

Learning to Retrieve for Environmental Knowledge Discovery: An Augmentation-Adaptive Self-Supervised Learning Framework

Shiyuan Luo^{1,†}, Runlong Yu^{2,†}, Chonghao Qiu¹, Rahul Ghosh³,
Robert Ladwig⁴, Paul C. Hanson⁵, Yiqun Xie⁶, Xiaowei Jia¹

¹University of Pittsburgh, ²University of Alabama, ³University of Minnesota,
⁴Aarhus University, ⁵University of Wisconsin-Madison, ⁶University of Maryland
{shl298, chq29, xiaowei}@pitt.edu, ryu5@ua.edu, ghosh128@umn.edu,
rladwig@ecos.au.dk, pchanson@wisc.edu, xie@umd.edu

Abstract—The discovery of environmental knowledge depends on labeled task-specific data, but is often constrained by the high cost of data collection. Existing machine learning approaches usually struggle to generalize in data-sparse or atypical conditions. To this end, we propose an *Augmentation-Adaptive Self-Supervised Learning* (A^2SL) framework, which retrieves relevant observational samples to enhance modeling of the target ecosystem. Specifically, we introduce a multi-level pairwise learning loss to train a scenario encoder that captures varying degrees of similarity among scenarios. These learned similarities drive a retrieval mechanism that supplements a target scenario with relevant data from different locations or time periods. Furthermore, to better handle variable scenarios, particularly under atypical or extreme conditions where traditional models struggle, we design an augmentation-adaptive mechanism that selectively enhances these scenarios through targeted data augmentation. Using freshwater ecosystems as a case study, we evaluate A^2SL in modeling water temperature and dissolved oxygen dynamics in real-world lakes. Experimental results show that A^2SL significantly improves predictive accuracy and enhances robustness in data-scarce and atypical scenarios. Although this study focuses on freshwater ecosystems, the A^2SL framework offers a broadly applicable solution in various scientific domains.

Code—<https://github.com/shiyuanlsy/A2SL>

I. INTRODUCTION

Environmental issues and sustainable development rank among the most critical global challenges of the 21st century. Advancing environmental knowledge discovery requires extracting insights from complex ecological systems [1], which relies on accurate and comprehensive data collection. However, environmental data collection is fraught with obstacles. The high costs associated with the deployment and maintenance of monitoring infrastructure, particularly in remote or inaccessible regions, and difficulties in data acquisition during specific periods (e.g., due to ice cover, cloud cover, or instrument failures) lead to substantial data sparsity and regional imbalances [2]–[4]. These pose considerable barriers to modeling and prediction efforts.

Traditional process-based models are widely used to simulate environmental variables, grounding their predictions in fundamental physical principles such as mass and energy

conservation. In freshwater ecosystems, for example, these models are commonly used to predict key water quality metrics, including water temperature and dissolved oxygen (DO) concentrations [5], [6], which are critical for assessing ecosystem health and ensuring water security. Similar approaches are also applied in fields such as geology, oceanography, atmospheric science, and climate research [4]. Despite their wide applicability, process-based models face significant limitations. They often require extensive parameterization or depend on oversimplified approximations, primarily due to incomplete physical understanding and the inherent complexity of modeling intricate environmental processes [7], [8]. These limitations can introduce inherent biases that ultimately compromise the accuracy and reliability of the predictions.

Data-driven approaches, particularly those that leverage machine learning (ML), have gained increasing traction in environmental modeling [2], [4]. ML methods excel at capturing complex patterns and nonlinear relationships within data, often surpassing traditional models when sufficient high-quality data is available [9]. However, environmental ecosystems often exhibit significant spatio-temporal heterogeneity, while observation data are typically collected in an imbalanced manner across different locations and time periods due to the high cost of data acquisition [10]. ML models are highly dependent on observation data. As a result, models trained on existing datasets often fail to generalize under atypical conditions, such as extreme weather events or rare ecological phenomena, due to overfitting or the lack of relevant training examples [11].

Recent advancements in generative AI have demonstrated the effectiveness of leveraging external information to augment context for target tasks [12]–[14]. Such data augmentation has been largely enabled by advanced pre-trained models, which can retrieve relevant content from large-scale auxiliary corpus data. Inspired by this success, we pose an important question: can environmental conditions similarly guide the retrieval of relevant data from different locations or time periods to supplement data-sparse scenarios? While promising, this approach presents two key technical challenges.

The first challenge is to accurately quantify the similarity of different environmental scenarios to support the retrieval

[†]Shiyuan Luo and Runlong Yu contributed equally to this research.

process. In NLP, the similarity or relevance of documents can be directly inferred by pre-trained models [15], which are effectively trained on large and representative datasets using objectives such as masked token prediction. However, these pre-training methods are not suitable for environmental modeling as they are not designed for capturing the complex dependencies between environmental input features (e.g., weather conditions) and target variables (e.g., ecosystem properties). Unlike structured text or images, environmental data is shaped by complex interactions among different variables and also contains substantial noise [16], [17]. As a result, directly using these existing methods can lead to the retrieval of scenarios that either fail to preserve key scenario characteristics or introduce extraneous variations, which hinders the effectiveness in enhancing the target task.

The second challenge lies in effectively determining when and where to apply data augmentation across different environmental scenarios. Environmental data contain both stable scenarios, which feature long-term and consistent trends, and variable scenarios, which exhibit distinct patterns (e.g., significant fluctuations), particularly under atypical or extreme conditions [18], [19]. Although data augmentation could help traditional ML models better generalize to data-scarce variable scenarios, applying it indiscriminately over all scenarios may disrupt the consistency of stable environments, potentially degrading model performance. Striking the right balance requires a selective augmentation strategy to identify scenarios that are likely to benefit from external knowledge without distorting inherent consistent environmental dynamics [20], [21].

To tackle these challenges, we propose an *Augmentation-Adaptive Self-Supervised Learning (A²SL)* framework. The idea is to dynamically switch between specialized models for capturing stable and variable scenarios, while selectively retrieving and integrating relevant data sharing similar environmental patterns to augment the learning for data-sparse, variable scenarios. In our framework, each scenario is defined by short-term environmental conditions, paired with simulated labels produced by process-based models. We create two predictive models: one specialized for stable scenarios, ensuring consistent long-term predictions (e.g., yearly trends); and another tailored for variable scenarios, improving short-term predictions (e.g., monthly behaviors) under atypical or extreme conditions where data augmentation is most needed. An augmentation-adaptive mechanism is proposed to determine which predictive model should be applied to a given scenario and whether augmentation is needed, ensuring an optimal prediction strategy for each scenario type. To effectively retrieve relevant scenarios for augmentation, we build a self-supervised scenario encoder that captures scenario similarities through a multi-level pairwise learning loss [22]. This loss guides the encoder in assessing the similarity of an anchor scenario with three types of other scenarios, i.e., positively augmented data scenarios, semi-positively augmented data scenarios, and negatively augmented data scenarios. The learned representation and similarities enable a retrieval mechanism to gather relevant data from different locations or time periods, supplementing

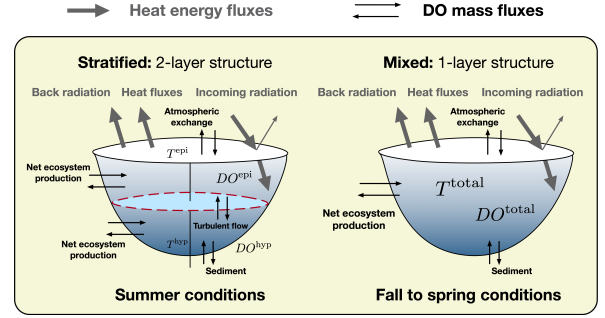


Fig. 1. Heat energy and DO mass fluxes in lake systems.

data-sparse variable scenarios. Finally, a decoder is employed to integrate information from both the current and retrieved scenarios to predict target ecosystem variables, improving the model’s accuracy and robustness.

We evaluate the proposed method using freshwater ecosystems as a case study, with a focus on modeling water temperature and DO dynamics in real-world lakes. These two variables are key indicators of water quality and ecosystem health [23]–[25]. Water temperature regulates metabolic rates, chemical reactions, and oxygen solubility, shaping overall ecosystem dynamics [26]. Similarly, DO concentrations are crucial for aquatic biodiversity and water safety. Oxygen fluctuations in lakes often reflect their ecological balance and overall health [27]. Our evaluation across 356 lakes in the Midwestern United States demonstrates the effectiveness of the A²SL framework in improving water temperature and DO predictions, even in data-sparse scenarios. Our contributions are summarized as follows:

- We propose a self-supervised learning approach that quantifies scenario similarity to retrieve relevant data across different locations and periods, enabling targeted retrieval and improving data availability in sparse settings.
- We develop an augmentation-adaptive mechanism that switches between models for stable and variable scenarios, improving robustness under diverse conditions.
- We validate A²SL in 356 real-world freshwater lakes in the Midwestern United States, showing significant gains in predicting water temperature and DO dynamics.

II. PROBLEM FORMULATION

The objective of this work is to predict daily water temperature and DO concentrations across numerous lake systems, spanning broad geographic regions and long time periods. As illustrated in Figure 1, water’s thermal expansion leads to stratification, forming a stable vertical density gradient. This stratification restricts vertical mixing, limiting nutrient and oxygen exchange between layers and reducing connectivity between bottom waters and the atmosphere, thereby hindering oxygen replenishment [28]. During summer, stratified lakes exhibit a vertical density difference exceeding 0.05 kg/m^3 between surface and bottom layers, an average water temperature above 4°C , and a distinct thermocline.

Water temperature prediction. During summer, we predict water temperature at two distinct depth layers: the epilimnion,

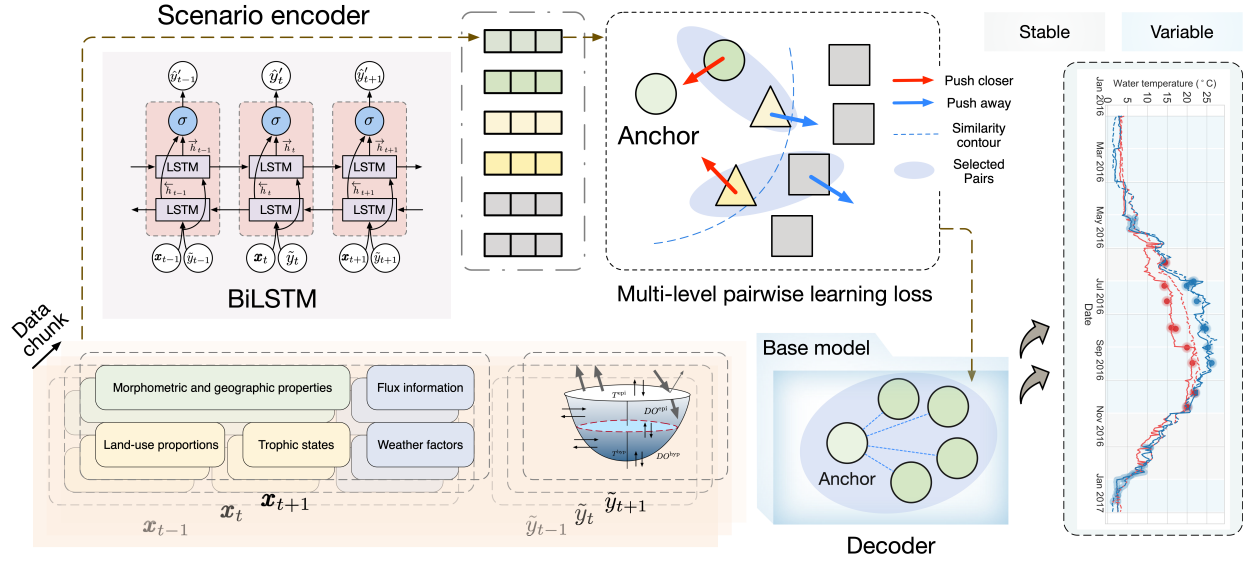


Fig. 2. The overall framework of A²SL.

a well-mixed surface layer, and the hypolimnion, a cooler, nutrient-rich, but light-limited deep layer. At each time step t , the epilimnion temperature is denoted as T_t^{epi} , while the hypolimnion temperature is denoted as T_t^{hyp} . From fall to spring, when the water column is typically well-mixed, we predict the overall water temperature of the entire lake, denoted as T_t^{total} .

DO concentration prediction. Similarly, during summer, DO concentrations are predicted separately for the epilimnion and hypolimnion at each time step t , denoted as DO_t^{epi} and DO_t^{hyp} , respectively. This distinction accounts for the differences in oxygen dynamics and metabolic processes in each layer. From fall to spring, when the water column undergoes full mixing, we predict the total DO concentration, DO_t^{total} , representing the oxygen distribution across the entire lake.

For each target lake, we have access to its environmental features at each time t , represented as $\mathbf{x}_t = \{x_t^1, \dots, x_t^m\}$, which influence lake temperature and DO. These features include morphometric and geographic properties such as lake area, depth, and shape; flux information such as ecosystem and sedimentation fluxes; weather factors including wind speed and air temperature; trophic states ranging from dystrophic to eutrophic; and land-use proportions, such as forests, wetlands, and other surrounding landscapes. In addition to these inputs, water temperature and DO concentrations are observed on a subset of days, denoted as y_t , where $y_t \in \{T_t^{\text{epi}}, T_t^{\text{hyp}}, T_t^{\text{total}}, DO_t^{\text{epi}}, DO_t^{\text{hyp}}, DO_t^{\text{total}}\}$. It is noteworthy that the observations are available only for a limited subset of time steps due to the cost of data collection. Also the number of observations can vary across different lakes.

III. METHODOLOGY

In this section, we present the *Augmentation-Adaptive Self-Supervised Learning* (A²SL) framework, as depicted in Fig. 2. We begin with the scenario encoder, followed by the self-supervised learning approach for quantifying scenario similarity using a multi-level pairwise learning loss. The encoder

is jointly trained with the decoder, which leverages these learned similarities to drive a retrieval mechanism, enriching sparse scenarios with relevant data from different locations or time periods. Finally, we introduce the augmentation-adaptive mechanism, which dynamically selects between two predictive models: one for stable scenarios and another for variable scenarios such as extreme or atypical conditions.

A. Scenario Encoder

1) *Scenario description:* A scenario S is defined as a short-term sequence of environmental conditions that characterize the state of a specific environmental ecosystem. As illustrated in Fig. 1, lake dynamics arises from complex interactions among meteorological conditions, hydrological processes, and biochemical fluxes, and often exhibit varying patterns in short periods (e.g., certain months). Hence, we represent each scenario as a *data chunk* $S = \{(\mathbf{x}_t, \tilde{y}_t)\}_{t=1}^n$, that captures the daily variations of relevant environmental conditions and flux variables in a lake over a one-month period. Here $\mathbf{x}_t$ denotes observed environmental features and $\tilde{y}_t$ represents simulated outputs generated by process-based models, which capture general physical relationships driving the dynamics of target variables. The index t represents each day within one month, forming a temporally ordered sequence¹. During training, the simulated series $\tilde{y}_t$ serves as an additional input feature.

For lake temperature simulation, we employ the General Lake Model (GLM) [5], a widely adopted process-based model that captures key heat fluxes regulating water temperature. These fluxes include heat input from terrestrial long-wave and incoming short-wave radiation, as well as heat loss through back radiation, sensible heat flux, and latent evaporative heat flux. A portion of short-wave and long-wave radiation is reflected by the lake surface, further shaping thermal dynamics. For DO concentration, we utilize a process-based

¹ For consistency, we fix each month to 30 days (thus $n = 30$ and 12 months per year); the remaining 5–6 calendar days are discarded.

model following [6], which simulates major DO mass fluxes, including atmospheric exchange, net ecosystem production, and oxygen consumption by sediment. During summer, it also accounts for DO entrainment fluxes driven by turbulent flow, thereby influencing thermocline depth and regulating oxygen transport between the epilimnion and hypolimnion.

2) *Self-supervised similarity learning*: We propose a self-supervised learning method to extract scenario representations that facilitate the retrieval of other scenarios governed by similar underlying environmental processes. Here we consider two scenarios to be *physically consistent* if they differ only in a few dimensions (e.g., year, month, or lake identity) that do not fundamentally alter their underlying physical or environmental processes, e.g., the data of the same lake taken from different years but during the same calendar month. We aim to ensure that these scenarios under similar environmental processes remain close in the representation space learned through the self-supervised learning.

Formally, we perform a multi-level pairwise learning [22], which captures the similarity relationships among an anchor scenario S_A , a *positively* augmented scenario S^+ , a *semi-positively* augmented scenario S^* , and a *negatively* augmented scenario S^- . To efficiently sample augmented scenarios, we employ a hierarchical sampling mechanism. First, we cluster all scenarios based on their morphometric and geographic properties. Within each cluster, we sample data chunks following predefined rules that ensure varying levels of physical and temporal consistency. Intuitively, the positively augmented scenarios are selected from the closest clusters and are physically consistent with the anchor scenario S_A , but varies in only a few dimensions. These include scenarios from the same lake and month but from a different year, or from a different lake that shares the same month and year with S_A . The semi-positively augmented scenarios come from relatively close clusters and deviate more significantly in temporal patterns while remaining largely consistent with S_A . In contrast, the negatively augmented scenarios are drawn from the farthest clusters and represent the least consistent scenario sample.

Based on the scenario representation generated by the encoder and derived similarity, we aim to impose the ordering

$$\Gamma: \quad \text{sim}(S_A, S^+) \geq \text{sim}(S_A, S^*) \geq \text{sim}(S_A, S^-), \quad (1)$$

which allows $\text{sim}(S_A, S^*)$ to vary between the bounds of $\text{sim}(S_A, S^+)$ and $\text{sim}(S_A, S^-)$, rather than enforcing strict consistency where it may not be physically valid.

3) *Encoder implementation*: The proposed A²SL framework is agnostic to the choice of encoder architecture, allowing the integration of various existing encoding methods. In this study, we employ a bidirectional long-short-term memory (BiLSTM) as the backbone of the scenario model [29], due to its effectiveness in explicitly modeling temporal dependencies in environmental ecosystems in both forward and backward directions. Let a scenario be $S = \{(\mathbf{x}_t, \tilde{\mathbf{y}}_t)\}_{t=1}^n$. At each time step t , we construct the concatenated input vector $[\mathbf{x}_t; \tilde{\mathbf{y}}_t]$, which is processed by forward and backward cells to produce

hidden states $\vec{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$. We form the scenario embedding by taking the mean of the final forward and backward states: $\mathbf{s} = \frac{\vec{\mathbf{h}}_n + \overleftarrow{\mathbf{h}}_1}{2}$. The BiLSTM inherently aggregates contextual information from the entire sequence in both temporal directions, which provides a robust scenario representation for subsequent self-supervised similarity learning.

4) *Multi-level pairwise learning loss*: To effectively capture the similarity order Γ (Eq. 1), we include multiple pairwise comparisons in the pre-training loss inspired by ranking-based learning methods. Unlike conventional pairwise loss functions that optimize a single similarity ranking at a time, our approach enforces two pairwise constraints simultaneously, allowing the model to learn a more structured similarity hierarchy among scenario augmentations.

Given the scenario embedding $\mathbf{s}_A$ for the anchor scenario S_A , and the embeddings of corresponding contrastive scenarios $\mathbf{s}^+$, $\mathbf{s}^*$, and $\mathbf{s}^-$ for the positively augmented, semi-positively augmented, and negatively augmented scenarios, respectively, the loss function is formulated as:

$$\mathcal{L}_{\text{En}} = -\log \sigma \left[\lambda (\text{sim}(\mathbf{s}_A, \mathbf{s}^+) - \text{sim}(\mathbf{s}_A, \mathbf{s}^*)) + (1 - \lambda) (\text{sim}(\mathbf{s}_A, \mathbf{s}^*) - \text{sim}(\mathbf{s}_A, \mathbf{s}^-)) \right], \quad (2)$$

where $\sigma(\cdot)$ denotes the Sigmoid function, and $\text{sim}(\cdot)$ computes the cosine similarity between two scenario embeddings. The tradeoff parameter $\lambda \in [0, 1]$ balances the two pairwise constraints, allowing flexibility in optimizing similarity rankings.

Here, $\mathbf{s}^+$, $\mathbf{s}^*$, and $\mathbf{s}^-$ represent the BiLSTM-encoded embeddings of S^+ , S^* , and S^- , respectively. These data chunks are sampled according to predefined rules. During this process, the anchor scenario S_A is explicitly excluded from the candidate pool, guaranteeing it is never re-selected as a positive, semi-positive, or negative example.

The multi-level pairwise learning loss refines the scenario representations by enforcing structured similarity relationships. Unlike contrastive losses that impose absolute constraints, it maintains relative similarities, ensuring that physically consistent scenarios remain closer in the learned space. This improves retrieval-based augmentation by guiding the encoder to focus on meaningful spatiotemporal relationships.

B. Retrieval-based Decoder

1) *Training the decoder*: The decoder is trained jointly with the scenario encoder, ensuring simultaneous updates to both the scenario embeddings and model parameters. This joint optimization allows the embeddings to capture scenario similarities effectively while improving predictive performance through retrieval-based augmentation. For an anchor scenario S_A , we compute its similarity with all scenarios in the dataset. The k most similar scenarios are selected and, together with S_A , are used as training inputs to a base predictive model.

Let $\mathbb{S}$ be the set of all scenarios and $\mathbf{s}_i$ the embedding of $S_i \in \mathbb{S}$. We first obtain the index set of the k scenarios most similar to the anchor S_A (excluding itself):

$$\mathcal{I}_k = \text{Top}_k \left(\{i \mid S_i \in \mathbb{S}, i \neq A\}, \text{sim}(\mathbf{s}_A, \mathbf{s}_i) \right). \quad (3)$$

The corresponding retrieval set and the final input batch are

$$\mathcal{S}_{\text{re}} = \{S_i \mid i \in \mathcal{I}_k\}, \quad \mathcal{S}_{\text{in}} = \mathcal{S}_{\text{re}} \cup \{S_A\}, \quad (4)$$

so that $\mathcal{S}_{\text{in}}$ always contains exactly $k+1$ *distinct* scenarios. Each scenario in the batch is fed to the base predictor $\mathcal{M}_\alpha$ to generate prediction:

$$\hat{y}_t^{(i)} = \mathcal{M}_\alpha(S_i, t), \quad S_i \in \mathcal{S}_{\text{in}}, \quad (5)$$

where $\hat{y}_t^{(i)}$ denotes the predicted target variable for scenario S_i at time step t . If fewer than two retrieved chunks contain observed targets, we incrementally increase k until the set includes at least two such chunks.

2) *Decoder implementation*: The decoder is designed to be model-agnostic so that different network backbones can be plugged in seamlessly. In this study, we adopt a long short-term memory (LSTM) network as the default backbone, given its strong performance on sparse hydrological and aquatic time series [30]; several alternative architectures are benchmarked in our experiments. We fix the sequence length at one month, although many days lack ground-truth observations. Let $m_t^{(i)} \in \{0, 1\}$ be an observation mask that equals 1 when the target variable for scenario S_i is observed at day t and 0 otherwise. For every scenario in the input batch $\mathcal{S}_{\text{in}}$ we still generate daily predictions $\hat{y}_t^{(i)}$, but only observed days ($m_t^{(i)} = 1$) contribute to the loss. To emphasise more relevant scenarios, each scenario's contribution is further weighted by its cosine similarity to the anchor:

$$w_i = \frac{\text{sim}(\mathbf{s}_i, \mathbf{s}_A)}{\sum_{S_j \in \mathcal{S}_{\text{in}}} \text{sim}(\mathbf{s}_j, \mathbf{s}_A)}. \quad (6)$$

The resulting similarity-weighted, masked mean-squared-error (MSE) objective is

$$\mathcal{L}_{\text{De}} = \sum_{S_A} \sum_{S_i \in \mathcal{S}_{\text{in}}} w_i \frac{1}{\sum_{t=1}^n m_t^{(i)}} \sum_{t=1}^n m_t^{(i)} (\hat{y}_t^{(i)} - y_t^{(i)})^2, \quad (7)$$

where $\mathcal{S}_{\text{in}}$ comprises the anchor scenario and its k retrieved neighbors. Because the weights $\{w_i\}$ sum to 1, the objective remains on the same scale as a conventional MSE while naturally giving greater influence to scenarios that are more similar to the anchor. Optimising Eq. (7) jointly updates decoder parameters and refines scenario embeddings, ensuring that retrieval-based augmentation is seamlessly integrated into the learning process.

3) *Tuning the decoder*: After the initial joint training stage (Eq. 7), we obtain a fully trained decoder $\mathcal{M}_\alpha$ that is shared across all scenarios. During the testing phase, the decoder is further fine-tuned to better adapt to the environmental conditions of each target period before generating the final predictions. In particular, for each target short-term period, we treat it as the anchor scenario S_A and optimize a scenario-specific decoder $\mathcal{M}_\beta^{(A)}$ by fine-tuning $\mathcal{M}_\alpha$ on the set of retrieved scenarios most relevant to S_A .

During fine-tuning, only the parameters of $\mathcal{M}_\beta^{(A)}$ are updated, while the scenario embeddings remain fixed. Freezing the embeddings preserves the similarity structure learned during joint training and prevents catastrophic forgetting, thereby enabling the decoder to focus on localized patterns

that are specific to S_A . Once this adaptation is complete, the customized decoder produces the final prediction for S_A as $\hat{y}_t^{(A)} = \mathcal{M}_\beta^{(A)}(S_A, t)$.

This two-stage procedure, which first performs global training and then applies scenario-specific fine-tuning, achieves a balance between generalization and adaptability. The fine-tuned decoder is especially useful for short-term forecasts, such as monthly predictions under atypical or extreme environmental conditions, where data augmentation provides significant gains in predictive accuracy.

C. Augmentation-Adaptive Mechanism

Environmental scenarios vary widely in their stability. Some exhibit stable, consistent long-term patterns, whereas others experience significant variability, especially under extreme conditions. Traditional predictive models often perform poorly in data-scarce variable scenarios, where augmentation can help improve accuracy. However, indiscriminate augmentation can introduce noise into stable scenarios, degrading performance.

To balance these trade-offs, we propose an *augmentation-adaptive mechanism* that dynamically selects the most appropriate predictor for each scenario. Specifically, we maintain two families of prediction networks: (i) a *yearly model* $\mathcal{M}_\gamma$, which ingests 360-day sequences and is trained independently of the scenario encoder to capture slow dynamics in stable scenarios; and (ii) *monthly models*, where the backbone $\mathcal{M}_\alpha$ operates on 30-day chunks, is jointly optimized with the scenario encoder, and serves as the common initialization for fine-tuning. For each variable scenario, we derive a scenario-specific decoder $\mathcal{M}_\beta^{(i)}$ (written as $\mathcal{M}_\beta$ when the index is clear) by fine-tuning $\mathcal{M}_\alpha$ on the scenario itself and its top- k retrieved neighbors. This adaptive switching enables a balanced augmentation strategy that enhances performance without compromising stability.

Let $\mathbb{S}$ be the full scenario set, and $m_t^{(i)} \in \{0, 1\}$ be the observation mask introduced in Eq. (7). For each scenario S_i with n days, we compute the masked MSE using the yearly model $\mathcal{M}_\gamma$ and fine-tuned monthly model $\mathcal{M}_\beta$:

$$\mathcal{E}_\gamma(S_i) = \frac{1}{\sum_{t=1}^n m_t^{(i)}} \sum_{t=1}^n m_t^{(i)} (\mathcal{M}_\gamma(S_i, t) - y_t^{(i)})^2, \quad (8)$$

$$\mathcal{E}_\beta(S_i) = \frac{1}{\sum_{t=1}^n m_t^{(i)}} \sum_{t=1}^n m_t^{(i)} (\mathcal{M}_\beta(S_i, t) - y_t^{(i)})^2. \quad (9)$$

A scenario is labeled as *variable* if the fine-tuned model yields a lower MSE:

$$\mathbb{P} = \{S_i \mid \mathcal{E}_\beta(S_i) < \mathcal{E}_\gamma(S_i)\}, \quad \mathbb{N} = \mathbb{S} \setminus \mathbb{P}, \quad (10)$$

where $\mathbb{P}$ and $\mathbb{N}$ denote the variable and stable subsets.

To automate this classification, we introduce a discriminative model $D_\omega(S_i)$, parameterized by ω , which outputs a probability in $[0, 1]$ indicating the likelihood of scenario S_i being variable. We use a predefined threshold $\tau \in (0, 1)$ to convert this probability into a binary decision. During training, the discriminator is optimized by maximizing the standard binary cross-entropy objective:

$$\max_{\omega} \mathbb{E}_{S_i \in \mathbb{P}} [\log D_{\omega}(S_i)] + \mathbb{E}_{S_i \in \mathbb{N}} [\log (1 - D_{\omega}(S_i))] . \quad (11)$$

The discriminator is trained to converge on a held-out validation set. For a new scenario S_i , we compute $p = D_{\omega}(S_i)$. The final prediction for each scenario is generated by exactly one model:

$$\hat{y}_t = \begin{cases} \mathcal{M}_{\gamma}(S_i, t), & p \leq \tau \\ \mathcal{M}_{\beta}(S_i, t), & p > \tau. \end{cases} \quad (12)$$

At inference time, the augmentation-adaptive mechanism ensures that retrieval-based augmentation is applied only when beneficial, thereby enhancing accuracy in variable scenarios while maintaining robust performance in stable scenarios.

IV. EXPERIMENTAL EVALUATION

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

- **RQ1.** How does the proposed A²SL framework compare to baseline methods in predictive performance?
- **RQ2.** Can A²SL serve as a general data augmentation strategy to enhance models under atypical conditions?
- **RQ3.** How well does A²SL capture temporal patterns in water temperature and DO concentration time series?
- **RQ4.** What types of data chunks does A²SL retrieve for a given anchor scenario?

A. Data Preparation

We evaluate the proposed A²SL framework for predicting water temperature and DO concentrations using a comprehensive dataset spanning 41 years (1979–2019). The dataset comprises ecological observations from 356 lakes across the Midwestern USA, as illustrated in Figure 3. In the figure, color intensity represents lake depth, while 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 [20], [31], [32]. Specifically, water temperature data were obtained from the U.S. Geological Survey [32], 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 [31]. Land use proportions within each lake’s watershed were extracted from the National Land Cover Database (NLCD). For model development, the dataset is segmented as follows: data collected up to 2011 form the training set, data spanning 2012 to 2015 constitute the validation set, and data from 2016 to 2019 are designated as the test set.

B. Baselines

To evaluate the effectiveness of our A²SL framework, we compare it against several baseline models commonly used in time series forecasting:

- **Physics-based** [5], [6]: As discussed in Section III-A1, these baselines represent state-of-the-art physics-based

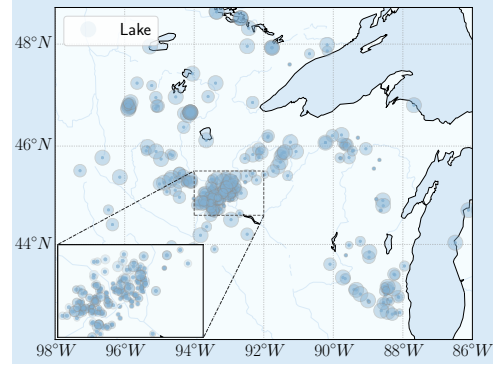


Fig. 3. Map of 356 tested lakes.

models. The flux features used by these models are obtained through calibration with observed data.

- **LSTM** [30]: A recurrent neural network architecture designed to capture long-term dependencies in sequential data, widely used for time series forecasting.
- **EA-LSTM** [33]: An Entity-Aware LSTM network that integrates hydrological behavior and distinguishes between similar dynamic behaviors across various entities, including distinct morphometric characteristics.
- **PT-LSTM** [34]: An enhanced LSTM model that incorporates time attributes via position encoding and a dedicated time gate, enabling better capture of temporal patterns.
- **iTransformer** [35]: A Transformer-based model designed for irregular time series data, incorporating adaptive attention mechanisms to better capture temporal dependencies.
- **TSMixer** [36]: 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** [37]: A framework that transforms one-dimensional time series into two-dimensional representations, allowing the use of 2D convolutional kernels to extract temporal patterns.
- **TimeMixer** [38]: A model that explores mixing-based architectures for time series forecasting, focusing on effectively capturing short- and long-term dependencies.

Among these baselines, we designate LSTM (w/o pre-training) as a baseline, trained solely on observed labels within the training set. In contrast, all other baseline models undergo pre-training on simulated labels before being fine-tuned on observed data. This pre-training step is essential, as these models are highly data-dependent, and their performance degrades significantly without sufficient pre-training. To assess the effectiveness of A²SL, we use LSTM as the base predictive model in the main evaluation. Additionally, we examine A²SL as a general data augmentation strategy by applying it to other advanced models, evaluating its ability to enhance predictive accuracy under atypical or extreme environmental conditions. Hyperparameters for all methods were carefully tuned via an extensive grid search. Optimal configurations were selected based on validation performance. We explored batch sizes ranging from 8 to 32, learning rates from 0.001 to 0.05, and

TABLE I
COMPARATIVE RMSE PERFORMANCE. LOWER VALUES ARE BETTER; GREY NUMBERS DENOTE STANDARD DEVIATION.

Algorithm	Water temperature ($^{\circ}\text{C}$)			DO concentration (g/m^3)		
	Summer (epi.)	Summer (hyp.)	Fall to spring	Summer (epi.)	Summer (hyp.)	Fall to spring
Physics-based	1.608 (0.000)	2.544 (0.000)	1.665 (0.000)	1.954 (0.000)	3.037 (0.000)	3.365 (0.000)
LSTM	1.330 (0.043)	1.815 (0.038)	1.039 (0.047)	1.943 (0.023)	3.460 (0.204)	2.474 (0.029)
EA-LSTM	1.617 (0.226)	2.778 (0.386)	1.266 (0.173)	1.969 (0.014)	2.883 (0.019)	2.580 (0.045)
PT-LSTM	1.295 (0.013)	1.726 (0.014)	1.038 (0.069)	1.896 (0.029)	2.830 (0.038)	2.398 (0.145)
iTransformer	1.841 (0.072)	2.181 (0.073)	2.088 (0.087)	2.020 (0.064)	3.044 (0.059)	2.435 (0.059)
TSMixer	1.753 (0.283)	1.676 (0.110)	1.533 (0.331)	2.058 (0.050)	2.947 (0.014)	2.519 (0.060)
TimesNet	1.812 (0.061)	1.723 (0.031)	2.140 (0.252)	1.966 (0.023)	2.923 (0.047)	2.445 (0.099)
TimeMixer	2.348 (0.252)	2.750 (0.732)	2.454 (0.551)	2.186 (0.228)	3.846 (0.836)	2.606 (0.137)
A ² SL	1.253 (0.023)	1.604 (0.019)	0.967 (0.033)	1.757 (0.013)	2.719 (0.036)	2.397 (0.120)

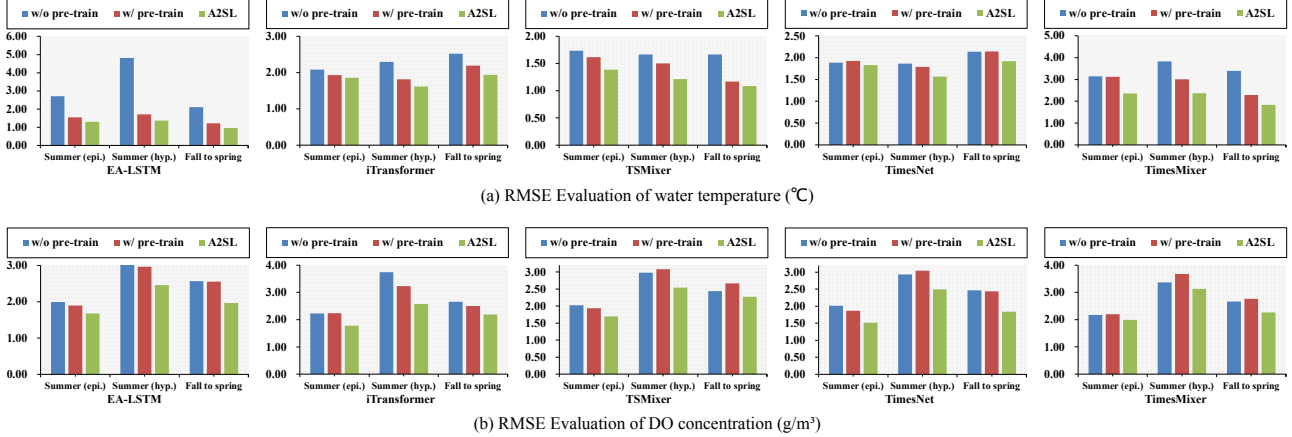


Fig. 4. Comparison under three training settings: (1) without pre-training, (2) with simulated-label pre-training, and (3) using our proposed A²SL framework.

hidden layer dimensions from 20 to 200. For Transformer-based models, the number of attention heads was varied between 4 and 16. For A²SL specifically, we also tuned the trade-off parameter $\lambda \in \{0.0, 0.1, \dots, 1.0\}$ and the discriminator threshold $\tau \in \{0.0, 0.1, \dots, 1.0\}$.

C. Experimental Results

1) *Performance comparison*: Table I compares the performance of A²SL integrated with a standard LSTM backbone against baselines. It measures performance based on the root mean square error (RMSE), including both mean and standard deviation, calculated over five runs. Key findings from this analysis are summarized below (RQ1):

- Physics-based models perform worse than most data-driven models across the majority of tasks, reflecting its limitations in capturing detailed thermal and biochemical variability, particularly within hypolimnion layers.
- A basic LSTM significantly outperforms the physics-based baseline. Enhanced models as EA-LSTM and PT-LSTM could offer further incremental improvements, particularly noticeable in hypolimnion DO predictions.
- iTransformer and Mixer-style models are competitive for temperature prediction tasks but noticeably weaker for DO concentration predictions. This weakness is likely due to their sensitivity to data sparsity.
- With identical LSTM capacity, A²SL consistently achieves the lowest RMSE across all six prediction tasks.

In the challenging summer hypolimnion setting, A²SL achieves an 11.6% RMSE reduction in water temperature prediction and a 21.4% reduction in DO concentration compared to LSTM. The improvement reaches up to 42.3% and 29.3% when compared against EA-LSTM and TimeMixer, respectively.

To assess the broader applicability of A²SL, we further integrate it with five predictive models and evaluate its effectiveness under atypical environmental conditions (Fig. 4). Key observations from this extended analysis include (RQ2):

- Pre-training with simulated labels consistently outperforms direct supervised training across nearly all tested models and tasks. This result underscores the effectiveness of weak supervision in data-scarce scenarios.
- A²SL consistently achieves the lowest RMSE across all tasks and demonstrates robust performance gains for both temperature and DO predictions, applicable across diverse model architectures. These results emphasize A²SL as a robust, model-agnostic augmentation strategy.

2) *Time-series analysis (RQ3)*: Fig. 5 illustrates one year of water temperature dynamics in a representative lake, while Fig. 6 presents the corresponding DO concentration series. Both figures utilize background shading to indicate the augmentation-adaptive mechanism's model choice: grey intervals represent stable periods handled by the yearly model, whereas light-blue intervals indicate variable condi-

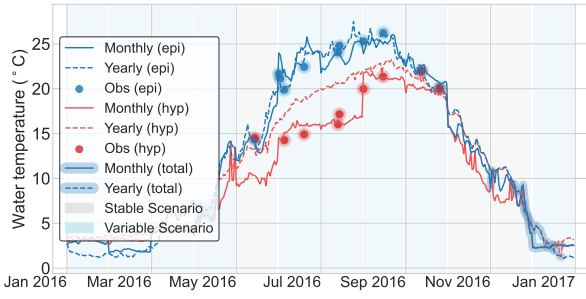


Fig. 5. Water temperature predictions with dynamic switching by A²SL.

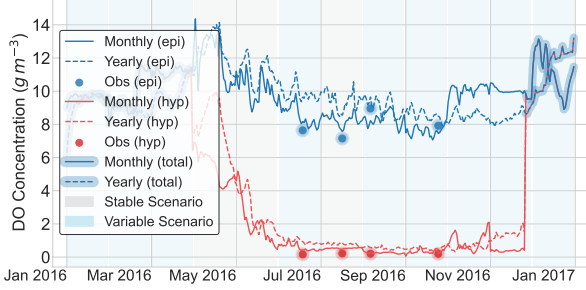


Fig. 6. DO concentration predictions with dynamic switching by A²SL.

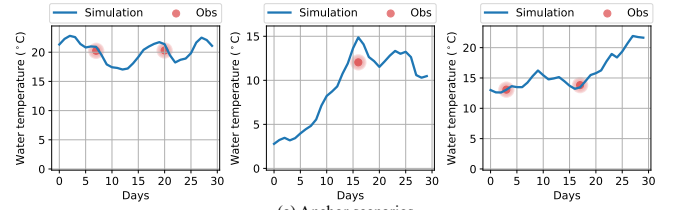
tions, prompting dynamic switching to the monthly model. Several key insights emerge from this analysis.

During rapid warming events, such as those observed between August and September, the augmentation-adaptive mechanism correctly identifies high variability. The benefit of dynamic model switching is particularly pronounced for hypolimnetic conditions. By switching to the monthly model, the predictions closely track the short-term temperature peaks, which the yearly model tends to smooth out and underestimate. During periods when oxygen-rich waters rapidly decrease, the yearly model consistently underestimates the decline rate. In contrast, the monthly model, activated selectively during variable intervals, accurately captures the step-like decreases and closely aligns with the sparse observational data.

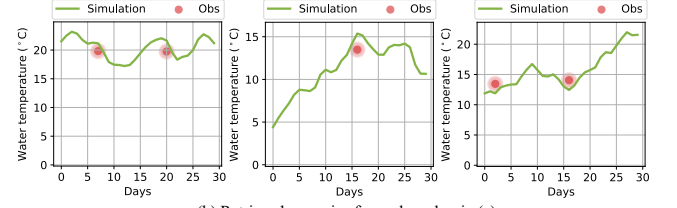
Overall, these results underscore the effectiveness of A²SL. The adaptive selection between yearly and monthly models significantly enhances the ability to capture short-term fluctuations and critical transitions in both water temperature and DO concentrations, thus providing a robust predictive advantage under dynamically changing environmental conditions.

3) *Retrieval analysis (RQ4)*: Fig. 7 and Fig. 8 illustrate the data-chunk retrieval process in A²SL. The top row in each figure presents three representative anchor scenarios (one-month windows) from the test set, shown as process-based simulations overlaid with sparse observations. The bottom row shows the corresponding nearest-neighbor scenarios retrieved by the encoder from different lakes or years.

Across both variables, the retrieved chunks exhibit strong temporal resemblance to their respective anchors. Rising limbs align with rising limbs, double-peak patterns are matched with similar structures, and sharp step-like drops are paired with equally abrupt transitions. This demonstrates that A²SL prioritizes short-term temporal dynamics when computing scenario similarity, rather than relying solely on raw magni-

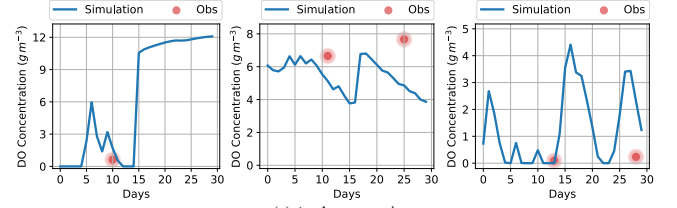


(a) Anchor scenarios

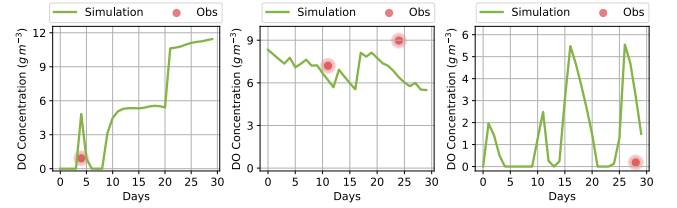


(b) Retrieved scenarios for each anchor in (a).

Fig. 7. A²SL retrieval for water temperature predictions.



(a) Anchor scenarios



(b) Retrieved scenarios for each anchor in (a).

Fig. 8. A²SL retrieval for DO concentration predictions.

tudes. In addition to shape similarity, the retrieved scenarios often include observations at time steps where the anchor is unobserved, effectively densifying the supervision signal during fine-tuning. These label-enriching and structurally aligned retrievals play a key role in boosting predictive accuracy, particularly under sparse or highly variable conditions, and explain the consistent performance gains observed across all tasks and model architectures.

4) *Discriminator threshold sensitivity test*: To assess the effectiveness of the augmentation-adaptive mechanism in A²SL, we analyze how prediction accuracy varies with changes in the discriminator threshold τ . As shown in Fig. 9, each curve represents the RMSE performance for a specific lake-year pair under different values of τ , which controls the dynamic switch between the yearly and monthly models. We observe that no universal threshold yields optimal performance across all scenarios. In some cases, lower values of τ (i.e., favoring the monthly model) produce better accuracy, while others perform better at higher τ values, indicating a preference for the yearly model. This variability highlights the importance of a learned, scenario-specific discriminator rather than a static threshold. Additionally, many curves exhibit flat RMSE regions beyond certain thresholds—either at the lower or upper end—indicating consistent model preference. For instance, a

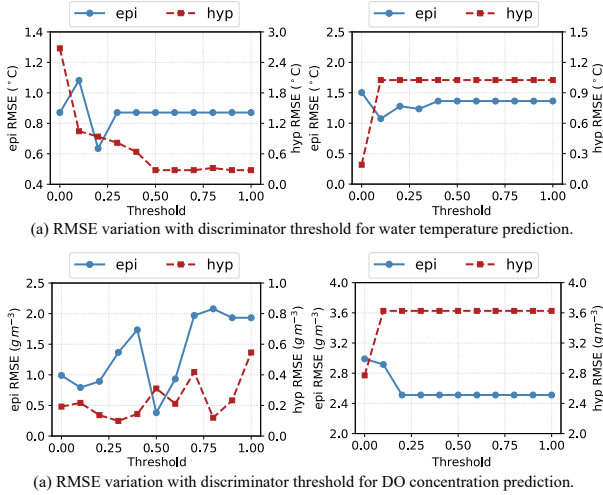


Fig. 9. Sensitivity of prediction performance to the discriminator threshold.

plateau beginning at $\tau = 0$ implies that the monthly model is generally superior for that lake and year, while a flat region at higher τ suggests the yearly model suffices. Some curves also display U-shaped behavior, where RMSE first decreases and then increases across the τ range, implying that adaptive switching between models outperforms relying on either alone.

V. RELATED WORK

Process-based modeling. Process-based models, also referred to as physics-based or theory-driven models, have long served as the foundation for environmental system modeling [16], [17]. These models are built upon first principles from physics, chemistry, and biology, using differential equations to represent key processes such as energy fluxes, hydrological transport, and biogeochemical cycling. They have been widely adopted in domains such as watershed hydrology [39], lake thermal regimes [5], and agroecosystem dynamics [40]. In freshwater systems, vertical one-dimensional aquatic ecosystem models (AEMs) such as GLM-AED [5], WET [41], and PCLake+ [42] integrate hydrodynamics with water quality and ecological processes. However, these models typically require extensive calibration, exhibit sensitivity to parameter choices, and suffer from equifinality, hindering robust generalization [43], [44]. Moreover, process-based models are calibrated under stable environmental regimes and struggle to capture short-term variability or rapid transitions [7].

Data-driven and hybrid models. Machine learning has significantly advanced predictive modeling in agriculture [45], [46], hydrology [33], and limnology [32], [47]. Modern architectures, including recurrent and attention-based networks, often outperform traditional process-based methods when ample observations are available [48], [49]. However, their performance can degrade under covariate shift or observation sparsity, reflecting their data-intensive and black-box nature [50]. To address these limitations, hybrid physics-machine learning models have emerged, integrating domain knowledge and physical principles into data-driven workflows. These hybrid paradigms improve generalization and ensure consistency with known ecological or hydrological dynamics by embedding

mechanistic constraints [20], [26], [51]. For instance, process-guided neural networks have been shown to yield improved long-term predictions in lake biogeochemistry by respecting stratification principles [20]. A common strategy in data-driven modeling is transfer learning [32]. In contrast to prior work that focuses solely on parameter transfer, A²SL actively retrieves physically analogous scenarios from simulation data to enrich training batches.

Self-supervised learning and retrieval augmentation. Self-supervised learning has become a foundational approach for representation learning in environmental sciences, enabling models to leverage vast amounts of unlabeled data through contrastive, masked-token, or predictive pretext tasks. However, most existing self-supervised approaches adopt a binary similarity formulation, which fails to capture the nuanced and continuous spectrum of physical similarity inherent in geoscientific systems. We address this limitation by proposing a multi-level pairwise learning loss that introduces an intermediate “semi-positive” class, enhancing the physical fidelity of learned embeddings. Meanwhile, Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for overcoming the limitations of static model memory by incorporating dynamic external knowledge at inference time [12], [13], [52]. Originally developed for language modeling, RAG has since expanded into multimodal and graph-based domains for more contextual and structured retrieval [14]. The environmental analogue shares conceptual underpinnings [43], but remains largely decoupled from learned neural encoders and self-supervised augmentation pipelines. Recent advances in curriculum retrieval, difficulty-aware sampling, and regime switching [20], [21] underscore the need for selective data augmentation. Our proposed framework extends these ideas. By enabling regime-aware augmentation, A²SL improves robustness and generalization across diverse geophysical regimes.

VI. CONCLUSION

In this paper, we presented an *Augmentation-Adaptive Self-Supervised Learning (A²SL)* framework for modeling freshwater ecosystems. A²SL retrieves physically consistent data chunks based on scenario similarity and dynamically selects between specialized models. Empirical evaluations of real-world freshwater systems demonstrate consistent gains over both process-based and ML baselines. As a general-purpose framework for adaptive augmentation, A²SL holds promise beyond freshwater systems, paving the way for scalable and robust environmental modeling across scientific domains.

VII. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) grants 2239175, 2316305, 2147195, 2425844, 2425845, 2430978, 2126474, 2530609, 2530610, and 2213549; the USGS awards G21AC10564 and G22AC00266; the NASA grants 80NSSC24K1061 and 80NSSC25K0013; and the NSF NCAR’s Derecho HPC system. This research was also supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

REFERENCES

- [1] R. Yu, Y. Xie, and X. Jia, "Environmental computing as a branch of science," *Communications of the ACM*, vol. 68, no. 7, pp. 92–94, 2025.
- [2] J. Willard *et al.*, "Integrating scientific knowledge with machine learning for engineering and environmental systems," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–37, 2022.
- [3] S. E. Hampton *et al.*, "Big data and the future of ecology," *Frontiers in Ecology and the Environment*, vol. 11, no. 3, pp. 156–162, 2013.
- [4] M. Reichstein *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.
- [5] 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.
- [6] 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*, vol. 67, no. 7, pp. 1502–1521, 2022.
- [7] J. J. McDonnell and K. Beven, "Debates—The future of hydrological sciences: A (common) path forward? A call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph," *Water Resources Research*, vol. 50, no. 6, 2014.
- [8] K. Beven, "A manifesto for the equifinality thesis," *Journal of hydrology*, vol. 320, no. 1–2, pp. 18–36, 2006.
- [9] A. Karpatne *et al.*, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE TKDE*, vol. 29, 2017.
- [10] C. Shen, "A transdisciplinary review of deep learning research and its relevance for water resources scientists," *Water Resources Research*, vol. 54, no. 11, pp. 8558–8593, 2018.
- [11] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, 2014.
- [12] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *arXiv preprint arXiv:2005.11401*, 2020.
- [13] M. M. Abootorabi *et al.*, "Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation," *arXiv preprint arXiv:2502.08826*, 2024.
- [14] B. Peng, Y. Zhu, Y. Liu *et al.*, "Graph retrieval-augmented generation: A survey," *arXiv preprint arXiv:2408.08921*, 2024.
- [15] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [16] R. Yu, S. Chen, Y. Xie, and X. Jia, "A survey of foundation models for environmental science," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2025, pp. 39–57.
- [17] R. Yu, S. Chen, Y. Xie, H. Yao, J. Willard, and X. Jia, "Foundation models for environmental science: A survey of emerging frontiers," *arXiv preprint arXiv:2504.04280*, 2025.
- [18] R. I. Woolway *et al.*, "Phenological shifts in lake stratification under climate change," *Nature communications*, vol. 12, no. 1, p. 2318, 2021.
- [19] A. P. Appling *et al.*, "Overcoming equifinality: Leveraging long time series for stream metabolism estimation," *Journal of Geophysical Research: Biogeosciences*, 2018.
- [20] 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.
- [21] P. Ravirathinam *et al.*, "Combining satellite and weather data for crop type mapping: An inverse modelling approach," in *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 2024, pp. 445–453.
- [22] R. Yu *et al.*, "Multiple pairwise ranking with implicit feedback," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1727–1730.
- [23] P. C. Wilson, "Water quality notes: Dissolved oxygen: SI1313/ss525, 1/2010," *EDIS*, vol. 2010, no. 2, 2010.
- [24] C. T. Solomon *et al.*, "Ecosystem respiration: drivers of daily variability and background respiration in lakes around the globe," *Limnology and Oceanography*, vol. 58, no. 3, pp. 849–866, 2013.
- [25] J. S. Phillips, "Time-varying responses of lake metabolism to light and temperature," *Limnology and Oceanography*, vol. 65, 2020.
- [26] R. Yu *et al.*, "Physics-guided foundation model for scientific discovery: An application to aquatic science," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, pp. 28 548–28 556, Apr. 2025.
- [27] E. A. Birge, "Gases dissolved in the waters of wisconsin lakes," *Transactions of the American Fisheries Society*, vol. 35, 1906.
- [28] J. S. Read *et al.*, "Derivation of lake mixing and stratification indices from high-resolution lake buoy data," *Environmental Modelling & Software*, vol. 26, no. 11, pp. 1325–1336, 2011.
- [29] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005.
- [30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [31] M. F. Meyer *et al.*, "National-scale remotely sensed lake trophic state from 1984 through 2020," *Scientific Data*, vol. 11, no. 1, p. 77, 2024.
- [32] 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*, 2021.
- [33] F. Kratzert *et al.*, "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets," *Hydrology and Earth System Sciences*, vol. 23, 2019.
- [34] Y. Yu *et al.*, "Pt-lstm: Extending lstm for efficient processing time attributes in time series prediction," in *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021*, 2021, pp. 450–464.
- [35] Y. Liu *et al.*, "itransformer: Inverted transformers are effective for time series forecasting," in *The Twelfth International Conference on Learning Representations*, 2024.
- [36] 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.
- [37] 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*.
- [38] S. Wang *et al.*, "Timemixer: Decomposable multiscale mixing for time series forecasting," in *The Twelfth International Conference on Learning Representations*.
- [39] S. L. Markstrom, *P2S-coupled Simulation with the Precipitation-Runoff Modeling System (PRMS) and the Stream Temperature Network (SNTemp) Models*. US Department of the Interior, US Geological Survey, 2012.
- [40] W. Zhou *et al.*, "Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for us midwestern agroecosystems," *Agricultural and Forest Meteorology*, vol. 307, p. 108521, 2021.
- [41] A. Nielsen, K. Bolding *et al.*, "An open source qgis-based workflow for model application and experimentation with aquatic ecosystems," *Environmental Modelling & Software*, 2017.
- [42] A. B. Janssen *et al.*, "Pelake+: A process-based ecological model to assess the trophic state of stratified and non-stratified freshwater lakes worldwide," *Ecological Modelling*, vol. 396, pp. 23–32, 2019.
- [43] A. B. Janssen, G. B. Arhonditsis *et al.*, "Exploring, exploiting and evolving diversity of aquatic ecosystem models: a community perspective," *Aquatic Ecology*, vol. 49, pp. 513–548, 2015.
- [44] L. Luo, D. Hamilton, J. Lan, C. McBride, and D. Trolle, "Autocalibration of a one-dimensional hydrodynamic-ecological model (dyresm 4.0-caedym 3.1) using a monte carlo approach: simulations of hypoxic events in a polymictic lake," *Geoscientific Model Development*, 2018.
- [45] G. Ghazaryan *et al.*, "Crop yield estimation using multi-source satellite image series and deep learning," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020.
- [46] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020.
- [47] F. Rahmani, K. Lawson *et al.*, "Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data," *Environmental Research Letters*, 2021.
- [48] D. Kochkov *et al.*, "Neural general circulation models for weather and climate," *Nature*, vol. 632, no. 8027, pp. 1060–1066, 2024.
- [49] N. Gupta, M. Sawhney, A. Daw, Y. Lin, and A. Karpatne, "A unified framework for forward and inverse problems in subsurface imaging using latent space translations," *arXiv preprint arXiv:2410.11247*, 2024.
- [50] T. Hey, S. Tansley, K. Tolle *et al.*, *The fourth paradigm: data-intensive scientific discovery*. Microsoft research Redmond, WA, 2009, vol. 1.
- [51] N. A. of Sciences Engineering & Medicine *et al.*, *Improving crop estimates by integrating multiple data sources*. National Academies Press, 2018.
- [52] 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.