LAKE WATER TEMPERATURE MODELING USING PHYSICS-INFORMED NEURAL NETWORKS

Trieu H. Vo[1]* **Cuong V. Nguyen**[2] **Dongsheng Luo**[1] **Leonardo Bobadilla**[1]

[1]KFSCIS, Florida International University, USA
[2]Department of Mathematical Sciences, Durham University, UK

ABSTRACT

Assessing water quality in bodies of water is important in evaluating the effects of climate change and its anthropogenic impacts. Such assessments often require good models of key indices such as water temperature, pH, or oxygen levels. In this work, we investigate time series models for lake water temperatures at multiple depths and develop a physics-informed neural network based on Koopman embeddings and LSTM that is capable of forecasting water temperatures in the long term. Experiment results show that our model can achieve a good performance and significantly outperforms the conventional LSTM model for this time series forecasting problem.

1 INTRODUCTION

Assessing water quality in bodies of water such as estuaries, coastal areas, and lakes is paramount in evaluating the effects of climate change and its anthropogenic impacts. Measuring key physical, chemical, and biological variables to develop and continuously maintain accurate models for these measurements can help generate early alerts and prevent several environmental problems such as harmful algal blooms, sediment runoff, water pollution, and oxygen decrease events (Reckhow, 1994; Sharma & Kansal, 2013).

Traditional water quality monitoring approaches have relied heavily on in-situ sensing and measurements. These approaches involve collecting water samples at specific points and conducting laboratory analyses of physical indices (e.g., temperature, turbidity, chromaticity, and electrical conductivity), chemical indices (e.g., pH, dissolved oxygen, chemical oxygen demand, biochemical oxygen demand, and total organic carbon), as well as microbiological indices (e.g., total bacteria and total coliforms) (Jähnig & Cai, 2010; Cloete et al., 2016). While in-situ monitoring provides direct measurements with high accuracy, they are often time-consuming, labor-intensive, and costly, as demonstrated in multiple studies (Cai et al., 2008; Giardino et al., 2010; Gholizadeh et al., 2016).

Furthermore, the point-based nature of in-situ sampling makes it challenging to achieve comprehensive spatial coverage, particularly for large water bodies. This limitation becomes especially problematic when monitoring extensive water systems or attempting to capture rapid changes in water quality parameters across different locations. In-situ sensing also has other disadvantages, such as the cost associated with sensors, communication components, and biofouling. Due to their cost, many water bodies may remain unmonitored for the foreseeable future.

The overarching aim of our research is to develop accurate models for several key measurements of water bodies using already available in-situ data combined with simulated data. In this preliminary work, we mainly focus on models for forecasting water temperatures at multiple depth levels. Our work is motivated by recent research on transfer learning of water temperature models from monitored sites to unmonitored ones (Willard et al., 2021) that tries to solve the limited in-situ data problem above.

In this paper, we investigate the use of physics-informed neural network models for forecasting lake water temperatures. We develop a model that combines Koopman embeddings (Geneva & Zabaras,

*Correspondence to: Trieu H. Vo (email: *tvo013@fiu.edu*).

2022) and LSTM (Schmidhuber & Hochreiter, 1997) to effectively forecast water temperatures over a long period of time. The Koopman embeddings have been shown to successfully capture the dynamics of different physical systems (Geneva & Zabaras, 2022). In our experiments, we also show that these embeddings can also improve the performance of time series models for the water temperature forecasting problem.

Other Related Work. Long-term multi-variate time series forecasting presents significant challenges due to the inherent complexity and high dimensionality of temporal data (Zheng et al., 2024). While conventional statistical approaches like Holt-Winters (Chatfield & Yar, 1988) and ARIMA (Contreras et al., 2003) demonstrate effectiveness in short-term predictions, they struggle with extended forecasting horizons. To address these limitations, researchers have explored machine learning techniques, including support vector machines (Huang et al., 2005), random forests (Kane et al., 2014), and gradient boosting approaches (Song & Chen, 2024), which offer improved capability in modeling non-linear patterns, although they necessitate substantial feature engineering. Deep learning also emerged as promising solutions, with various architectures (e.g., RNNs, GRUs, and LSTMs), and attention-based models (Ni et al., 2024) that demonstrate superior performance in capturing long-range dependencies. Contemporary approaches have further advanced the field through innovative architectures: FEDformer (Zhou et al., 2022b) and FiLM (Zhou et al., 2022a) leverage frequency domain transformations, while PatchTST (Nie et al., 2023) and SparseTSF (Lin et al., 2024) introduce novel attention mechanisms and sparse computation strategies. Additionally, hybrid models that integrate statistical methods with deep learning techniques have shown enhanced predictive accuracy across various forecasting tasks (Júnior et al., 2019).

2 Materials and Methods

2.1 Dataset

In this study, we use the data originally published by Willard et al. (2021) that includes water temperature data and lake characteristics for 450 lakes across the Midwestern United States. The dataset contains various attributes such as maximum lake depth, surface area, simulated water temperatures at multiple depths, and in situ temperature observations. The data span 40 years from January 1st, 1980, to December 31st, 2019. For our work, we utilize a subset of this dataset that contains 14,600 daily records of temperatures for a single lake over the entire period. This specific lake is chosen due to its extensive records of water temperatures at different depths and in situ observations.

The temperature data for the above lake were generated using the PB0 model—a process-based lake temperature model to estimate thermal dynamics based on meteorological inputs and lake attributes (Willard et al., 2021). While the dataset contains temperature profiles at 50 different depths for each recorded day, in this preliminary study, we only use the water temperatures at depths of 0, 1, and 2 meters to demonstrate our method. We will use this multi-depth temperature data from the first 32 years to build our model. Data for the next 4 years will be used as a validation set, while data for the last 4 years will be used for testing our model.

2.2 Physics-Informed Neural Networks

We propose to approach the above lake water temperature prediction problem using physics-informed neural network models. In particular, we will transform the multi-depth temperature data to a higher dimensional time series using a Koopman embedding encoder (Geneva & Zabaras, 2022). This time series will then be combined with periodic date features to train an LSTM model (Schmidhuber & Hochreiter, 1997) capable of predicting the next multi-depth temperatures given a previous window of temperature embeddings.

By leveraging the Koopman embeddings, our model can capture the dynamics of the time series, reinforcing relationships between states that occur close together in time or follow similar dynamic patterns. Furthermore, the model's ability to transform nonlinear systems into linear representations enables more efficient training and analysis, which is crucial for studying complex dynamical systems like lake water temperature variation across depths. As shown later in our experiments, our approach allows the LSTM to accurately predict future temperatures at multiple depths, making it an effective tool for forecasting lake temperature patterns with improved accuracy and efficiency.

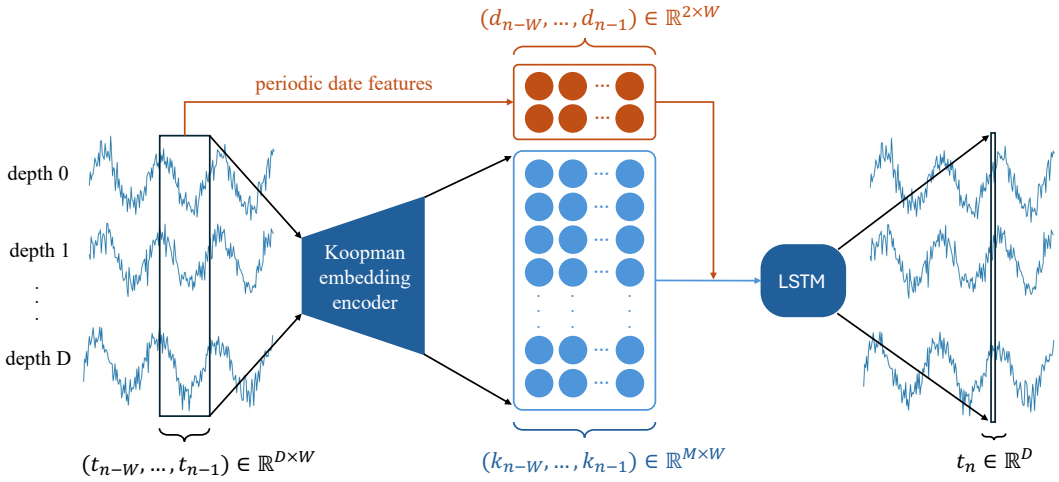


Figure 1: An illustration of our physics-informed model for multi-depth lake temperature prediction.

Model. Formally, let $\boldsymbol{t} = (t_1, t_2, \ldots, t_N)$ be our multi-depth temperatures of a lake over N time steps, where $t_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,D}) \in \mathbb{R}^D$ contains the water temperatures at D depth levels that are measured at time step i. A Koopman embedding encoder is a neural network K that transforms any t_i into a higher dimensional embedding vector $k_i = K(t_i) \in \mathbb{R}^M$, where M is the length of an embedding vector and can be tuned as a hyper-parameter. Assuming we have already trained this encoder, we can utilize it to convert the original data into a higher dimensional time series and then train an LSTM with this time series.

An illustration of our model is shown in Figure 1. The aim of this model is to predict, at any time step n, the multi-depth temperatures t_n using a window of previous W multi-depth temperatures $t_{n-W:n-1} = (t_{n-W}, \ldots, t_{n-1}) \in \mathbb{R}^{D\times W}$. For this purpose, the model first uses the pre-trained Koopman embedding encoder K to convert $t_{n-W:n-1}$ into an embedding matrix $(k_{n-W}, \ldots, k_{n-1}) \in \mathbb{R}^{M\times W}$, where $k_i = K(t_i)$. Then this matrix will be combined with periodic date features $(d_{n-W}, \ldots, d_{n-1}) \in \mathbb{R}^{2\times W}$ to form the full feature matrix $x_{n-W:n-1} = (x_{n-W}, \ldots, x_{n-1}) \in \mathbb{R}^{(M+2)\times W}$, where $x_i = (d_i, k_i)$ and d_i consists of the sine and cosine of the date of year at time step i. The full feature matrix will be used as the input to an LSTM model M_θ with parameters θ, which will return the prediction vector $\hat{t}_n = M_\theta(x_{n-W:n-1}) \in \mathbb{R}^D$ for the next time step.

Training and prediction. Given a training time series $(t_1, t_2, \ldots, t_N)$, we train our model by minimizing the average squared Euclidean distance between $\hat{t}_i$ and t_i. That is, we minimize the loss function $\mathcal{L}(\theta) = \sum_{i=W+1}^{N} \|\hat{t}_i - t_i\|^2/(N-W)$, which can be solved easily using an off-the-shelf optimizer such as SGD or Adam. After training the model, we can use it to predict a new time series $(\tilde{t}_1, \tilde{t}_2, \ldots)$ by sequentially predicting the value $\tilde{t}_i$ using the previous W predictions. For our system, the first W values of the target time series need to be given as seed values.

Koopman embedding encoder. This encoder is part of the Koopman embedding model (Geneva & Zabaras, 2022), which includes an encoder and a decoder neural networks. The model transforms physical state data into high-dimensional embedding vectors, which can be subsequently used by deep learning models. This method effectively models the evolution of a dynamical system over time using the infinite-dimensional linear Koopman operator, which can convert a nonlinear system into a linear one, allowing for more complex state representations while simplifying the system. In this model, the encoder maps the physical state space (e.g., multi-depth lake water temperatures) into an M-dimensional embedding space that captures the underlying dynamics of the system. The decoder then reconstructs the original physical states from this embedding space. In our work, we pre-train the whole Koopman embedding model with our training data but only use the encoder in our physics-informed LSTM model.

Table 1: Comparison of test MSEs of the LSTM baseline and our model. The results show that our model significantly outperforms the baseline on predictions at all depth levels.

Depth (m)	LSTM	LSTM+Koopman (ours)
0	36.306	11.120
1	35.871	10.890
2	34.869	10.390
Overall	35.682	10.800

3 Experiments and Results

Experiment settings. In our experiments, we construct a training time series consisting of the lake water temperature data in the first 32 years (from 1980 to 2011), a validation time series consisting of data in the next 4 years (from 2012 to 2015), and a test time series consisting of data in the last 4 years (from 2016 to 2019). Each time series has 365 time steps per year, with the 29th of February removed from leap years to ensure consistency in the data. The time step size is set to 1 day to allow for a uniform temporal resolution across all samples. Thus, the lengths of the time series are 11,680 for training, 1,460 for validation, and 1,460 for testing.

We first train the Koopman embedding model using the same procedure as Geneva & Zabaras (2022). The model consists of an encoder that maps a 3-dimensional input ($D = 3$) to a 32-dimensional embedding space ($M = 32$) and a decoder that reconstructs the original state. The encoder comprises two fully connected layers (with 500 and 32 units respectively), each of which uses the ReLU activation, layer normalization and dropout. The decoder mirrors this structure, mapping the embeddings back to the original dimensions. The training process minimizes a loss function comprising three components: a reconstruction loss, a Koopman dynamics loss, and a decay term. The model is trained using the Adam optimizer for 70 epochs with a learning rate decay strategy. During training, we monitor the performance of the model on the validation set.

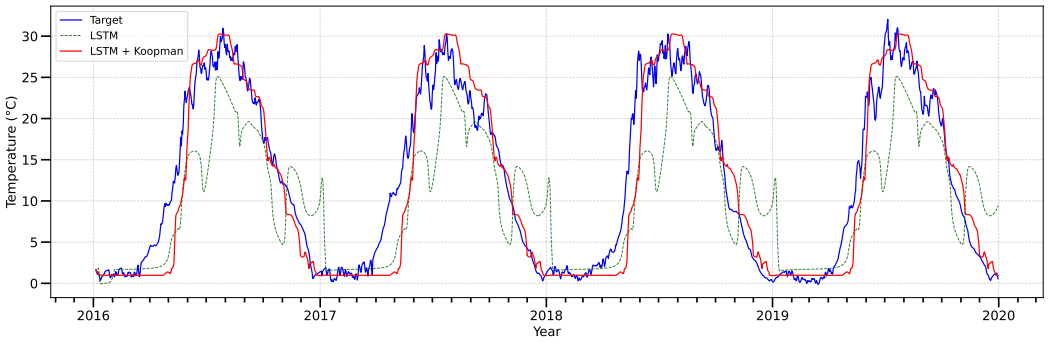
After training the Koopman embedding model, we freeze its encoder network and use it to train the LSTM component of our model (see Figure 1). The LSTM comprises four layers, each containing 64 hidden nodes. The LSTM layers' output is then passed through a fully connected linear layer mapping the 64-dimensional hidden state to a 3-dimensional output, which corresponds to the desired predictions. We train the LSTM with a window size $W = 4$ by running the Adam optimizer with learning rate 0.01 for 250 epochs. After training, we employ the model to predict the whole test time series, using the first 4 days' temperatures as seed data.

To evaluate the effectiveness of our approach, we compare it with a baseline method that trains an LSTM model without the Koopman embeddings. This LSTM baseline is trained using the same network architecture and number of epochs as in our method, but with the learning rate set to 0.001 to ensure a good performance. Both the baseline (named **LSTM**) and our method (named **LSTM+Koopman**) are evaluated using the mean squared error (MSE) on the test time series.

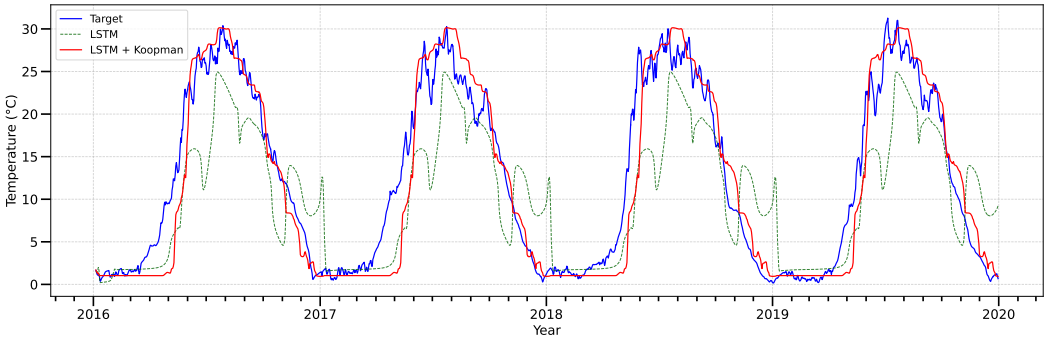
Results. In Table 1 and Figure 2, we show the results of our experiments. From the results, it is clear that our LSTM+Koopman method significantly outperforms the LSTM baseline on the predictions at all three depth levels. Overall, our method reduces the test MSE from 35.682 to 10.800, a nearly 70% improvement. These results confirm that using physics-informed features, particularly Koopman embeddings in this case, can improve the performance of the LSTM model for the lake water temperature prediction problem.

4 Conclusion and Future Work

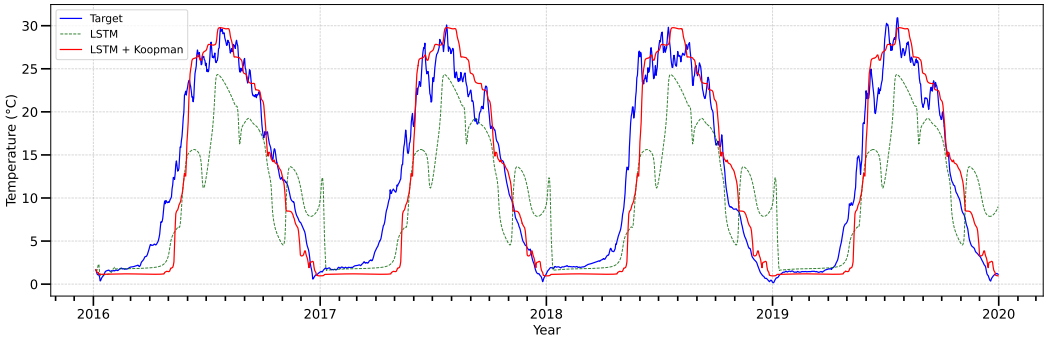
In this paper, we presented a new approach for predicting lake water temperatures using physics-informed neural networks. Using the Koopman embeddings, our method can capture the temperature dynamics efficiently, reducing the mean squared errors of predictions by nearly 70%. Our findings confirm the effectiveness of using physics-informed features to enhance the performance of deep learning models for the prediction of lake water temperature.



(a) 0 meter (surface level).



(b) 1 meter.



(c) 2 meters.

Figure 2: The target and predicted time series of the LSTM baseline and our method at different depth levels.

For future work, we will expand our research by combining both in-situ and simulated data to further improve the model's prediction ability. We will also explore more powerful models such as transformers (Waswani et al., 2017) and broaden our research to other water quality indices such as pH or oxygen levels to provide more comprehensive solutions for monitoring and managing lake resources efficiently.

ACKNOWLEDGEMENTS

This work is partially funded by grants from the National Science Foundation (IIS-2024733, IIS-2331908), the Office of Naval Research (N00014-23-1-2789), the U.S. Department of Homeland Security (23STSLA00016-01-00), the U.S. Department of Defense (78170-RT-REP), and the Florida Department of Environmental Protection (INV31).

REFERENCES

L Cai, P Liu, and C Zhi. Discussion on remote sensing based on water quality monitoring methods. *Geomatics & Spatial Information Technology*, 31(4):68–73, 2008.

Chris Chatfield and Mohammad Yar. Holt-Winters forecasting: Some practical issues. *Journal of the Royal Statistical Society Series D: The Statistician*, 37(2):129–140, 1988.

Niel Andre Cloete, Reza Malekian, and Lakshmi Nair. Design of smart sensors for real-time water quality monitoring. *IEEE Access*, 4:3975–3990, 2016.

Javier Contreras, Rosario Espinola, Francisco J Nogales, and Antonio J Conejo. ARIMA models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18(3):1014–1020, 2003.

Nicholas Geneva and Nicholas Zabaras. Transformers for modeling physical systems. *Neural Networks*, 146:272–289, 2022.

Mohammad Haji Gholizadeh, Assefa M Melesse, and Lakshmi Reddi. A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors*, 16(8):1298, 2016.

Claudia Giardino, Mariano Bresciani, Paolo Villa, and Angiolo Martinelli. Application of remote sensing in water resource management: the case study of Lake Trasimeno, Italy. *Water Resources Management*, 24:3885–3899, 2010.

Wei Huang, Yoshiteru Nakamori, and Shou-Yang Wang. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10):2513–2522, 2005.

Sonja C Jähnig and Qinghua Cai. River water quality assessment in selected Yangtze tributaries: Background and method development. *Journal of Earth Science*, 21(6):876–881, 2010.

Domingos S de O Santos Júnior, João FL de Oliveira, and Paulo SG de Mattos Neto. An intelligent hybridization of ARIMA with machine learning models for time series forecasting. *Knowledge-Based Systems*, 175:72–86, 2019.

Michael J Kane, Natalie Price, Matthew Scotch, and Peter Rabinowitz. Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*, 15:1–9, 2014.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. SparseTSF: Modeling long-term time series forecasting with 1k parameters. In *International Conference on Machine Learning*, 2024.

Zelin Ni, Hang Yu, Shizhan Liu, Jianguo Li, and Weiyao Lin. BasisFormer: Attention-based time series forecasting with learnable and interpretable basis. In *Advances in Neural Information Processing Systems*, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Kenneth H Reckhow. Water quality simulation modeling and uncertainty analysis for risk assessment and decision making. *Ecological Modelling*, 72(1-2):1–20, 1994.

Jürgen Schmidhuber and Sepp Hochreiter. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Deepshikha Sharma and Arun Kansal. Assessment of river quality models: A review. *Reviews in Environmental Science and Bio/Technology*, 12:285–311, 2013.

Xuefei Song and Zhong Shuo Chen. Enhancing financial time series forecasting in the shipping market: A hybrid approach with Light Gradient Boosting Machine. *Engineering Applications of Artificial Intelligence*, 136:108942, 2024.

A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Jared D Willard, Jordan S Read, Alison P Appling, Samantha K Oliver, Xiaowei Jia, and Vipin Kumar. Predicting water temperature dynamics of unmonitored lakes with meta-transfer learning. *Water Resources Research*, 57(7), 2021.

Xu Zheng, Tianchun Wang, Wei Cheng, Aitian Ma, Haifeng Chen, Mo Sha, and Dongsheng Luo. Parametric augmentation for time series contrastive learning. In *International Conference on Learning Representations*, 2024.

Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. FiLM: Frequency improved Legendre memory model for long-term time series forecasting. In *Advances in Neural Information Processing Systems*, 2022a.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 2022b.