Toward Scientific Foundation Models for Aquatic Ecosystems

Abhilash Neog

PhD Computer Science, Virginia Tech

H33D: Advancing Water Science Through Artificial Intelligence: Lessons, Strategies, and New Frontiers

AGU 2025, New Orleans, LA

Motivation

Aquatic Ecosystem 1

Motivation

Aquatic Ecosystem 1

$$\left\{ \underbrace{\text{Air Temp, Shortwave}}_{\text{1D}}, \ldots, \right.$$

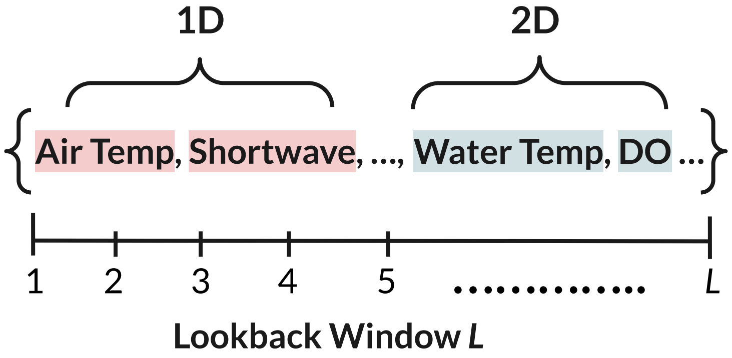
Motivation

Aquatic Ecosystem 1

$$\left\{ \underbrace{\text{Air Temp, Shortwave}}^{\text{1D}}, \ldots, \overbrace{\text{Water Temp, DO}}^{\text{2D}} \ldots \right\}$$

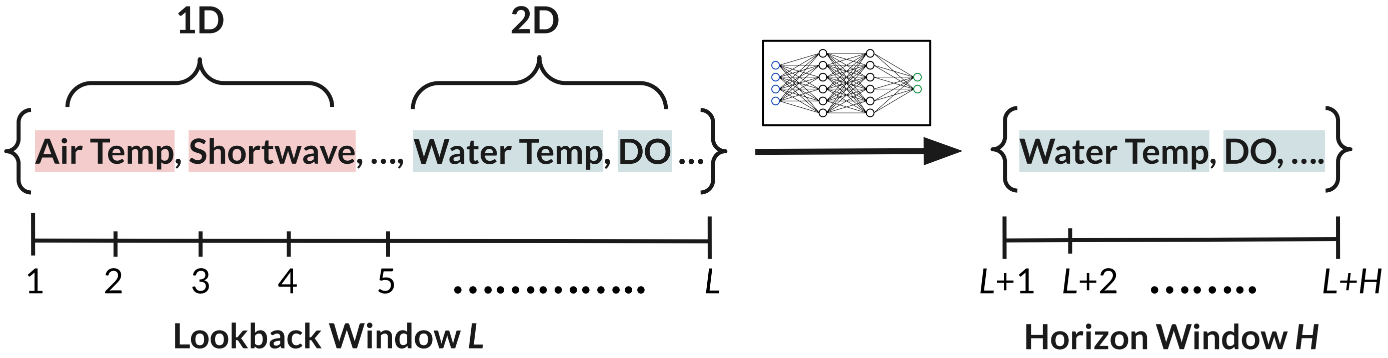
Motivation

Aquatic Ecosystem 1



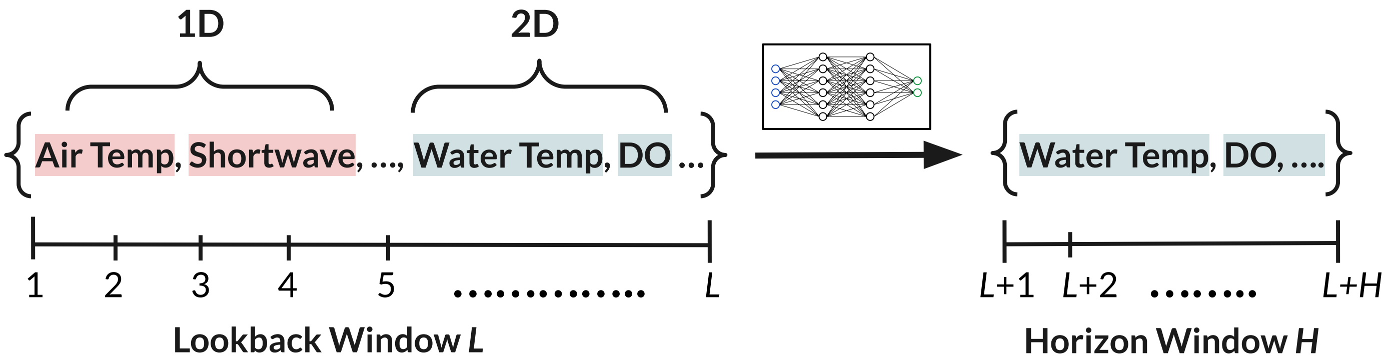
Motivation

Aquatic Ecosystem 1



Motivation

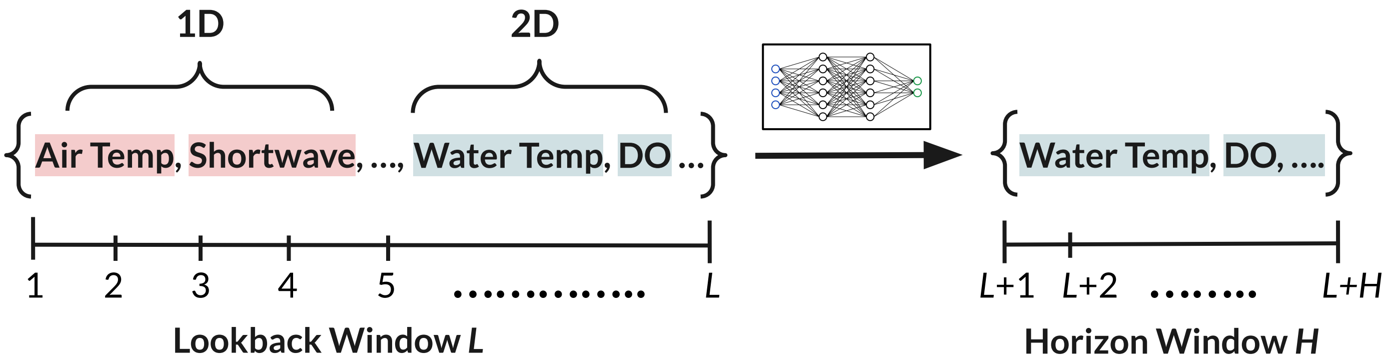
Aquatic Ecosystem 1



Different subset of variables available in different ecosystems

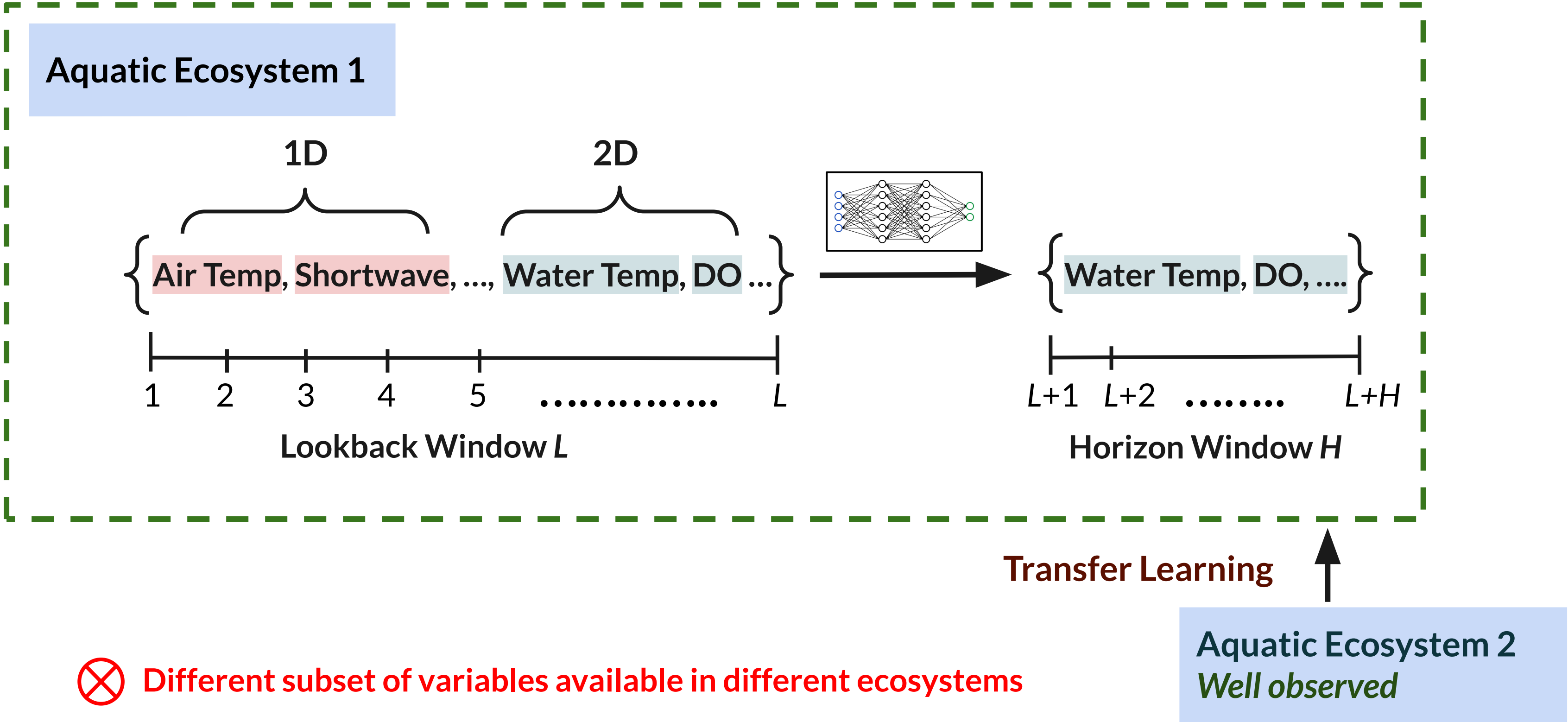
Motivation

Aquatic Ecosystem 1



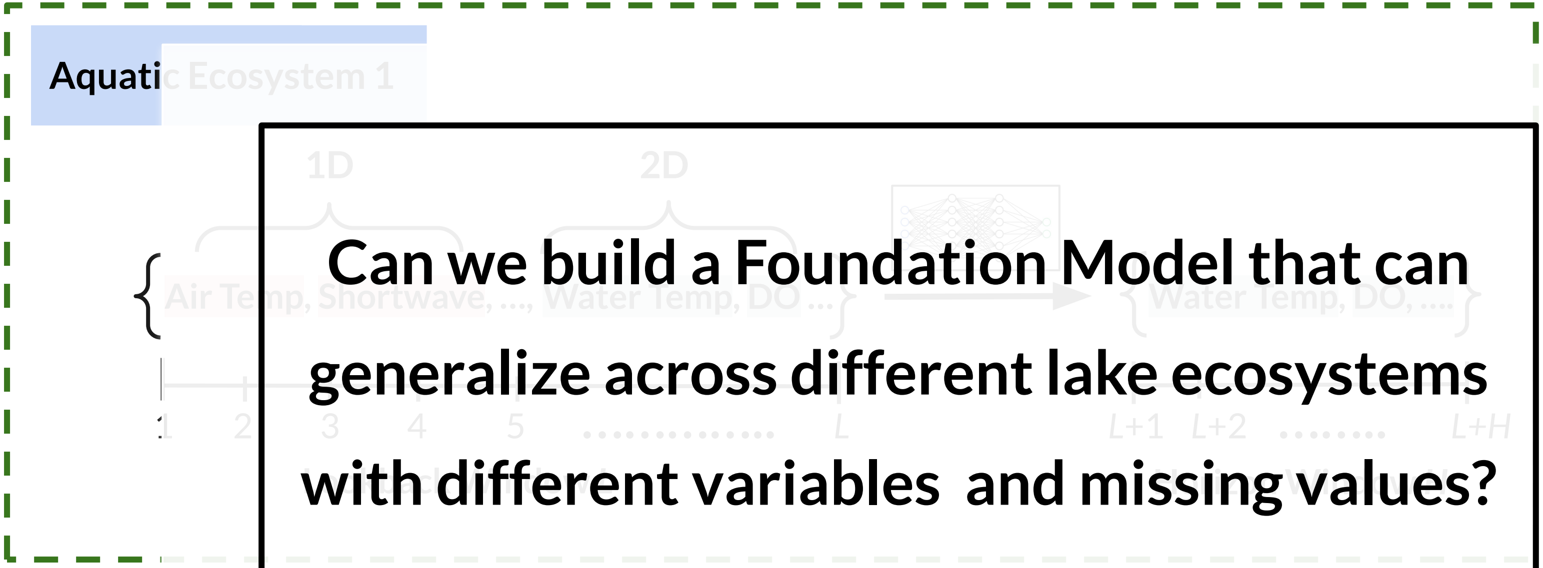
- Different subset of variables available in different ecosystems
- Large amounts of missing data (*e.g., Falling Creeks Reservoir, VA, has 50% missing data, 2017-04 to 2022-10*)

Motivation

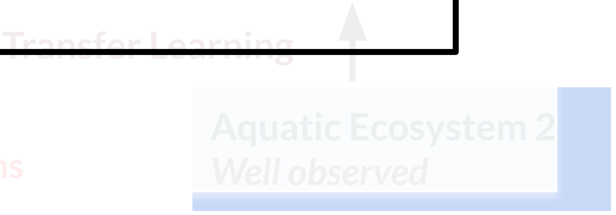


- ⊗ Different subset of variables available in different ecosystems
- ⊗ Large amounts of missing data (*e.g., Falling Creeks Reservoir, VA, has 70% missing data, 2017-04 to 2022-10*)

Motivation

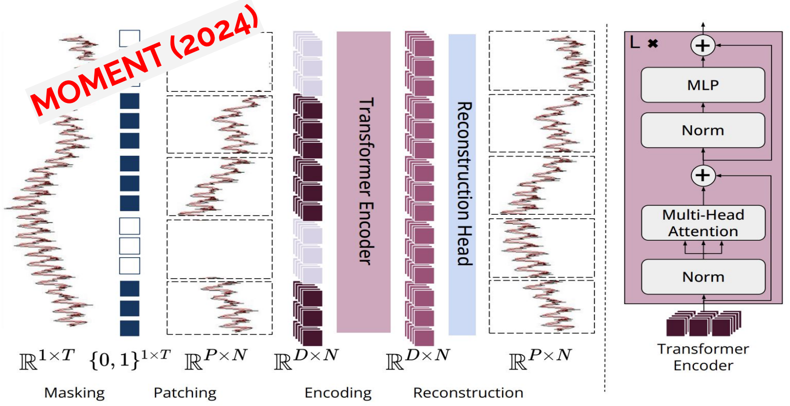


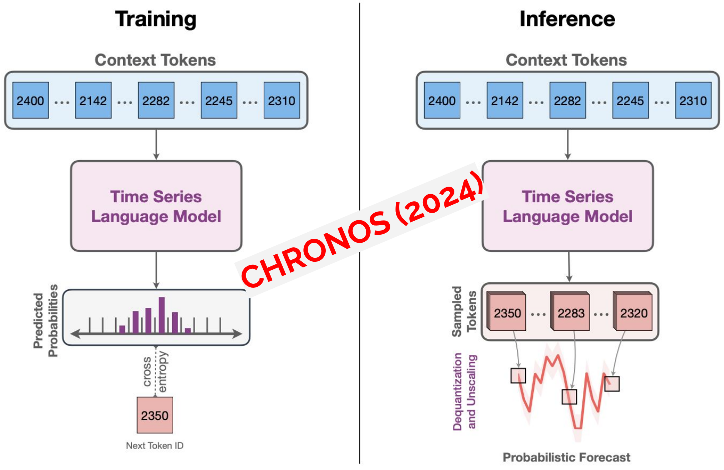
Can we build a Foundation Model that can generalize across different lake ecosystems with different variables and missing values?

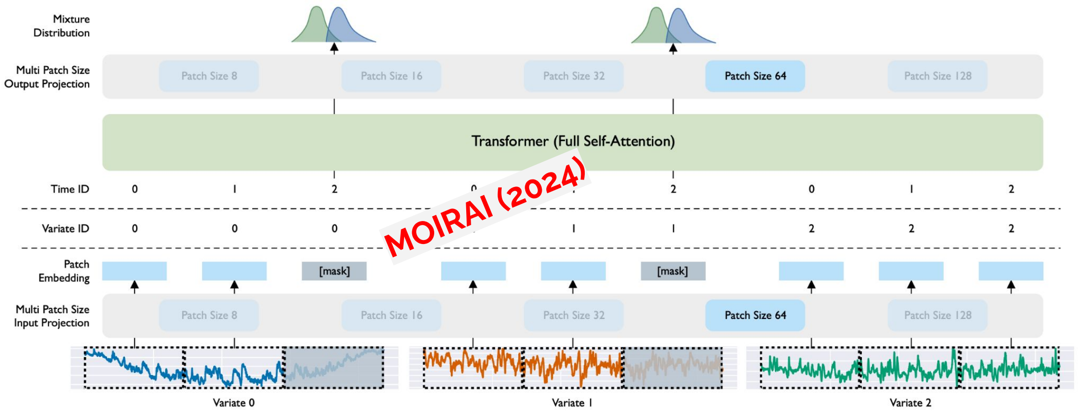


- Different subset of variables available in different ecosystems
- Large amounts of missing data (*e.g., Falling Creeks Reservoir, VA, has 70% missing data, 2017-04 to 2022-10*)

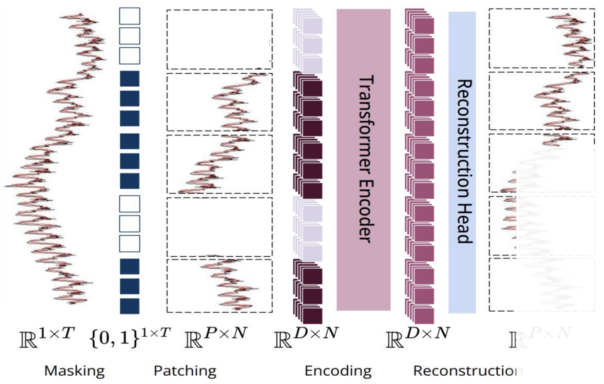
Existing Foundation Models

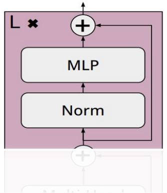


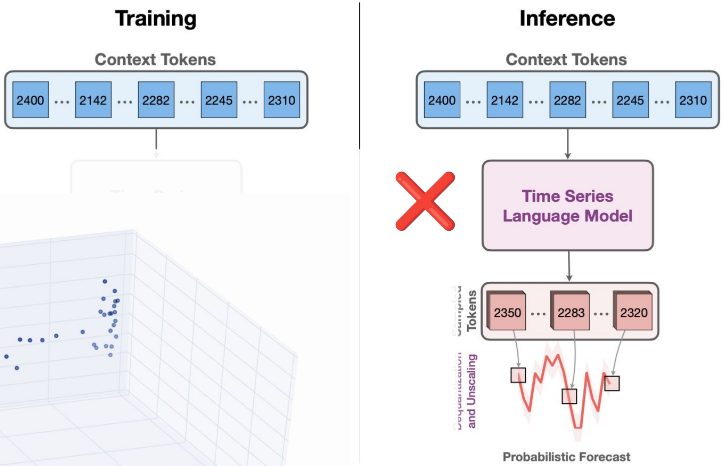


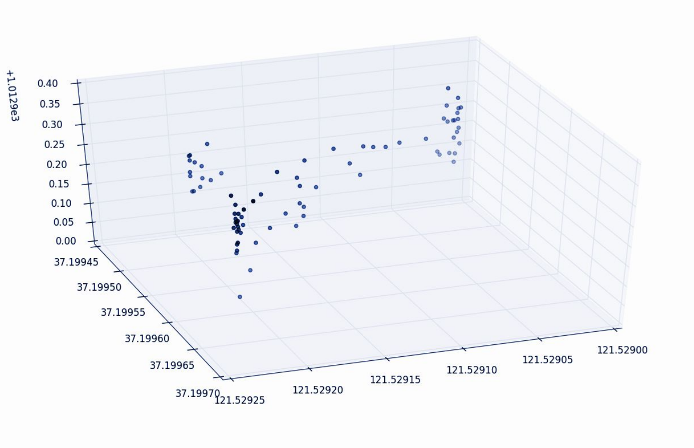


Challenges with existing Foundation Models





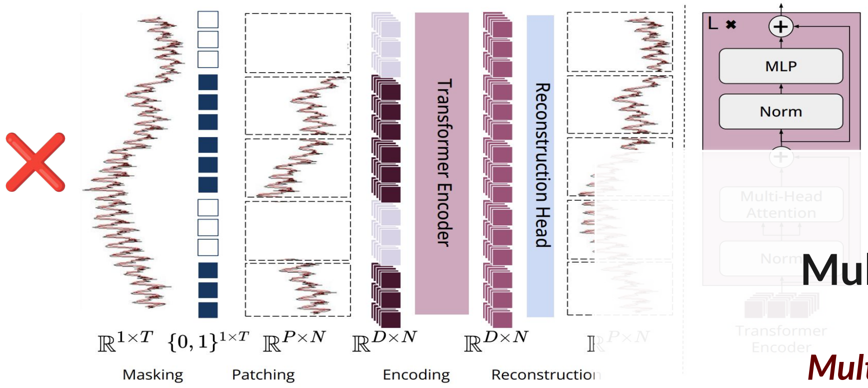


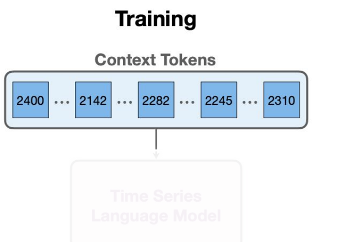


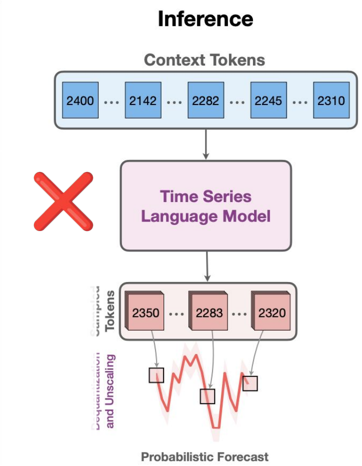
Sparse and Irregular Data

All existing FMs assume fully observed data

Challenges with existing Foundation Models

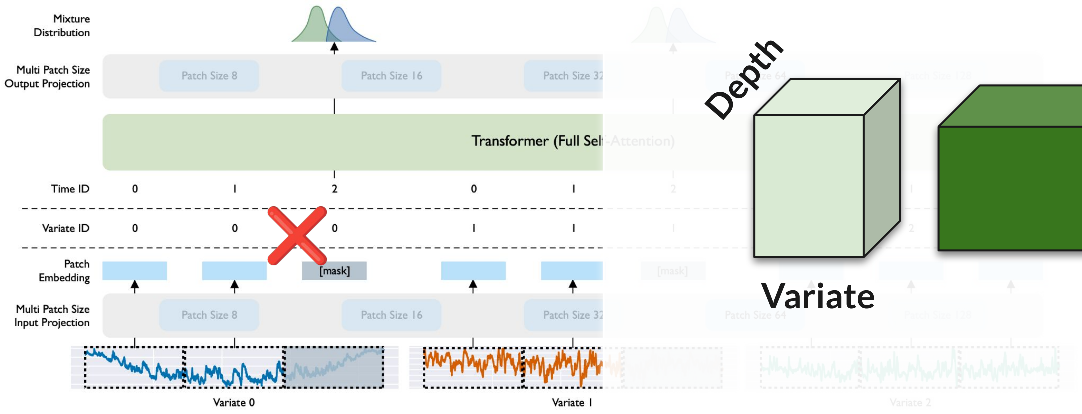


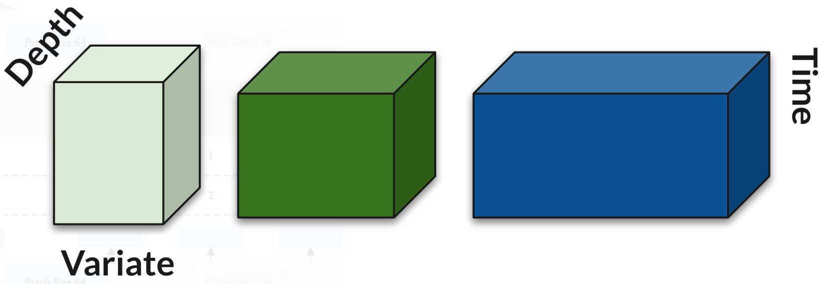




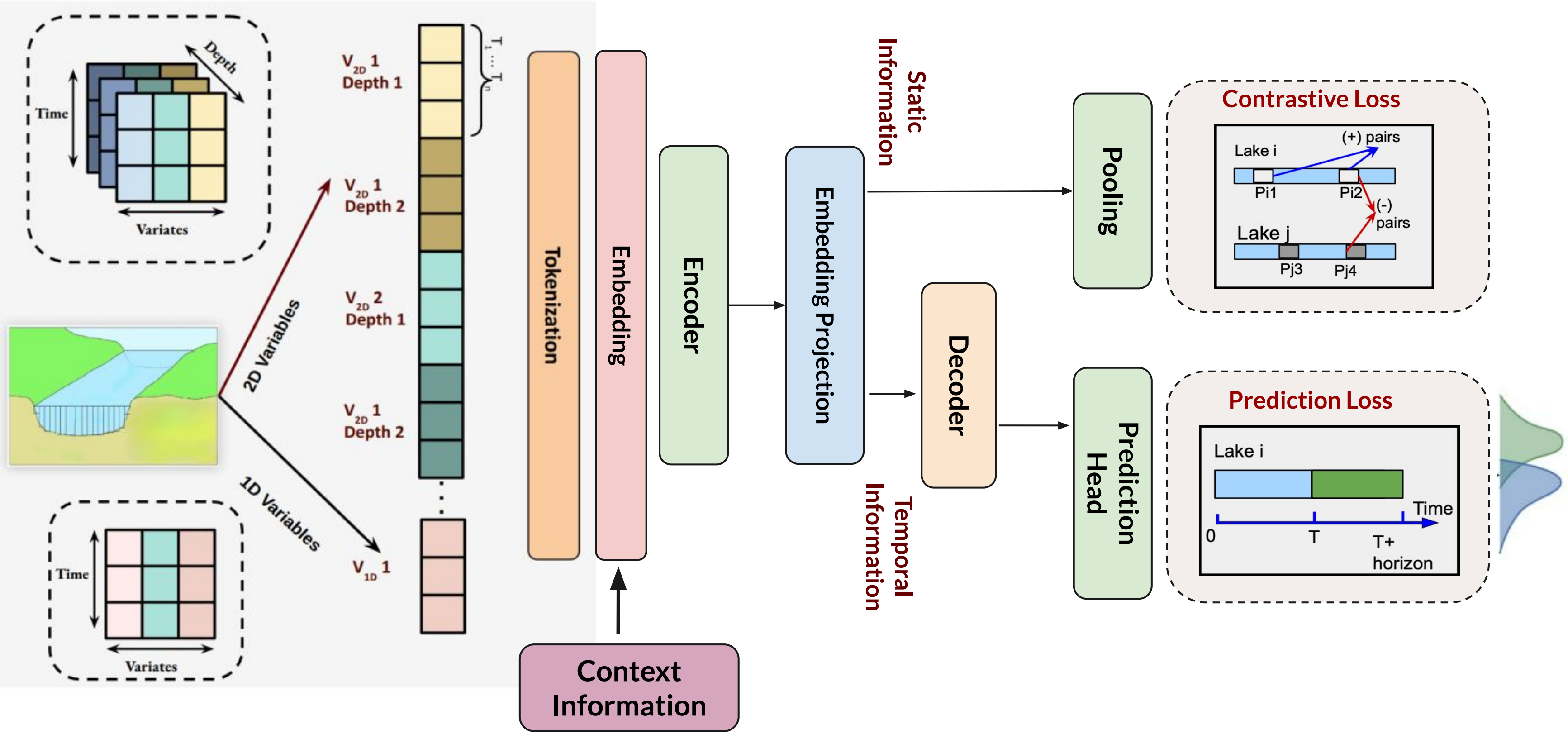
Multivariate and multi-depth

Multivariate FMs treat each depth as independent variable

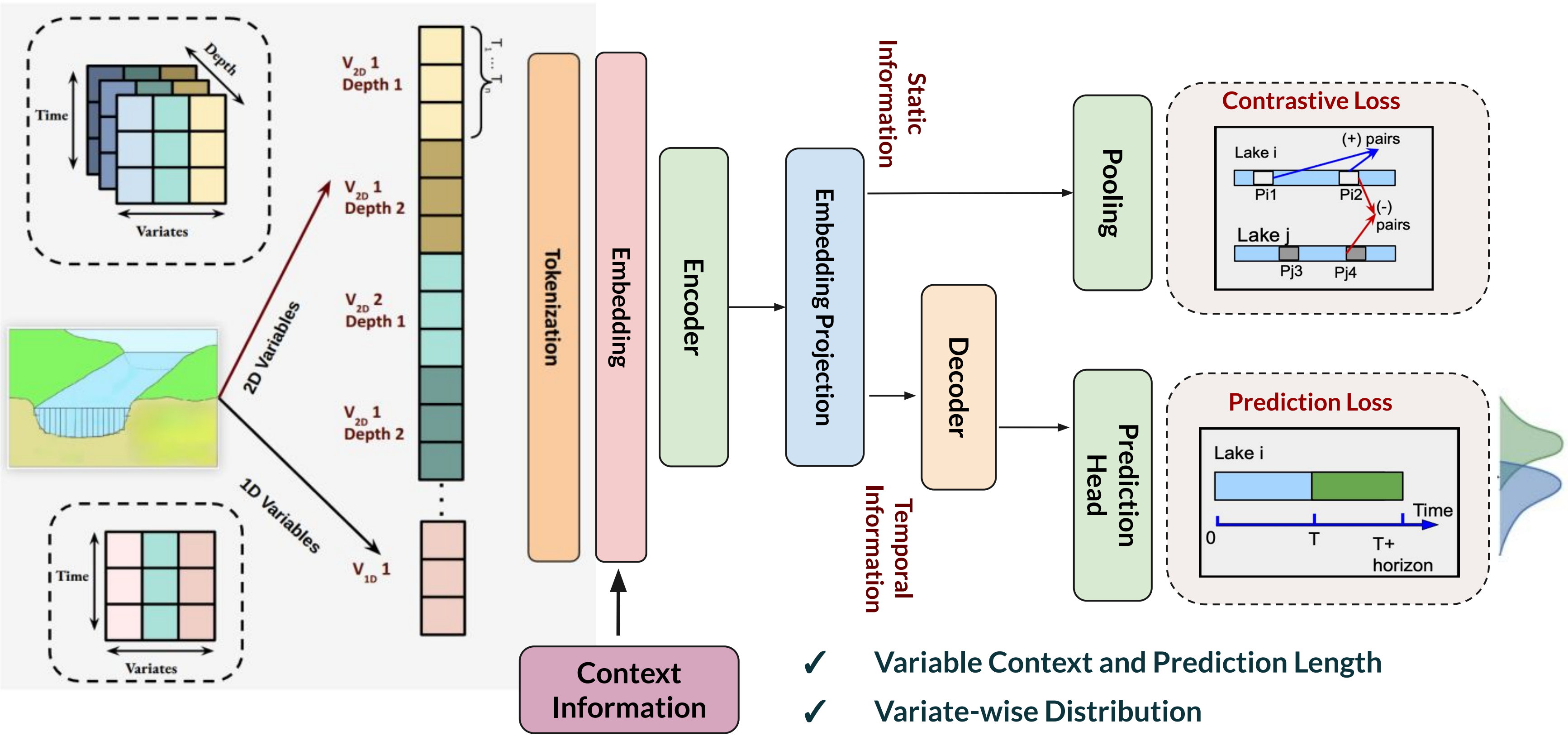




Lake Foundation Model (LakeFM) - An Overview

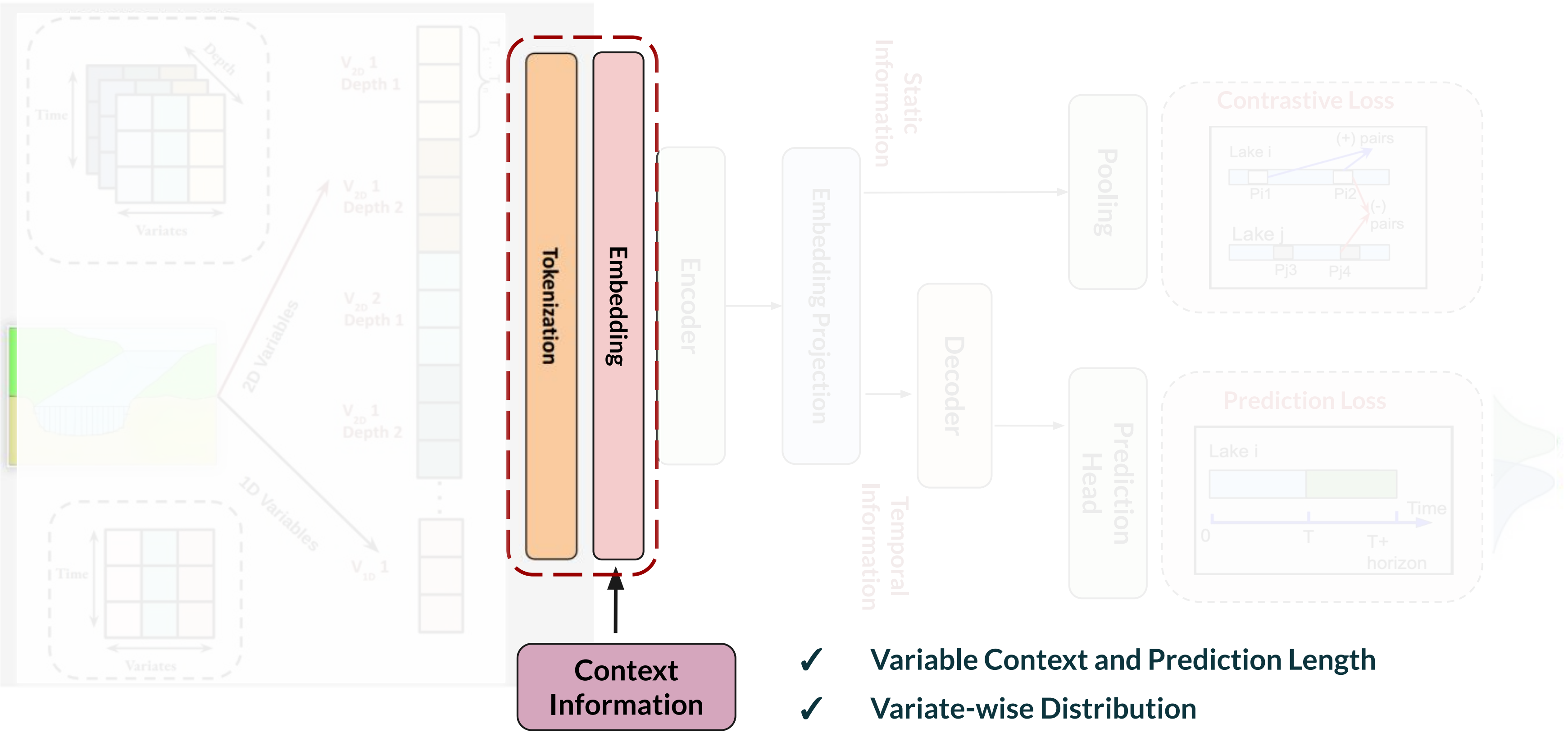


Lake Foundation Model (LakeFM) - An Overview



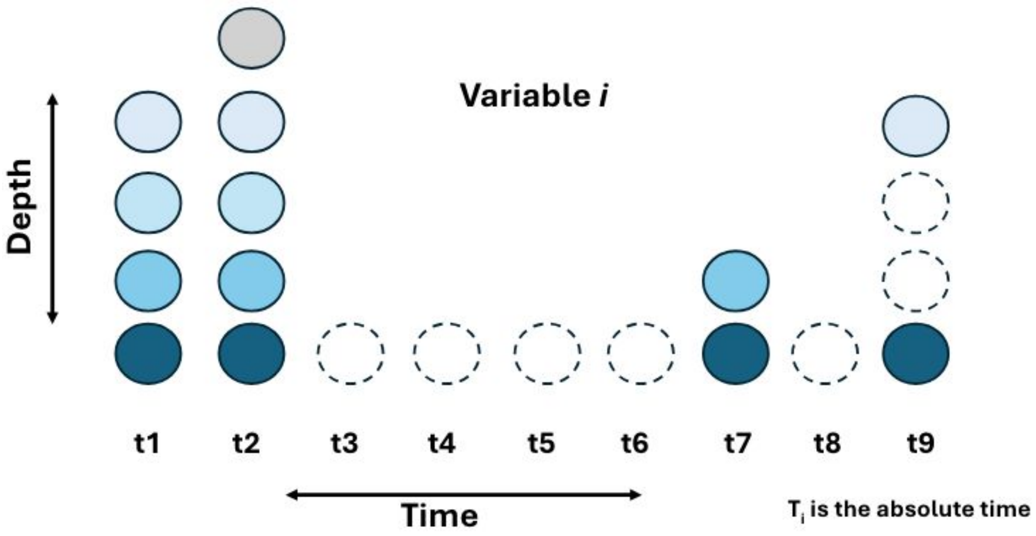
- ✓ Variable Context and Prediction Length
- ✓ Variate-wise Distribution
- ✓ Any-variate any-depth prediction

Lake Foundation Model (LakeFM) - An Overview

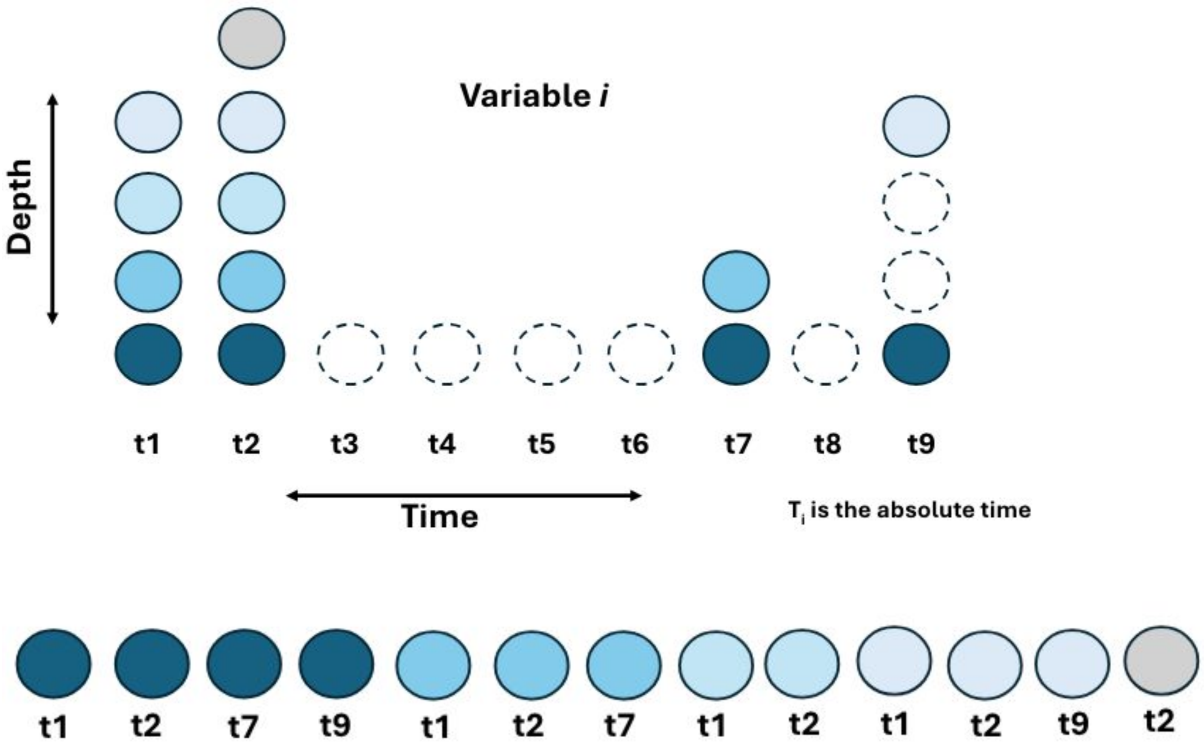


- ✓ Variable Context and Prediction Length
- ✓ Variate-wise Distribution
- ✓ Any-variate any-depth prediction

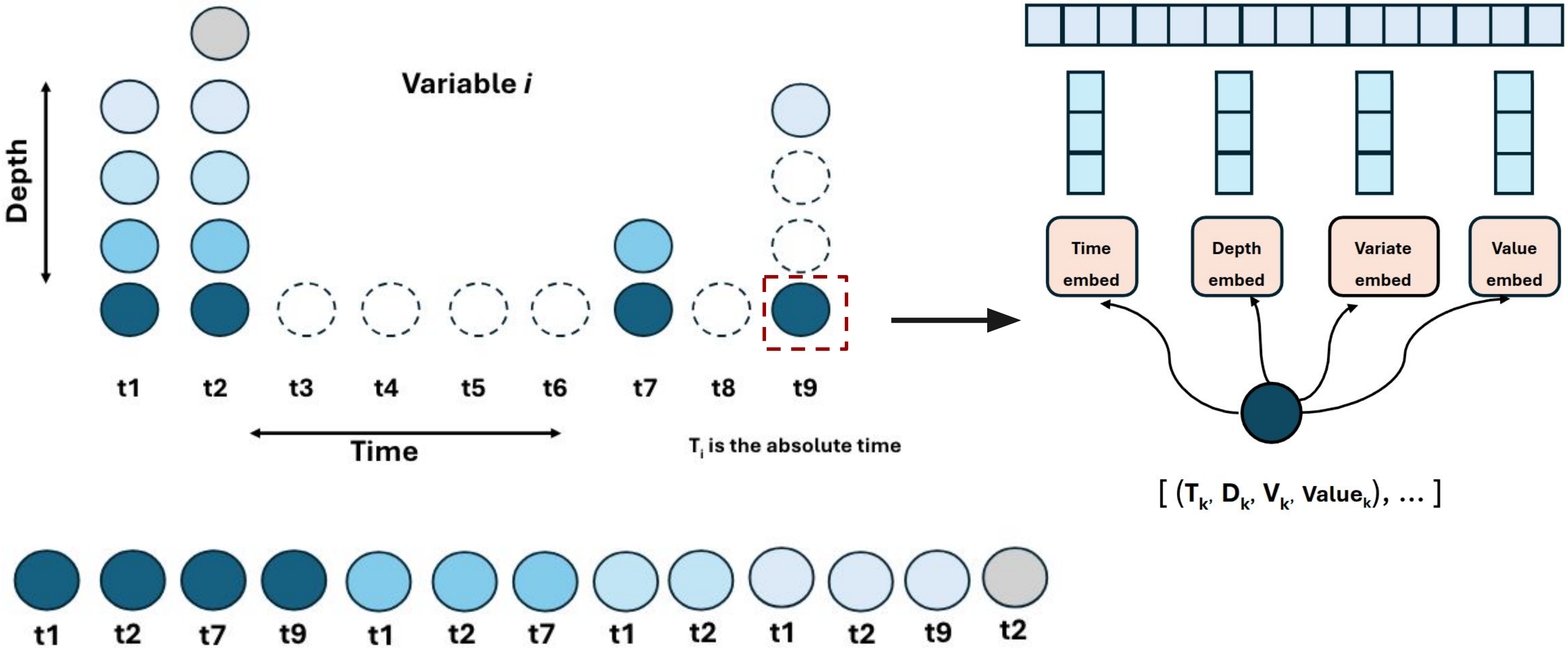
Lake Foundation Model (LakeFM) - Tokenization & Embedding



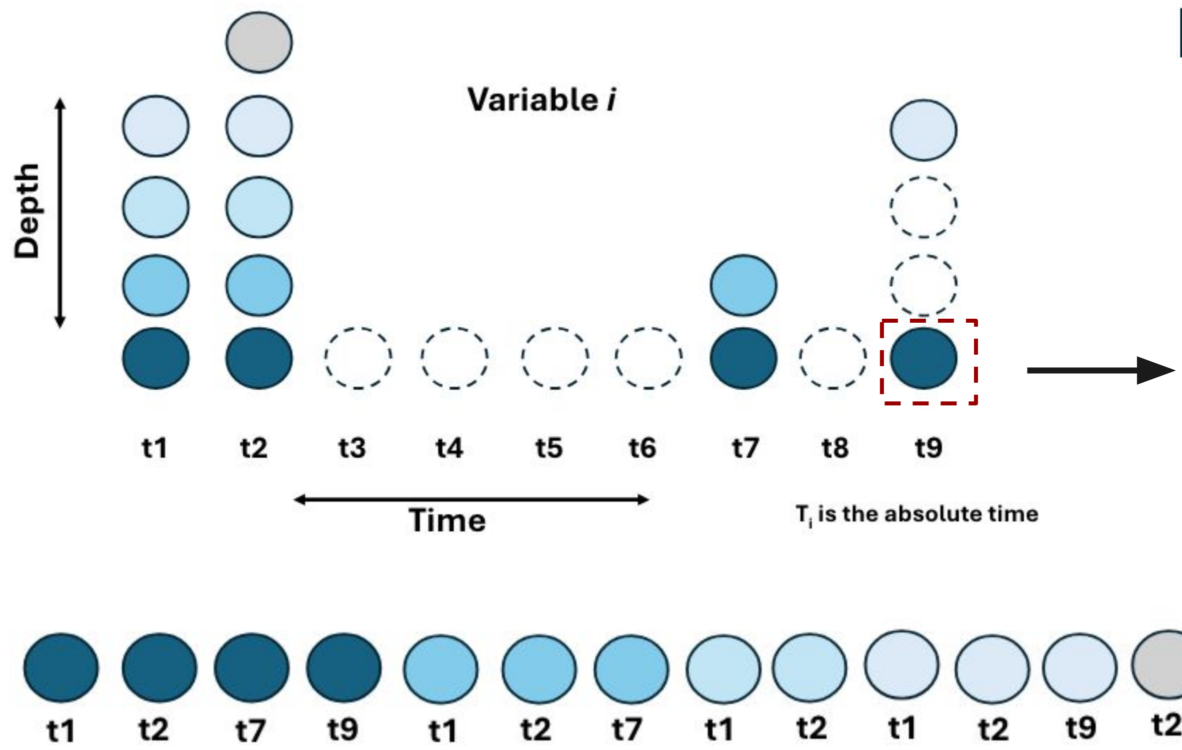
Lake Foundation Model (LakeFM) - Tokenization & Embedding

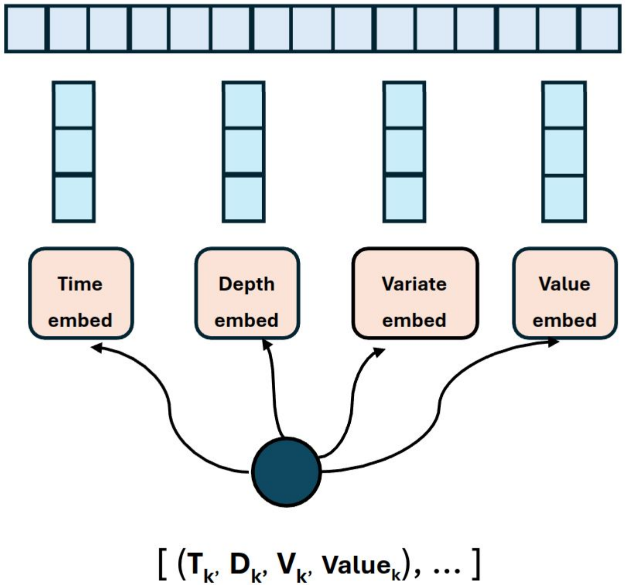


Lake Foundation Model (LakeFM) - Tokenization & Embedding



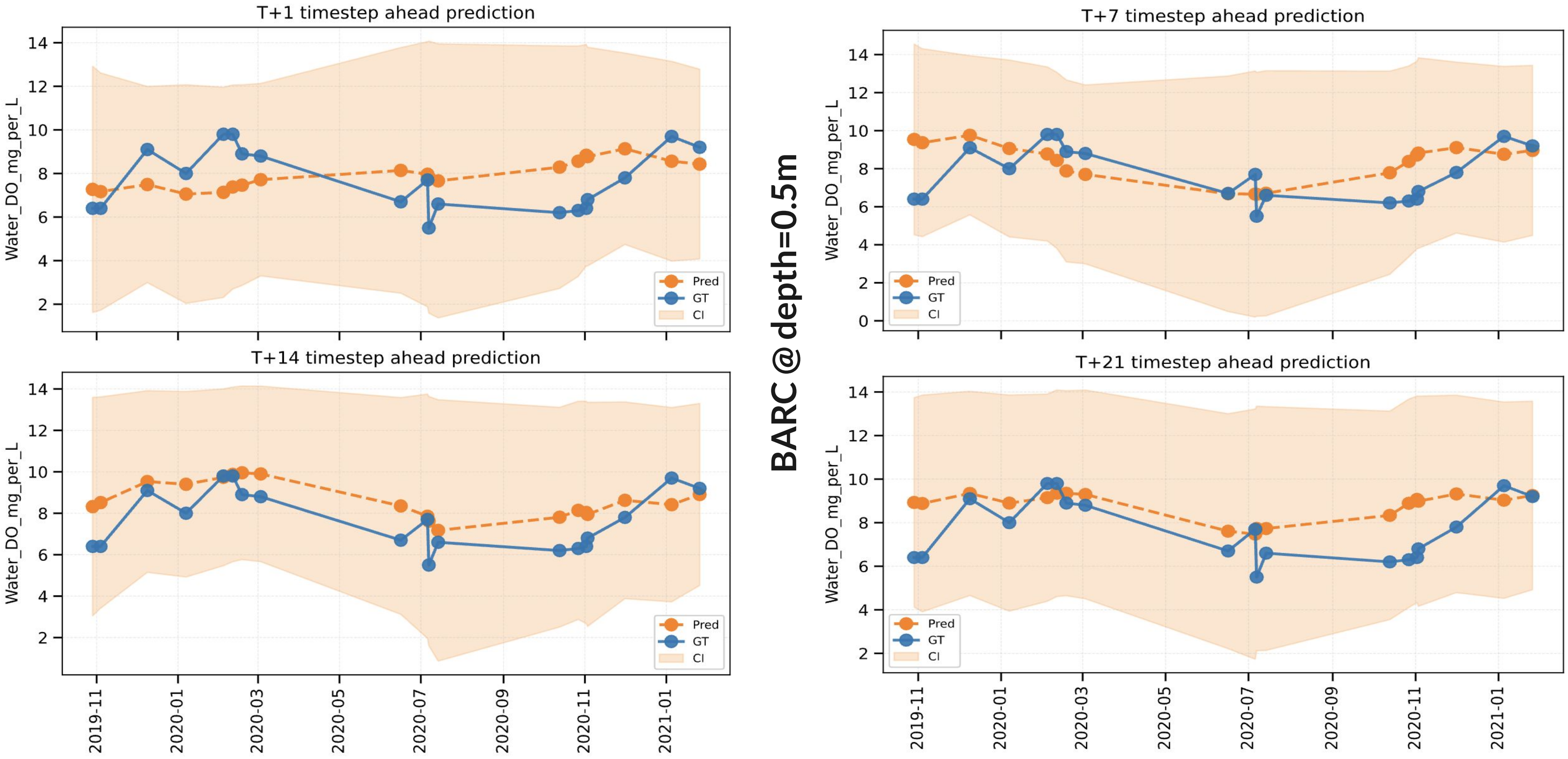
Lake Foundation Model (LakeFM) - Tokenization & Embedding



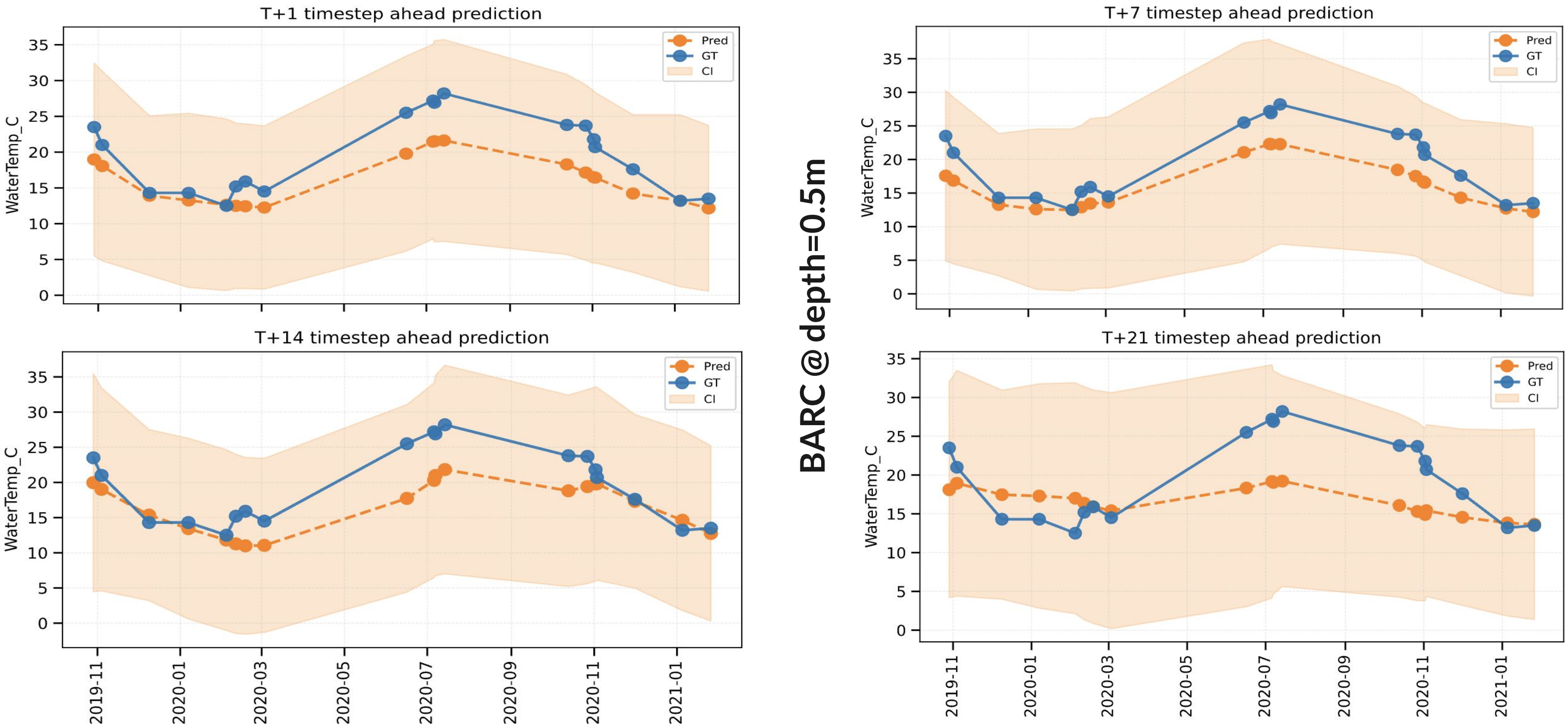


✓ Irregular Grid in depth, time & variates with missing values

Lake Foundation Model (LakeFM) - Findings (I) - Increasing Horizon Length

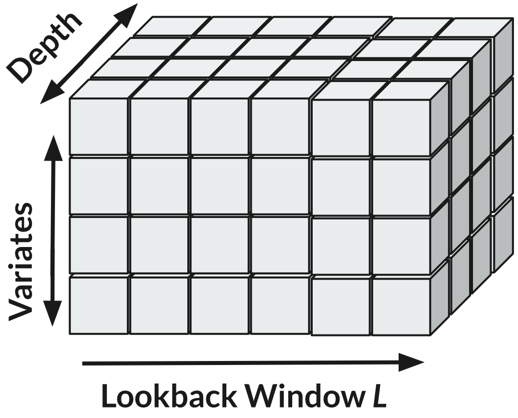


Lake Foundation Model (LakeFM) - Findings (I) - Increasing Horizon Length

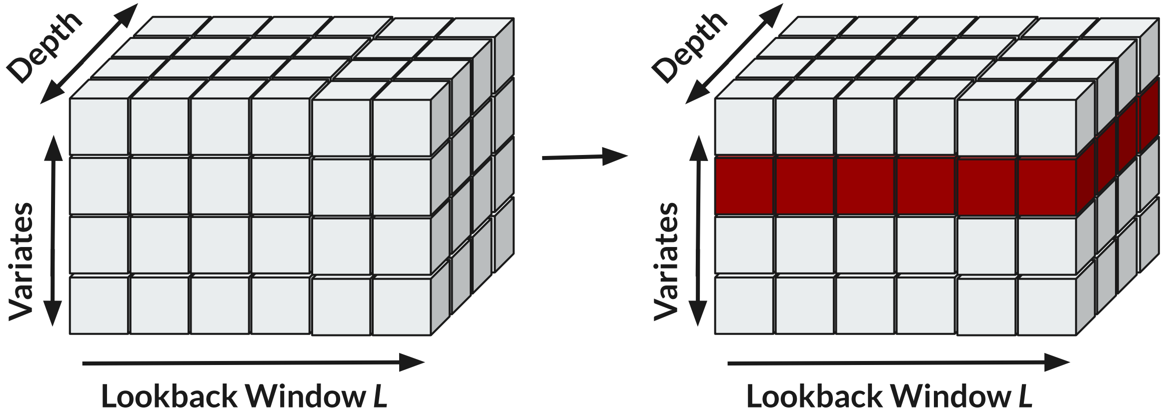


Key Observation : Overall, LakeFM maintains stable performance across increasing horizon lengths

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

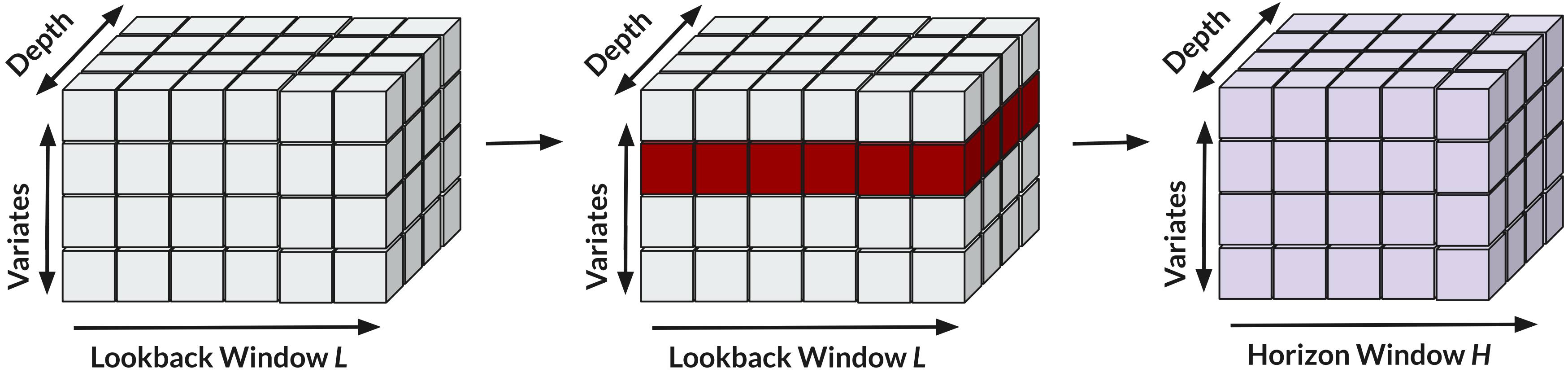


Lake Foundation Model (LakeFM) - Findings (II) - Incomplete Data (Variables)



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

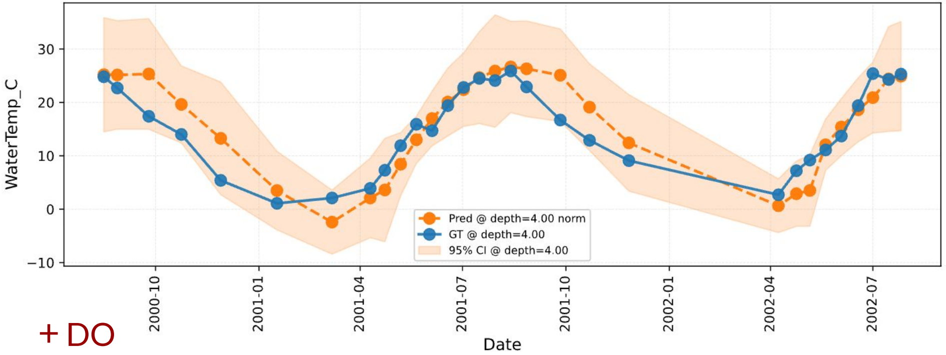
Lake Foundation Model (LakeFM) - Findings (II) - Incomplete Data (Variables)

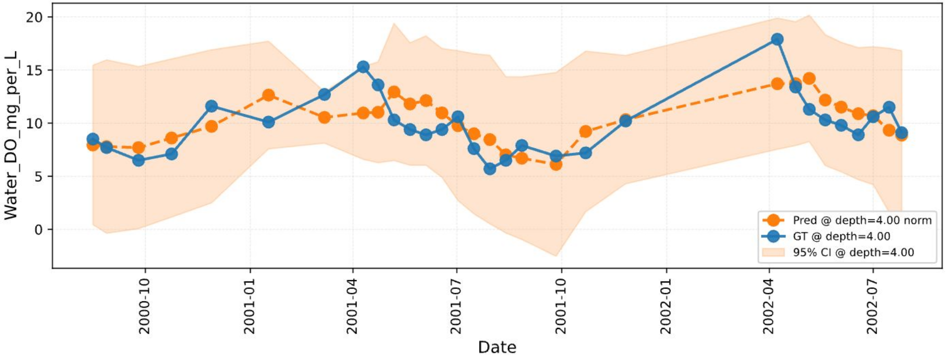


{Air Temp, Shortwave, ..., ~~Water Temp~~, DO ...} → {Air Temp, Shortwave, ..., Water Temp, DO ...}

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

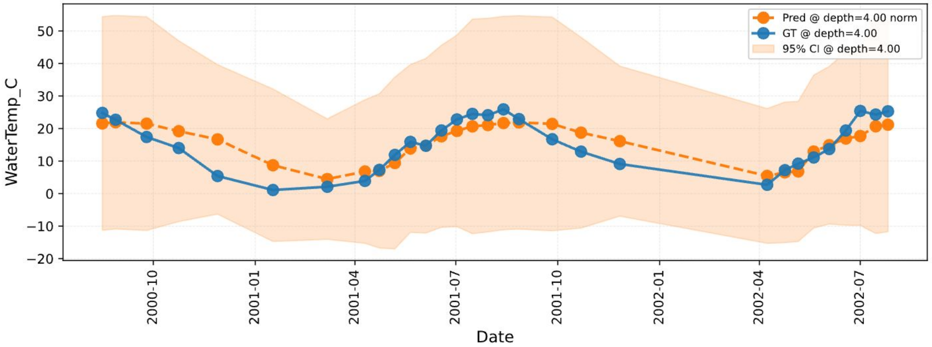
ME : 30 timesteps ahead forecast, at Depth 4.00m (shaded = 95% CI)

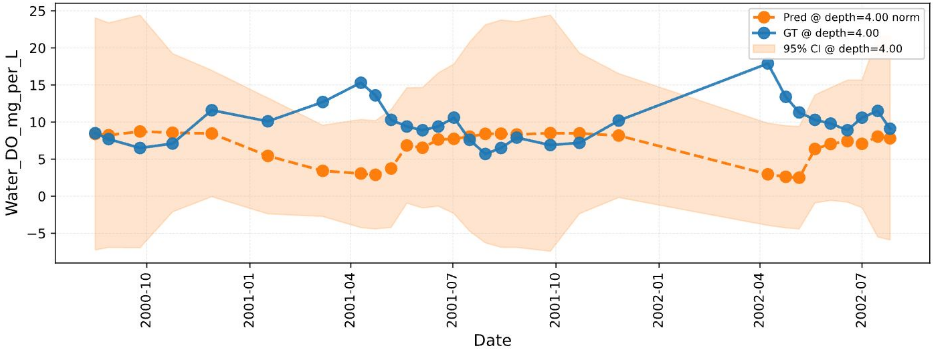




+ DO

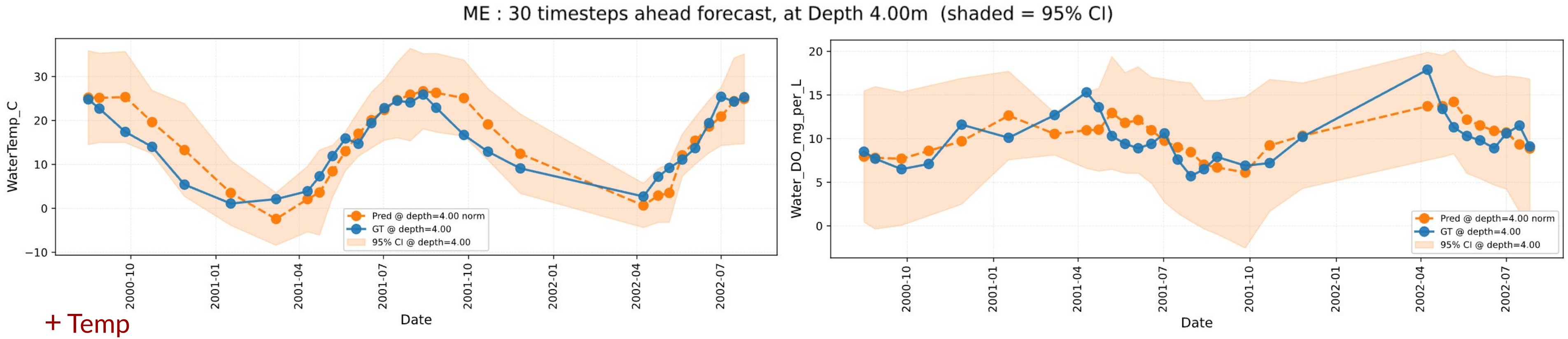
- DO





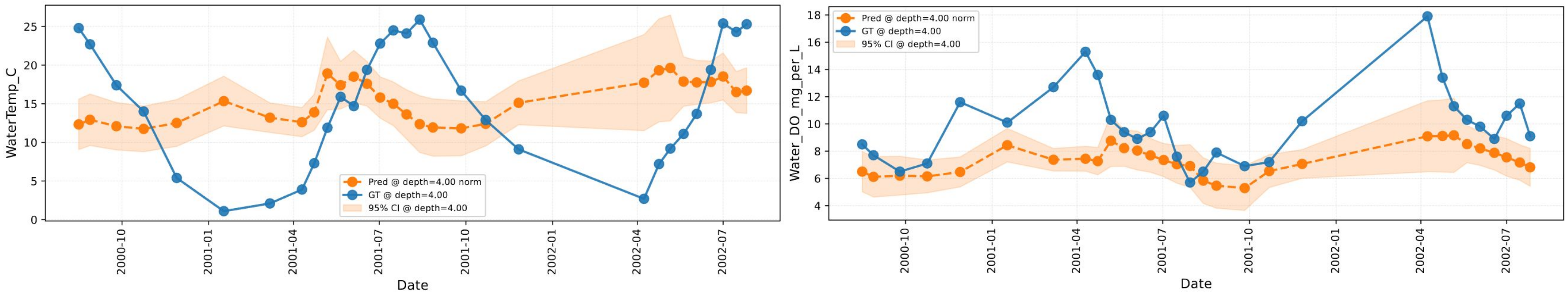
Key Observation : Removing DO from inputs increases uncertainty in predictions of water temperature

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



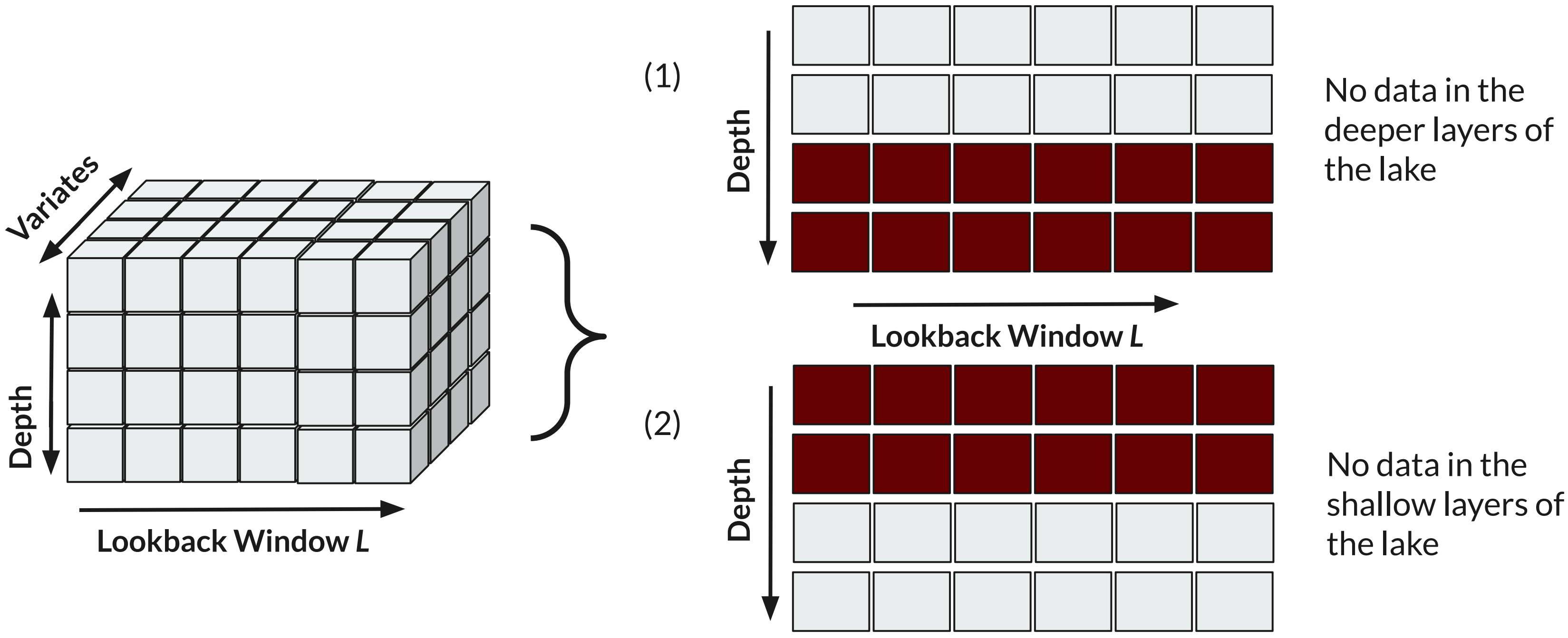
+ Temp

- Temp



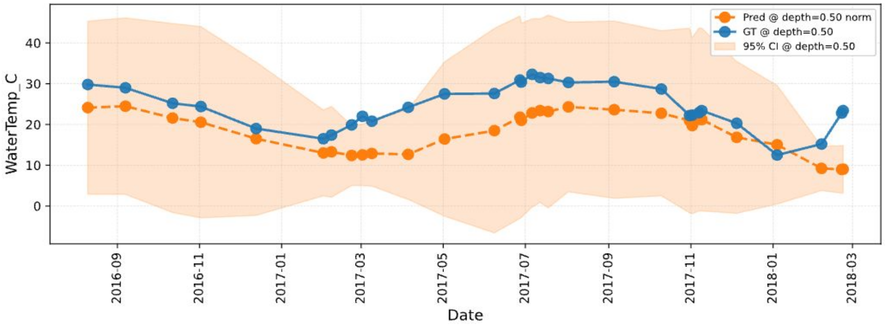
Key Observation : Water temperature is a critical variable. Removing it degrades all predictions.

Lake Foundation Model (LakeFM) - Findings (III) - Incomplete Data (Depth)

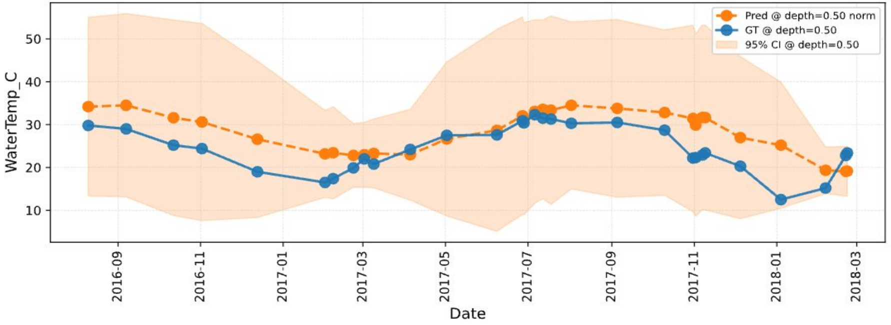


Lake Foundation Model (LakeFM) - Findings (III) - Incomplete Data (Depth)

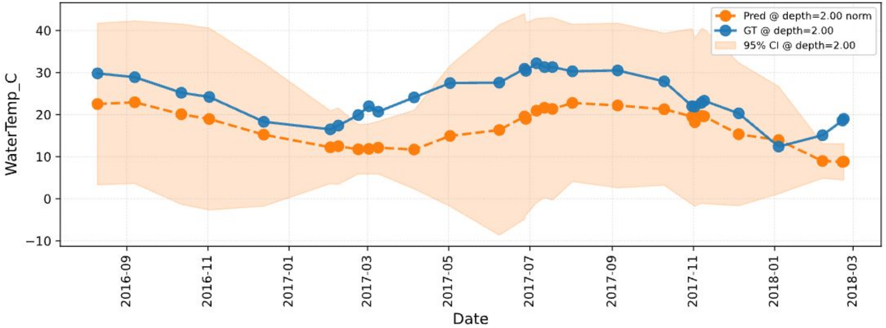
Water Temp Predictions @ 0.5 m using full-depth history



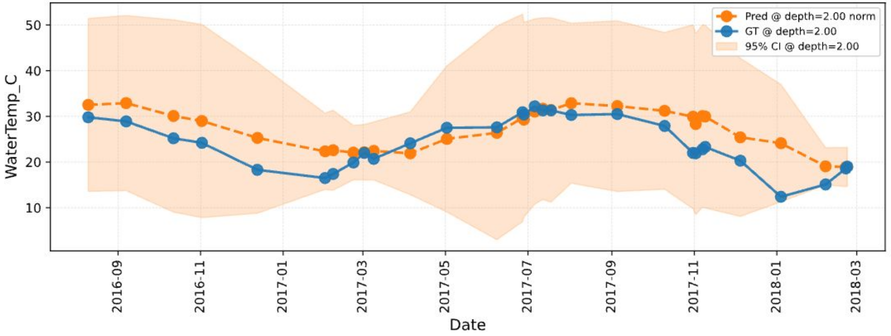
Water Temp Predictions @ 0.5 m using only deeper-depth history



Water Temp Predictions @ 2.0 m using full-depth history



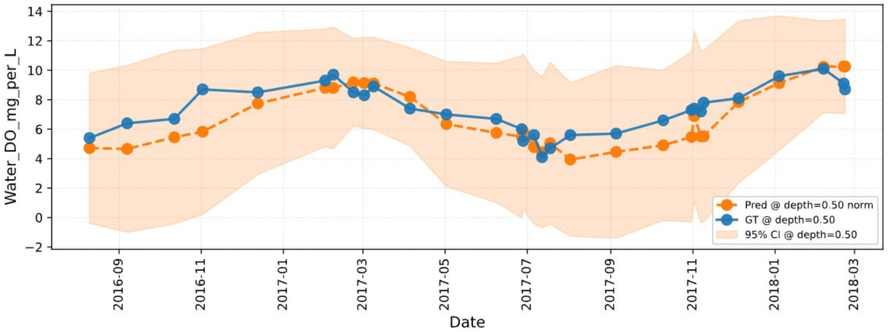
Water Temp Predictions @ 2.0 m using only deeper-depth history

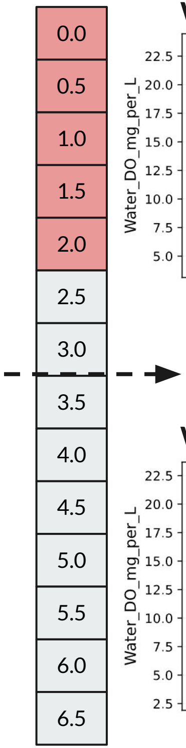


Key Observation : Water temperature predictions remain stable even without shallow-layer variables

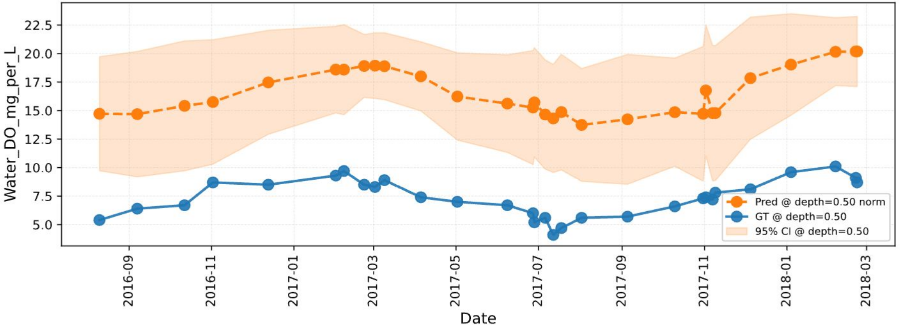
Lake Foundation Model (LakeFM) - Findings (III) - Incomplete Data (Depth)

Water DO Predictions @ 0.5 m using full-depth history

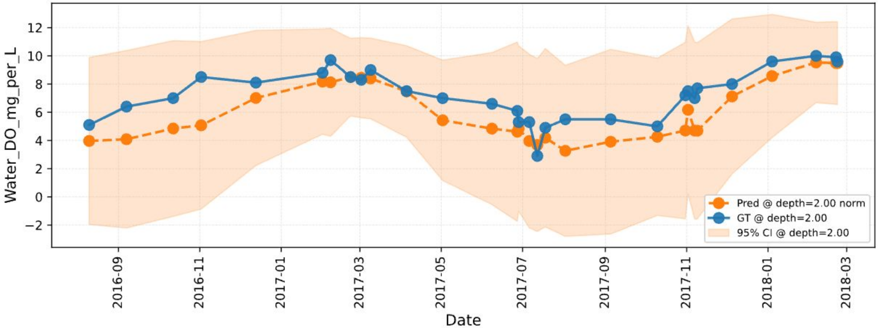




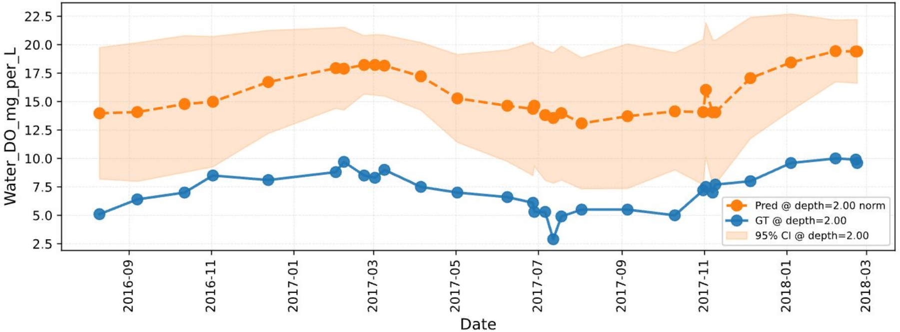
Water DO Predictions @ 0.5 m using only deeper-depth history



Water DO Predictions @ 2.0 m using full-depth history



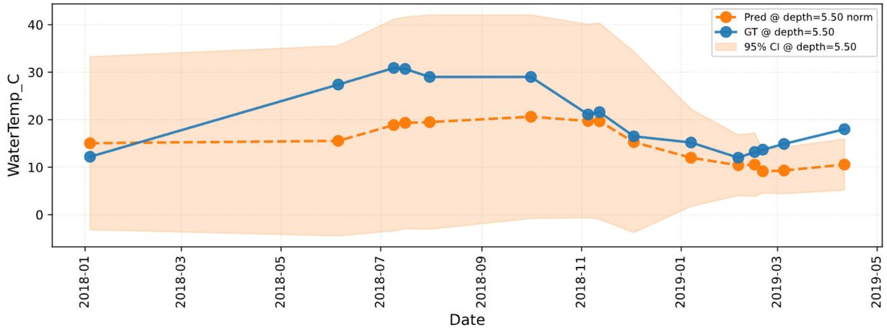
Water DO Predictions @ 2.0 m using only deeper-depth history

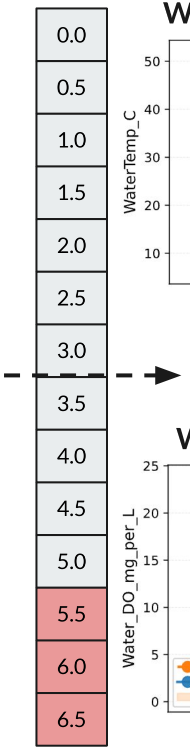


Key Observation : In contrast, DO predictions cannot rely on deeper-layer variables, indicating stronger vertical variability along the water column

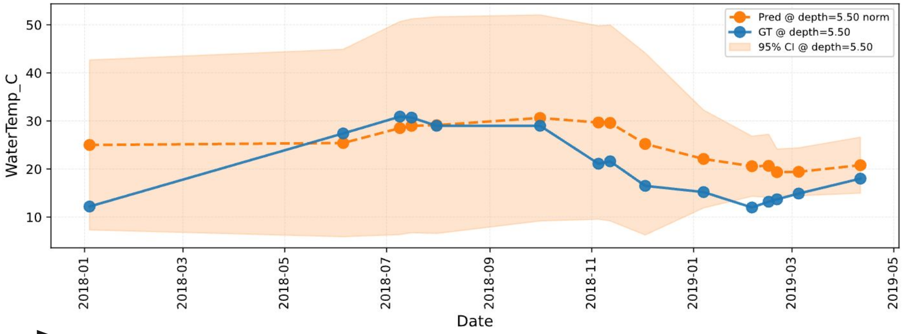
Lake Foundation Model (LakeFM) - Findings (III) - Incomplete Data (Depth)

Water Temp Predictions @ 5.5 m using full-depth history

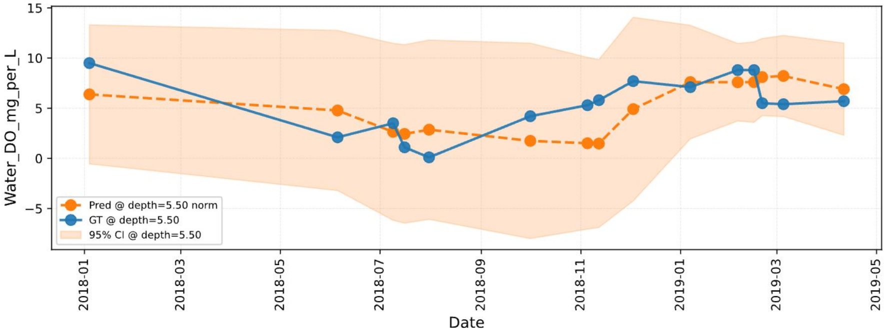




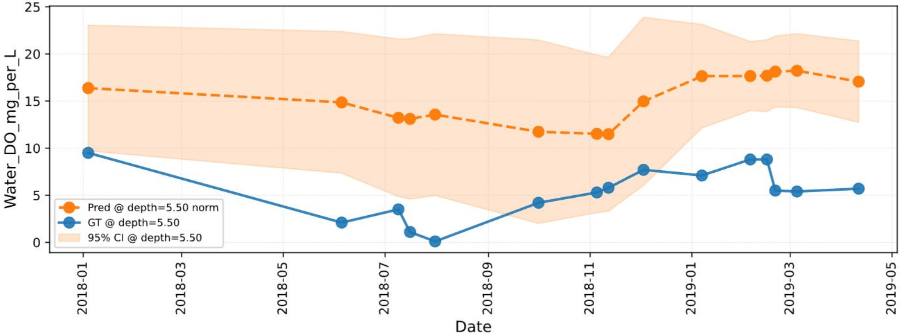
Water Temp Predictions @ 5.5 m using only shallow-depth history



Water DO Predictions @ 5.5 m using full-depth history



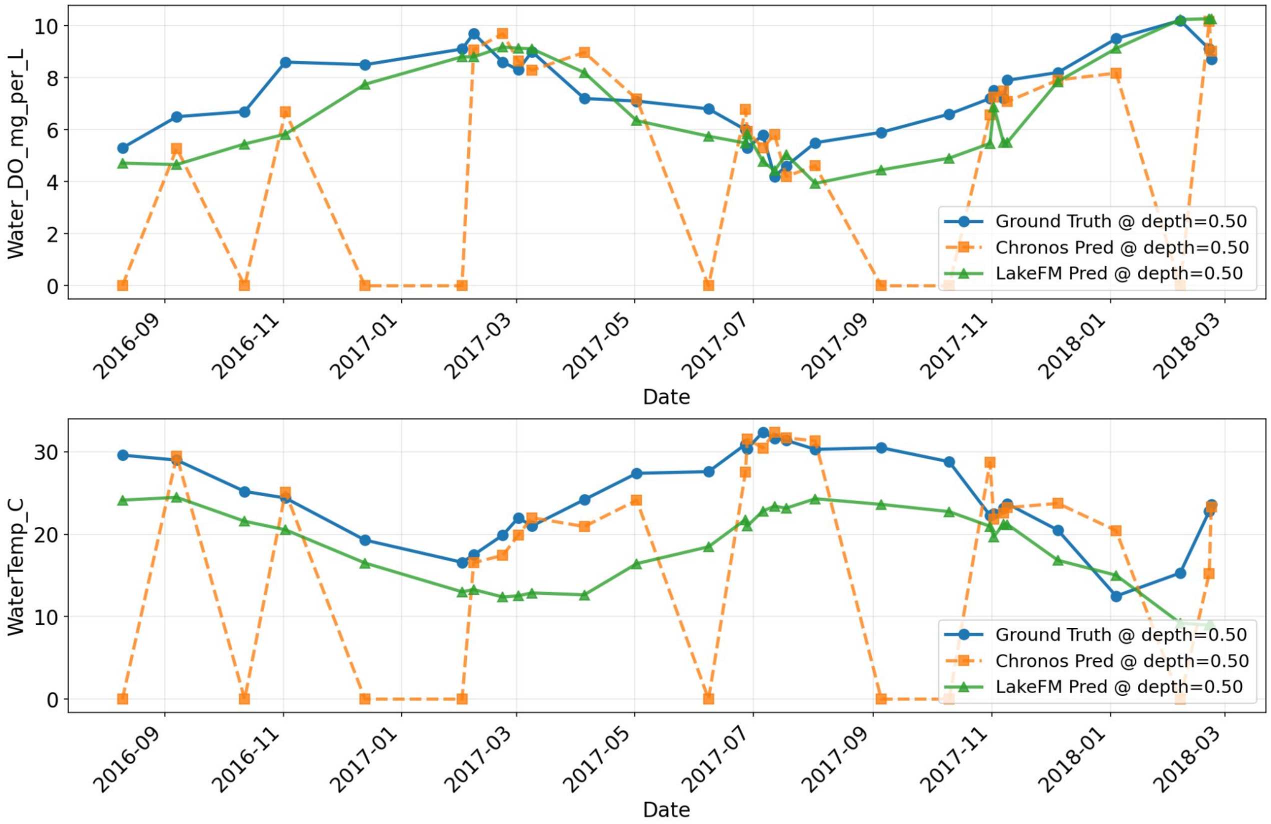
Water DO Predictions @ 5.5 m using only shallow-depth history



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.

Lake Foundation Model (LakeFM) - Findings (IV) - Performance Comparison

Comparing LakeFM predictions (on a horizon window of 30 timesteps) with Chronos Foundation Model on Lake BARC at Depth 0.5m



Key Observation :

Chronos struggles with missing data, leading to context-dependent instability, whereas LakeFM remains stable under the same conditions

Lake Foundation Model (LakeFM) - Ongoing Work

Lake representation analysis

- ❖ Visualization of learned lake embeddings
- ❖ Analyzing seasonal clustering patterns in lake representations
- ❖ Analyzing temporal trajectories of lake representations over time
- ❖ *Geographic structure emerging in embedding space*

Variable representation analysis

- ❖ Visualization of learned variable embeddings
- ❖ Analyzing variable similarities inferred from embedding clusters
- ❖ Empirical verification of embedding-based variable similarities

Thank you