data and use them to adapt a scenario-specific predictor. The core challenge is defining relevance in a way that accounts for both temporal similarity and physical consistency with the target system's dynamics.

2.2 Physics-Derived Data. In many environmental domains, process-based models are available that simulate system behavior based on established physical laws. Although limited by simplified parameterizations [10], these models produce additional physics data that PIER can leverage: (1) globally calibrated simulated labels that compensate for sparse observations, and (2) intermediate physics-derived flux features for each system, which serve as inputs to local dynamics models described in §3.2.1. These flux features are distinct from the environmental input features $\mathbf{x}_t$: while input features describe external conditions (e.g., meteorological forcing), flux features capture the system's internal physical response to those conditions.

3 Method. In the following, we first provide an overview of the PIER framework, and then describe the physics-aware retrieval enhancement process and the retrieval-augmented adaptation process.

3.1 Framework Overview. The PIER framework is built on two complementary ideas: (1) environmental systems of the same type share universal underlying patterns of input-response relationships, and (2) each system also exhibits unique dynamics, as certain variables in the universal relationships are not observable and thus cannot be included as input features. PIER consists of a global predictive model and a retrieval-augmented adaptation mechanism. The global predictive model $\mathcal{F}_{\theta^*}$ is trained across all available scenarios using environmental input features, while mapping each scenario to a latent representation. The retrieval could be performed based on the representation similarity, which captures the shared response patterns learned from available data. However, relying on data embeddings alone does not guarantee that the retrieved systems share similar underlying physical processes.

As shown in Fig. 1, PIER addresses this through two key mechanisms: **(1) Physics-aware retrieval enhancement:** local verifiers trained on physics-derived features score candidates by whether they share flux-response dynamics consistent with the target system. The *representation similarity stream* and *physics-aware stream* produce a merged candidate pool that is both broadly relevant and physically grounded. **(2) Retrieval-augmented scenario adaptation**: because the physics-aware stream's reliability varies across scenarios, a learned gate adaptively balances the two streams for the target scenario, and the top-ranked scenarios are used to adapt a scenario-specific predictor.

3.2 Physics-Aware Retrieval Enhancement. Rather than relying on a single retrieval criterion, we draw candidates from both streams and merge them. We use q to denote the *query scenario*, and i to index *candidate scenarios*. Top-$K(\cdot)$ means selecting the K candidates with the highest scores, where K is a hyperparameter controlling the size of the retrieval set. For each query q, this process can be expressed as

$$\begin{aligned} \mathcal{C}_q^{\text{rep}} = \text{Top-}K\big(s_{q,i}^{\text{rep}}\big), \qquad \mathcal{C}_q^{\text{phys}} = \text{Top-}K\big(s_{q,i}^{\text{phys}}\big), \\ \mathcal{U}_q = \mathcal{C}_q^{\text{rep}} \cup \mathcal{C}_q^{\text{phys}}, \end{aligned} \tag{3.1}$$

where $\mathcal{C}_q^{\text{rep}}$ contains the top candidates ranked by representation similarity $s_{q,i}^{\text{rep}} = \cos(e_q, e_i)$, with e_q being the global model's hidden-state representation for the scenario q; and $\mathcal{C}_q^{\text{phys}}$ contains the top candidates ranked by a physical-consistency score $s_{q,i}^{\text{phys}}$ defined below. The representation-similarity stream provides broad, stable candidates, preferring those that share general patterns but may lack physical relevance; the physics stream identifies candidates that are locally grounded in the target's underlying dynamics. We now describe how the physical-consistency score $s_{q,i}^{\text{phys}}$ is constructed.

3.2.1 Local dynamics verifier bank. For each system $\ell \in \{1, \ldots, L\}$, we create a local verifier g_ℓ using only the physics-derived flux variables from that system. Unlike the global model, which is trained using the full observation data across all systems, each local dynamics verifier is trained independently for each system. The input to this model includes only those flux variables that are relevant to the final target variable. These flux variables are generated by an independent process-based model. While relying on such a small set of selected flux variables may reduce the ability of model g_ℓ to fit observation data as closely as the pure data-driven model, it can better capture system-specific underlying patterns and may be less susceptible to overfitting.

3.2.2 Physical-consistency scoring. Ideally, if a candidate scenario shares similar physical dynamics with the target scenario, then the target system's local verifier should be able to predict the candidate's behavior well. Conversely, if the candidate is governed by different relationships, the target's model will produce poor predictions. Given a target scenario from system ℓ_q, we score each candidate scenario i by applying the target system's local verifier to the candidate's flux features and measuring how well the prediction aligns with the candidate's observed targets:

$$s_{q,i}^{\text{phys}} = \exp\left(-\frac{1}{\tau} \sqrt{\frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \left(g_{\ell_q}(\mathbf{x}_i^{\text{flux}})_t - y_t^{(i)} \right)^2} \right), \tag{3.2}$$

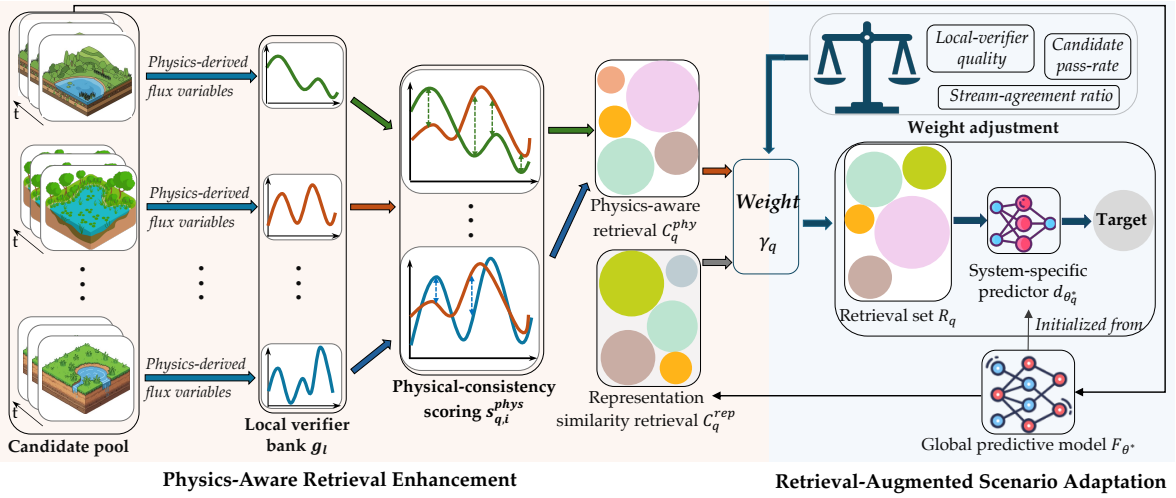


Figure 1: Overview of PIER.

where $\mathbf{x}_i^{\text{flux}}$ denotes the physics-derived flux features for candidate scenario i and τ is a scaling parameter, and $\mathcal{T}_i$ denotes the set of time steps for which ground-truth observations are available for candidate scenario i. The exponential mapping converts the prediction error into normalized similarity range $s_{q,i}^{\text{phys}} \in (0, 1]$, ensuring that small errors yield scores close to 1 and large errors decay smoothly toward 0, which is preferable to a hard threshold. A candidate whose dynamics are well reproduced by the target's local verifier receives a high score, whereas poor predictability indicates physical inconsistency.

3.2.3 Model-quality safeguard. While the local verifiers g_ℓ are intended to capture system-specific patterns, some may be less reliable due to limited observations for certain systems and biases in process-based model estimates of input flux variables. We assess each model's quality by defining $\mu_\ell = \mathcal{L}_{\text{sim},\ell} - \mathcal{L}_{\text{model},\ell}$, where $\mathcal{L}_{\text{sim},\ell}$ and $\mathcal{L}_{\text{model},\ell}$ are the prediction errors of the process-based simulator and the learned model on system ℓ's held-out validation data, respectively. When $\mu_\ell > 0$, the learned model outperforms the simulator, indicating reliable local dynamics have been captured. Otherwise, we disable the physical stream for that target scenario and rely on representation similarity alone.

3.3 Retrieval-Augmented Scenario Adaptation. The physics-aware stream enhances retrieval by introducing candidates that are grounded in the target's local dynamics. However, each environmental system is unique, and this stream is not equally reliable across all scenarios. For a system with abundant observations and a well-fitted local verifier, physical-consistency scores are reliable signals and should carry substantial weight in reranking. For a system with a poorly calibrated local verifier, those scores become noisy, and relying on them can degrade retrieval quality. Uniformly weighting candidates in the merged pool cannot accommodate this variability. PIER addresses this by adaptively determining how much the physics-aware stream contributes to the target scenario through a dynamic scoring mechanism. Then it uses the reranked retrieval set to create a scenario-specific predictor.

3.3.1 Combined retrieval score. For each candidate i in $\mathcal{U}_q$, we combine the representation similarity $s_{q,i}^{\text{rep}}$ with the physical-consistency $s_{q,i}^{\text{phys}}$:

$$s_{q,i}^{\text{comb}} = (1 - \gamma_q)\, s_{q,i}^{\text{rep}} + \gamma_q\, s_{q,i}^{\text{phys}}, \quad \gamma_q \in [0, 1]. \tag{3.3}$$

where γ_q controls how much the physics-aware stream influences the final ranking for this scenario. When $\gamma_q = 0$, retrieval relies entirely on representation similarity; as γ_q increases, physical consistency plays a larger role. The top-K candidates ranked by $s_{q,i}^{\text{comb}}$ form the retrieval set $\mathcal{R}_q$. The weight parameter γ_q can be automatically adjusted using the diagnostic features as described later.

3.3.2 Scenario-adaptive prediction. The final predictions for each testing scenario are generated by a system-specific predictor $d_{\theta_q^\star}$ initialized from the global predictive model and fine-tuned using the retrieved data. Starting from the global model (parameterized by $\theta^\star$ trained over all samples) provides a strong starting point that already captures general environmental dynamics. Fine-tuning is performed using sparse observations from retrieved scenarios, as follows:

$$\mathcal{L}_{\text{adapt}} = \sum_{i \in \mathcal{R}_q} w_i \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \left(\hat{y}_t^{(i)} - y_t^{(i)} \right)^2, \tag{3.4}$$

where $\mathcal{T}_i$ denotes the set of time steps with available observations for retrieved scenario i, the weight w_i measures the contribution made by each retrieved scenario to the target scenario q. Given a retrieval set $\mathcal{R}_q$ with associated weights, each retrieved scenario contributes in proportion to its normalized combined score:

$$w_i = \frac{s_{q,i}^{\text{comb}}}{\sum_{j \in \mathcal{R}_q} s_{q,j}^{\text{comb}}}. \tag{3.5}$$

The global predictive model $\mathcal{F}_{\theta^*}$ and system-specific predictor $d_{\theta_q^\star}$ are designed to be model-agnostic so that different network backbones can be plugged in seamlessly. In our main experiments, we use a long short-term memory (LSTM) network as the default backbone, given its strong performance on sparse hydrological and aquatic time series [7]. We verify that the framework generalizes across architectures in §4.4.2.

3.3.3 Weight adjustment using diagnostic features. The quality of retrieval is ultimately measured by the quality of the adapted predictor. This naturally gives rise to a bilevel optimization: the *inner level* adapts the predictor $d_{\theta_q^\star}$ on the retrieval set (Eq. 3.4), while the *outer level* selects the degree of physics enhancement γ_q that yields the best adapted predictor. A choice of γ_q determines the combined scores (Eq. 3.3), which determine the retrieval set $\mathcal{R}_q$ and the weights $\{w_i\}$, which determine the inner-level adaptation, and ultimately the prediction quality on the target scenario.

$$\mathcal{L}_{\text{outer}} = \frac{1}{|\mathcal{T}_q|} \sum_{t \in \mathcal{T}_q} \left(d_{\theta_q^*(\gamma_q)}(\mathbf{x_q})_t - y_t^{(q)} \right)^2, \tag{3.6}$$

where $\theta_q^\star(\gamma_q)$ denotes the predictor parameters obtained after the inner-level adaptation on the retrieval set induced by γ_q.

Directly optimizing γ_q by differentiating through the inner-level adaptation would require backpropagating through an entire training procedure, which requires unrolling the optimization or computing second-order gradients. We instead adopt a two-phase approach. In the first phase, for each training scenario q, we evaluate the full bilevel pipeline across a discrete set of γ values: for each candidate γ, we run the complete inner-level adaptation and compute $\mathcal{L}_{\text{outer}}$. We record the value that achieves the lowest outer-level loss as the per-scenario target:

$$\gamma_q^\star = \arg\min_{\gamma \in \mathcal{G}} \ \mathcal{L}_{\text{outer}}(\gamma), \tag{3.7}$$

where $\mathcal{G} \subset [0, 1]$ is a discrete grid. In the second phase, we train a small MLP gate h_ϕ to predict these optimal weight values. This two-phase design sidesteps the cost of differentiating through the inner loop while still optimizing the gate for downstream prediction quality. At test time, the gate produces $\hat{\gamma}_q$ in a single forward pass, making the full pipeline efficient at inference. Specifically, the weight parameter is predicted from diagnostic features $\mathbf{z}_q$:

$$\hat{\gamma}_q = h_\phi(\mathbf{z}_q), \qquad \mathcal{L}_{\text{gate}} = \left(\hat{\gamma}_q - \gamma_q^\star\right)^2, \tag{3.8}$$

where the diagnostic feature vector $\mathbf{z}_q$ captures signals about each stream's reliability for the current scenario: *Local-verifier quality*: the model-quality margin μ_{ℓ_q} from §3.2.3, measuring how well the target's local verifier performs on its own system. *Candidate pass-rate*: the fraction of candidates in $\mathcal{U}_q$ for which the target's local verifier g_{ℓ_q}, applied to the candidate's flux features, predicts the candidate observations with lower error than the simulation, measuring whether the local verifier useful for scoring the current candidates. *Stream-agreement ratio*: the fraction of overlap between $\mathcal{C}_q^{\text{rep}}$ and $\mathcal{C}_q^{\text{phys}}$. High overlap suggests both streams agree and the physics enhancement is reinforcing the representation-similarity ranking.

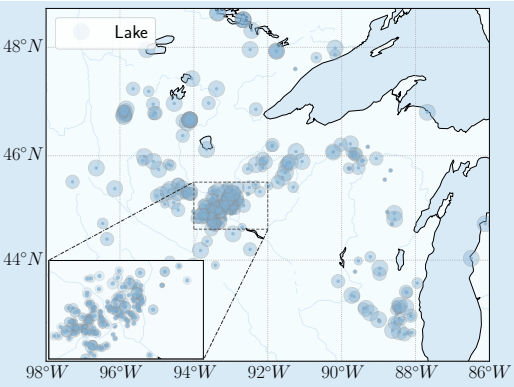


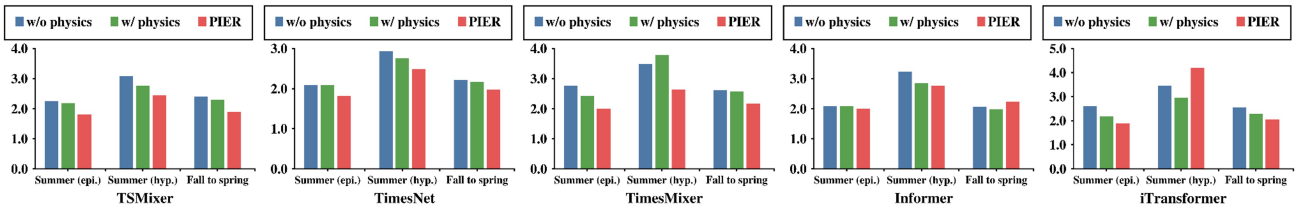
Figure 2: Map of 356 tested lakes.

4 Experimental Evaluation. We conduct extensive experiments across various lakes in the Midwestern United States to address research questions:

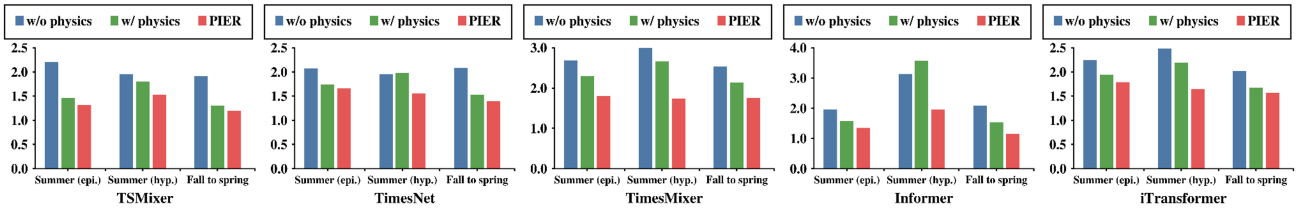
- **RQ1.** How does PIER compare to other temporal prediction methods in predictive performance?
- **RQ2.** Can PIER serve as a general augmentation strategy that improves diverse model backbones?
- **RQ3.** How does physics-aware retrieval enhance capturing temporal patterns in water temperature and DO concentration?
- **RQ4.** What is the contribution of each component in PIER's pipeline?

Table 1: Comparative RMSE performance. Lower values are better; grey numbers denote standard deviation.

Algorithm	DO concentration (g/m³)			Water temperature (°C)		
	Summer (epi.)	Summer (hyp.)	Fall to spring	Summer (epi.)	Summer (hyp.)	Fall to spring
Physics	2.1018 (0.000)	2.8793 (0.000)	2.7438 (0.000)	1.6349 (0.000)	2.7821 (0.000)	1.5436 (0.000)
LSTM	1.9841 (0.011)	2.9404 (0.106)	2.2172 (0.024)	1.5043 (0.036)	1.8524 (0.117)	1.2679 (0.067)
Informer	2.0903 (0.016)	3.2290 (0.034)	2.0687 (0.040)	1.9545 (0.033)	3.1358 (0.305)	2.0768 (0.163)
iTransformer	2.5965 (0.416)	3.4620 (0.060)	2.5482 (0.288)	2.2468 (0.179)	2.4872 (0.094)	2.0218 (0.109)
TSMixer	2.2486 (0.042)	3.0824 (0.014)	2.3996 (0.062)	2.2038 (0.034)	1.9521 (0.064)	1.9115 (0.067)
TimeMixer	2.7571 (0.101)	3.4909 (0.035)	2.6167 (0.107)	2.6866 (0.131)	3.2544 (0.162)	2.5377 (0.137)
TimesNet	2.0879 (0.064)	2.7629 (0.029)	2.1690 (0.015)	2.0766 (0.059)	1.9501 (0.060)	2.0817 (0.023)
PIER	**1.8888** (0.004)	**2.5395** (0.002)	**2.0605** (0.040)	**1.2486** (0.010)	**1.7020** (0.012)	**1.0673** (0.039)



(a) RMSE Evaluation of DO concentration (g/m³)



(b) RMSE Evaluation of water temperature (°C)

Figure 3: Comparison under three training settings.

4.1 Data Preparation. We evaluate PIER's predictive performance for water temperature and DO concentrations using a comprehensive dataset spanning 41 years (1979–2019), with data up to 2011 for training, 2012–2015 for validation, and 2016–2019 for testing. The dataset comprises ecological observations from 356 lakes across the Midwestern USA. The spatial distribution of these lakes is shown in Fig. 2, where color intensity represents lake depth, and marker size corresponds to surface area. The dataset contains approximately 1.75 million daily records, each including 47 environmental features such as morphometric attributes, meteorological conditions, trophic states, and land use characteristics. Data sources are described in [17, 31, 27]. Specifically, water temperature data were obtained from the U.S. Geological Survey [27], with 476,215 observed measurements recorded across depths on 57,156 distinct days. DO concentration data were retrieved from the Water Quality Portal (WQP), encompassing 23,192 days of observations. Lake residence time was sourced from the HydroLAKES dataset. Trophic state probabilities were derived from [17]. Land use proportions within each lake's watershed were extracted from the National Land Cover Database (NLCD).

Domain context: Lakes undergo seasonal stratification driven by thermal expansion of water. During summer, a stable vertical density gradient separates the water column into the epilimnion (a warm, well-mixed surface layer) and the hypolimnion (a cooler, denser bottom layer) [20]. This stratification restricts vertical mixing, limiting oxygen exchange between layers. From fall through spring, the water column becomes well-mixed, behaving as a single layer. This seasonal structure motivates our prediction targets: during summer, we predict temperature and DO separately for the epilimnion (**epi.**) and hypolimnion (**hyp.**); from fall to spring, when the water column is mixed, we predict the total DO concentration throughout the lake.

Physics-derived data: For water temperature, we employ the General Lake Model [6], which simulates the principal heat energy fluxes governing lake thermal dynamics, including incoming short-wave radiation, long-wave radiation, back radiation, sensible heat flux, and latent evaporative heat flux. For DO concentration, we

use the process-based model [12], which simulates atmospheric exchange, net ecosystem production, sediment oxygen demand, and turbulent entrainment fluxes between layers during summer stratification. These models produce globally calibrated simulation labels for pretraining the global predictive model and physics-derived flux features for training the local verifiers.

4.2 Baselines. To evaluate the effectiveness of our PIER framework, we compare it against several baseline models commonly used in time series forecasting: **Physics** [6, 12]: It represent state-of-the-art physics-based models. **LSTM** [7]: A recurrent neural network architecture designed to capture long-term dependencies in sequential data, widely used for time series forecasting. **Informer** [36]: A forecasting model for very long time-series data that keeps the Transformer's ability to learn long-range patterns while making it fast and memory-efficient enough to predict long future sequences in one shot. **iTransformer** [14]: A Transformer-based model designed for irregular time series data, incorporating adaptive attention mechanisms to better capture temporal dependencies. **TSMixer** [2]: An all-MLP architecture for time series forecasting that integrates historical data, future known inputs, and static contextual information through conditional feature mixing and mixer layers. **TimesNet** [29]: A framework that transforms one-dimensional time series into two-dimensional representations, allowing the use of 2D convolutional kernels to extract temporal patterns. **TimeMixer** [24]: A model that explores mixing-based architectures for time series forecasting, focusing on effectively capturing short- and long-term dependencies.

4.3 Implementation Details. Hyperparameters for all methods were tuned via grid search, with optimal configurations selected on validation performance. We explored batch sizes in $\{8, 16, 32\}$, learning rates in $[0.001, 0.05]$, and hidden dimensions in $[20, 200]$. For Transformer-based models, the number of attention heads was varied in $\{4, 8, 16\}$. For PIER, we additionally tuned the retrieval set size K, the scaling parameter *tau* in Eq. 3.2. We set $K = 50$ and $\tau = 10$.

4.4 Experimental Results

4.4.1 Performance comparison. Table 1 reports the RMSE performance across all baselines for both DO concentration and water temperature prediction under different seasonal regimes. PIER achieves the best performance across all six settings, consistently outperforming both physics-based and data-driven approaches:

(a) Increasing model complexity does not help. Advanced architectures designed for complex temporal patterns, such as iTransformer, TSMixer, etc, all underperform the simpler LSTM on this dataset. These models are designed for data-rich settings and appear to struggle with the severe sparsity characteristic of environmental observations.

(b) PIER consistently improves over all baselines across all six settings, with the largest gains in the most challenging cases such as hypolimnetic DO in summer. This suggests the bottleneck lies in how data are used rather than in the model's complexity. PIER addresses this through a data-centric view, retrieving and reusing informative scenarios filtered through physical-consistency constraints.

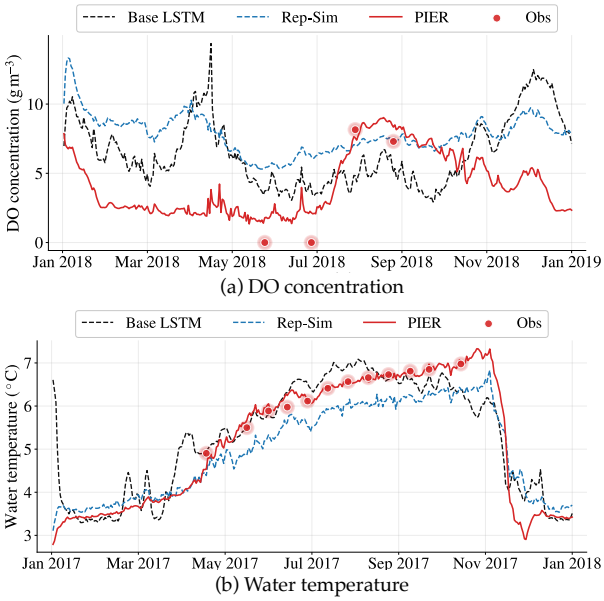
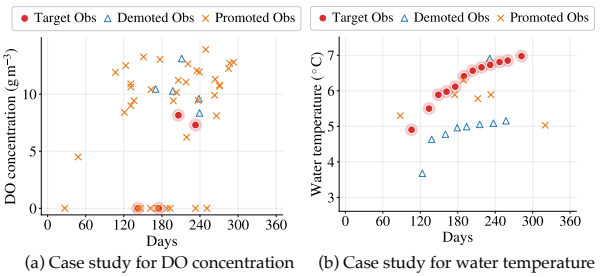


Figure 4: Time-series predictions for two case study scenarios. (a) Epilimnion DO concentration for *Long Lake*, Year 2018. (b) Hypolimnion water temperature for *Weaver Lake*, Year 2017.

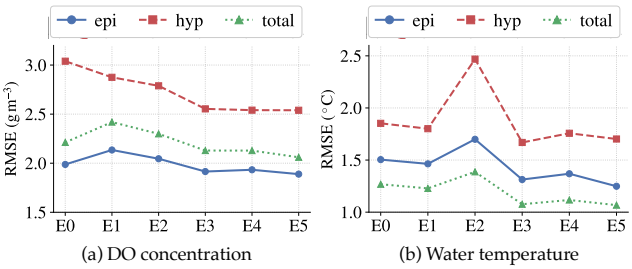
4.4.2 PIER as a general augmentation strategy. To assess the broader applicability of PIER, we further integrate it with five predictive models and compare each model's performance under three settings (RQ2): (1) trained only on observed labels with standard environmental features, (2) pretrained on the simulation labels produced by the process-based model and fine-tuned on observations, with physics-derived flux features concatenated to the standard inputs, and (3) the full PIER framework. Fig. 3 reports the results.

(a) Physics enhancement outperforms direct training across the majority of tested models and tasks,



(a) Case study for DO concentration (b) Case study for water temperature

Figure 5: Observations of promoted and demoted candidates alongside target observations. Promoted candidates are shown in orange; demoted candidates in blue. Observations are plotted against day of year.



(a) DO concentration (b) Water temperature

Figure 6: Results of ablation study.

confirming that the simulation labels and flux features are useful resources for addressing data sparsity challenge in environmental systems. In a few cases, physics enhancement does not help, as the physics flux features, designed to capture system-specific physical responses, may not be fully exploitable when simply concatenated as additional inputs. This suggests PIER's local verifiers provide a more principled way to leverage them.

(b) PIER achieves the lowest RMSE across almost all tasks, demonstrating its applicability across diverse model architectures. While physics-derived data improves models by leveraging better data sources, the retrieval mechanism provides additional gains. In tasks such as DO prediction with iTransformer, PIER does not improve over the physics-enhanced setting, as the DO hyp. dynamics are particularly challenging for this backbone, and fine-tuning on retrieved scenarios can amplify rather than correct the base model's weaknesses.

4.5 Case Study. To understand how PIER's retrieval mechanism works in practice and where the improvements come from, we examine two representative target scenarios in detail: one for DO concentration and one for water temperature. For each case, we present time-series predictions by Base LSTM, Representation-Similarity Retrieval, and PIER in Fig. 4, and observations of promoted candidates and demoted candidates alongside target observations in Fig. 5. Demoted candidates are those ranked high by representation similarity but down-weighted by PIER; promoted candidates are those elevated by the physics-aware stream.

4.5.1 Case 1. Fig. 4(a) and Fig. 5(a) show epilimnion DO for *Long Lake* in 2018, where the observed DO drops to near zero around July. Base LSTM and representation-similarity retrieval both predict 7–14 g/m^3 throughout summer (RMSE 3.33 and 4.14, respectively); retrieval actually worsens the base model because the demoted candidates are dominated by the target lake's own historical years, which reflect typical high-DO behavior. Embeddings encode statistically typical patterns from training data and fail to generalize when test conditions deviate from what the model has seen. The physics-aware stream promotes candidates from other lakes (*Bashaw Lake*-2009/2015, *Parley Lake*-2006) with compatible flux-response dynamics and similarly low DO, reducing error to 1.52.

4.5.2 Case 2. Fig. 4(b) and Fig. 5(b) show hypolimnion water temperature for *Weaver Lake* in 2017. Here, representation-similarity retrieval selects lakes with shared general temporal structure but mismatched thermal regimes, since the representations capture shared statistical structure from training but not physical compatibility. PIER removes these and promotes scenarios such as *Bavaria Lake-2013* and *Upper Nemahbin Lake-2008*, both within the target's narrow range. Notably, none of these appeared in the representation-similarity retrieval set.

4.6 Ablation Study. Ablation experiments are designed to evaluate the contribution of each component while holding everything else fixed: *E0: Base LSTM.* No simulation and no physics flux features, no retrieval. The model is trained solely on observed labels. *E1: Simple retrieval.* Retrieval without simulated labels or physics flux features. *E2: Physics-aware stream only.* Full physics-aware retrieval enhancement pipeline, but ranking uses only the physical-consistency scores ($\gamma = 1$). *E3: Fixed equal weight.* Physics-aware stream and representation-similarity stream contribute equally with a fixed global weight ($\gamma = 0.5$). This tests whether a naive combination of the two streams already improves over either stream alone. *E4: Representation-similarity stream only.* The simulated labels and physics flux features are used, but ranking uses only the representation similarity scores ($\gamma = 0$). *E5: PIER.* Fixed γ is replaced with the learned weight adjustment. Fig. 6 presents the results:

(a) Simple retrieval (E1) yields mixed results: it slightly improves water temperature predictions but actually degrades DO predictions. Without any physics components, retrieval can introduce misleading scenarios that hurt rather than help, particularly for complex DO dynamics.

(b) When the full physics components are used but ranking relies only on representation similarity (E4), performance improves substantially compared to E1. The improvement comes entirely from the physics-derived data that produces better representations to make retrieval more effective.

(c) The Physics-aware retrieval alone (E2) improves over E1 for DO tasks but performs poorly on water temperature, as it can be noisy when used as the sole ranking criterion. Notably, equal weighting (E3) improves over both E2 and E4, confirming the two streams provide complementary information.

(d) PIER's learned weight adjustment (E5) achieves the best performance on all tasks, confirming that scenario-specific weight adjustment is necessary: no single fixed γ is best across all settings, and neither is optimal for any individual scenario.

5 Conclusion. In this paper, we presented PIER, a physics-informed retrieval framework that augments embedding-based retrieval with a physics-aware stream grounded in local verifiers, and adaptively balances the two streams per scenario through a weight adjustment mechanism. Experiments on 356 lakes over 41 years show consistent gains over physics-based and data-driven baselines for water temperature and dissolved oxygen prediction, and demonstrate that PIER serves as a general augmentation strategy across diverse backbones. Future work includes extending PIER to other environmental domains, such as agriculture, and exploring additional ways to integrate physical processes into machine learning methods.

References

[1] E. A. Birge, *Gases dissolved in the waters of wisconsin lakes*, Transactions of the American Fisheries Society, 35 (1906).

[2] S.-A. Chen, C.-L. Li, N. Yoder, S. O. Arik, and T. Pfister, *Tsmixer: An all-mlp architecture for time series forecasting*, arXiv preprint arXiv:2303.06053, (2023).

[3] S. Gupta, R. Ranjan, and S. N. Singh, *A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions*, arXiv preprint arXiv:2410.12837, (2024).

[4] S. Han, S. Lee, M. Cha, S. O. Arik, and J. Yoon, *Retrieval augmented time series forecasting*, arXiv preprint arXiv:2505.04163, (2025).

[5] P. C. Hanson, D. L. Bade, S. R. Carpenter, and T. K. Kratz, *Lake metabolism: Relationships with dissolved organic carbon and phosphorus*, Limnology and Oceanography, 48 (2003), pp. 1112–1119.

[6] M. R. Hipsey, L. C. Bruce, et al., *A general lake model (glm 3.0) for linking with high-frequency sensor data from the global lake ecological observatory network (gleon)*, Geoscientific Model Development, (2019).

[7] S. Hochreiter et al., *Long short-term memory*, Neural computation, 9 (1997), pp. 1735–1780.

[8] Y. Hu and Y. Lu, *Rag and rau: A survey on retrieval-augmented language model in natural language processing*, arXiv preprint arXiv:2404.19543, (2024).

[9] X. Jia, A. Khandelwal, D. J. Mulla, P. G. Pardey, and V. Kumar, *Bringing automated, remote-sensed, machine learning methods to monitoring crop landscapes at scale*, Agricultural Economics, 50 (2019), pp. 41–50.

[10] X. Jia, J. Willard, A. Karpatne, J. Read, J. Zwart, M. Steinbach, and V. Kumar, *Physics guided rnns for modeling dynamical systems: A case study in simulating lake temperature profiles*, in Proceedings of the 2019 SIAM international conference on data mining, SIAM, 2019, pp. 558–566.

[11] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, *Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets*, Hydrology and Earth System Sciences, 23 (2019), pp. 5089–5110.

[12] R. Ladwig, A. P. Appling, A. Delany, H. A. Dugan, Q. Gao, N. Lottig, J. Stachelek, and P. C. Hanson, *Long-term change in metabolism phenology in north temperate lakes*, Limnology and Oceanography, 67 (2022), pp. 1502–1521.

[13] J. Liu, L. Yang, H. Li, and S. Hong, *Retrieval-augmented diffusion models for time series forecasting*, Advances in Neural Information Processing Systems, 37 (2024), pp. 2766–2786.

[14] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, *itransformer: Inverted transformers are effective for time series forecasting*, in The Twelfth International Conference on Learning Representations, 2024.

[15] S. Luo, C. Qiu, R. Yu, Y. Xie, and X. Jia, *Great: Generalizable representation enhancement via auxiliary transformations for zero-shot environmental prediction*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, 2026, pp. 38998–39006.

[16] S. Luo, R. Yu, C. Qiu, R. Ghosh, R. Ladwig, P. C. Hanson, Y. Xie, and X. Jia, *Learning to retrieve for environmental knowledge discovery: An augmentation-adaptive self-supervised learning framework*, arXiv preprint arXiv:2509.14563, (2025).

[17] M. F. Meyer, S. N. Topp, T. V. King, R. Ladwig, R. M. Pilla, H. A. Dugan, J. R. Eggleston, S. E. Hampton, D. M. Leech, I. A. Oleksy, et al., *National-scale remotely sensed lake trophic state from 1984 through 2020*, Scientific Data, 11 (2024), p. 77.

[18] B. Peng, Y. Zhu, Y. Liu, et al., *Graph retrieval-augmented generation: A survey*, arXiv preprint arXiv:2408.08921, (2024).

[19] J. S. Phillips, *Time-varying responses of lake metabolism to light and temperature*, Limnology and Oceanography, 65 (2020), pp. 652–666.

[20] J. S. Read et al., *Derivation of lake mixing and stratification indices from high-resolution lake buoy data*, Environmental Modelling & Software, 26 (2011), pp. 1325–1336.

[21] M. Reichstein et al., *Deep learning and process understanding for data-driven earth system science*, Nature, 566 (2019), pp. 195–204.

[22] C. T. Solomon et al., *Ecosystem respiration: drivers of daily variability and background respiration in lakes around the globe*, Limnology and Oceanography, 58 (2013), pp. 849–866.

[23] P. A. Staehr, D. Bade, M. C. Van de Bogert, G. R. Koch, C. Williamson, P. Hanson, J. J. Cole, and T. Kratz, *Lake metabolism and the diel oxygen technique: state of the science*, Limnology and Oceanography: Methods, 8 (2010), pp. 628–644.

[24] S. Wang et al., *Timemixer: Decomposable multiscale mixing for time series forecasting*, in The Twelfth International Conference on Learning Representations.

[25] Z. Wang, L. Ma, G. Hurtt, X. Jia, Y. Li, R. Li, Z. Li, S. Xu, and Y. Xie, *Carbonglobe: A global-scale, multi-decade dataset and benchmark for carbon forecasting in forest ecosystems*, in The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.

[26] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, *Integrating scientific knowledge with machine learning for engineering and environmental systems*, ACM Computing Surveys, 55 (2022), pp. 1–37.

[27] J. D. Willard, J. S. Read, A. P. Appling, S. K. Oliver, X. Jia, and V. Kumar, *Predicting water temperature dynamics of unmonitored lakes with meta-transfer learning*, Water Resources Research, 57 (2021), p. e2021WR029579.

[28] P. C. Wilson, *Water quality notes: Dissolved oxygen: Sl313/ss525, 1/2010*, EDIS, 2010 (2010).

[29] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, *Timesnet: Temporal 2d-variation modeling for general time series analysis*, in The Eleventh International Conference on Learning Representations.

[30] Y. Xie, E. He, X. Jia, H. Bao, X. Zhou, R. Ghosh, and P. Ravirathinam, *A statistically-guided deep network transformation and moderation framework for data with*

spatial heterogeneity, in 2021 IEEE International Conference on Data Mining (ICDM), IEEE, 2021, pp. 767–776.

[31] R. Yu et al., *Adaptive process-guided learning: An application in predicting lake do concentrations*, in 2024 IEEE International Conference on Data Mining (ICDM), IEEE, 2024, pp. 580–589.

[32] R. Yu, S. Luo, R. Ghosh, L. Li, Y. Xie, and X. Jia, *Rag for geoscience: What we expect, gaps and opportunities*, arXiv preprint arXiv:2508.11246, (2025).

[33] R. Yu, C. Qiu, R. Ladwig, P. Hanson, Y. Xie, and X. Jia, *Physics-guided foundation model for scientific discovery: An application to aquatic science*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 2025, pp. 28548–28556.

[34] M. Zhang, N. Liu, R. Harper, Q. Li, K. Liu, X. Wei, D. Ning, Y. Hou, and S. Liu, *A global review on hydrological responses to forest change across multiple spatial scales: Importance of scale, climate, forest type and hydrological regime*, Journal of Hydrology, 546 (2017), pp. 44–59.

[35] S. Zhong, K. Zhang, M. Bagheri, J. G. Burken, A. Gu, B. Li, X. Ma, B. L. Marrone, Z. J. Ren, J. Schrier, et al., *Machine learning: new ideas and tools in environmental science and engineering*, Environmental science & technology, 55 (2021), pp. 12741–12754.

[36] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, *Informer: Beyond efficient transformer for long sequence time-series forecasting*, in Proceedings of the AAAI conference on artificial intelligence, vol. 35, 2021, pp. 11106–11115.

[37] J. Zhou, P. R. Leavitt, K. C. Rose, X. Wang, Y. Zhang, K. Shi, and B. Qin, *Controls of thermal response of temperate lakes to atmospheric warming*, Nature Communications, 14 (2023), p. 6503.

[38] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, *A comprehensive survey on transfer learning*, Proceedings of the IEEE, 109 (2020), pp. 43–76.