Toward Scientific Foundation Models for Aquatic Ecosystems

Abhilash Neog [1] **Medha Sawhney** [1] **Kazi Sajeed Mehrab** [1] **Sepideh Fatemi** [1] **Mary E. Lofton** [1] **Amartya Dutta** [1] **Aanish Pradhan** [1] **Bennett J. McAfee** [2] **Emma Marchisin** [2] **Robert Ladwig** [3] **Arka Daw** [4] **Cayelan C. Carey** [1] **Paul Hanson** [2] **Anuj Karpatne** [1]

Abstract

Understanding and forecasting lake dynamics is essential for monitoring water quality and ecosystem health in lakes and reservoirs. While machine learning models trained on ecological time-series data have shown promise, they tend to be task-specific and struggle with generalization across diverse aquatic environments. Current research is limited to single-lake single-variable models, inconsistent observation frequencies, and a lack of foundation models that can generalize across ecosystems, hindering reproducibility and transferability. To address these challenges, we introduce LAKEFM, a foundation model for lake ecosystems, pre-trained on multi-variable and multi-depth data drawn from a combination of simulated and observational lake datasets. Through empirical results and qualitative analysis, we demonstrate that LAKEFM learns meaningful representations spanning both fine-grained variable-level dynamics and broader lake-level patterns. Furthermore, it achieves competitive—and in some cases superior—forecasting performance compared to existing time-series foundation models

1. Introduction

Limnology, the study of inland aquatic systems such as lakes and reservoirs, focuses on understanding their complex physical-biogeochemical dynamics that evolve over multiple temporal scales and depth layers. With the recent availability of multi-variable, multi-depth data from sensor deployments, ML has shown promise in enabling data-driven prediction of lake dynamics. Physics-guided RNNs (Jia et al., 2018), Modular Compositional Learning (Ladwig et al., 2024) and lake-specific deep networks have improved temperature prediction, but their tight coupling to individual variables and sites hampers transfer to lakes that differ in morphometry, climate, or sampling cadence. However, modeling lake systems at scale remains difficult due to the heterogeneity in variable number and types, and data sparsity across sites, making it hard to develop general-purpose models that transfer well.

At the same time, the broader ML community has made significant progress in developing foundation models that learn task-agnostic representations from large, heterogeneous corpora: CLIP (Radford et al., 2021) for vision–language alignment, Chronos (Ansari et al., 2024) and Moment (Goswami et al., 2024) for generic time-series forecasting, and domain-specific backbones such as PAPAGEI (Pillai et al., 2024) for photoplethysmography signals. In contrast, limnology still lacks an analogous model capable of unifying multiple lakes and variables observed with irregular frequencies and depths, leaving cross-ecosystem synthesis as an open challenge. Moreover, most generic TS foundation models either focus solely on univariate signals or assume clean, densely sampled data—assumptions that are rarely valid in limnology, where data is multivariate and are inherently sparse across both time and depth dimensions. While recent efforts such as PGFM (Yu et al., 2025) have begun exploring foundation models for lake systems, they remain limited in scope, being restricted to a small number of variables and lacking the ability to generalize across diverse measurement depths.

Motivated by this gap, we ask the following questions. *(a) Can we build a single model that can capture generic lake processes, encompassing multiple lake ecosystems and variables, while retaining site-specific nuances? (b) Can treating scientific variables (temperature, chlorophyll, oxygen, . . .) as tokens reveal their functional relationships and potentially be applicable as feature extractors for understanding more complex dynamical systems? (c) Can we encode lake characteristics that reveal novel insights about the structure of ecosystems?* To answer these questions, we introduce LAKEFM, a foundation model pre-trained on simulated as well as observed lakes with irregular, multi-depth records. LAKEFM flattens each variable–depth pair into

[1]Virginia Tech [2]University of Wisconsin-Madison [3]Aarhus University [4]Oak Ridge National Lab. Correspondence to: Abhilash Neog <abhilash22@vt.edu>, Anuj Karpatne <karpatne@vt.edu>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

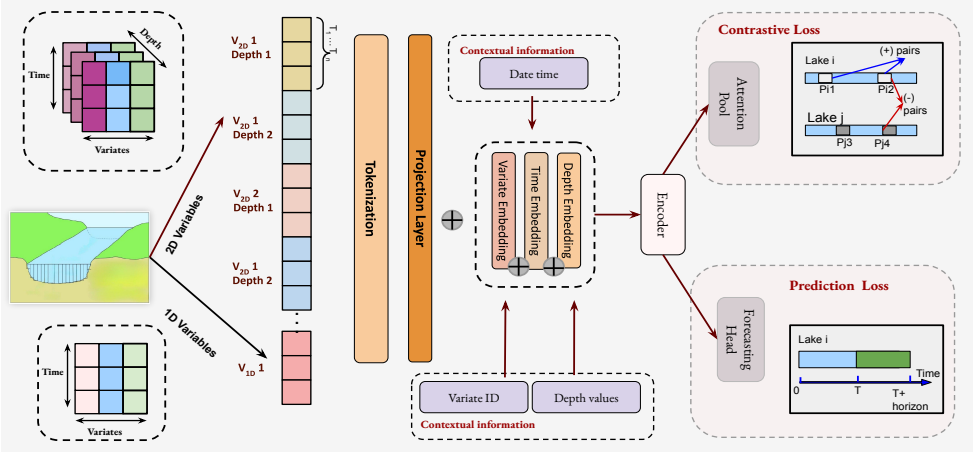


Figure 1. Overview of the proposed LAKEFM model.

a token sequence and learns representations via multi-step forecasting loss, augmented by a weighted contrastive term that encourages—but does not force—samples from the same lake to align. Overall, LAKEFM attempts to establish a practical step towards scalable and generalizable modeling of lake ecosystems. Our main contributions are as follows.

1. A unified pre-training framework that can ingest multi-variable, multi-depth lake observations and produce generalizable representations, enabling zero-shot transfer to unseen lakes and improving performance on downstream ecological forecasting tasks.
2. Learning variable-aware embeddings that capture the semantic roles of physical and bio-geochemical drivers, in contrast to existing time-series foundation models that treat input variables as unstructured features. By learning representations grounded in variable identity and behavior, the model opens up pathways for interpretability, enabling insights into variable interactions
3. Learning lake-level embeddings that capture site-specific characteristics, enabling discovery of shared patterns and analyzing lake similarity and clustering.

2. Methodology

Background and Notations. Let $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ denote a collection of N lakes, where each lake $\mathcal{D}_i$ contains a multivariate, multi-depth time series: $\mathcal{D}_i = \left\{(\mathbf{x}_t^{(i)}, \mathbf{m}_t^{(i)}, \ell_i)\right\}_{t=1}^{T_i}$, where $\mathbf{x}_t^{(i)} \in \mathbb{R}^{V \times D}$ represents observations of V scientific variables (e.g., temperature, oxygen) at D depths for timestep t in lake i, and $\mathbf{m}_t^{(i)} \in \{0,1\}^{V \times D}$ is a binary mask indicating missing values. ℓ_i denotes the lake identifier, used for contrastive training. Time intervals are irregular and vary across lakes. We define an encoder f_θ that maps a context window of L timesteps into a latent representation: $\mathbf{z}_i = f_\theta\left(\{\mathbf{x}_t^{(i)}\}_{t=1}^{L}\right), \quad \mathbf{z}_i \in \mathbb{R}^d$, where d is the dimension of the learned embedding. Now, given a context window of L timesteps from a time series $\{\mathbf{x}_t^{(i)}\}_{t=t_0-L}^{t_0-1}$, the **forecasting task** aims to predict the next H steps: $\hat{\mathbf{x}}_{t_0:t_0+H-1}^{(i)}$. We optimize the model to minimize the mean squared error (MSE) between predictions and observed values, $\mathcal{L}_{\text{forecast}} = \frac{1}{H}\sum_{h=0}^{H-1}\left\|\hat{\mathbf{x}}_{t_0+h}^{(i)} - \mathbf{x}_{t_0+h}^{(i)}\right\|_2^2$.

2.1. Model Architecture

As illustrated in Figure 3, LAKEFM is built upon a masked transformer encoder, drawing inspiration from the MOIRAI-style modeling paradigm (Woo et al.). The architecture is composed of three key components: (i) contextual/metadata embeddings, (ii) a transformer-based encoder, and (iii) dual task-specific heads for forecasting and clustering. The transformer encoder incorporates a binary attention bias (Woo et al.) to differentiate intra- and inter-variate interactions, enabling it to learn structured attention patterns across variables. For positional encoding, we adopt Rotary Position Embeddings (RoPE) (Su et al., 2024) to model relative temporal dependencies. The encoder output is fed into two parallel heads: (i) a forecasting head, which applies a feed-forward network over each context length L to predict future values, and (ii) an attention pooling head, which aggregates the encoded sequence into a fixed-length representation for contrastive learning. The pooled representation captures the point-level summary of the window and serves as a lake-specific embedding for representation-level objectives.

Input Representation. The input consists of spatiotemporal sequences over a context window $\{t_o - L, \ldots, t_o - 1\}$, where L denotes the lookback window length. At each time step t, we observe a set of two-dimensional (depth-varying) lake variables $\mathbf{x}_t^{(2D)} \in \mathbb{R}^{V_{2D} \times D}$ and one-dimensional meteorological drivers $\mathbf{x}_t^{(1D)} \in \mathbb{R}^{V_{1D}}$, where V_{2D} and V_{1D} denote the number of variables in each group, and D is the number of depth levels. To unify these heterogeneous

signals, we flatten each $\mathbf{x}_t^{(2D)}$ into a sequence of $V_{2D} \times D$ tokens and each $\mathbf{x}_t^{(1D)}$ into V_{1D} tokens, resulting in a total of $S = L \cdot (V_{2D} \cdot D + V_{1D})$ tokens per input sequence,

$$\mathbf{x} = \left[\mathbf{x}_{t,d}^{(2D,v)} \mid v \in V_{2D},\ d \in D,\ t \in [t_o - L, t_o - 1]\right] \cup \left[\mathbf{x}_t^{(1D,v)} \mid v \in V_{1D},\ t \in [t_o - L, t_o - 1]\right]$$

After flattening, the two-dimensional lake variables and one-dimensional meteorological drivers are combined into a single unified sequence. Instead of using separate encoders, we model them jointly through a shared transformer encoder to capture their inter-dependencies—meteorological drivers often influence lake dynamics, and decoupling their encoding would ignore important interactions.

Contextual Information. Each token in the sequence is enriched with contextual embeddings: variable (learned from a fixed vocabulary, akin to word embeddings in language), depth (via Fourier feature projections), and time (using sinusoidal embeddings). Specifically, depth embeddings are generated using Fourier feature encoding, where each scalar depth d is projected to a vector of sinusoidal components. Specifically, we apply K frequency bands to produce $[\sin(\omega_1 d), \cos(\omega_1 d), \ldots, \sin(\omega_K d), \cos(\omega_K d)]$, where $\omega_k = \frac{2^k \pi}{\texttt{max\_resolution}}$ for $k = 0, \ldots, K-1$ frequency bands and `max_resolution` is the max value of input used to scale frequencies. Optionally, the raw input d can be prepended to the encoding. Time embeddings are constructed using 2D sinusoidal features derived from the month-of-year index, offering a lightweight, parameter-free encoding of seasonal (here, monthly) patterns.

Rather than summing these embeddings with the input token representation, we concatenate them, $\mathbf{e}_i = [\mathbf{x}_i \parallel \mathbf{v}_i \parallel \mathbf{d}_i \parallel \mathbf{t}_i]$, where $\mathbf{x}_i$ is the raw token embedding and $\mathbf{v}_i$, $\mathbf{d}_i$, and $\mathbf{t}_i$ are the variable, depth, and time embeddings respectively. Empirically, we find that concatenation preserves the semantic distinction between different embedding types and allows the model to attend over heterogeneous subspaces independently—whereas summation tends to blur these roles in a shared latent space.

2.2. Pre-training

LAKEFM is pre-trained to optimize two tasks/objectives - prediction/forecasting loss and contrastive loss. In the first case, given a context window $\{\mathbf{x}_t\}_{t=1}^{L}$, we aim to predict the next H steps. The objective is to minimize the prediction loss (i.e., MSE), $\mathcal{L}_{\text{forecast}} = \sum_{h=1}^{H} \|\hat{\mathbf{x}}_{t+h} - \mathbf{x}_{t+h}\|_2^2$

To encourage lake-specific representations, we adopt a hard contrastive learning objective. Given a batch of B samples with corresponding representations $\{\mathbf{z}_1, \ldots, \mathbf{z}_B\}$ and lake identifiers $\{\ell_1, \ldots, \ell_B\}$, we treat samples from the same lake as positives and those from different lakes as negatives. Each representation is ℓ_2-normalized, and the contrastive loss is computed using the standard InfoNCE (Oord et al., 2018) formulation (here, τ is a temperature hyperparameter),

$$\mathcal{L}^{(i)} = -\sum_j w_{ij} \left(\frac{z_i^\top z_j}{\tau} - \log \sum_k \exp(z_i^\top z_k / \tau) \right) / \sum_j w_{ij}, \quad i = 1, \ldots, B$$

$$\mathcal{L}_{\text{contrast}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}^{(i)}.$$

The final pretraining objective combines forecasting and contrastive learning: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forecast}} + \lambda \mathcal{L}_{\text{contrast}}$, where λ balances the weight of contrastive loss.

Table 1. MSE comparison on in-distribution LakeBeD-US data. Best performance is shown in **bold**. Second-best performance is shown in underline.

Lake	Baseline	Water_DO_mg_per_L	WaterTemp_C	Water_Secchi_m	par	Inflow_cms	Lake_MSE
BARC	Chronos	2.3253	1.4375	1.6224	–	–	1.8578
	LPTM	2.2901	1.4458	1.5937	–	–	1.843
	MOMENT	5.0759	1.7661	2.0752	–	–	3.2987
	LakeFM	**1.1866**	**1.0513**	**1.193**	–	–	**1.1257**
BM	Chronos	1.0758	1.1338	1.4098	1.3189	–	1.1941
	LPTM	**0.853**	1.053	1.3402	1.1891	–	1.0555
	MOMENT	0.8765	1.0748	1.3271	1.1827	–	1.0664
	LakeFM	1.0384	**1.0254**	**1.0652**	**1.055**	–	**1.0414**
LIRO	Chronos	1.2263	2.7699	1.5377	–	–	1.9562
	LPTM	**0.7012**	3.1816	2.0241	–	–	1.9489
	MOMENT	38.3198	13.5536	4.4672	–	–	23.9849
	LakeFM	1.3815	**1.1645**	**1.4198**	–	–	**1.2859**
SUGG	Chronos	1.2379	1.7081	1.3756	–	–	1.4622
	LPTM	1.0839	1.4955	1.1533	–	–	1.2746
	MOMENT	3.4189	1.8863	1.1905	–	–	2.4902
	LakeFM	**1.1008**	**1.0316**	**0.9157**	–	–	**0.6143**
TOOK	Chronos	0.9848	1.4319	–	–	1.2535	1.2148
	LPTM	1.1142	1.6768	–	–	1.2604	1.3762
	MOMENT	1.3364	7.518	–	–	3.6585	4.3174
	LakeFM	**1.0208**	**1.0263**	–	–	**1.1086**	**1.0435**

3. Experiments

We train our model using both simulated and real-world data, but evaluate primarily on the latter. The real-world dataset comprises 21 lakes from LakeBeD-US (McAfee et al., 2025). Additional details on the datasets and experimental setup are provided in Appendix B, with implementation specifics in Appendix C.

We evaluate on two test settings: (a) **In-distribution**, and (b) **Leave-out set**. In the in-distribution setting, each model is evaluated on lakes for which a portion of the data was seen during training, with the remaining held out for testing. As shown in Table 1 (full results in Appendix E), LAKEFM consistently achieves the lowest MSE across most lake-variable combinations, outperforming Chronos (Ansari et al., 2024), LPTM (Prabhakar Kamarthi & Prakash, 2024), and MOMENT (Goswami et al., 2024) on lakes "seen" during training.

In the leave-out setting, we withhold five lakes entirely from training and evaluate models in a zero-shot manner. Despite this stringent challenge, LAKEFM maintains competitive accuracy relative to the best-performing baselines. Table 2 highlights LAKEFM's ability to generalize across diverse lake systems.

Table 2. MSE comparison on leave-out LakeBeD-US data. Best performance is shown in **bold**. Second-best performance is shown in underline.

Lake	Baseline	Water_DO_mg_per_L	WaterTemp_C	Water_Secchi_m	par	Lake_MSE
AL	**Chronos**	0.9313	**0.6046**	1.0721	1.1562	0.9516
	LPTM	**0.8908**	0.7019	1.0098	1.1082	**0.94**
	MOMENT	1.2584	0.8964	0.9977	1.3618	1.1741
	LakeFM	1.0049	1.0258	**0.9482**	**0.9797**	0.9942
BVR	**Chronos**	1.3908	1.3591	1.8854	–	1.4521
	LPTM	1.5393	0.6034	1.5801	–	**0.9221**
	MOMENT	1.5929	**0.5437**	7.9697	–	1.9562
	LakeFM	**1.0053**	0.6703	**1.0792**	–	1.011
CRAM	**Chronos**	0.9715	0.9187	1.3632	–	0.9831
	LPTM	**0.8866**	**0.624**	1.0824	–	**0.7851**
	MOMENT	2.7135	0.7601	**0.9776**	–	1.6678
	LakeFM	1.0033	0.9173	1.1038	–	0.9733
FI	**Chronos**	1.1408	1.0244	1.1578	–	1.0894
	LPTM	1.3029	**0.7445**	**0.9595**	–	**1.0179**
	MOMENT	1.2853	0.9222	1.1916	–	1.1117
	LakeFM	**1.0645**	1.1013	1.0759	–	1.082
MO	**Chronos**	1.1404	**0.8638**	1.1849	–	1.0282
	LPTM	1.0748	0.8829	1.1997	–	**1.0104**
	MOMENT	1.3892	0.8747	1.3523	–	1.1634
	LakeFM	**1.0451**	1.0776	**1.0712**	–	1.0627

3.1. Ablations

3.1.1. Pretraining Strategies

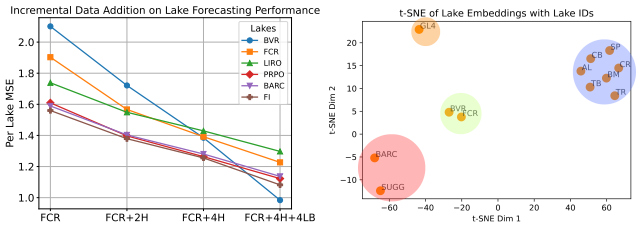
We conduct an ablation study to compare three different pretraining strategies (see Table 3). First, simulation-only pretraining ($\text{LakeFM}_{\text{SimOnly}}$) trains exclusively on synthetic Hanson and FCR datasets and is evaluated "zero-shot" on LakeBeD US, this approach yields moderate MSEs but struggles to fully bridge the simulation to real world domain gap. Second, Sim $\rightarrow$ Real Fine-tune ($\text{LakeFM}_{\text{Sim2RealFT}}$) first pretrains on the same simulations and then fine-tunes on real LakeBed measurements; by adapting to real-world variability, it achieves a substantial reduction in error compared to simulation-only. Finally, Joint Sim+Real (CL) ($\text{LakeFM}_{\text{JointCL}}$) trains simultaneously on both simulated and real data using a contrastive loss to align their representations; this approach yields the lowest MSEs of all three and were used for all the LakeFM results on this paper. Together, these results demonstrate that while simulation-only pretraining provides a useful initialization, incorporating real observations significantly enhances predictive performance on LakeBeD.

3.1.2. Incremental Inference

Figure 2a shows an incremental-inference ablation on LakeBeD, quantifying how progressive expansions of the training set affect per-lake MSE. We begin with a model trained exclusively on FCR data and then fine-tune it by adding two Hanson lakes (FCR + 2 Hanson). Next, we incorporate all four Hanson lakes (FCR + 4 Hanson) before finally introducing four LakeBed lakes (FCR + 4 Hanson + 2 LakeBed). Each augmentation yields a consistent reduction in MSE, with the largest drop occurring upon the initial inclusion of Hanson data. Subsequent gains from adding more Hanson data and real LakeBed observations are smaller but still meaningful, demonstrating that progressively enriching the training corpus steadily enhances predictive accuracy.

Table 3. Mean squared error (MSE) across five lakes for three different LakeFM pretraining strategies: Simulation-only, Sim→Real fine-tuning, and Joint Sim+Real (contrastive)

Pretraining Strategy	AL	FCR	TOOK	SP	GL4
$\textbf{LakeFM}_{\textbf{SimOnly}}$	1.5626	1.4180	1.5973	1.5167	1.5691
$\textbf{LakeFM}_{\textbf{Sim2RealFT}}$	1.0065	1.1137	1.1026	1.0644	1.2081
$\textbf{LakeFM}_{\textbf{JointCL}}$	0.9942	1.0889	**1.0435**	**1.0284**	**1.1704**



(a) Incremental training data (b) Lake embeddings t-SNE

Figure 2. (a) Improvement in lake forecasting performance upon incrementally increasing training data. (b) Lake embedding clusters learned by the model. Red: Lakes in Florida; Green: Lakes in Virginia; Blue: Lakes in Wisconsin; Orange: Lakes in Colorado.

3.1.3. Insight on Lake Clustering

We visualize the learned lake-level representations using t-SNE in Figure 2b. The embeddings reveal some interesting and clear spatial structure, with lakes from similar geographic regions forming distinct clusters. This suggests that the model is able to capture meaningful lake-specific characteristics and encode latent similarities driven by regional climate, morphology, or variable dynamics—even though geographic information was not explicitly provided during training. These emergent clusters demonstrate the model's potential for cross-site generalization and transfer across ecosystems.

4. Conclusion

In this work, we introduced LakeFM, a foundation model for lake ecosystems that learns generalizable representations from multi-variable, multi-depth time-series data across thousands of lakes. By unifying variable-level semantics and site-level dynamics within a single framework, LakeFM enables zero-shot transfer to unseen lakes and improves downstream ecological forecasting. A key limitation in this domain lies in the sparsity and limited scale of available ecological observations—both in temporal coverage and variable diversity. While our model is designed to inherently handle sparse inputs, the performance continues to improve with data volume, suggesting that larger, more comprehensive datasets could yield even stronger foundation models. Future work could explore pretraining on large simulation datasets and further leveraging the learned variable embeddings for scientific discovery and interpretability.

Acknowledgments

We thank the Advanced Research Computing Center at Virginia Tech and the Pittsburgh Supercomputing Center for GPU resources. This work was supported in part by NSF awards IIS-2239328 and DEB-2213550. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).

References

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., and Li, D. Contiformer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, 36:47143–47175, 2023.

Du, W., Côté, D., and Liu, Y. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, June 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.119619. URL http://dx.doi.org/10.1016/j.eswa.2023.119619.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Hanson, P. C., Ladwig, R., Buelo, C., Albright, E. A., Delany, A. D., and Carey, C. C. Legacy phosphorus and ecosystem memory control future water quality in a eutrophic lake. *Journal of Geophysical Research: Biogeosciences*, 128(12):e2023JG007620, 2023. doi: 10.1029/2023JG007620.

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., Hanson, P. C., Read, J. S., de Sousa, E., Weber, M., and Winslow, L. A. 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*, 12 (1):473–523, 2019.

Jia, X., Karpatne, A., Willard, J., Steinbach, M., Read, J., Hanson, P. C., Dugan, H. A., and Kumar, V. Physics guided recurrent neural networks for modeling dynamical systems: Application to monitoring water temperature and quality in lakes. *arXiv preprint arXiv:1810.02880*, 2018.

Ladwig, R., Daw, A., Albright, E. A., Buelo, C., Karpatne, A., Meyer, M. F., Neog, A., Hanson, P. C., and Dugan, H. A. Modular compositional learning improves 1d hydrodynamic lake model performance by merging process-based modeling with deep learning. *Journal of Advances in Modeling Earth Systems*, 16(1):e2023MS003953, 2024.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

McAfee, B. J., Pradhan, A., Neog, A., Fatemi, S., Hensley, R. T., Lofton, M. E., Karpatne, A., Carey, C. C., and Hanson, P. C. Lakebed-us: a benchmark dataset for lake water quality time series and vertical profiles. *Earth System Science Data Discussions*, 2025:1–43, 2025.

Neog, A., Daw, A., Khorasgani, S. F., and Karpatne, A. Masking the gaps: An imputation-free approach to time series modeling with missing data. *arXiv preprint arXiv:2502.15785*, 2025.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pillai, A., Spathis, D., Kawsar, F., and Malekzadeh, M. Papagei: Open foundation models for optical physiological signals. *arXiv preprint arXiv:2410.20542*, 2024.

Prabhakar Kamarthi, H. and Prakash, B. A. Large pre-trained time series models for cross-domain time series analysis tasks. *Advances in Neural Information Processing Systems*, 37:56190–56214, 2024.

Pradhan, A., McAfee, B. J., Neog, A., Fatemi, S., Lofton, M. E., Carey, C. C., Karpatne, A., and Hanson, P. C. LakeBeD-US: Computer Science Edition - a benchmark dataset for lake water quality time series and vertical profiles, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shukla, S. N. and Marlin, B. M. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers. arxiv 2024. *arXiv preprint arXiv:2402.02592*.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, 117(D3), 2012. ISSN 2156-2202. doi: 10.1029/2011JD016048.

Yu, R., Qiu, C., Ladwig, R., Hanson, P., Xie, Y., and Jia, X. Physics-guided foundation model for scientific discovery: An application to aquatic science. *arXiv preprint arXiv:2502.06084*, 2025.

A. Related Works and Discussion

Time-series forecasting models, including statistical approaches and deep learning architectures such as PatchTST (Nie et al., 2022), iTransformer (Liu et al., 2023), have shown strong performance on benchmark datasets. However, these models are typically domain- or dataset-specific. As a result, they struggle to generalize across ecosystems or variable configurations, limiting their applicability in scientific domains with high structural heterogeneity. Scientific datasets, particularly in ecology and environmental modeling, introduce unique challenges: missing values, irregular sampling, and multi-resolution measurements across time and depth. Models like mTAN (Shukla & Marlin, 2021) and ContiFormer (Chen et al., 2023) attempt to address these issues through neural ODEs, temporal embeddings, or attention over irregular grids, however, these methods are often task-specific, rely on carefully engineered architectures, and do not scale well to large multi-lake or multi-variable ecosystems. While MissTSM (Neog et al., 2025) provide a model agnostic approach to handle missing values, it is not very computationally scalable. In contrast, our approach incorporates simulation and real-world data by flattening multivariate, multi-depth signals into a unified representation, facilitating model training under partial observations while maintaining generalizability.

Recent Time Series Foundation Models (TSFM) aim to generalize across diverse time-series tasks by learning from large corpora of univariate or multivariate signals. However, univariate TSFMs lack the capacity to model inter-variable interactions, while multivariate TSFMs tend to treat each variable—including depth-specific versions—as independent features, ignoring structured dependencies such as how a single variable (e.g., temperature) behaves across the vertical depth column of a lake. Moreover, existing TSFMs generally lack variable semantics—they do not encode or exploit the meaning or identity of each variable, reducing interpretability and weakening scientific generalization.

LakeFM is designed to address these limitations by explicitly encoding the behavior of each variable across temporal and depth dimensions. Instead of relying purely on temporal pattern recognition, LakeFM attempts to learns semantically meaningful variable embeddings that capture how scientific signals evolve across lake ecosystems. This allows the model to align its predictions for a variable from any lake with the learned behavior of that variable across all lakes. Additionally, LakeFM incorporates contrastive pretraining across lake systems, enabling two levels of alignment: *(a) First, inter-lake alignment, where the model learns to associate a test lake with similar ecosystems based on observed dynamics; (b) Second, intra-variable alignment, where predictions for each variable are grounded in its globally learned behavior.*

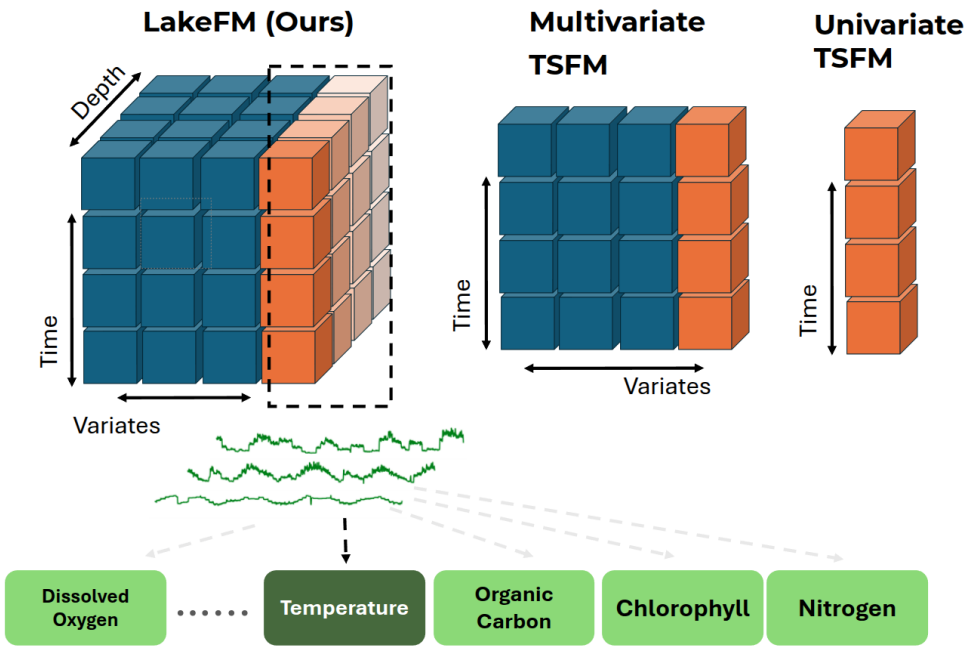


Figure 3. Understanding the differences between LAKEFM and Time Series Foundation Models

B. Dataset Description

We pre-train and evaluate LAKEFM on three complementary datasets that together span both observed and process-based simulated lake dynamics. We use the first 80% of each dataset for training. For evaluation, 20% of the LakeBeD-US dataset and 10% of each WQHansonSim and FcrSimPhy datasets are held out as test data. To assess out-of-distribution (OOD) generalization, we exclude 5 lakes from the LakeBeD-US dataset entirely during training and use them as an unseen test set. Each dataset contributes unique strengths to the modeling framework, as described below.

B.1. LakeBeD-US

Our primary observational dataset is LakeBeD-US (McAfee et al., 2025; Pradhan et al., 2024), consisting of over 500 million unique lake water quality observations collected between 1981 and 2024. The data span 21 U.S. lakes and include both high- and low-frequency measurements. The dataset features 17 variables organized into three categories: (1) static attributes, such as lake morphology and geographic location; (2) one-dimensional (1D) variables that vary over time (e.g., Secchi depth, inflow); and (3) two-dimensional (2D) variables that vary over both time and depth. This rich observational dataset captures diverse temporal and spatial lake dynamics.

B.2. WQHansonSim simulation

The WQHansonSim dataset is a synthetic lake water quality simulation covering four lakes: Green Lake, Lake Mendota, Prairie Lake, and Trout Lake. The synthetic data were created using a process-based water quality model (Hanson et al., 2023) driven by meteorological forcing data from the second phase of the North American Land Data Assimilation System (NLDAS-2; Xia et al., 2012). Each simulation underwent a 60-year burn-in period to allow slow-changing ecosystem states to reach dynamic equilibrium, followed by a 20-year simulation period. The outputs are structured as daily time series, with each row representing a unique date-depth combination.

Each record includes six core water quality variables: water temperature, dissolved oxygen, dissolved organic carbon, particulate organic carbon, total phosphorus, and depth, alongside the corresponding date. Depths are lake-specific and selected to reflect stratification layers, representing both the epilimnion and hypolimnion (e.g., 5 m and 23 m for Trout Lake)—allowing for realistic modeling of thermal and chemical compositions among layers of the lake.

B.3. FcrSimPhy: simulations at Falling Creek Reservoir

The FcrSimPhy dataset was generated using the General Lake Model coupled with the AED water quality module (GLM-AED; Hipsey et al., 2019), and comprises 1,000 process-based model runs at Falling Creek Reservoir (FCR), VA, spanning daily resolution from December 1, 2016, to December 31, 2020. Each run represents a distinct ecological scenario defined by a unique set of phytoplankton trait parameters, sampled using Latin hypercube sampling. Six parameters were varied across three phytoplankton groups—cyanobacteria, green algae, and diatoms—including group-specific growth rates and sinking rates. Model outputs include five key water quality variables: water temperature, soluble reactive phosphorus (SRP), dissolved inorganic nitrogen (DIN), chlorophyll-a (Chla), and the light attenuation coefficient (Kd). These are reported at seven depths (0.1, 1.6, 3.8, 5, 6.2, 8, and 9 m), corresponding to observational depths in FCR. Additionally, meteorological driver variables (e.g., AirTemp, Shortwave, Inflow) are included. Each row represents a specific date and depth, enabling detailed analysis of how phytoplankton trait variation influences ecosystem dynamics, particularly nutrient-light-temperature interactions and emergent biogeochemical patterns.

C. Implementation details

C.1. LakeFM

LakeFM employs a transformer encoder with 6 layers of grouped-query self-attention, each having 4 attention heads. The model's hidden dimensiona (d_model) is set to 128. Embedding dimensions are set to 128, 32, and 16 for the variate, depth, and input features, respectively. Temporal information is encoded using a 2-dimensional embedding. For grouped query attention we use a group size of 4. Dropout is applied to the attention heads with a rate of 0.02, while the overall model dropout rate is set to 0.0. The feedforward network dimension is set to 2048 for the attention layers, and the SwigLU activation function is used for the feedforward networks. Rotary Positional Embedding (RoPE) is used to incorporate relative positional encodings. The model utilizes a scalar tokenization strategy with a patch size of 3. For contrastive learning, the

projection dimension is set to 64, and attention pooling is used. During training, we implement a warmup phase with 10,000 warmup steps.

Hyperparameter tuning We perform hyper-parameter sweeps involving the following parameters: *enc_layers*, *num_heads*, *weight_decay*, *warmup_iterations*, *embed_dim*, *attention_dropout*, *head_dropout*, *variate_embed_dim*, *depth_embed_dim*, *contrastive_loss_weight*. These were optimized using a validation split derived from the WQHansonSim simulation dataset.

Contrastive Sampling Strategy. We adopt a custom balanced sampling strategy, built on top of PyTorch's DistributedSampler, to construct batches for contrastive pretraining. Each batch consists of multiple anchor-positive groups, where each anchor is paired with `P_pos = 4` positive samples from the same lake. For e.g., for a total `batch_size = 64`, this allows up to 12 such anchor-positive sets per batch, with the remaining slots filled by negative samples drawn from different lakes. Positive and negative pools are precomputed per lake for efficiency, and sampling is performed with deterministic seeding to support reproducibility across distributed processes. This sampling strategy ensures within-lake similarity and across-lake contrast, enabling the model to learn lake-discriminative representations.

Hardware. We use a combination of NVIDIA H100 and A100 GPUs for pretraining and carrying out the experiments

C.2. Baselines

For our baselines, we evaluate the zero-shot forecasting performance of three well-established time-series foundation models: Chronos (Ansari et al., 2024), MOMENT (Goswami et al., 2024), and LPTM (Prabhakar Kamarthi & Prakash, 2024). Our implementation leverages the Samay Time-series Foundational Models Library for Python (Prabhakar Kamarthi & Prakash, 2024). For Chronos, we use the *amazon/chronos-t5-small* variant. For MOMENT, we use the *AutonLab/MOMENT-1-large* variant. For all models, we use a context length of 42, a prediction length of 21, and a stride of 1. Prior to feeding the data into the models, we standardize each attribute in our datasets to ensure consistent scaling across all features.

Since the baseline methods cannot operate on sparse, non-imputed data, we first impute all missing entries in the LakeBeD dataset using SAITS (Du et al., 2023), a self-attention–based imputation model, so that each baseline receives a fully dense time series for evaluation.

D. Ecological Variables Modeled by LAKEFM

Table 4. Overview of available 2D and 1D variables for each lake across all datasets that forms the vocabulary of LAKEFM. In addition to the variables shown in this table, WQHansonSim also includes the following 1D variables that are modeled by LAKEFM: Longwave, Elevation, Precipitation, Discharge, and TOC.

Dataset	Lake ID	Chl a	DOC	DO	DRP	NO3	POC	PAR	TP	Temp	DIN	Kd	Inflow	Secchi	Air Temp	Shortwave
LakeBedUS	AL	✓		✓				✓		✓				✓		
	BVR		✓	✓	✓				✓	✓				✓		
	CRAM			✓						✓				✓		
	FI			✓						✓				✓		
	MO			✓						✓				✓		
	BARC			✓						✓				✓		
	BM	✓		✓				✓		✓				✓		
	CB	✓		✓				✓		✓				✓		
	CR	✓		✓				✓		✓				✓		
	FCR		✓	✓	✓				✓	✓				✓		
	GL4			✓		✓		✓		✓				✓		
	LIRO			✓						✓				✓		
	ME			✓						✓				✓		
	PRLA			✓						✓				✓		
	PRPO			✓						✓				✓		
	SP	✓		✓				✓		✓				✓		
	SUGG			✓						✓				✓		
	TB	✓		✓				✓		✓				✓		
	TOOK			✓						✓			✓	✓		
	TR			✓				✓		✓				✓		
	WI			✓						✓				✓		
WQHansonSim	All		✓	✓			✓		✓	✓				✓	✓	✓
FcrSimPhy	All	✓								✓	✓	✓	✓		✓	✓

E. Full Results

Table 5. Performance comparison on in-distribution LakeBeD-US data

Lake	Baseline	Water_DO_mg_per_L	SRP_ugL	WaterTemp_C	Water_TP_mg_per_L	Water_Secchi_m	par	Inflow_cms	no3	Lake_MSE
BARC	**Chronos**	2.3253	–	1.4375	–	1.6224	–	–	–	1.8578
	LPTM	2.2901	–	1.4458	–	1.5937	–	–	–	1.843
	MOMENT	5.0759	–	1.7661	–	2.0752	–	–	–	3.2987
	LakeFM	1.1866	–	1.0513	–	1.193	–	–	–	1.1257
BM	**Chronos**	1.0758	–	1.1338	–	1.4098	1.3189	–	–	1.1941
	LPTM	0.853	–	1.053	–	1.3402	1.1891	–	–	1.0555
	MOMENT	0.8765	–	1.0748	–	1.3271	1.1827	–	–	1.0664
	LakeFM	1.0384	–	1.0254	–	1.0652	1.055	–	–	1.0414
CB	**Chronos**	1.3767	–	1.5903	–	1.5466	1.5991	–	–	1.5476
	LPTM	1.0053	–	0.9939	–	1.0003	1.0173	–	–	1.0085
	MOMENT	1.0087	–	1.0378	–	1.2727	1.192	–	–	1.1326
	LakeFM	1.028	–	1.0268	–	1.0872	1.0382	–	–	1.0388
CR	**Chronos**	1.4407	–	1.4127	–	1.327	1.2205	–	–	1.3556
	LPTM	0.8465	–	0.8389	–	1.0045	0.941	–	–	0.8854
	MOMENT	0.7886	–	0.8602	–	0.89	0.8525	–	–	0.8381
	LakeFM	1.0514	–	1.0302	–	1.0692	1.0211	–	–	1.0372
FCR	**Chronos**	0.8032	1.2797	0.7211	1.0751	1.2216	–	–	–	0.9464
	LPTM	1.07	1.0743	0.7634	0.8943	0.9704	–	–	–	0.9437
	MOMENT	1.0464	1.8066	0.7187	1.7656	1.449	–	–	–	1.2216
	LakeFM	1.2669	1.0639	1.2711	1.0788	1.099	–	–	–	1.0889
GL4	**Chronos**	1.5296	–	1.411	–	1.4306	–	–	2.0672	1.6543
	LPTM	1.4388	–	1.2985	–	1.0462	–	–	1.5282	1.3984
	MOMENT	7.8317	–	2.0117	–	1.3582	–	–	10.7023	6.5054
	LakeFM	1.2302	–	1.1901	–	0.985	–	–	1.1321	1.1704
LIRO	**Chronos**	1.2263	–	2.7699	–	1.5377	–	–	–	1.9562
	LPTM	0.7012	–	3.1816	–	2.0241	–	–	–	1.9489
	MOMENT	38.3198	–	13.5536	–	4.4672	–	–	–	23.9849
	LakeFM	1.3815	–	1.1645	–	1.4198	–	–	–	1.2859
ME	**Chronos**	1.3861	–	1.4473	–	1.1576	–	–	–	1.3797
	LPTM	0.9472	–	0.9141	–	1.0349	–	–	–	0.9455
	MOMENT	0.845	–	0.8544	–	1.3372	–	–	–	0.9194
	LakeFM	1.037	–	1.0627	–	1.085	–	–	–	1.0548
PRLA	Chronos	1.3852	–	0.9938	–	1.5742	–	–	–	1.2245
	LPTM	1.2558	–	0.9893	–	0.9589	–	–	–	1.1077
	MOMENT	2.4559	–	1.1458	–	0.9971	–	–	–	1.7278
	LakeFM	1.0034	–	1.0244	–	0.8489	–	–	–	0.9998
PRPO	**Chronos**	1.3668	–	0.875	–	2.2169	–	–	–	1.2426
	LPTM	1.1841	–	1.0609	–	1.1318	–	–	–	1.1235
	MOMENT	2.9214	–	1.3513	–	0.906	–	–	–	1.9996
	LakeFM	1.1834	–	1.0691	–	1.0924	–	–	–	1.1221
SP	**Chronos**	1.2051	–	1.1932	–	1.2328	1.2574	–	–	1.2197
	LPTM	0.9903	–	0.9431	–	1.0868	1.1291	–	–	1.0259
	MOMENT	1.014	–	0.9876	–	1.0706	1.004	–	–	1.0072
	LakeFM	1.0402	–	1.0276	–	1.0384	1.0143	–	–	1.0284
SUGG	**Chronos**	1.2379	–	1.7081	–	1.3756	–	–	–	1.4622
	LPTM	1.0839	–	1.4955	–	1.1533	–	–	–	1.2746
	MOMENT	3.4189	–	1.8863	–	1.1905	–	–	–	2.4902
	LakeFM	0.1008	–	1.0316	–	0.9157	–	–	–	0.6143
TB	**Chronos**	1.0906	–	0.8741	–	1.0942	1.2872	–	–	1.1459
	LPTM	1.0215	–	0.9735	–	1.0175	1.2697	–	–	1.1356
	MOMENT	0.996	–	1.0177	–	1.1382	1.245	–	–	1.139
	LakeFM	1.0232	–	1.0157	–	1.0692	1.0018	–	–	1.0173
TOOK	**Chronos**	0.9848	–	1.4319	–	–	–	1.2535	–	1.2148
	LPTM	1.1142	–	1.6768	–	–	–	1.2604	–	1.3762
	MOMENT	1.3364	–	7.518	–	–	–	3.6585	–	4.3174
	LakeFM	1.0208	–	1.0263	–	–	–	1.1086	–	1.0435
TR	Chronos	1.2538	–	1.32	–	1.1681	1.0741	–	–	1.2123
	LPTM	1.0031	–	0.9493	–	1.0116	0.958	–	–	0.9733
	MOMENT	1.1394	–	0.978	–	1.1833	0.8093	–	–	0.9915
	LakeFM	1.0403	–	1.0327	–	0.9996	0.9649	–	–	1.0125
WI	**Chronos**	1.1943	–	1.5961	–	1.3185	–	–	–	1.3882
	LPTM	1.0338	–	1.5774	–	1.2681	–	–	–	1.3022
	MOMENT	0.9228	–	1.501	–	1.2912	–	–	–	1.2191
	LakeFM	0.9584	–	1.0445	–	1.027	–	–	–	1.0043