Toward Robust Foundation Models for Scientific Systems: An Application to Lake Sciences

Abhilash Neog

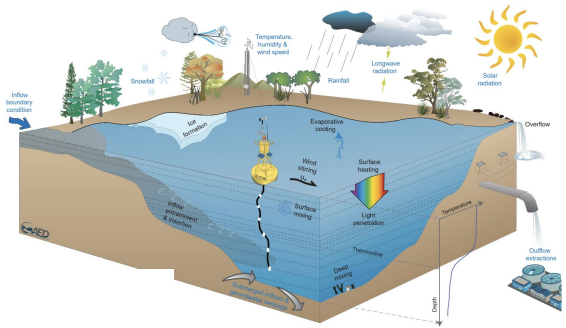
PhD Computer Science,
Virginia Tech

SS009B: The Next Frontier in Aquatic Sciences: Linking remote sensing, data science, modeling, and open science to better understand aquatic ecosystems

ASLO 2026, Montreal

Ecological Forecasting as a Machine Learning Problem

Aquatic Ecosystem 1

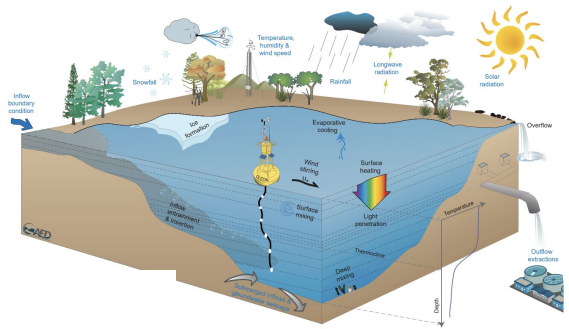


Ecological Forecasting as a Machine Learning Problem

Aquatic Ecosystem 1

1D

Air Temp, Shortwave, ...,

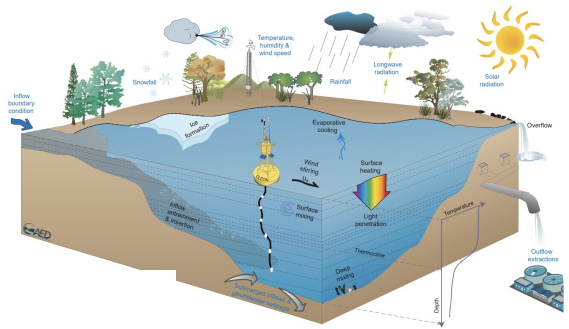


Ecological Forecasting as a Machine Learning Problem

Aquatic Ecosystem 1

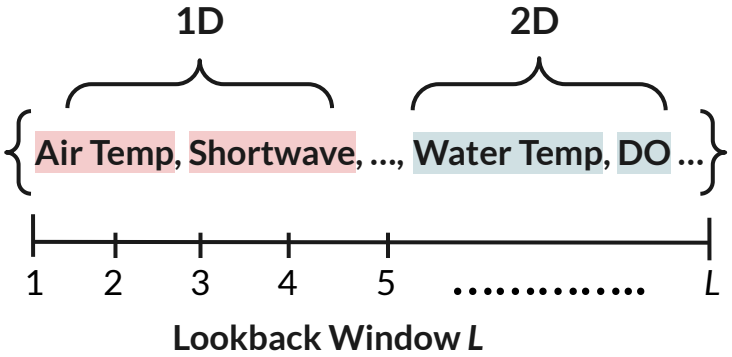
1D 2D

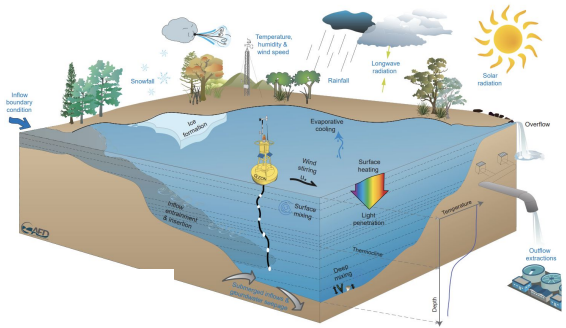
{ Air Temp, Shortwave, ..., Water Temp, DO ... }



Ecological Forecasting as a Machine Learning Problem

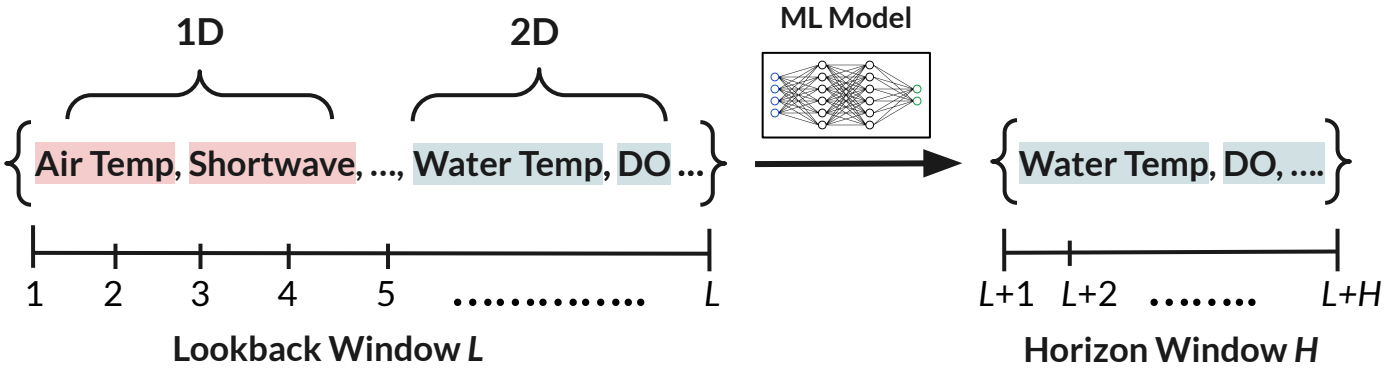
Aquatic Ecosystem 1

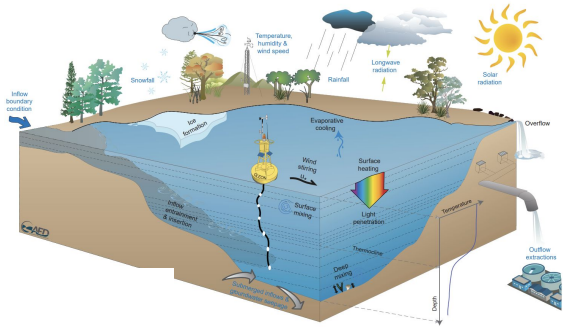




Ecological Forecasting as a Machine Learning Problem

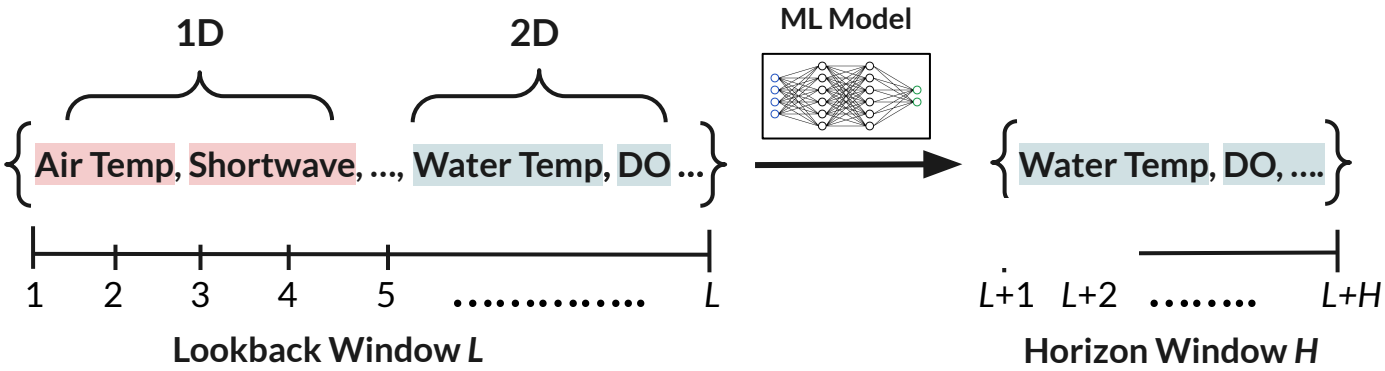
Aquatic Ecosystem 1





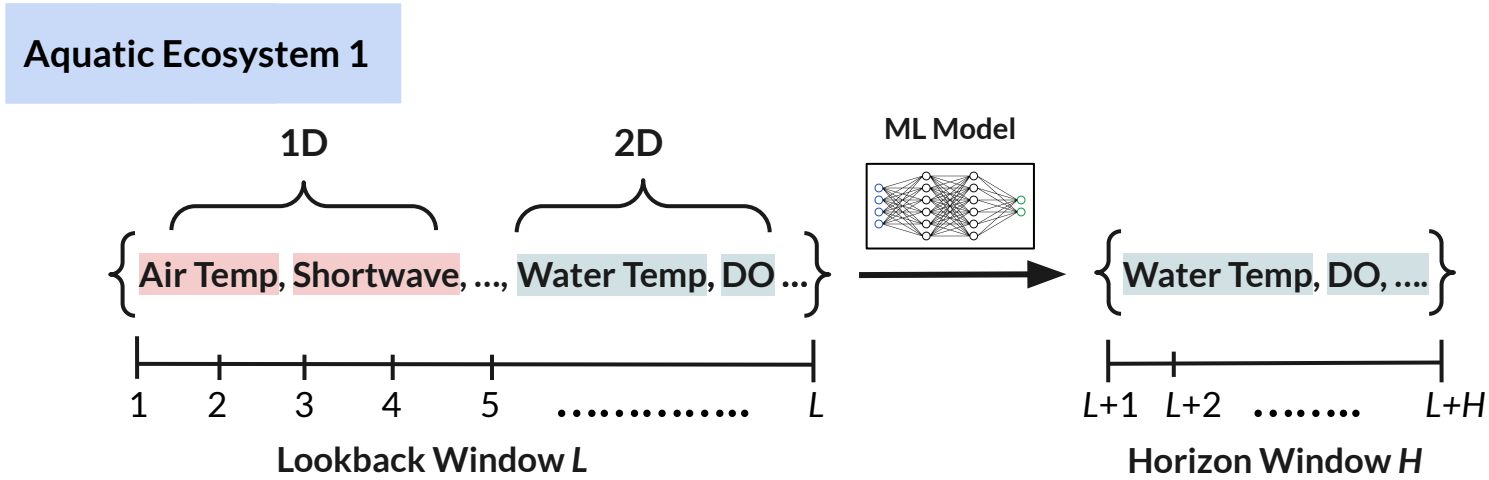
Ecological Forecasting as a Machine Learning Problem

Aquatic Ecosystem 1



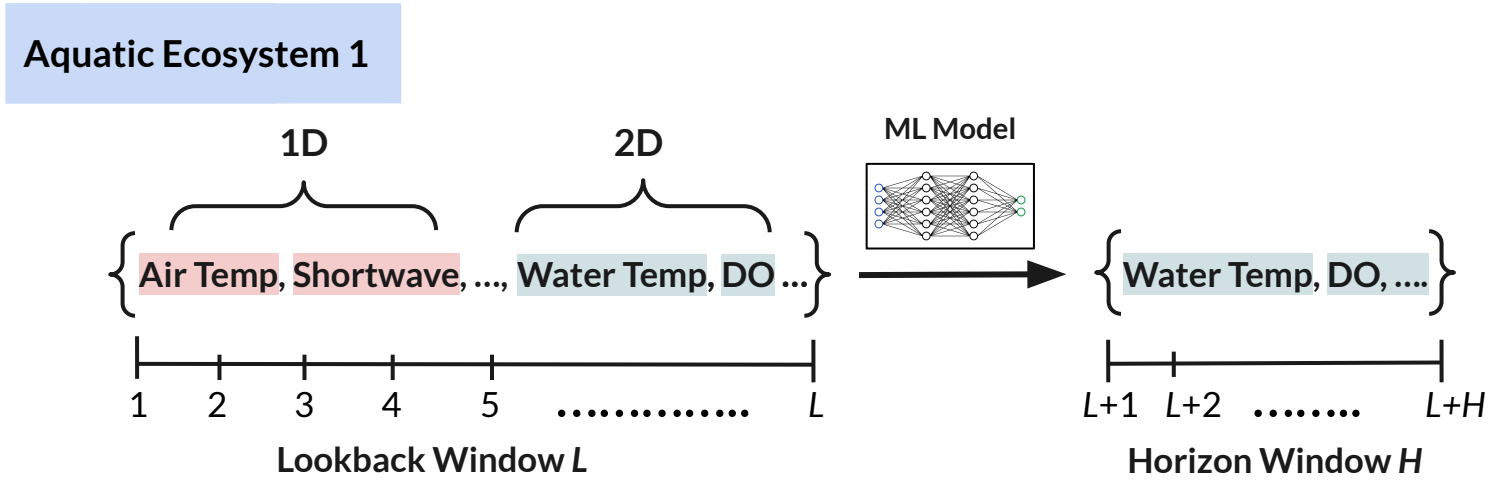
Ecological forecasting provides a natural machine learning formulation: learning temporal dependencies across multivariate lake variables.

Challenges in Building ML Models for Aquatic Systems



⊗ Different subset of variables available in different ecosystems

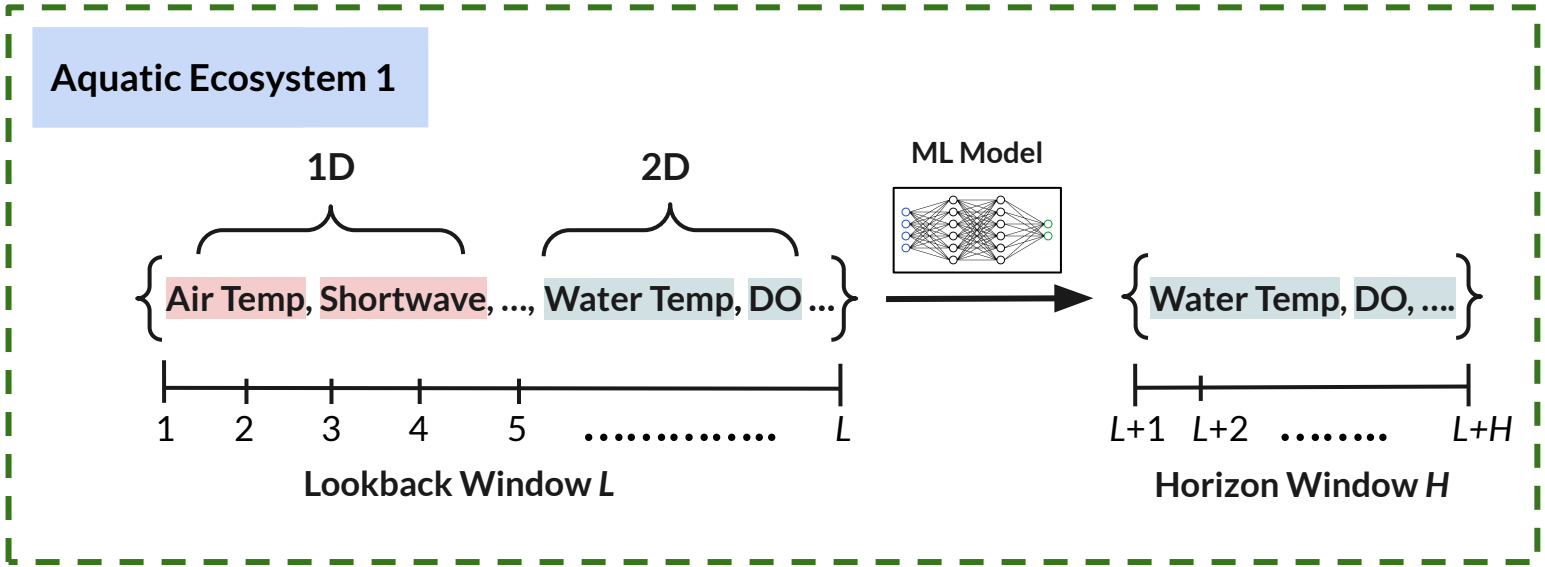
Challenges in Building ML Models for Aquatic Systems



⊗ Different subset of variables available in different ecosystems

⊗ Large amounts of missing data (*e.g., Falling Creeks Reservoir, VA, has ~50% missing data, 2017-04 to 2022-10*)

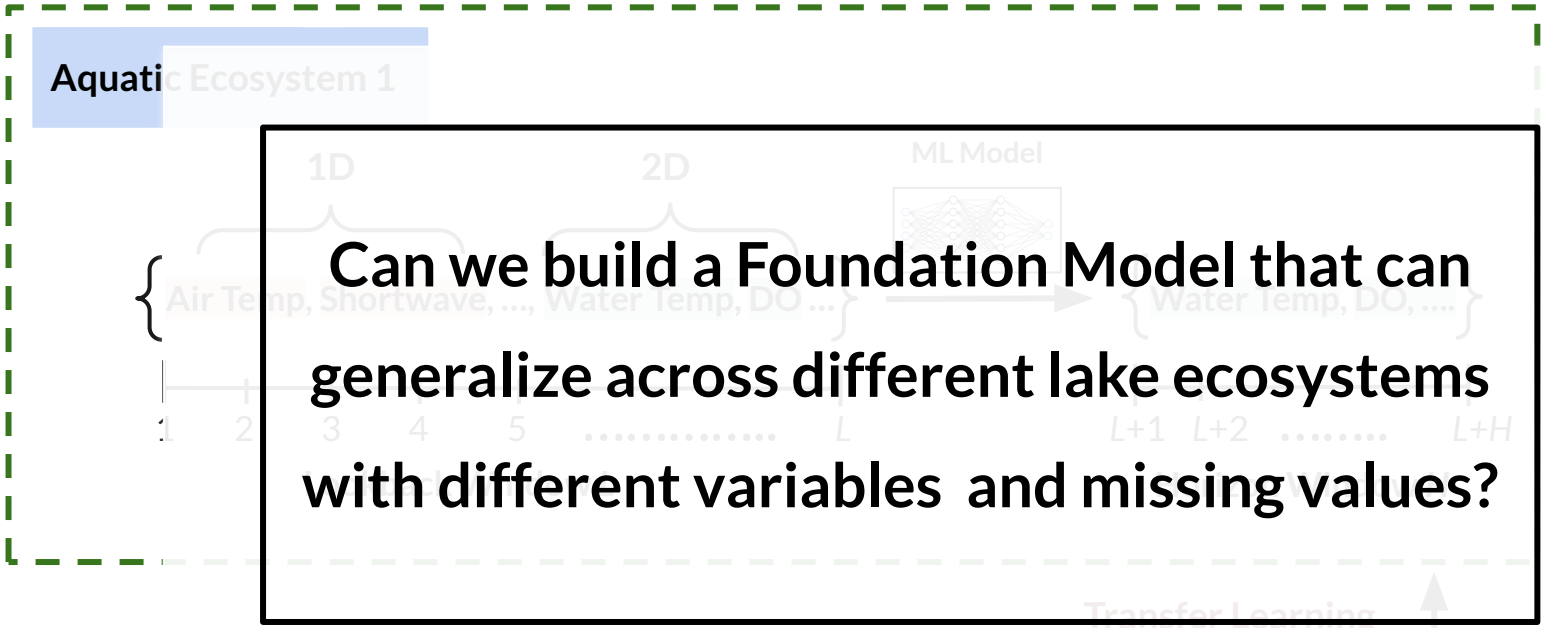
Transfer Learning as a Path Toward Generalizable Aquatic Models



⊗ Different subset of variables available in different ecosystems

⊗ Large amounts of missing data (*e.g., Falling Creeks Reservoir, VA, has 70% missing data, 2017-04 to 2022-10*)

Transfer Learning as a Path Toward Generalizable Aquatic Models

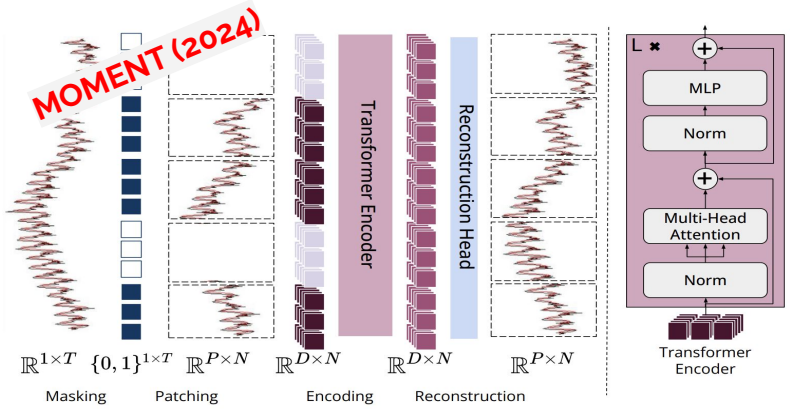


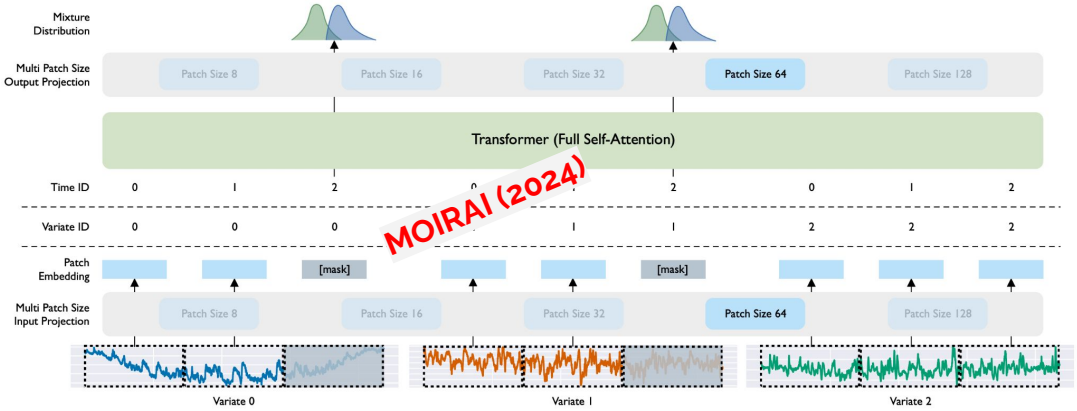
Can we build a Foundation Model that can generalize across different lake ecosystems with different variables and missing values?

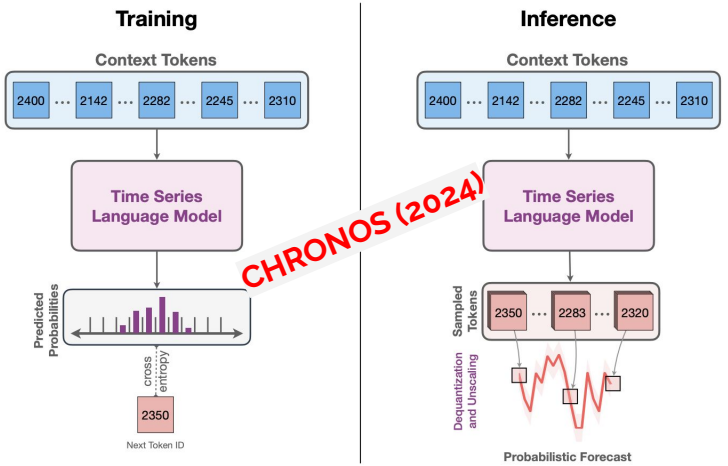
⊗ Different subset of variables available in different ecosystems

⊗ Large amounts of missing data (*e.g., Falling Creeks Reservoir, VA, has 70% missing data, 2017-04 to 2022-10*)

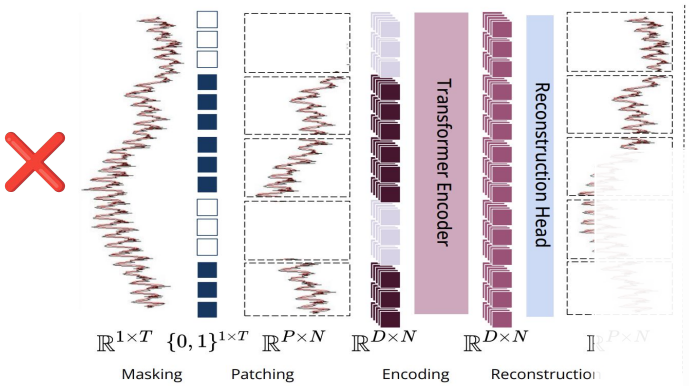
Existing Time Series Foundation Models (TSFM)

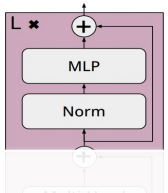


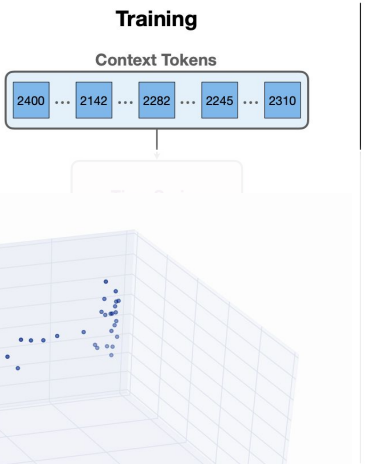


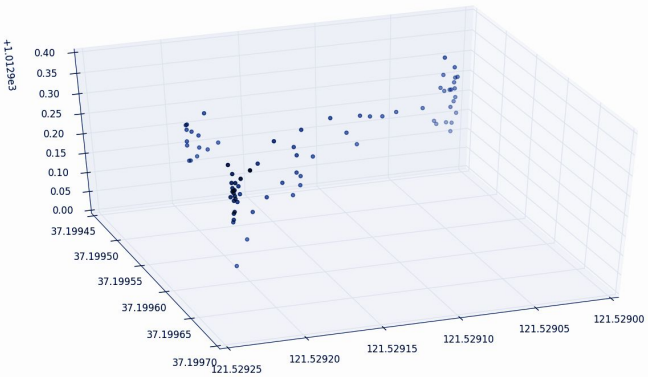


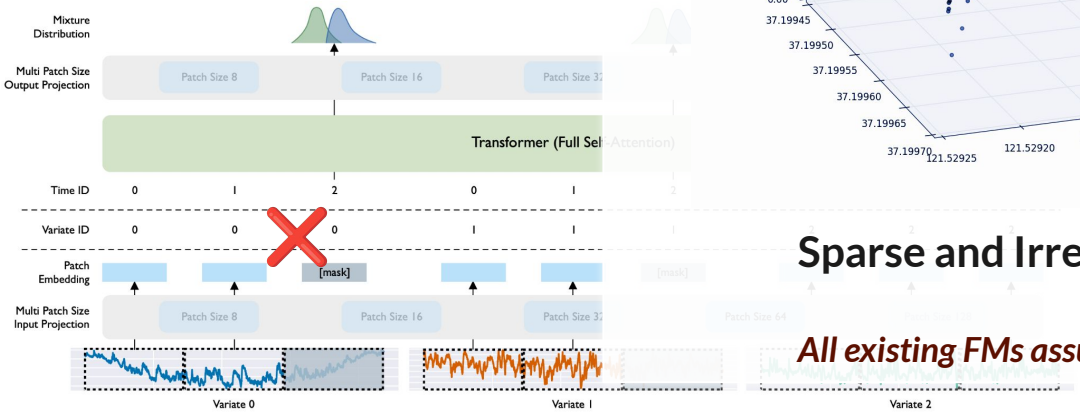
Why Existing TSFMs Fall Short for Aquatic Systems







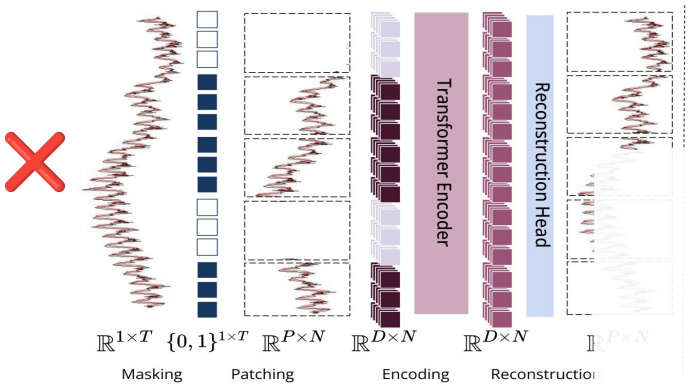


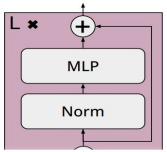


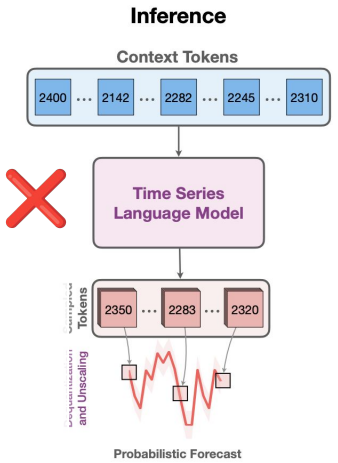
Sparse and Irregular Data

All existing FMs assume fully observed data

Why Existing TSFMs Fall Short for Aquatic Systems

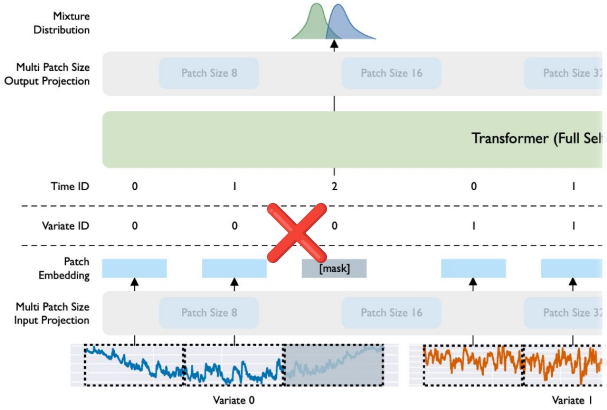


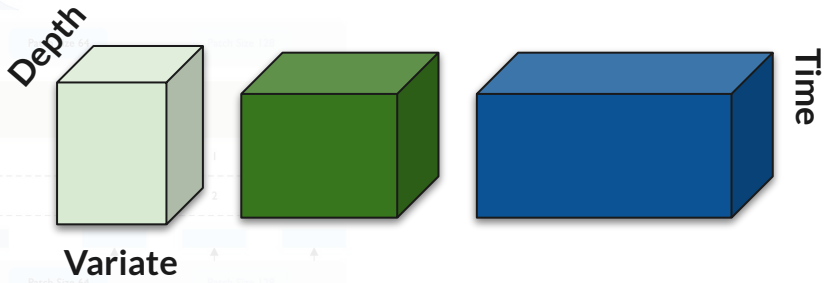




Multivariate and multi-depth

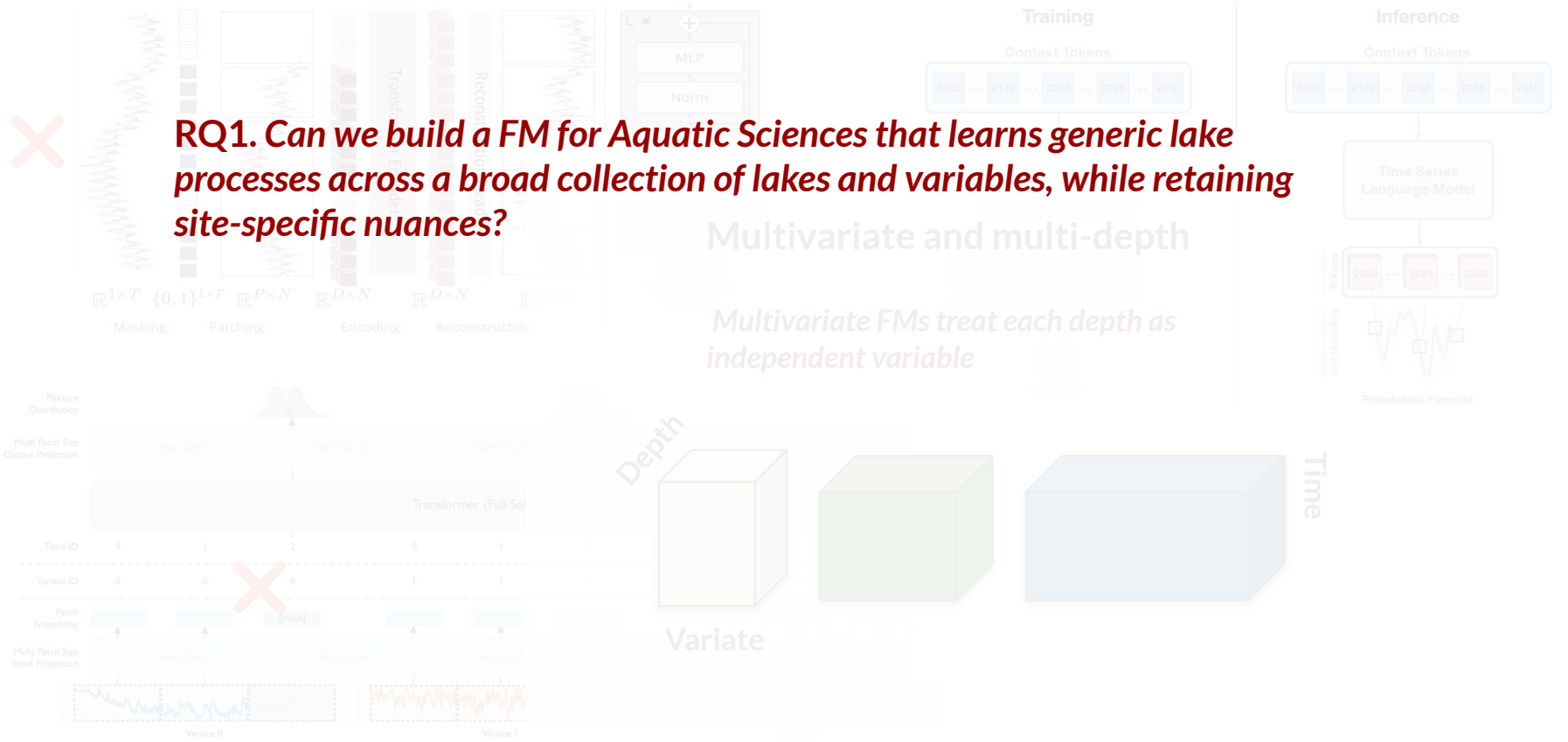
Multivariate FMs treat each depth as independent variable





Research Questions

RQ1. ***Can we build a FM for Aquatic Sciences that learns generic lake processes across a broad collection of lakes and variables, while retaining site-specific nuances?***



Research Questions

RQ1. ***Can we build a FM for Aquatic Sciences that learns generic lake processes across a broad collection of lakes and variables, while retaining site-specific nuances?***

RQ2. ***Can we use such a FM to forecast lake dynamics using any subset of variables available at a lake with irregular observations across time and depth?***

Research Questions

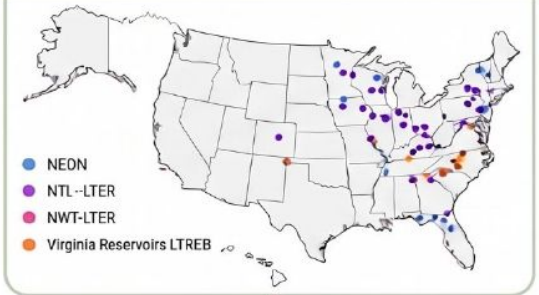
RQ1. ***Can we build a FM for Aquatic Sciences that learns generic lake processes across a broad collection of lakes and variables, while retaining site-specific nuances?***

RQ2. ***Can we use such a FM to forecast lake dynamics using any subset of variables available at a lake with irregular observations across time and depth?***

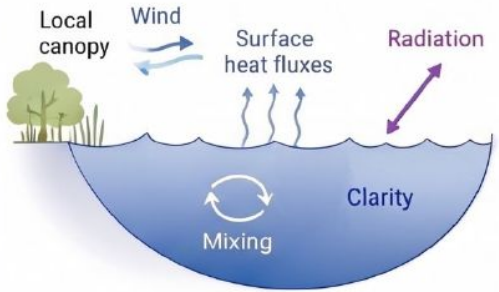
RQ3. ***Can we extract feature representations of lakes that capture their static and time-varying characteristics, revealing novel information about their similarity and temporal evolution at macro-system scales?***

Challenges with existing Foundation Models

Lake Data (US)



Ecology Knowledge

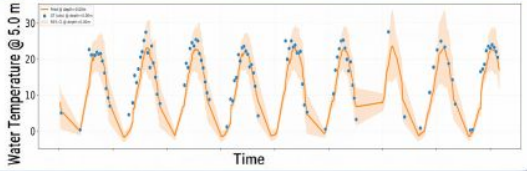


- ✓ Handles missing values in time and depth
- ✓ Forecasts using any available subset of driver variables at a new lake

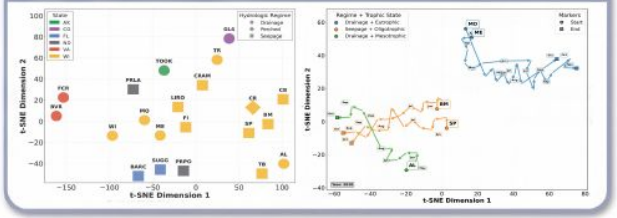
Lake Foundation Model (LakeFM)

Eco-Knowledge-guided Machine Learning

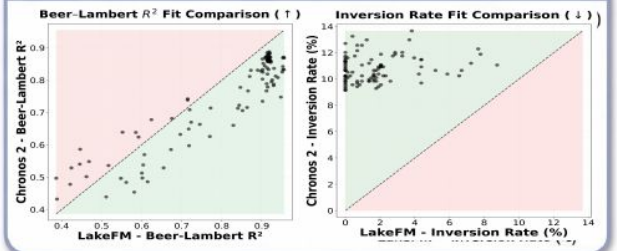
Forecasting Water Quality



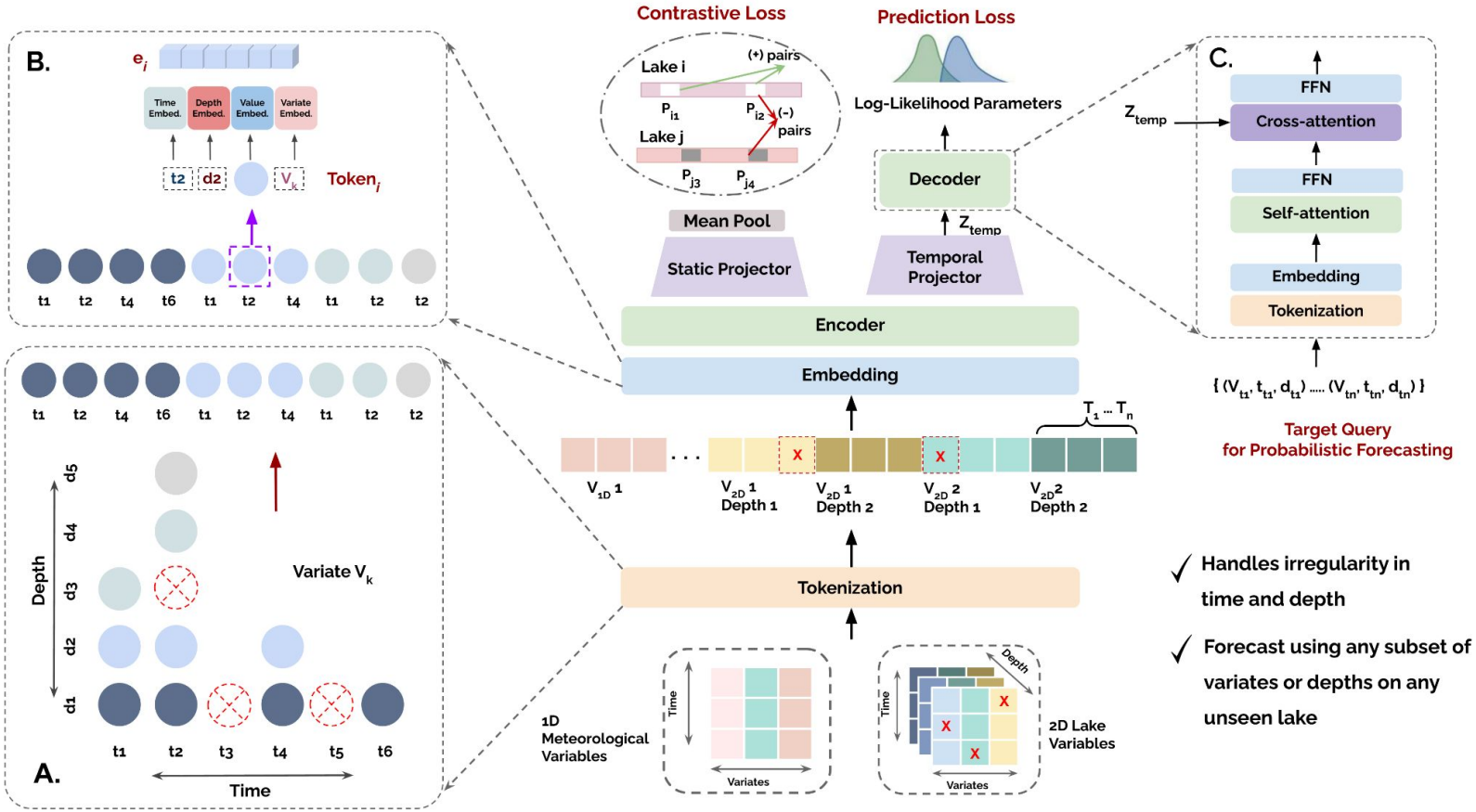
Learning Lake Representations



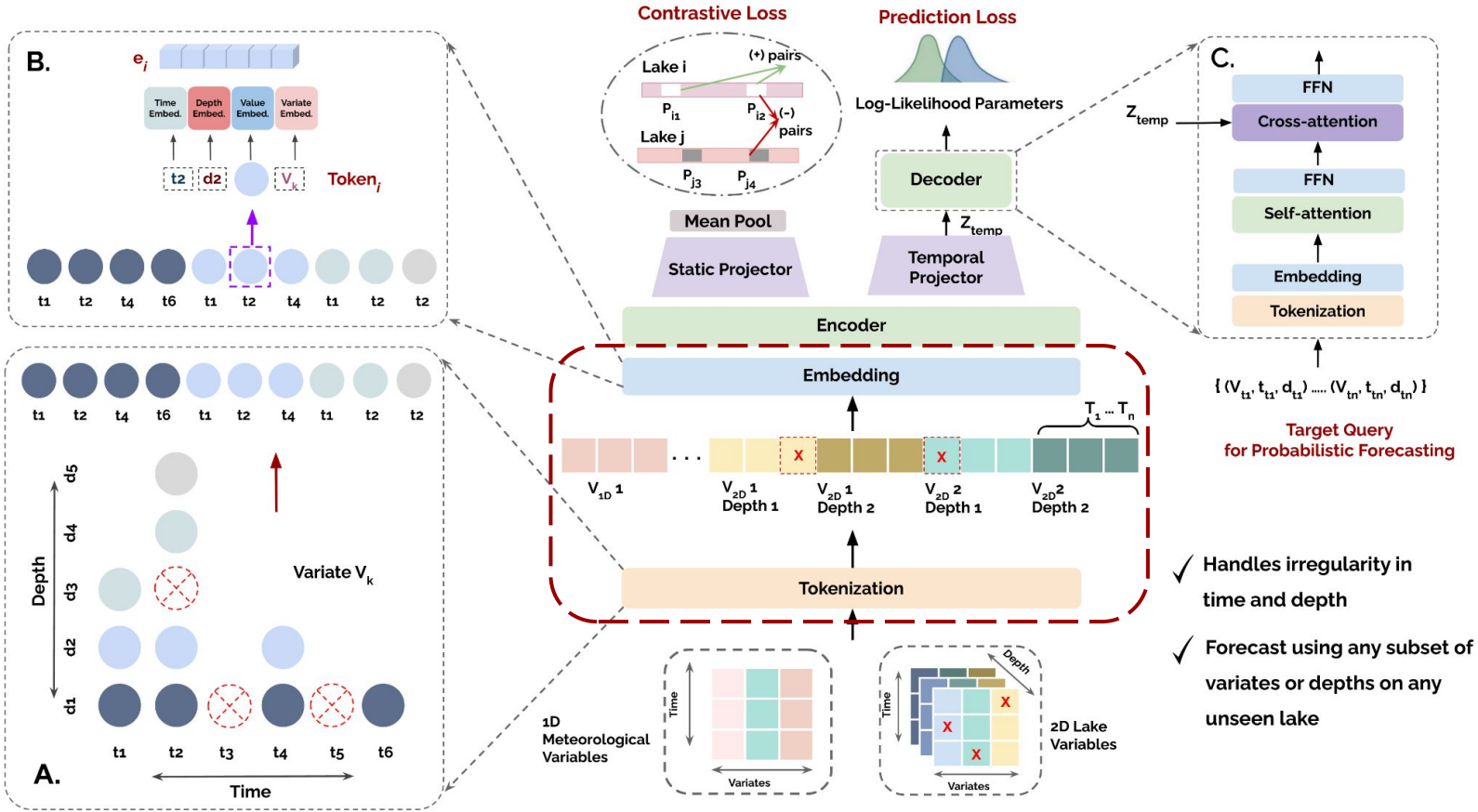
Verifying Physical Consistency



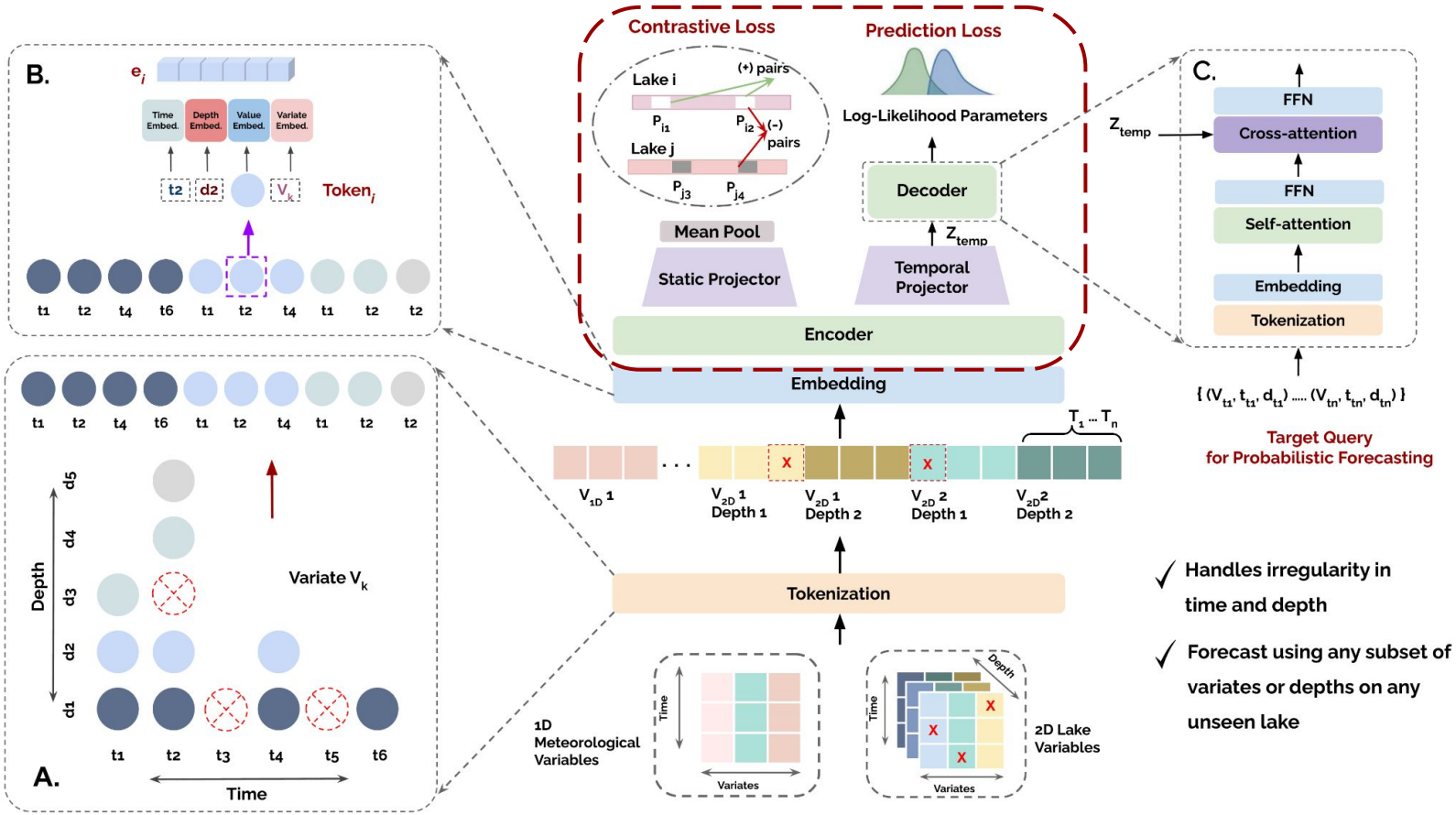
LakeFM: A Foundation Model for Aquatic Forecasting



LakeFM: A Foundation Model for Aquatic Forecasting



LakeFM: A Foundation Model for Aquatic Forecasting

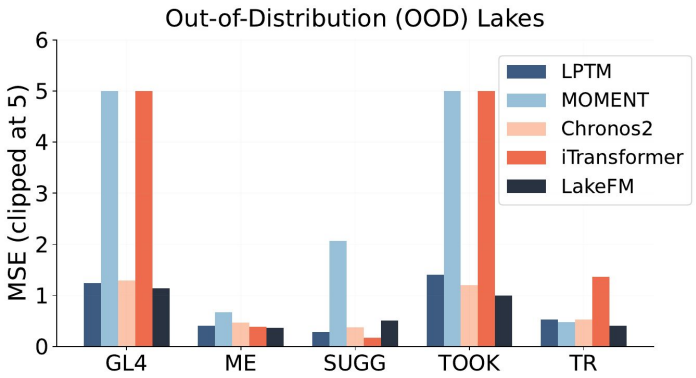


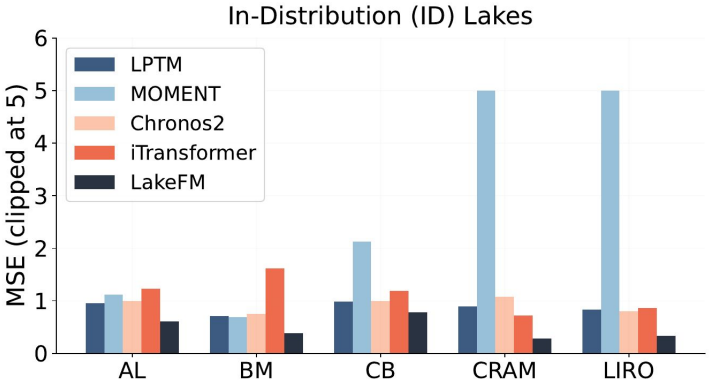
Experiments and Results

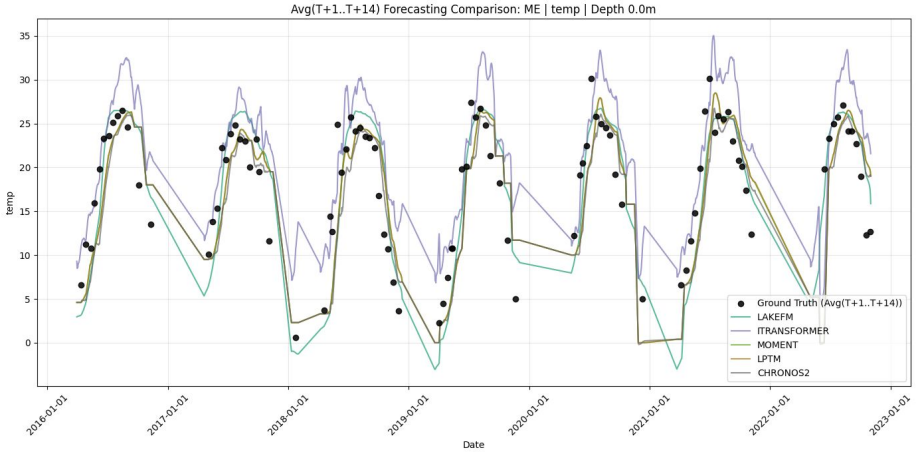
LakeFM (**5M**) is pre-trained on a large-scale ecological dataset containing over **1.5 million** samples:

- **LakeBeD-US** [McAfee et al., 2025] — 500 million unique observations spanning 21 lakes across the United States, covering 17 variables with 60–70% sparsity on average.
- **WQHanson Simulations** [Hanson et al., 2023] — 4 simulation lakes generated using the process-based water quality model.
- **FCR Simulations** [Hipsey et al., 2019] — 1,000 simulations generated using the GLM-AED process-based model.

Findings (I) - Forecasting Performance

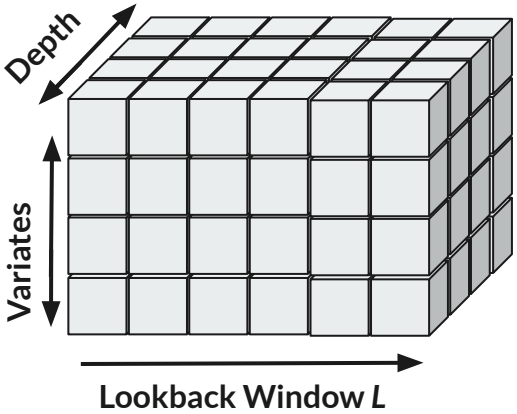




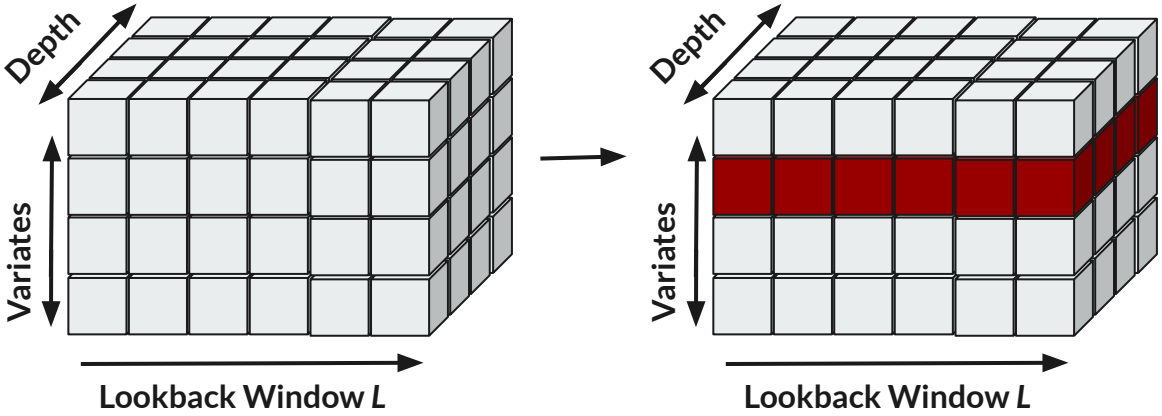


LakeFM achieves a **best overall rank of 1.86 across all In-Distribution lakes** and **2.17 across all Out-of-distribution lakes** in terms of lake-wise MSE

Findings (II) - Incomplete Data (Variables)

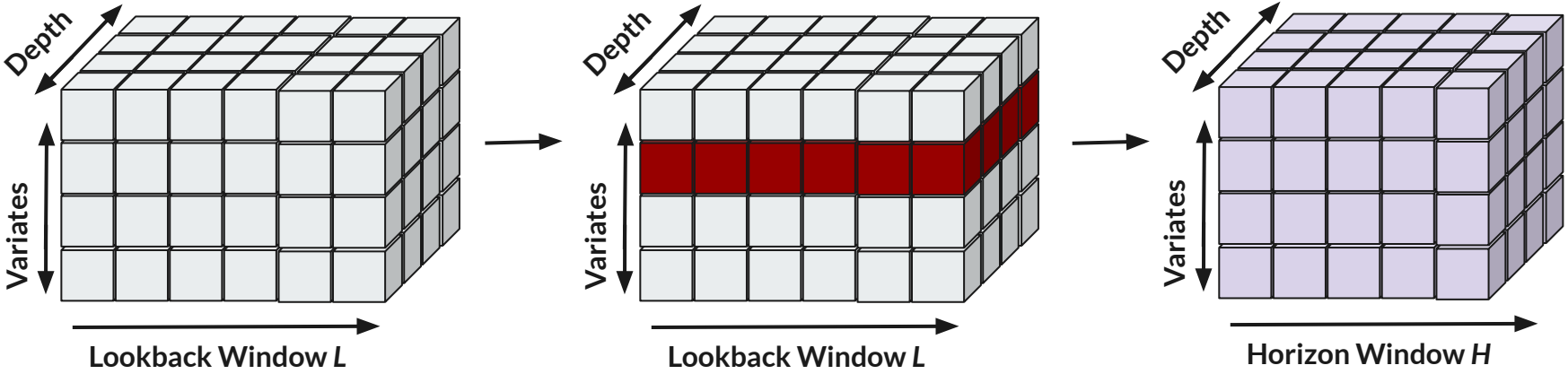


Findings (II) - Incomplete Data (Variables)



{ Air Temp, Shortwave, …, ~~Water Temp~~, DO … }

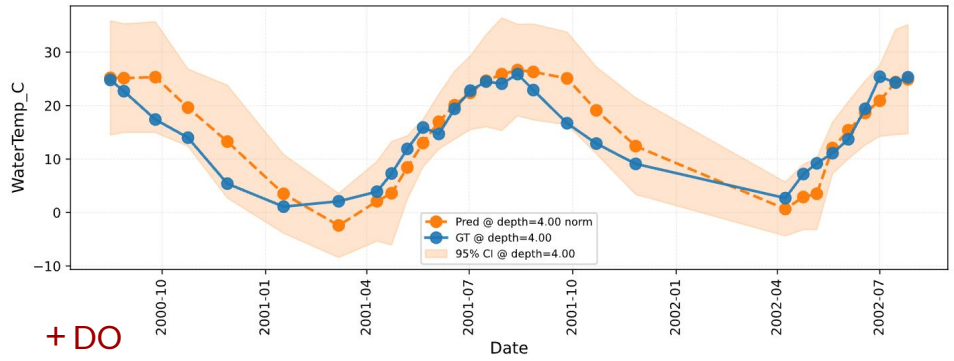
Findings (II) - Incomplete Data (Variables)

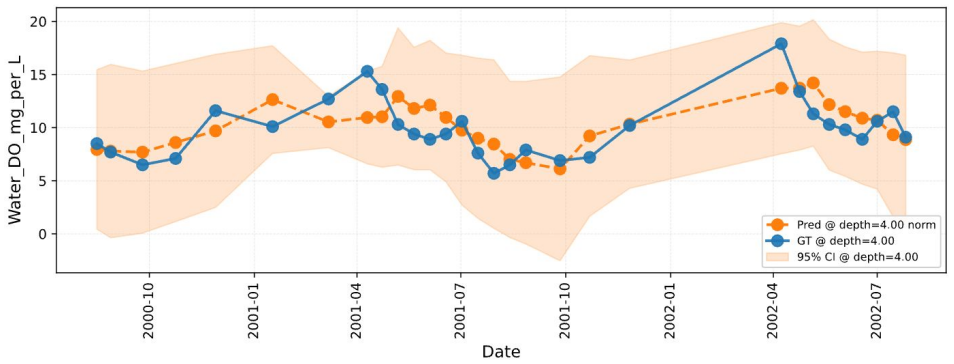


{ Air Temp, Shortwave, ..., ~~Water Temp~~, DO ... } → { Air Temp, Shortwave, ..., Water Temp, DO ... }

Findings (II) - Incomplete Data (Variables)

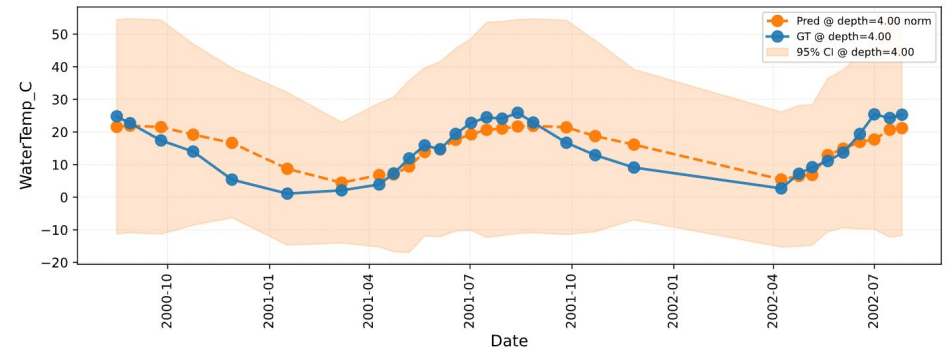
ME : 30 timesteps ahead forecast, at Depth 4.00m (shaded = 95% CI)

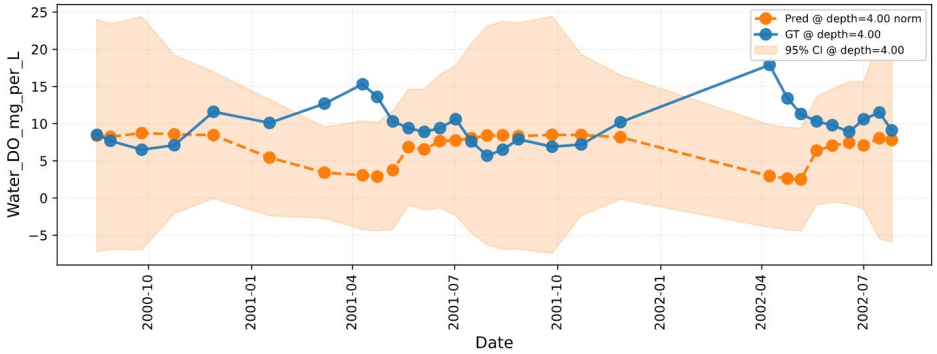




\+ DO

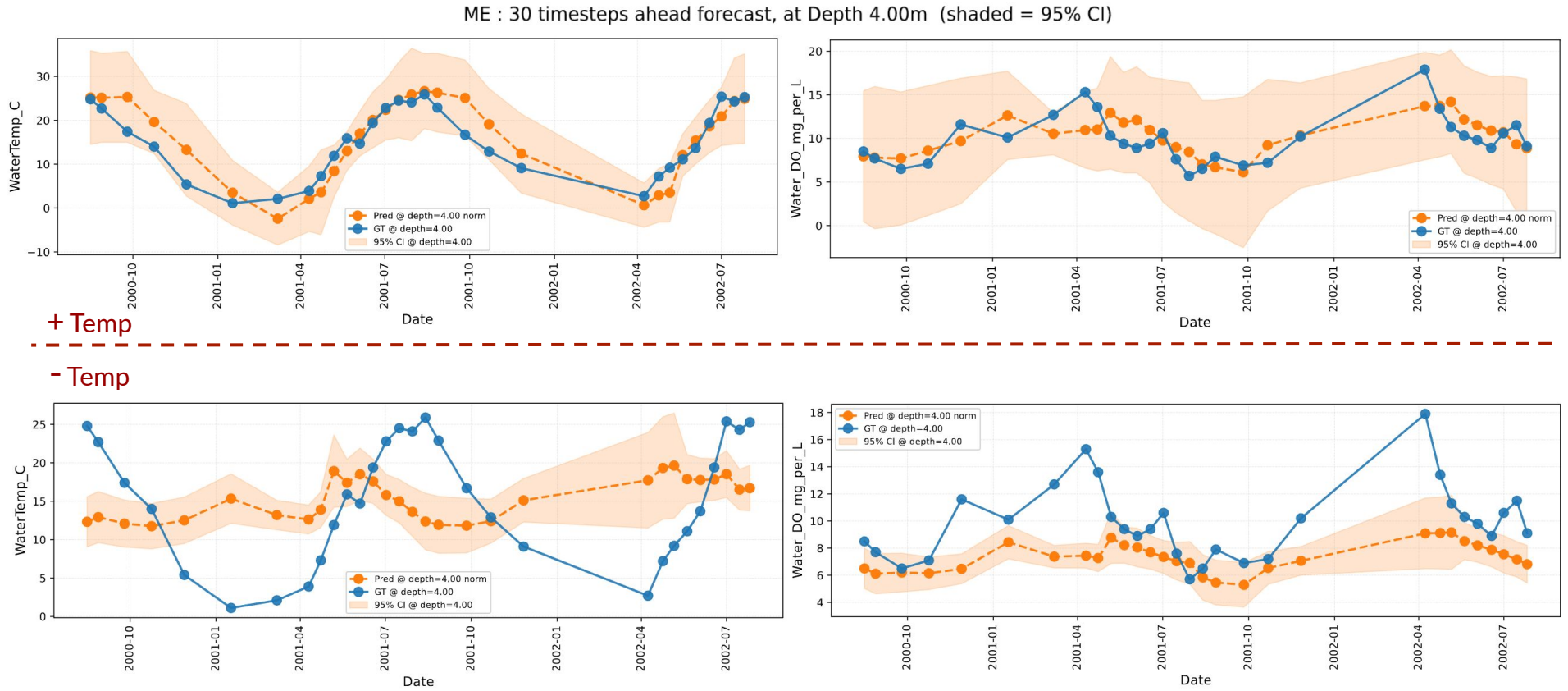
- DO





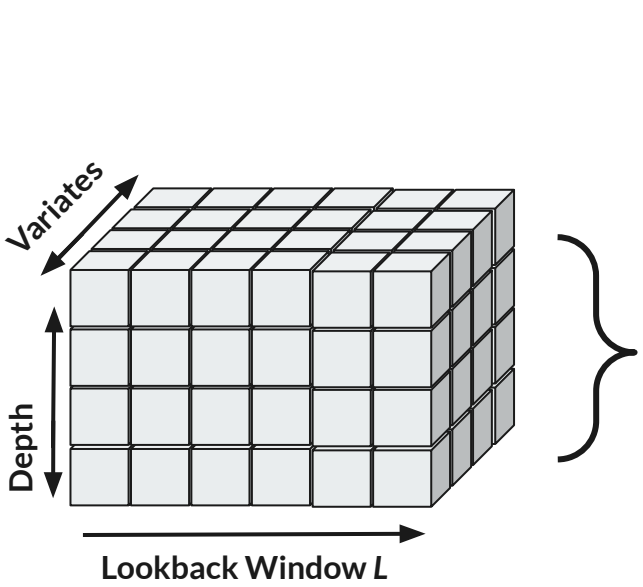
Key Observation : Removing DO from inputs increases uncertainty in predictions of water temperature

Lake Foundation Model (LakeFM) - Findings (II) - Incomplete Data (Variables)

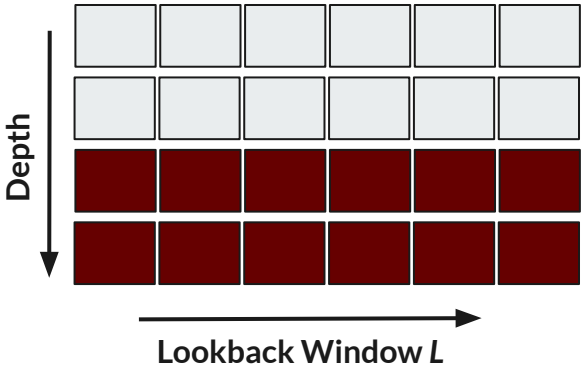


Key Observation : Water temperature is a critical variable. Removing it degrades all predictions.

Findings (III) - Incomplete Data (Depth)



(1)



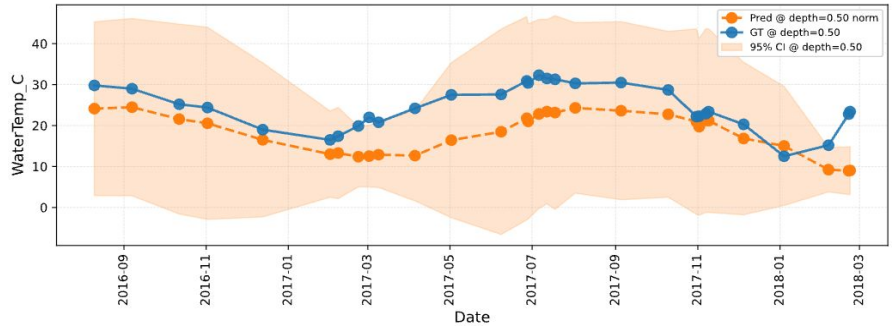
No data in the deeper layers of the lake

(2)

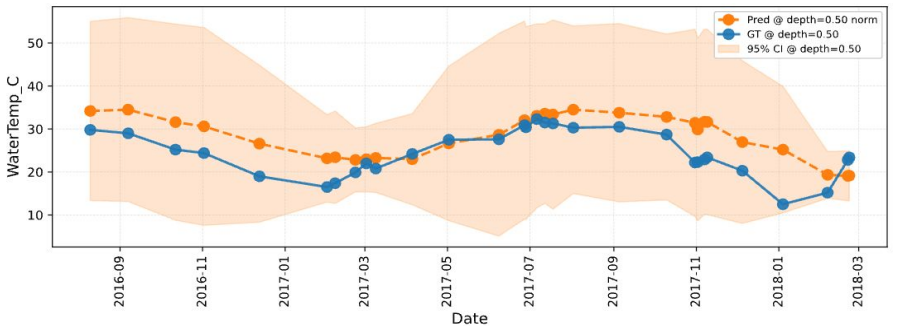
No data in the shallow layers of the lake

Findings (III) - Incomplete Data (Depth)

Water Temp Predictions @ 0.5 m using full-depth history

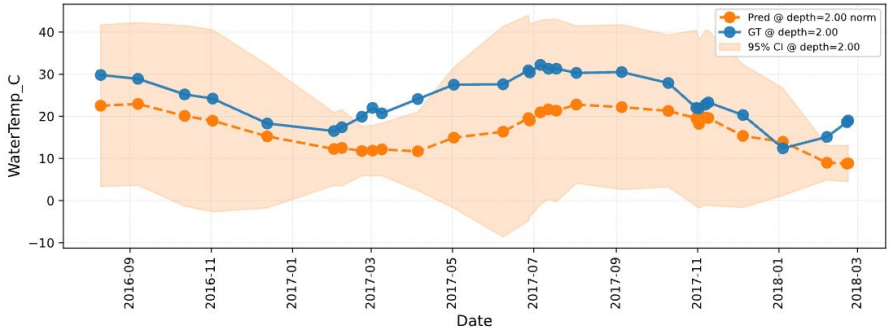


Water Temp Predictions @ 0.5 m using only deeper-depth history

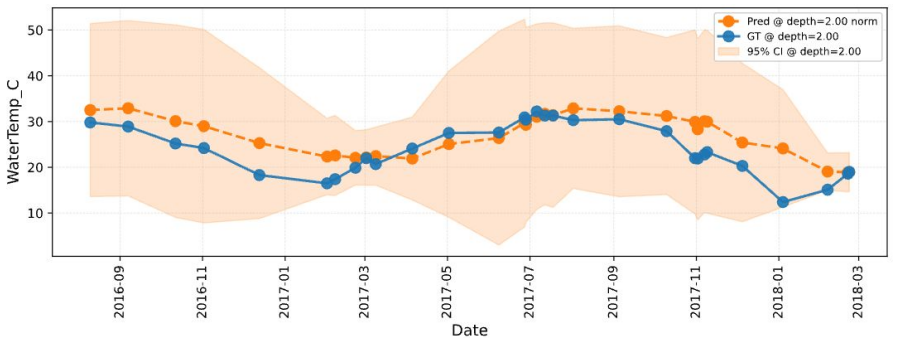


0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5

Water Temp Predictions @ 2.0 m using full-depth history

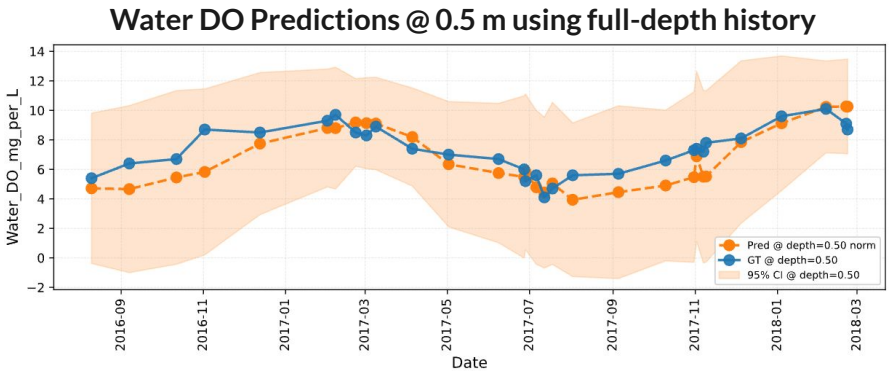


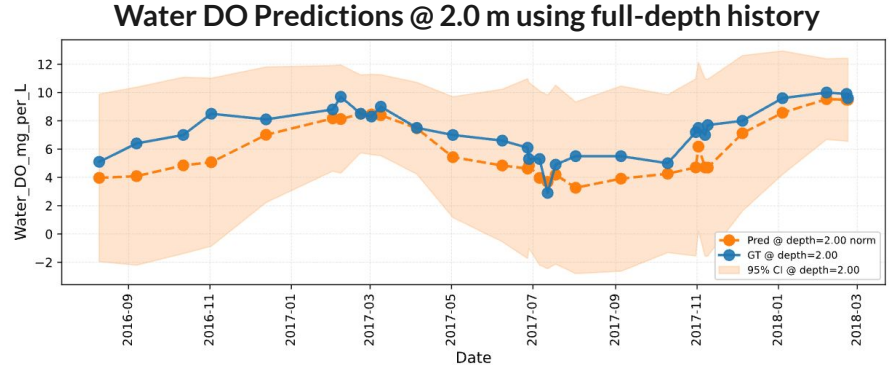
Water Temp Predictions @ 2.0 m using only deeper-depth history



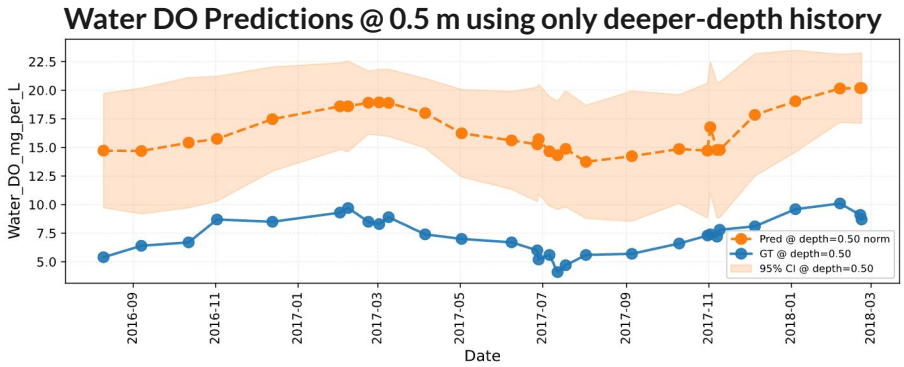
Key Observation : Water temperature predictions remain stable even without shallow-layer variables

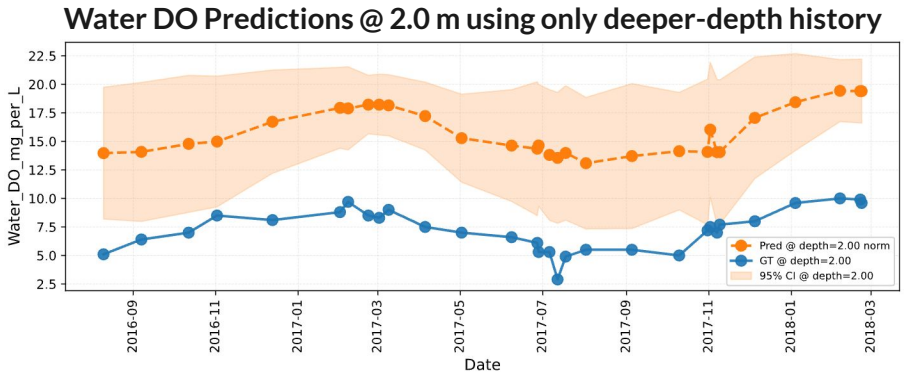
Findings (III) - Incomplete Data (Depth)





0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5

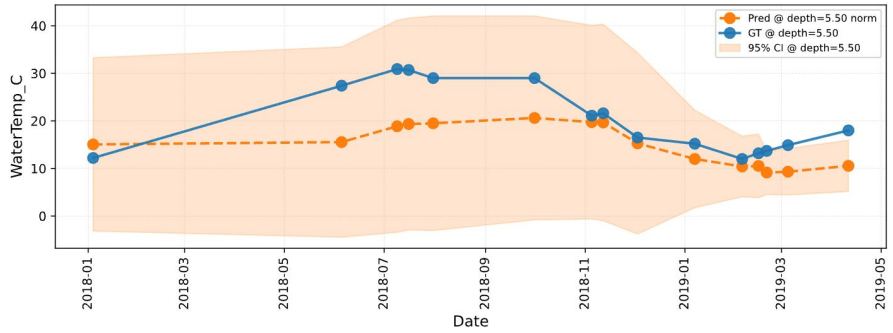




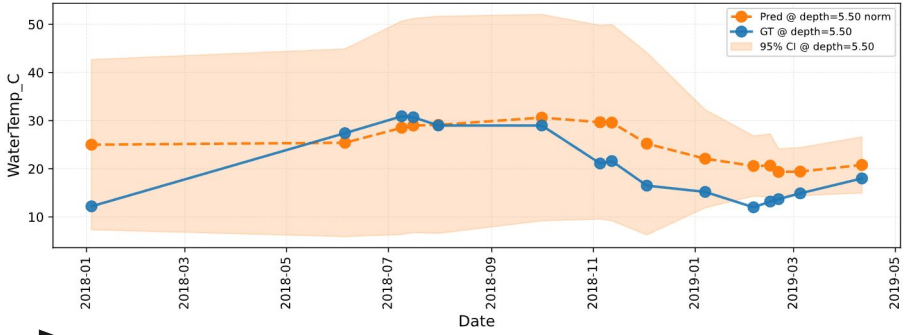
Key Observation : In contrast, DO predictions cannot rely on deeper-layer variables, indicating stronger vertical variability along the water column

Findings (III) - Incomplete Data (Depth)

Water Temp Predictions @ 5.5 m using full-depth history

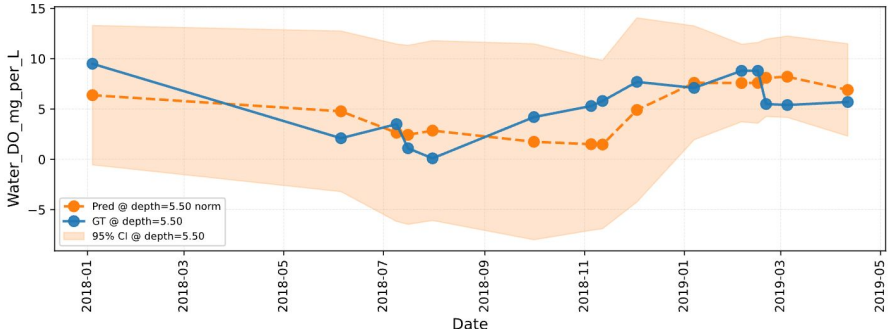


Water Temp Predictions @ 5.5 m using only shallow-depth history

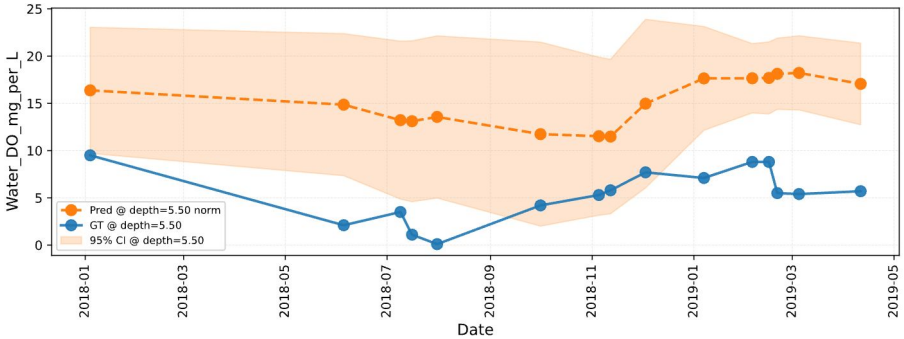


0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5

Water DO Predictions @ 5.5 m using full-depth history

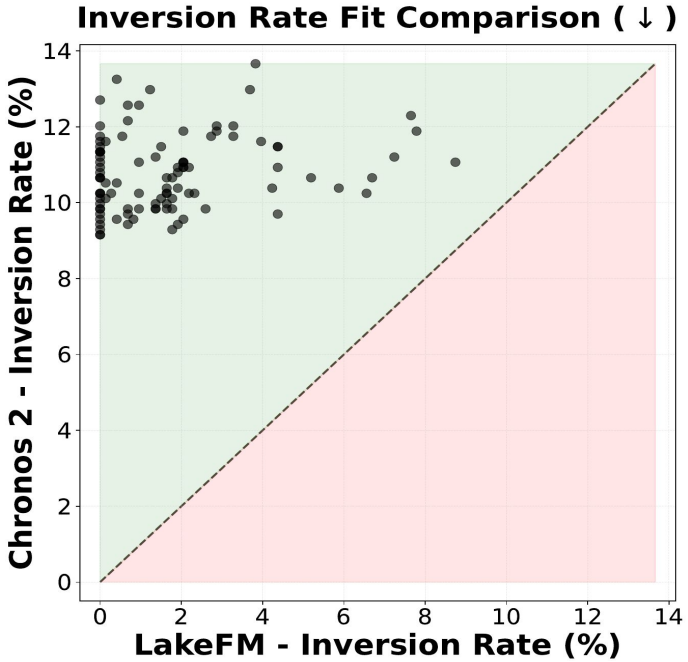


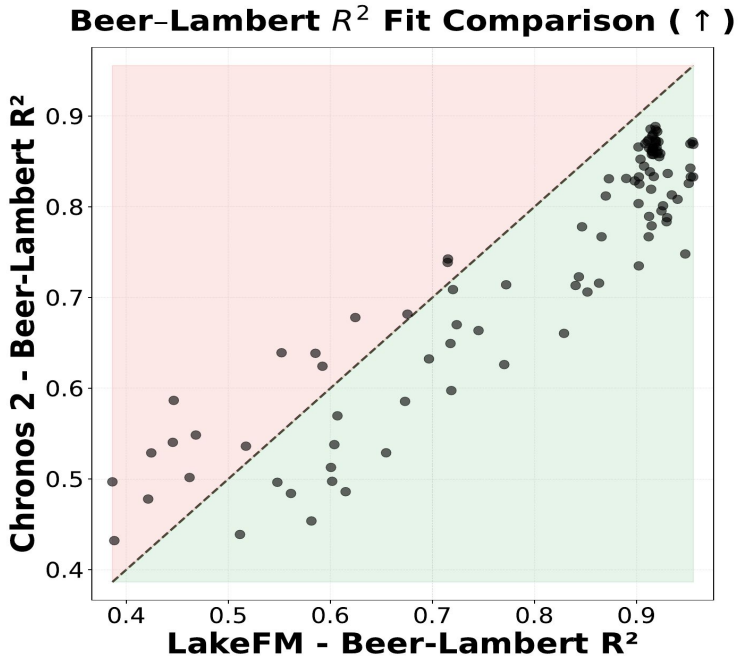
Water DO Predictions @ 5.5 m using only shallow-depth history



Key Observation : In general, water temperature remains stable and is predictable using either shallow or deeper layers, while DO dynamics are tightly coupled to the local depth.

Findings (IV) - Physical Consistency





Key Observation : LakeFM demonstrates emergent physical consistency - without any explicit physical supervision, it adheres better than Chronos 2 to both thermal stratification and Beer-Lambert laws across a large majority of 100 unseen simulated lakes

Findings (V) - Learned Embeddings

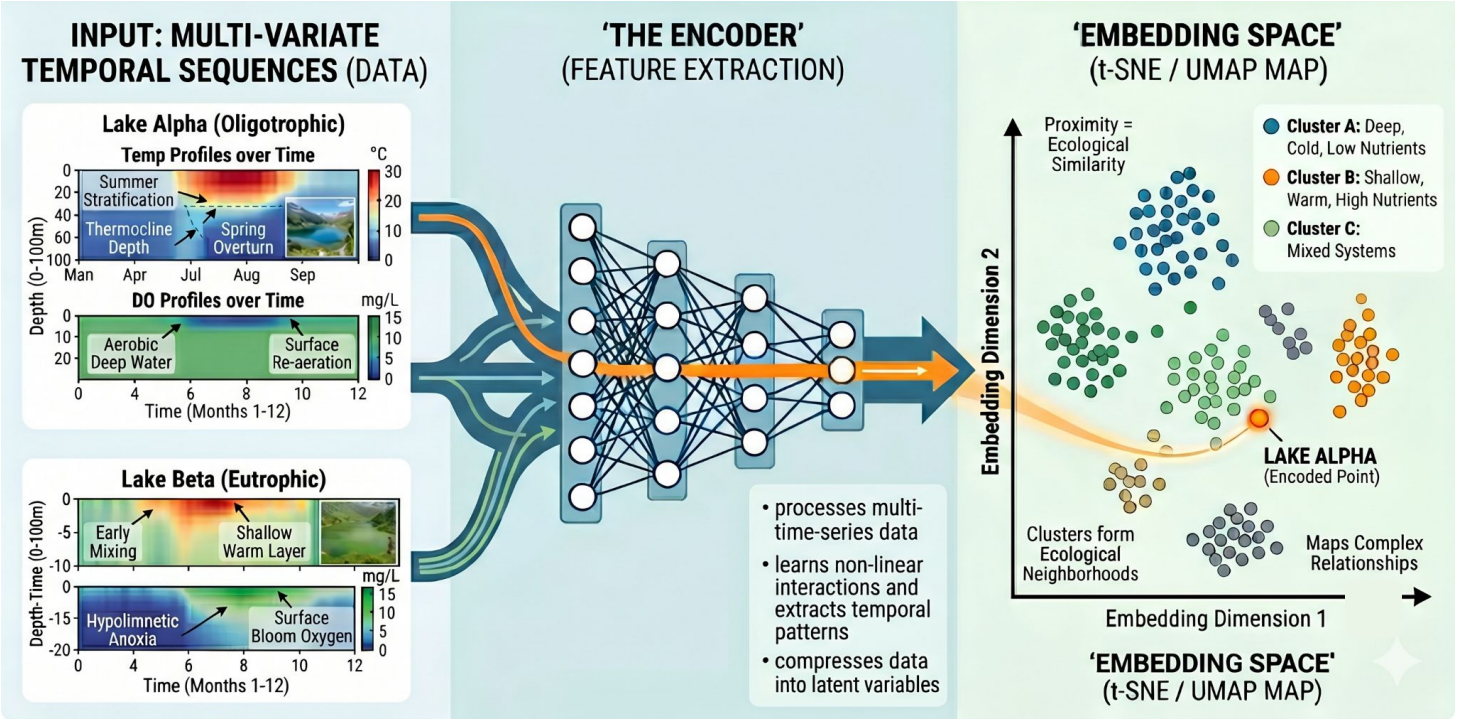
What are embeddings?

What are embeddings?

~ learned ecological fingerprint

Traditional limnology uses measured variables (depth, Secchi disk, phosphorus) to describe a lake. An embedding is the model's way of condensing all those complex, non-linear relationships into a single point in a "conceptual space"

Overview of Embeddings

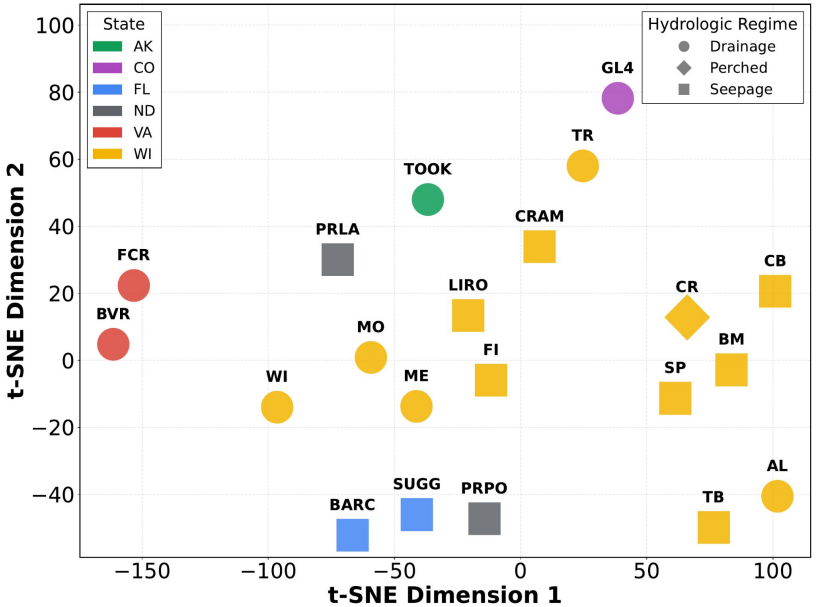


Translation: Converting raw data into a language the model understands.

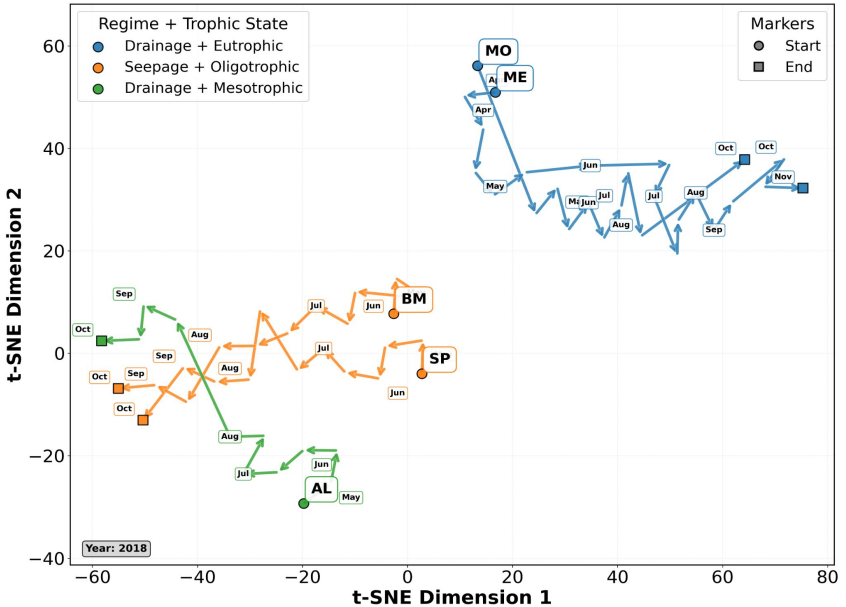
Compression: Retaining the most important ecological "signals"

Clustering: Lakes that function similarly live in the same "neighborhood" in this space.

Findings (V) - Learned Embeddings - Static and Temporal Embeddings

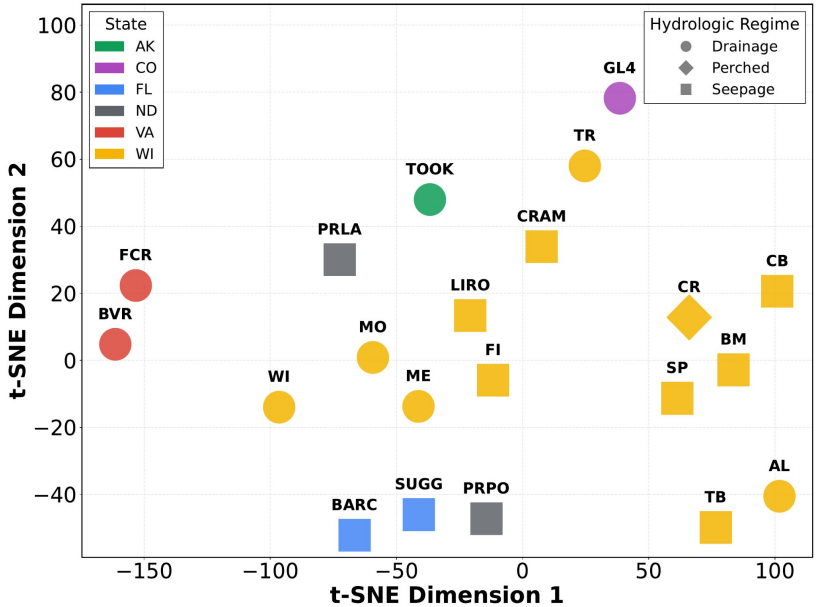


Key Observation: LakeFM's static embeddings spontaneously cluster lakes by geographic location and hydrologic regime - even separating drainage from seepage lakes within Wisconsin



Key Observation: LakeFM's dynamic embeddings trace distinct seasonal trajectories for lakes that differ in trophic state and hydrologic regime, with eutrophic drainage lakes (MO, ME) following closely aligned paths while oligotrophic seepage lakes (SP, BM) form a separate cluster.

Findings (V) - Learned Embeddings - Static Embeddings



Key Observation: LakeFM's static embeddings spontaneously cluster lakes by geographic location and hydrologic regime - even separating drainage from seepage lakes within Wisconsin

What do they reflect? The static embeddings are learned by the LakeFM model to represent the static (time-invariant) characteristics of a lake.

Learned through Contrastive Loss Optimization, these characteristics learnt could be trophic state and/or geographical, etc.

What does the plot show? This plot shows the embedding space learned by the model, with each lake's embedding represented as a point.

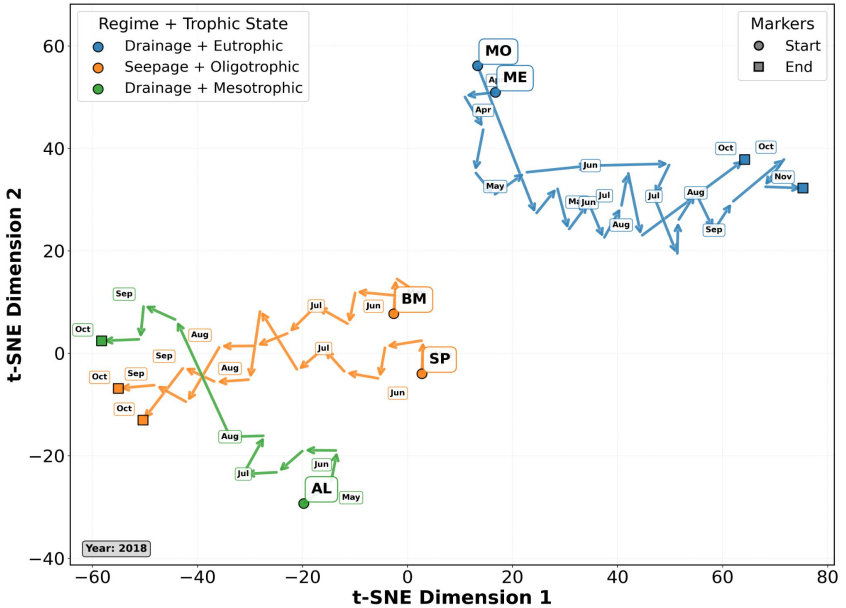
We categorize each lake by the known geographical location and hydrologic regime, to verify the similarities in the embedding space with the known ecological axes

Findings (V) - Learned Embeddings - Temporal Embeddings

What do they reflect? These are more dynamic. They represent the **snapshot of the system state,** with respect to time.

How are they different from the static embeddings? These embeddings are not just about what the lake IS, rather, how they evolve over time.. A point moving through this space represents a season changing.

What does the plot show? The plot shows the embedding space of different lakes within Wisconsin. Each lake is represented as a trajectory, showing how their embeddings evolve over time.



Key Observation: LakeFM's dynamic embeddings trace distinct seasonal trajectories for lakes that differ in trophic state and hydrologic regime, with eutrophic drainage lakes (MO, ME) following closely aligned paths while oligotrophic seepage lakes (SP, BM) form a separate cluster.

Acknowledgement

LakeFM Team

Abhilash Neog, *Virginia Tech*

Sepideh Fatemi, *Virginia Tech*

Medha Sawhney, *Virginia Tech*

Kazi Sajeed Mehrab, *Virginia Tech*

Mary E. Lofton, *Virginia Tech*

Aanish Pradhan, *Virginia Tech*

Bennett J. McAfee, *Annis Water Resources Institute*

Emma Marchisin, *UW-Madison*

Robert Ladwig, *Aarhus University*

Arka Daw, *Oak Ridge National Lab*

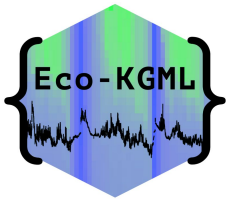
Cayelan C. Carey, *Virginia Tech*

Paul C. Hanson, *UW-Madison*

Anuj Karpatne, *Virginia Tech*

KGML

Knowledge-guided Machine Learning



Funding

NSF #2213549, #2213550

Compute Resources

NAIRR pilot award #240161

NSF Access

Virginia Tech ARC

Thank you !

LakeFM Project Page

Eco-KGML Website

Contact:

Abhilash Neog,
Email: abhilash22@vt.edu

Dept. of Computer Science,
Virginia Tech

Personal Site