Predictive Understanding of Lake Water Temperature and Dissolved Oxygen Profiles Across the Red River Basin Through Interpretable Machine Learning

Isabela Suaza Sierra
University of Texas at El Paso

Follow this and additional works at: https://scholarworks.utep.edu/open_etd

Part of the Artificial Intelligence and Robotics Commons, Environmental Sciences Commons, Geology Commons, and the Hydrology Commons

Recommended Citation
https://scholarworks.utep.edu/open_etd/4151

This is brought to you for free and open access by ScholarWorks@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of ScholarWorks@UTEP. For more information, please contact lweber@utep.edu.
PREDICTIVE UNDERSTANDING OF LAKE WATER TEMPERATURE AND DISSOLVED OXYGEN PROFILES ACROSS THE RED RIVER BASIN THROUGH INTERPRETABLE MACHINE LEARNING

ISABELA SUAZA SIERRA

Master’s Program in Geological Sciences

APPROVED:

Hernan A. Moreno, Chair, Ph.D.

Deana D. Pennington, Ph.D.

Thomas M. Neeson, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School
PREDICTIVE UNDERSTANDING OF LAKE WATER TEMPERATURE AND DISSOLVED OXYGEN PROFILES ACROSS THE RED RIVER BASIN THROUGH INTERPRETABLE MACHINE LEARNING

by

ISABELA SUAZA SIERRA

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Earth, Environmental Resource Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

May 2024
Acknowledgements

I would like to express my gratitude to my advisor, Dr. Hernan Moreno at The University of Texas at El Paso, for his support and trust in my abilities. I also thank the other members of my committee, Dr. Deana D. Pennington at The University of Texas at El Paso and Dr. Thomas M. Neeson at the University of Oklahoma. Their suggestions, comments, and additional guidance were invaluable in completing this work.

I would like to acknowledge the Department of Earth, Environmental, and Resource Sciences, the College of Science, and The Graduate School for their support and financial assistance. A special thanks to Dr. Maryam Zarei Chaleshtori, who mentored me throughout this master’s process and never ceased to support and advise me.

I extend special thanks to the wonderful people who became more than just colleagues during my time at the University of Texas at El Paso. To Steph, Nuri, Brenda, Mayra, Vale, Betul, Jorge, Santi, Andres, Rosa, Jayanga, Kat, Marc, Carmel, Paola, and all my DEERS peers, your support and friendship have been invaluable.

I would like to thank Joseph Vaska for being my constant support and comfort throughout this journey. His love, kindness, patience, and encouragement have helped me through even the most challenging times. I am also grateful to Patty Vaska for her endless optimism and unwavering belief in me.

A genuine thanks to Dr. David Yates, Dr. Thomas Hopson, Jerry Cycone, Benjamin Fellman, Ye Mu, Juan Colón, Paulina Soto, Stephanie Ortiz, Daniel Bonilla, Nish Silva, Auliya McCauley-Hartner, Harvana Laing, and all NCAR people who made one of the best experiences of my life possible.

Doy gracias a lo más importante de mi vida: Mi familia en Colombia. Ellos han sido el motor y la inspiración de cada paso que he tomado. A mi abuela, mamá y tías, las mujeres berracas que me criaron y me enseñaron todo. A mis amigos. A los que no están más en este planeta tierra.
Abstract

Accurately predicting lake water temperature (LWT) and dissolved oxygen (DO) is crucial for determining threshold values of fish survivability under warmer global conditions, with recreational fishing in reservoirs significantly contributing to regional economies, such as $779$ million and $1,891$ million annually to the economies of Oklahoma and Texas, respectively. Current mathematical models for temperature and oxygen profiles, which incorporate multi-layer and turbulent mixing equations, are complex and challenging to parameterize, particularly due to uncertainties in acquiring sufficient data for training and validation. Leveraging the flexibility and information extraction power of machine learning (ML) methods, this master thesis aimed to set up and test ML and deep learning (DL) models to predict LWT and DO across 12 lakes within the Red River Basin of the South in the United States, using historical spatially distributed measurements. Five ML approaches, including Random Forest (RF), Gradient Boosting Extreme (XGBoost), Tree-Boosting with Gaussian Process and Mixed Effects Model (GPBoost), Support Vector Machine (SVM), and Deep Learning (DL), were assessed using numerical k-fold cross-validation metrics. The results highlight GPBoost as the most effective method for predicting LWT and DO, which is attributed to their incorporation of interpretable physical variables. Notably, GPBoost exhibited robust performance under various lake conditions, while RF, XGBoost, and SVR showed signs of overfitting. Comparisons with traditional 1-D numerical approaches underscore the potential of ML algorithms for faster and more precise results, offering valuable insights into the dynamics of lake ecosystems and emphasizing the need for alternative methods to capture their complexities effectively.
# Table of Contents

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>v</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>List of Acronyms</td>
<td>xi</td>
</tr>
</tbody>
</table>

## Chapter

1. **Introduction** .................................................. 1  
2. **Methods** ..................................................... 10  
   2.1 Study Area .................................................. 10  
   2.2 Lake Description and Historical Data Reports ........... 11  
   2.3 Input Feature Selection Criteria .......................... 15  
   2.4 Data .......................................................... 18  
      2.4.1 LWT Observation Dataset ............................... 18 
      2.4.2 DO Observation Dataset ................................. 21 
      2.4.3 Meteorological and Hydrological Datasets ............ 22 
   2.5 Data Processing and Preparation ........................... 22 
   2.6 Data-driven Modeling ........................................ 24  
      2.6.1 Random Forest .......................................... 25 
      2.6.2 Extreme Gradient Boosting ............................. 26 
      2.6.3 Combining Tree-Boosting and Mixed Effects Models ... 26 
      2.6.4 Artificial Neural Network and Deep Learning ......... 27 
      2.6.5 Support Vector Regression ............................. 28 
      2.6.6 ML Hyperparameter Optimization ....................... 29
2.6.7 Feature Ranking and Importance ........................................... 29
2.7 Processed-based modeling ..................................................... 31
2.8 Performance Evaluation ...................................................... 32
2.9 Synthesized Working Framework .......................................... 33
3 Results ..................................................................................... 35
3.1 Exploratory Data Analysis .................................................... 35
3.2 LWT Modeling ................................................................. 37
   3.2.1 Hyperparameter tuning .................................................. 37
   3.2.2 Cross-validation comparison ......................................... 37
   3.2.3 Overall testing performance comparison ......................... 39
   3.2.4 Understanding the drivers ............................................ 44
   3.2.5 Model Performance Analysis in Stratified and Non-Stratified Water Column Conditions .................................................. 48
3.3 GLM ................................................................................... 49
3.4 DO Modeling ................................................................. 53
   3.4.1 Hyperparameter tuning .................................................. 53
   3.4.2 Cross-validation comparison ......................................... 53
   3.4.3 Overall testing performance comparison ......................... 55
   3.4.4 Understanding the drivers ............................................ 60
4 Discussion ............................................................................... 64
5 Limitations and Future Work ................................................... 70
6 Conclusions ............................................................................ 73
Curriculum Vitae ........................................................................ 89
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>List of RRB Reservoirs and lakes from west to east. The figure highlights the</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>12 study lakes and reservoirs, along with the number of LWT vertical profile</td>
<td></td>
</tr>
<tr>
<td></td>
<td>samples collected per season and average number of samples per profile.</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>Input features used by the ML models to predict LWT and DO. Note that the</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>input features drive the lake temperature and DO in some way by providing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>encoded information about the relations between inputs and outputs.</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Selected hyperparameter values after a random search strategy for each of the</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>test ML algorithms at predicting LWT</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Model evaluation using 5-fold cross-validation for LWT</td>
<td>39</td>
</tr>
<tr>
<td>3.3</td>
<td>Test Model Performance for LWT</td>
<td>39</td>
</tr>
<tr>
<td>3.4</td>
<td>LWT Model Performance under Stratified Water Column Conditions</td>
<td>49</td>
</tr>
<tr>
<td>3.5</td>
<td>LWT Model Performance under Non-Stratified Water Column Conditions</td>
<td>49</td>
</tr>
<tr>
<td>3.6</td>
<td>Selected hyperparameter values after a random search strategy for each of the</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>test ML algorithms at predicting DO</td>
<td></td>
</tr>
<tr>
<td>3.7</td>
<td>Model evaluation using 5-fold cross-validation for DO</td>
<td>55</td>
</tr>
<tr>
<td>3.8</td>
<td>Test DO Model Performance</td>
<td>55</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Drainage basin of the RRB. The main channel is shown in bold blue and major tributaries in light blue. The principal dams are represented by filled circles. Red circles illustrate the set of reservoirs studied in this thesis with available measurements of LWT and DO. Inset map gives the location of the basin in the south-central region of the United States. 11

2.2 Precipitation climatology over the RRB 12

2.3 Landcover map of the RRB 13

2.4 Lake Texoma. Coordinates are in WGS 1984 14

2.5 Temporal variation of temperature measurements across lakes 18

2.6 A summary framework for this master’s thesis 34

3.1 Correlation matrix of input features to predict LWT 36

3.2 Lake Water Temperature ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values. 40

3.2 (Continued) Lake Water Temperature ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values. 41

3.2 (Continued) Lake Water Temperature ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values. 42

3.3 Probability density function, observed temperature vs LWT ML models 43

3.4 Feature importance for the LWT ML models 45

3.4 (Continued) Feature importance for the LWT ML models 46

3.5 SHAP additive explanations for the Deep Learning model of LWT 48
List of Acronyms

AI  Artificial Intelligence

ANN  Artificial Neural Networks

DD  Data Driven

DO  Dissolved Oxygen

DOY  Day Of the Year

GLM  General Lake Model

**GPBoost**  Tree-Boosting with Gaussian Process and Mixed Effects Model

LWT  Lake Water Temperature

ML  Machine Learning

PB  Physics Based

RRB  Red River Basin

RF  Random Forest

SVR  Support Vector Regression

**XGBoost**  Extreme Gradient Boosting
Chapter 1

Introduction

Lake Water Temperature (LWT) and Dissolved Oxygen (DO) are crucial for understanding and managing freshwater ecosystems, as they directly affect aquatic organisms' physiology, behavior, and distribution (Caissie et al. 2007); among the environmental factors limiting survival and growth of fish, LWT and DO concentrations are considered the two most significant (H. G. Stefan et al. 1996; Coutant 1987; Christie et al. 1988; Magnuson 1990). They also serve as a proxy for climate change, as temperature and oxygen variations can indicate changes in atmospheric and radiative patterns (He et al. 2019; Yang et al. 2020). Furthermore, density differences in water bodies contribute to chemical variations with profound implications for living organisms in lakes. The temperature and dissolved substances play decisive roles in generating density differences in the water. The atmosphere imparts a temperature forcing on the lake surface, leading to the establishment of thermal stratification during the warm season in sufficiently deep lakes. Conversely, in the cold period, surface cooling prompts vertical circulation of water masses, resulting in the removal of gradients of water properties (Boehrer et al. 2008). A lake is considered stratified when the temperature difference between the epilimnion (the upper layer of lakes) and the hypolimnion (the deeper layers of lakes) exceeds one-degree Celsius (1 °C) (H. Stefan et al. 1996; Foley et al. 2012).

There are measurable indicators in the lake ecosystems that are useful for monitoring climate change due to their sensitivity and ability to capture atmospheric and catchment changes. These indicators include water temperature, dissolved organic carbon, and plankton composition. Understanding lake processes and their physical, chemical, and biological characteristics is crucial to assessing the effectiveness of lakes as climate change sentinels (Adrian et al. 2009). One such variable is the lake water level, particularly in non-regulated
lakes, that reflects the dynamic equilibrium between water inputs (i.e., precipitation and runoff) and losses (i.e., evaporation or infiltration) across timescales ranging from hours to centuries (Adrian et al. 2009; Argyilan et al. 2003; Ghanbari et al. 2008). Climate-related drivers influencing this property include precipitation, air temperature, water temperature, wind speed, cloud cover, and relative humidity (Adrian et al. 2009; Jansen et al. 2004; Rosenzweig et al. 2007; ACIA 2004; Rodionov 1994).

Another important response variable is water temperature, particularly in the epilimnion. It is a valuable and direct indicator of climatic conditions in lake ecosystems due to its swift and straightforward response to the available atmospheric energy, making it a reliable marker for detecting climate-related trends (Adrian et al. 2009). In contrast, hypolimnetic temperatures in the deeper layers of lakes display more complex behaviors influenced by factors such as lake morphometry (Gerten et al. 2001) and seasonality (Robertson et al. 1990; Straile et al. 2003). These vertical temperature changes significantly impact the density gradient within the water column, leading to alterations in vertical mixing, stratification, thermal stability, and thermocline depth (David M Livingstone 2003; David M. Livingstone et al. 2001; Coats et al. 2006). Long-term shifts in thermal structure can signal changes in mixing regimes, carrying implications for nutrient distribution, oxygen levels, and the composition of aquatic life. While detecting long-term alterations in thermal structure may necessitate precise measurements in the water column, they stand as robust indicators of climate change due to their direct and sensitive response to climatic forcing, with far-reaching consequences for lake ecosystems (Adrian et al. 2009).

According to Schmid et al. (2022), the energy budget of most lakes is primarily influenced by heat fluxes at the lake surface, especially shortwave radiation and incoming and outgoing longwave radiation. Changes in these heat fluxes and the consequent alterations in the lake’s thermal structure have the most direct impact on the effect on biotic components. The dynamics of these heat fluxes determine LWT and stratification, which, in turn, impact water quality and biota.

First, the shortwave radiation component represents the direct solar energy, including
visible light, near-ultraviolet, and near-infrared radiation. Its variations are influenced by latitude, cloud cover, elevation, and landscape features. Some incoming radiation is reflected as albedo, which varies by location, time of day, year, and surface water conditions (Schmid et al. 2022; Hipsey et al. 2014). The incoming longwave radiation component accounts for the radiative energy flux from the atmosphere. The magnitude of this flux is influenced by atmospheric temperature, humidity, and cloud cover. Most of this energy is absorbed by the lake’s surface, with the remaining energy being reflected at the lake’s surface (Schmid et al. 2022; Flerchinger et al. 2009). The outgoing longwave radiation is the largest energy loss term in a lake’s heat budget and is emitted from the lake, depending on surface water temperature (Schmid et al. 2022).

Another important component is the Latent heat flux at the lake surface, which quantifies the amount of energy gained or lost due to state changes in water (evaporation or condensation). It depends on prevailing meteorological conditions and the surface temperature and contributes to water addition or removal from the lake (Schmid et al. 2022; Winter et al. 2003). Sensible heat flux represents the energy transferred via conduction between the lake surface and the atmosphere. Depending on meteorological conditions and temperature differences, it can either add or remove energy from the lake. The net heat flux at the lake control volume is determined by energy inputs and outputs, including radiative, advected, conducted, and convected, and its storage depends on the heat capacity and density. The seasonality of this heat budget is primarily driven by the strong seasonality of incoming shortwave radiation and can vary with latitude and climate zones. Increasing air temperature can elevate longwave radiation and modify sensible heat exchange between the lake and the atmosphere (Schmid et al. 2022; R Iestyn Woolway et al. 2015).

On the other hand, the lake’s DO content is an important variable for water quality as it directly indicates an aquatic resource’s ability to support aquatic life: generally, DO levels less than 5 mg/L are considered stressful for fish, and levels less than three mg/L are too low to support fish. DO levels below one mg/L are considered hypoxic and usually devoid of life (U.S. Environmental Protection Agency 2023). DO concentrations depend
on many factors besides ecosystem metabolism, such as water temperature, gas exchange with the atmosphere, abiotic chemical reactions, and inputs in precipitation, groundwater, and surface water (H. G. Stefan et al. 1996; Hanson et al. 2006). LWT long-term physical changes would have severe consequences for nutrient and oxygen concentrations and the vertical distribution and composition of the biota; high water temperatures commonly lead to the drop of oxygen in water (Hanson et al. 2006).

The Intergovernmental Panel on Climate Change (IPCC) identified water temperature as one of the most sensitive variables to climate change, with expected increases in the coming decades (IPPC 2021). This increase will significantly impact aquatic ecosystems, including changes in biodiversity, species distribution, and water quality. In light of these findings, accurate prediction of LWT and DO in lakes and reservoirs is essential for effective ecosystem management, resource conservation, and climate change mitigation.

The physics-based modeling of LWT involves complex interactions of heat (shortwave and longwave radiation, evaporation, and conduction), matter (water vapor, inflows, and outflows), and mechanical energy (wind stress). Additionally, simulations consider internal mixing and stratification influenced by energy exchanges, chemical and thermal gradients, and biological activity. Process-based lake ecosystem models comprehensively address these factors and prove most effective when the system is well understood, and interest lies in studying its behavior beyond the range of historical observations (Robson 2014). One significant issue is that traditional methods for predicting water temperatures, such as physical models and empirical equations, possess inherent accuracy, complexity, and efficiency limitations, particularly when applied to multiple lakes and varying environmental conditions. The General Lake Model (GLM), a well-known physics-based model for addressing LWT (and other variables), computes vertical temperature, salinity, and density profiles by considering inflows/outflows, mixing, and surface heating and cooling. This model pairs with the Framework for Aquatic Biogeochemical Models (FABM) to integrate lake and reservoir water quality and ecosystem health simulations. However, similar to other physics-based models, the GLM has recognized limitations stemming from simplified representations of the mod-
eled physical processes and challenges in selecting and calibrating appropriate parameters (Xiaowei Jia et al. 2021). These limitations become more pronounced when utilizing such models for climate change projections (e.g., Representative Concentration Pathways - RCPs) due to the scarcity of data available in climate change scenarios.

Sarović et al. (2022) introduced a simplified 1-D energy budget model (SIMO) that provides a straightforward method for predicting the vertical temperature profile in a small, warm, monomictic lake. Importantly, this model relies solely on routinely observed surface meteorological data, including air temperature, relative humidity, atmospheric pressure, wind speed, and precipitation, excluding ultraviolet B radiation. However, the application of this simplified method was limited to a single small, warm, monomictic lake, making it unsuitable for broader application to diverse lakes and reservoirs under investigation, which exhibits varying volume capacities, surface areas, and stratification processes. Prats et al. (2019) developed a semi-empirical Ottosson-Kettle (OK) model, categorizing the water column into two layers: the upper epilimnion and the denser hypolimnion below. It computes temperatures for these layers in inland water bodies using meteorological variables (mean daily air temperature and mean daily solar radiation) along with lake characteristics (depth, surface area, volume, altitude, and latitude). The model maintains the two-layer distinction even when the temperature is uniform throughout the water column. The seven model parameters were determined by utilizing official monitoring data and satellite temperature data from the LakeSST dataset for French water bodies (Prats et al. 2018). The OK method calculates the epilimnion temperature as follows:

$$T_{e,i} = A + Bf \left( T_{a,i}^* \right) + C S_i$$  \hspace{1cm} (1.1)$$

where $T_{e,i}$ is the epilimnion temperature, $S$ is the Solar radiation (W/m2), $i$ is the day number, $A$, $B$, $C$ are calibration parameters. The variable $T_{a,i}^*$ is defined as:

$$T_{a,i}^* = T_{a,i} - MAAT$$  \hspace{1cm} (1.2)$$

where $T_{a,i}$ is air temperature and $MAAT$ is the mean annual air temperature for the study
period or for a relevant reference period.

The method calculates the epilimnion temperature using equation 1.2:

\[ T_{h,i} = A \cdot D + E \cdot g(T_{e,i}) \]  \hspace{1cm} (1.3)

where \( T_{h,i} \) is the hypolimnion temperature, \( D \) and \( E \) are calibration parameters, and the function \( g(*) \) is a simple exponential smoothing of \( T_e \).

The proposed OK parametrization is adapted to metropolitan France, considering the geographical features such as latitude and altitude and morphological characteristics like depth, surface area, and volume of water bodies. Metropolitan France is situated around 46°N latitude, while the study area of this master’s thesis is approximately around 34°N, which is the median value for both. The difference in latitude implies a difference in the angle at which solar radiation reaches the Earth’s surface. Lower latitudes generally receive more direct sunlight throughout the year than higher latitudes. This variation in solar radiation can significantly impact water temperature patterns in lakes. Therefore, applying the OK model to the study area may not be appropriate.

On the other hand, DO predictions in lakes are commonly based on mass-balance principles. Like water temperature physics-based models, these predictions involve simplified representations and assumptions about the underlying physical processes. For instance, the Lake Ontario oxygen model (Snodgrass et al. 1985) assumes a two-layer structure for the lake, consisting of the hypolimnion and the epilimnion. This type of two-layer oxygen model has been in use for more than two decades (See Equations 1.4 and 1.5). The DO models typically share a similar structure with slight variations in emphasis and parameterization.

Equations 1.3 and 1.4 illustrate the general framework of the Lake Ontario oxygen model.

\[ \frac{\partial C_{O_2, epi}}{\partial t} = -\nabla \cdot J_{O_2, epi} + R_{O_2, epi} \]  \hspace{1cm} (1.4)
Where:

\[ C_{O_2, epi} \] : Concentration of DO in the epilimnion  
\[ t \] : Time  
\[ \nabla \] : Divergence operator  
\[ J_{O_2, epi} \] : DO flux in the epilimnion  
\[ R_{O_2, epi} \] : Rate of oxygen generation or consumption in the epilimnion

\[
\frac{\partial C_{O_2, hypo}}{\partial t} = -\nabla \cdot J_{O_2, hypo} + R_{O_2, hypo} \quad (1.5)
\]

Where:

\[ C_{O_2, hypo} \] : Concentration of DO in the hypolimnion  
\[ t \] : Time  
\[ \nabla \] : Divergence operator  
\[ J_{O_2, hypo} \] : DO flux in the hypolimnion  
\[ R_{O_2, hypo} \] : Rate of oxygen generation or consumption in the hypolimnion

Deterministic and stochastic models are employed to model changes in DO concentration and uncover patterns from measured data. However, these models demand substantial data to accurately represent DO patterns, considering the complexity of the task (Latif et al. 2020; Najah et al. 2014). Many factors influence the DO concentration in a reservoir or lake, as discussed in the preceding paragraphs, leading to non-linear patterns. This challenges statistical models, as they often assume a linear relationship between DO and other parameters (Latif et al. 2020; Afifah et al. 2014). As in LWT, the solution to these types of DO models requires intensive data (including biotic activity) and parameterizations commonly unavailable at the appropriate spatial resolution and across multiple lakes, as proposed in this master’s thesis.
Conversely, machine learning (ML) models have emerged as a promising alternative for LWT and DO forecasting, as they can handle large datasets, discern intricate relationships between input and output variables, capture the nonlinearity in complex systems, and adapt to evolving conditions over time. ML models such as Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Deep Learning (DL), Tree-Boosting with Gaussian Process, and Mixed Effects Model (GPBoost) have been successfully applied to water temperature prediction in different contexts and regions worldwide. For example, Read et al. (2019) used a process-guided deep learning (PGDL) hybrid modeling framework with a use-case of predicting depth-specific LWT. The PGDL model performed well when extended to 68 lakes, with a median Root Mean Square (RMSE) error of 1.65 °C during the test period. In addition, Quan et al. (2020) used the SVR model and optimized the parameters using Genetic Algorithms (GA) for water temperature prediction in high-altitude reservoirs in western China; the study demonstrated that the prediction model M-GASVR could reasonably predict the vertical water temperature and water temperature structure in the reservoir area. For DO studies, Ziyad et al. (2022) proposed an artificial intelligence model with a simple architecture and achieved a high-performance level in predicting DO concentrations. They used water temperature, biological oxygen demand, iron, and total organic carbon as predictors for DO due to the availability of these parameters. The results revealed that the proposed model effectively captured the nonlinearity of DO with an acceptable level of accuracy, as evidenced by an $R^2$ value of 0.98. Furthermore, Moghadam et al. (2021) used four input parameters—water temperature, specific conductance, streamflow discharge, pH, and DO concentration—as input variables to predict DO concentration for three different lead times (“$t + 1$,” ”$t + 3$,” and ”$t + 7$”). Also, Heddam et al. (2016) developed a model for predicting DO concentration using an optimally pruned extreme learning machine (OPELM). This study found that OPELM provided a reasonable estimate of DO.

Some studies have suggested that ML can outperform physics-based models in certain scenarios (Gumière et al. 2020; D. Jiang et al. 2022). Others have shown that the performance
of ML models can be comparable to physics-based models (Zhu et al. 2019). Additionally, there are instances where physics-guided machine learning models have been developed to enhance prediction accuracy while ensuring consistency with physical laws (X. Jia et al. 2019). It has also been observed that ML methods can offer more robust predictions by incorporating additional features in the prediction process (Abdi et al. 2021). However, it is important to note that ML models, when provided with insufficient training data, may sometimes yield physically inconsistent results (Xiaowei Jia et al. 2021).

While ML models have shown promise for LWT and DO predictions, few studies have evaluated their performance and suitability for multiple lakes and reservoirs within a single basin or region. Therefore, this Master’s thesis aims to identify the most effective ML model for predicting LWT and DO in multiple lakes and reservoir, as well as to understand their drivers across the Red River Basin (RRB), a major watershed in the south-central USA. The RRB exhibits a significant spatial gradient in water availability from west to east and is home to more than 3 million people, according to the US Census Bureau, who rely heavily on water resources for economic activities.

This study will compare five ML models (RF, SVR, DL, GPBoost, and XGBoost) using different performance metrics to select the model that demonstrates superior accuracy, efficiency, and robustness in predicting LWT and DO across a set of lakes within the Red River Basin. Despite the GLM requires substantial data for its parameter calibration, a task beyond the scope of this master’s thesis, it will be executed in its uncalibrated form. The purpose is to compare its results with those obtained via ML models, assessing the potential advantages of using a smaller dataset in ML versus employing more data and parameters for a physical model, all within comparable time frames.

Additionally, this study aims to identify and rank the main drivers of LWT and DO, as suggested by each ML method. The general objective is to develop a model that incorporates relevant variables and can be applied as a reliable tool for predicting LWT and DO under changing climatic conditions. This will enable informed decision-making regarding ecosystem management, infrastructure investments, and water planning policy.
Chapter 2

Methods

2.1 Study Area

The region of interest is the Red River Basin (RRB) of the south in the United States, as illustrated in Figure 2.1. Originating in New Mexico and Texas, the river flows along the Texas-Oklahoma border, crosses Arkansas, and drains into the Mississippi River in Louisiana. The Red River is the fourth longest river in the United States, with a principal stream extending approximately 2,076 km (Christman et al. 2018). The basin experiences diverse climatic conditions, with a marked gradient in water availability from west (driest) to east (wettest), see Figure 2.2. Based on data from 1981-2010, the average annual precipitation varies across the basin, ranging from 870 mm in Texas, 920 mm in Oklahoma, 1,260 mm in Arkansas, and 1,590 mm in Louisiana (Data 2019). The RRB has a watershed area of approximately 233,000 km$^2$ (Christman et al. 2018).

The study area contains tributaries and major dams, as Figure 2.1 shows. Significant tributaries to the main river include the Washita River in Oklahoma, the Wichita River in Texas, the Little River in Louisiana, the Sulfur River in Texas, Big Cypress Bayou in Texas, and the Ouachita River in Arkansas and Louisiana. Within the study area, there are a total of 38 major reservoirs. The upper basin, characterized by drier conditions, primarily employs these reservoirs for water storage and meeting year-round water demands. In contrast, with a wetter climate, the lower basin primarily utilizes these reservoirs for flood control (Roland 2023).
Figure 2.1: Drainage basin of the RRB. The main channel is shown in bold blue and major tributaries in light blue. The principal dams are represented by filled circles. Red circles illustrate the set of reservoirs studied in this thesis with available measurements of LWT and DO. Inset map gives the location of the basin in the south-central region of the United States.

Figure 2.3 shows a landcover map of the study area, which was created using the ESA-CCI-LC version 2.0.7 land cover dataset from 2010 (the default landcover map for the hydrological model WEAP) and processed in ArcGIS Pro. The map illustrates that the western part of the basin primarily consists of cropland, grassland, and shrublands, while the eastern region features more forested and swampy areas.

2.2 Lake Description and Historical Data Reports

Among the major reservoirs/lakes in the RRB, 12 have available data on LWT and DO, as indicated in Table 2.1, provided by The Water Quality Portal (WQP). Some of these lakes have associated reports within state documents, offering an understanding of the stratification processes within these lakes and reservoirs.
The first report considered is Oklahoma’s Beneficial Use Monitoring Program – Lakes Sampling report for 2006-2007 (Oklahoma Water Resources Board 2007). This report covers several reservoirs, including Lake Texoma (See Figure 2.4), the primary water body in the basin. It is an 88,000-acre reservoir managed by the US Army Corps of Engineers (USACE) within the USACE Tulsa District. Constructed in 1944, it serves various purposes, such as flood control, water supply, hydroelectric power generation, flow regulation, navigation, and recreation. Classified as eutrophic, Lake Texoma exhibited stratification during the spring, with anoxic conditions accounting for 5 to 15% of the water column between 18 and 19 meters. Strong thermal stratification in the summer led to anoxic conditions at varying depths, affecting 8 to 76% of the water column. Also, the Arbuckle Reservoir is included in the report, and it showed no thermal stratification during the fall, winter, or spring quarters. DO levels remained above 6.0 mg/L in the fall and above 9 mg/L in the winter. However, thermal stratification and anoxic conditions occurred in the summer, with stratification between 6 and 7 meters, resulting in anoxic conditions affecting 65% of the water column.
In the same report, the Fort Cobb Reservoir, located in Caddo County and owned by the Bureau of Reclamation, demonstrated hypereutrophic conditions. While lacking thermal stratification during the fall, winter, and spring quarters, the lake exhibited strong stratification in the summer. At depths between 9 and 10 meters, anoxic conditions prevailed, persisting below 1.0 mg/L to the lake bottom at 11.7 meters. Similarly, Hugo Lake displayed consistent mixing without thermal stratification during the fall, winter, and spring, with DO levels above 5.0 mg/L. In the summer, stratification between 7 and 9 meters led to anoxic conditions affecting 38 and 31% of the water column. Pine Creek Lake maintained DO levels above 3.0 mg/L without thermal stratification during fall and winter quarters. In the spring, weak stratification occurred at 12 to 13 meters, while strong thermal stratification during the summer affected various depths, leading to anoxic conditions between 2 and 3 meters.

Additionally, Sardis Lake, a 13,610-acre reservoir, exhibited no stratification during winter or spring, but in the summer, stratification between 5 and 6 meters resulted in anoxic conditions for 60% of the water column. In the same report, Tom Steed Reservoir, a 6,400-acre reservoir constructed by the Bureau of Reclamation in Kiowa County in 1975, demonstrated
no thermal stratification due to its shallow nature during the sampling periods. Finally, Waurika Lake, a 10,100-acre reservoir constructed by the USACE in 1977 in Jefferson County, showed no stratification during the fall, winter, and spring intervals. In the summer, thermal stratification became apparent, resulting in anoxic conditions below the thermocline.

Lastly, the second report considered is Kemp Lake: Water Quality Objectives Attainment (2016-2017) (British Columbia Ministry of Environment and Climate Change Strategy 2017). Lake Kemp presented relatively warmer temperatures at all depths during the summer months, but visible thermal stratification was observed at around 5-6m. In February 2017, during the spring turnover, the water was well-mixed, and temperatures were uniform at all depths.
2.3 Input Feature Selection Criteria

The accurate selection of features or variables for training and predicting outcomes (LWT and DO) is fundamental for the development of reliable and accurate ML models. In this study, the choice of input variables is guided by two key considerations. Firstly, it takes into account the environmental, hydrologic, geomorphologic, and atmospheric factors that influence the processes governing lake temperature and oxygen dynamics. Secondly, the selection is influenced by the availability of these factors, derived from prospective climate simulations (CMIP-6; 2020-2100) obtained through the integration of general circulation models with hydrologic models (Fovargue et al. 2021; Zamani Sabzi et al. 2019). Based on the described drivers of lake temperature (see introduction), the following variables are included for the training and validation of the ML models:

**Same day and 7-day antecedent average air temperature:** Air temperature is a primary driver of LWT. It influences the heat exchange between the atmosphere and the lake surface (Schmid et al. 2022). It is important to note that lakes, especially large ones, have thermal inertia (Piccolroaz et al. 2024). This means they can take time to respond to changes in air temperature (lagged response) and other external factors, so the air temperature on the same day and 1-week average antecedent values were added as driving input features.

**Same day and 7-day antecedent average wind speed:** When wind speed increases, water surface mixing in the epilimnion increases due to increased surface momentum (Piccolroaz et al. 2024). As a result, warmer surface water and cooler subsurface water can mix more efficiently. While wind promotes mixing in the epilimnion, its effect on deeper layers (hypolimnion) can be limited, especially in deep lakes. When the wind is calm (decrease in wind speed), it will result in less latent and sensible heat loss due to reduced evaporation and convection, and thus warming at the lake surface (R. Iestyn Woolway et al. 2019). Under these circumstances, temperature stratification can occur, with warmer water at the surface and cooler water below. In such conditions, the air temperature appears to exert less
influence the lake’s surface temperature. The reason to consider both the 1-day and 1-week average values is the cumulative effect of many days under consecutive wind-forced water mixing and their effect on LWT values’ persistence (or change).

**Same day and 7-day antecedent cumulative precipitation**: Direct precipitation onto the lake’s surface affects LWT by introducing cooler water from rainfall, hail, or snow (Fujisaki-Manome et al. 2020). It also influences the lake’s thermal structure, especially when near-frozen water or snow forms a surface layer of different temperatures. Since LWT can respond to precipitation in a delayed manner, 24-hour and 1-week average antecedent precipitation values were added as driving input features.

**Day of the year**: It conveys information on several driving factors such as seasonal solar radiation variability, monthly precipitation, runoff regimes, spring snowmelt periods, seasonal atmospheric cycles driving winds, cloudiness, atmospheric pressures, and evaporation (Weng 2012).

**Lake volume**: The amount of water stored in each lake influences the thermal inertia, currents, and dynamic structure of the lake (Weng 2012). Larger lakes with higher volumes tend to change temperature more slowly than smaller ones.

**Water inflow**: A substantial water inflow (via streamflow), coupled with a short residence time (refers to the average amount of time water remains within a lake before flowing out through its outlet), means that the new water mixes relatively quickly with the lake’s existing water, leading to rapid adjustments in temperature. This can result in more significant temperature fluctuations, making the lake highly responsive to external temperature changes (Christianson et al. 2020; Hipsey et al. 2014). Conversely, a limited water inflow, combined with a longer residence time, creates a more stable thermal environment, as the lake retains its heat for an extended period and is less sensitive to short-term temperature variations. Thus, the interplay between water inflow and residence time governs a lake’s ecosystem’s thermal resilience and responsiveness to changing environmental conditions (Christianson et al. 2020; Hipsey et al. 2014).

**Water measurement depth**: Depth controls the air-water interactions including solar
radiation, wind speed and precipitation. Depth is also related to water currents the location of the thermocline and biological activity.

**Surface area-to-maximum depth ratio:** This ratio provides information about the shape and dimensions of the water body, which in turn affects how heat is exchanged with the surrounding environment (Magee et al. 2017a).

The geographical coordinates of a lake (latitude and longitude) inherently play a key role in influencing the angle and intensity of solar radiation received at the site. However, in this study, the explicit inclusion of latitude and longitude as independent variables is intentionally omitted. This deliberate exclusion is driven by the recognition that essential information encapsulated by these geographical coordinates is already embedded within the meteorological variables considered. This approach aims to prevent potential biases arising from location-specific influences, ensuring a more nuanced and unbiased exploration of the complex interplay between meteorological factors and LWT dynamics.

For the DO models, the same variables were used to train and validate the models, with the addition of the LWT variable, which is crucial for lakes’ physical, chemical (e.g., dissolved oxygen concentration), and biological dynamics (e.g., fish growth) (Piccolroaz et al. 2024). Air temperature was selected because its changes can directly impact LWT, consequently affecting DO concentrations. Furthermore, daily runoff and the 7-day average can influence oxygen levels through dilution and nutrient runoff (Luo et al. 2024). Additionally, wind speed influences surface mixing (Piccolroaz et al. 2024), which can either enhance or deplete oxygen levels, depending on whether it encourages oxygen transfer from the atmosphere or the surrounding water. Moreover, the volume of water and the inflow of fresh, oxygen-rich water into the lake system can significantly affect DO concentrations (WETZEL 2001). Finally, DO levels in lakes change with depth, with deeper parts of the lake, especially the hypolimnion, tending to have lower oxygen levels due to their reduced influence by surface oxygen exchange and photosynthetic activity (WETZEL 2001).
2.4 Data

2.4.1 LWT Observation Dataset

In this study, in situ measurements of LWT profiles were utilized to train and test all models. Data spanning from 1996 to 2020 were obtained from 12 lakes in the RRB (refer to Figure 2.1 and 2.5), sourced from the National Water Quality Monitoring Council (NWQMC). These data primarily consist of discrete LWT profile measurements, as none of the lakes were equipped with buoys for continuous temperature monitoring throughout specific periods of the year. In instances where multiple discrete measurements were taken on a given lake-day (considering that a lake may have different measurement sites), these observations were treated independently.

![Figure 2.5: Temporal variation of temperature measurements across lakes](image)

A total of 10,386 data points were collected, representing 674 LWT profiles across 12 lakes and reservoirs within the RRB (refer to Figure 2.5). It is a prerequisite that each profile includes a minimum of 4 measurements to ensure data reliability. Table 2.1 illustrates the
number of profile samples per season for each of the 12 study lakes and the average number of samples per profile. Lake Texoma stands out among these water bodies with 490 profiles. Fort Cobb Reservoir and Waurika Lake also have relatively large numbers of profiles, with 40 and 33, respectively. On the other hand, certain lakes, such as Pat Mayse, Lake Kemp, and Wright Patman Lake, have only one or two profiles each, indicating more limited data availability. The variation in the number of measured profiles per water body is attributed to the typical annual monitoring schedule.

Among the 674 LWT profiles in the dataset, we observe distinct seasonal distributions, with 173 profiles corresponding to the fall season (25.6%), 140 profiles to the spring season (20.7%), 181 profiles to the summer season (26.8%), and 180 profiles to the winter season (26.7%). Within this dataset, 218 profiles, constituting approximately 32% of the total profiles, exhibit thermal stratification. A lake is considered stratified when the temperature difference between the epilimnion and the hypolimnion exceeds 1°C (H. Stefan et al. 1996; Foley et al. 2012). Also, as discussed by H. Stefan et al.(1996), the thermocline depth (estimated from the water density gradient profile) is another indicator of strong stratification. The thermocline is a water layer within the temperature stratification that is characterized by a rapid decrease in temperature with depth. This study determines thermal stratification for each profile using the PyLake Python package, which calculates meaningful physical properties in aquatic sciences. The function "meta.depths" is utilized to compute the top and bottom depths of the metalimnion in a stratified lake. The metalimnion is the water stratum surrounding the thermocline in a stratified lake with the steepest thermal gradient, demarcated by the bottom of the epilimnion and the top of the hypolimnion. If the function does not find distinct metalimnion top and bottom depths, it indicates that the profile is not stratified.
Table 2.1: List of RRB Reservoirs and lakes from west to east. The figure highlights the 12 study lakes and reservoirs, along with the number of LWT vertical profile samples collected per season and average number of samples per profile

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Avg. Samples per profile</th>
<th>Profile samples per season</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Buffalo Reservoir</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>2</td>
<td>Greenbelt Reservoir</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>3</td>
<td>Lake Altus-Lugert</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>4</td>
<td>Foss Reservoir</td>
<td>12</td>
<td>Spring 5, Summer 10, Fall 9, Winter 5</td>
</tr>
<tr>
<td>5</td>
<td>Lake Kemp</td>
<td>14</td>
<td>Summer 2</td>
</tr>
<tr>
<td>6</td>
<td>Tom Steed Reservoir</td>
<td>7</td>
<td>Spring 3, Summer 4, Fall 7, Winter 5</td>
</tr>
<tr>
<td>7</td>
<td>Lake Diversion</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>8</td>
<td>Lake Kickapoo</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>9</td>
<td>Fort Cobb Reservoir</td>
<td>11</td>
<td>Spring 10, Summer 10, Fall 10, Winter 10</td>
</tr>
<tr>
<td>10</td>
<td>Lake Arrowhead</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>11</td>
<td>Ellsworth Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>12</td>
<td>Waurika Lake</td>
<td>10</td>
<td>Spring 5, Summer 10, Fall 10, Winter 9</td>
</tr>
<tr>
<td>13</td>
<td>Lake of the Arbuckles</td>
<td>15</td>
<td>Summer 5, Fall 5, Winter 5</td>
</tr>
<tr>
<td>14</td>
<td>Lake Texoma</td>
<td>16</td>
<td>Spring 113, Summer 125, Fall 115, Winter 137</td>
</tr>
<tr>
<td>15</td>
<td>Atoka Reservoir</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>16</td>
<td>McGee Creek Reservoir</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>17</td>
<td>Jim Chapman Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>18</td>
<td>Pat Mayse Lake</td>
<td>12</td>
<td>Summer 1</td>
</tr>
<tr>
<td>19</td>
<td>Hugo Lake</td>
<td>10</td>
<td>Spring 4, Summer 5, Fall 9</td>
</tr>
<tr>
<td>20</td>
<td>Sardis Lake</td>
<td>14</td>
<td>Fall 5, Winter 5</td>
</tr>
<tr>
<td>21</td>
<td>Pine Creek Lake</td>
<td>12</td>
<td>Summer 10, Fall 4, Winter 5</td>
</tr>
<tr>
<td></td>
<td>Lake Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---------------------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>22</td>
<td>Lake Bob Sandlin</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>23</td>
<td>Broken Bow Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>24</td>
<td>Lake O' The Pines</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>25</td>
<td>Dequeen Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>26</td>
<td>Gillham Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>27</td>
<td>Wright Patman Lake</td>
<td>4</td>
<td>Fall: 1</td>
</tr>
<tr>
<td>28</td>
<td>Dierks Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>29</td>
<td>Millwood Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>30</td>
<td>Caddo Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>31</td>
<td>Lake Greeson</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>32</td>
<td>Lake Bistineau</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>33</td>
<td>Lake Ouachita</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>34</td>
<td>DeGray Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>35</td>
<td>Lake Hamilton</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>36</td>
<td>Bayou D’arbonne Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>37</td>
<td>Catahoula Lake</td>
<td>No available data</td>
<td>No available data</td>
</tr>
<tr>
<td>38</td>
<td>Lake Jack Lee</td>
<td>No available data</td>
<td>No available data</td>
</tr>
</tbody>
</table>

### 2.4.2 DO Observation Dataset

The DO dataset consists of 3,059 points representing 235 DO vertical profiles in various lakes and reservoirs. Lake Texoma stands out among these water bodies with 73 DO profiles. Waurika Lake and Foss Reservoir also have relatively large numbers of profiles, with 33 and 27, respectively. On the other hand, certain lakes, such as Pat Mayse, Lake Kemp, and Wright Patman Lake, have only one profile each, indicating more limited data availability. Among the 235 DO profiles in the dataset, we observe distinct seasonal distribution, with 74 profiles corresponding to the fall season (31.4%), 34 profiles to the spring season (14.6%), 62 profiles to the summer season (26.3%), and 65 profiles to the winter season (27.7%).
2.4.3 Meteorological and Hydrological Datasets

In this study, GridMet (Abatzoglou 2013) served as the gridded surface meteorological dataset, offering high spatial resolution (4 km) daily surface fields of temperature, precipitation, winds, humidity, and radiation across the contiguous United States from 1979. This dataset integrates high-resolution spatial data from PRISM (Oregon State University 2014) with the high-temporal-resolution data from the National Land Data Assimilation System (Xia et al. 2012) to generate spatially and temporally continuous fields suitable for additional land surface modeling. For each lake, atmospheric forcing data from GridMet were downloaded, representing the mean value in the lake surface area. The daily volume and water inflow data for the lakes and reservoirs were acquired through the US Army Corps of Engineers (USACE 2023) platform.

Table 2.2 provides a summary of the atmospheric and hydrological predictors used for the data-driven estimation of LWT. These drivers include gridMET daily mean air temperature, the average of the last 7 days’ air temperature, daily accumulated precipitation, the average of the last 7 days’ accumulated precipitation, daily wind speed, and the average of the last 7 days’ wind speed. Additionally, the models incorporate the day-of-the-year. Further, the USACE daily reservoir inflow and volume are included as supplementary dynamic input features. In addition to these dynamic features, contextual data for each lake, such as mean depth, is added, which serves as a steady feature for each reservoir/lake throughout time.

2.5 Data Processing and Preparation

In the initial stages following data collection, it is essential to conduct exploratory data analysis (EDA) to comprehend the underlying processes and address issues such as missing values or inadequate data, which are unsuitable for model training. As noted by KASKI (1997), ”Most EDA techniques are graphical in nature with a few quantitative techniques.” The reliance on graphics in EDA is justified by its primary role of exploration, providing analysts with a powerful means to gain insights into the data.
Table 2.2: Input features used by the ML models to predict LWT and DO. Note that the input features drive the lake temperature and DO in some way by providing encoded information about the relations between inputs and outputs.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Abbrev.</th>
<th>Source</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same day air temperature</td>
<td>air_temp</td>
<td>Gridmet UC Merced</td>
<td>4 km, daily</td>
</tr>
<tr>
<td>7-day antecedent average air temperature</td>
<td>air_temp7d</td>
<td>Gridmet UC Merced</td>
<td>4 km, daily</td>
</tr>
<tr>
<td>Same day precipitation</td>
<td>prcp</td>
<td>Gridmet UC Merced</td>
<td>4 km, daily</td>
</tr>
<tr>
<td>7-day antecedent cumulative precipitation</td>
<td>prcp_cum7</td>
<td>Gridmet UC Merced</td>
<td>4 km, daily</td>
</tr>
<tr>
<td>Same day wind speed</td>
<td>wind</td>
<td>Gridmet UC Merced</td>
<td>4 km, daily</td>
</tr>
<tr>
<td>7-day antecedent average wind speed</td>
<td>wind_avg7</td>
<td>Gridmet UC Merced</td>
<td>4 km, daily</td>
</tr>
<tr>
<td>Day of the year</td>
<td>doy</td>
<td>Calendar</td>
<td>Daily</td>
</tr>
<tr>
<td>Lake volume</td>
<td>vol_lake</td>
<td>USACE</td>
<td>Lake, daily</td>
</tr>
<tr>
<td>Water inflow</td>
<td>inflow_lake</td>
<td>USACE</td>
<td>Lake, Daily</td>
</tr>
<tr>
<td>Measurement depth</td>
<td>depth_meas</td>
<td>NWQDMC</td>
<td>Point, Instant</td>
</tr>
<tr>
<td>Surface area max depth ratio</td>
<td>surf_area_depth</td>
<td>Calculated</td>
<td>Point, Instant</td>
</tr>
</tbody>
</table>

After the EDA step, preprocessing steps, including standardization and normalization, become necessary when input values exhibit differing units and ranges, as such discrepancies can lead to biased or inaccurate predictions (Maharana et al. 2022). These preprocessing steps are critical for ensuring data readiness for ML models, aiming for accurate and reliable results (Raju et al. 2020). Consequently, post-EDA, the data will undergo processing to make it suitable for ML models, involving the normalization of all inputs for use as features. This study will leverage the StandardScaler method to standardize the input data (Equation 2.1). The StandardScaler method transforms data to have a zero mean and unit variance, centering the data around a mean of zero with a standard deviation of one (Ahsan et al. 2021). Notably, the StandardScaler method operates without assuming any specific input
data distribution, establishing its robustness as a preprocessing technique applicable to a diverse array of datasets (Ahsan et al. 2021).

\[
Z_{scaled} = \frac{(X - \mu)}{\sigma}
\]

(2.1)

where \(Z_{scaled}\) represents the scaled value of the input feature, \(X\) represents the original input feature value, \(\mu\) represents the mean of the input feature values, and \(\sigma\) represents the standard deviation of the input feature values.

After preparing the dataset for training and testing, it is ready for use in ML. The final dataset will have variables represented by the abbreviations listed in Figure 2.2, consolidating all the previously described data.

### 2.6 Data-driven Modeling

This section will briefly describe the ML models that will be used in this study to predict LWT and DO in lakes and reservoirs in the RRB. The data is randomly divided into a training set (60%) and a testing set (40%), with RMSE used as the cost function for all five models.

The consolidated dataset for training and testing the ML models can be described as follows: each distinct measurement profile collected presents unique depth and temperature measurements. However, the meteorological and hydrological variables exhibit replications across the entire day and lake profile. The modeling of such grouped data can be approached through four alternatives: (i) Disregard the grouping structure and allow the model to distinguish it autonomously. (ii) Model each group separately, i.e., each lake, although this is rarely advisable due to the typically limited number of measurements per group relative to the multitude of distinct groups. (iii) Integrate the grouping variable (e.g., lake name or ID) into the chosen model and treat it as a categorical variable. In such instances, the model must learn numerous parameters (one for each group) based on relatively sparse data, impeding efficient learning. Additionally, for tree models, high cardinality categorical variables may
pose challenges. (iv) Model the grouping variable using random effects in a mixed-effects model. This often represents a sensible compromise between the approaches mentioned above. Specifically, as elucidated below and in (Sigrist 2020), this proves advantageous compared to other methods, particularly in tree-boosting.

This study will apply option (i) to four classical ML models (RF, SVR, DL, and XGBoost). In contrast, option (iv) will be implemented using GPBoost: a fusion of Tree-Boosting and Mixed Effects Models. These models will be described in the following subsections.

2.6.1 Random Forest

RF is an ensemble model that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It works by constructing a forest of decision trees, where each tree is trained using a random subset of input variables and samples. RF can handle complex and nonlinear relationships between input and output variables and has been used for LWT prediction tasks, such as predicting LWT in rivers and lakes (Tyralis et al. 2019). The basic Equation for RF is presented in Equation (2.2), where $f(x)$ is the predicted output, $T(x; \theta_b)$ is the prediction of the $b^{th}$ tree, $B$ is the total number of trees, and $\theta_b$ is the set of parameters for the $b^{th}$ tree. During training, the model optimizes the parameters of each tree to maximize accuracy and minimize error.

$$f(x) = \frac{1}{B} \sum_{b=1}^{B} T(x; \theta_b) \quad (2.2)$$

The algorithm has several parameters that can be optimized to improve its performance. Some of the key parameters include (1) maximum depth of trees, (2) minimum samples required at leaf nodes, (3) maximum number of features considered for splitting and (4) number of trees.
2.6.2 Extreme Gradient Boosting

XGBoost is a tree-boosting algorithm that was developed based on the already existing concept of boosting, which was further enhanced to increase efficiency and scalability and reduce overfitting. Like bagging, boosting methods combine the prediction of an ensemble of weak learners with improving prediction accuracy. However, while bagging ensemble members are trained in parallel, boosting iteratively trains new ensemble members and adds them to the existing ensemble (Chen et al. 2016). The basic expression for a gradient boosting model is presented in Equation (2.3), where $y_i$ is the target variable for the $i^{th}$ observation, $f_j$ is the $j^{th}$ regression function, $x_i$ is the input vector for the $i^{th}$ observation, $m$ is the number of regression functions, and $\epsilon_i$ is the residual error. During training, the model optimizes the weights assigned to each tree to minimize the error in the training data.

$$y_i = \sum_{j=1}^{m} f_j(x_i) + \epsilon_i \quad (2.3)$$

The algorithm has several parameters that can be optimized to improve its performance. Some of the key parameters include: (1) Gamma or the minimum loss reduction required to make a further partition on a leaf node of the tree, (2) the subsample ratio of columns by tree; (3) the maximum depth of trees, (4) the minimum child weight, (5) the learning rate and (6) the number of estimators.

2.6.3 Combining Tree-Boosting and Mixed Effects Models

The GPBoost algorithm allows for combining tree-boosting and mixed effects models. It is an extension of linear mixed effects models where the fixed-effects are learned using tree-boosting (Sigrist 2020). The GPBoost algorithm operates under the assumption that the response variable, denoted as $y$, is the summation of a potentially non-linear mean function $F(X)$ and random effects $Zb$: 
\[ y = F(X) + Zb + e \]

where \( y \) is the response variable (label), \( X \) contains the predictor variables (features), \( F() \) represents an ensemble of trees (non-linear function), \( Zb \) are random effects assumed to follow a multivariate normal distribution, and \( e \) is an error term.

The GPBoost model is trained iteratively, involving the learning of covariance parameters (hyper-parameters) for the random effects and the regression function \( F(X) \) using a tree ensemble. The estimation (or prediction) of random effects \( Zb \) occurs after the model has been learned. Notably, GPBoost stands out from existing boosting algorithms by accounting for data dependency due to clustering and by learning the covariance parameters of random effects. The algorithm employs gradient and/or Newton boosting steps to learn covariance parameters and expand the ensemble of trees. The (co-)variance parameters can be learned through accelerated gradient descent or Fisher scoring, while the tree learning process utilizes the LightGBM library. It is worth noting that the GPBoost library provides access to the full functionality of LightGBM (Sigrist 2020).

### 2.6.4 Artificial Neural Network and Deep Learning

The Artificial Neural Network (ANN) model mimics the structure and function of the human brain, composed of interconnected layers of nodes that process and transform input data. The ANN model can learn complex relationships between input and output variables and can handle nonlinear and high-dimensional data. The ANN has been applied to various prediction tasks, including LWT forecasting (Maier et al. 2000). The basic framework for an ANN is presented in Equation (2.4), where \( y \) is the output, \( f \) is the activation function, \( w_i \) are the weights, \( x_i \) are the inputs, \( b \) is the bias, and \( n \) is the number of inputs. The weights and biases are adjusted during training to optimize the model’s performance on the training data.
\[ y = f(\sum_{i=1}^{n} w_i x_i + b) \] (2.4)

The algorithm has several parameters that can be optimized to improve its performance. Some of the key parameters include (1) number of layers; (2) number of neurons; (3) batch size; (4) epochs, (5) dropout rate and (6) learning rate. DL is a neural network with three or more layers.

2.6.5 Support Vector Regression

SVR is a regression model that uses a kernel function to map the input data into a high-dimensional space, where a linear regression model is trained to predict the output variable. SVR can handle nonlinear and high-dimensional data and is robust to noise and outliers. SVR has been used in various LWT prediction tasks, such as predicting LWT in lakes and rivers (Dingguo Jiang et al. 2022). The basic expression for SVR is presented in Equation (2.5) where \( y \) is the predicted output, \( \alpha_i \) are the Lagrange multipliers, \( x_i \) are the training inputs, \( K \) is the kernel function, \( b \) is the bias term, and \( n \) is the number of training inputs. Unlike traditional regression models, SVR introduces a "margin of error" on either side of the hyperplane, allowing some deviation from the predicted values. The goal of the SVR algorithm is to minimize the error while maximizing the margin.

\[ y = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \] (2.5)

The algorithm has several parameters that can be optimized to improve its performance. Some of the key parameters include (1) Kernel type, (2) regularization factor, (3) epsilon-insensitive loss parameter.
2.6.6 ML Hyperparameter Optimization

A sensitivity analysis for all hyperparameters will be conducted to ensure that the models were accurately calibrated. Once the range of acceptable values for each hyperparameter is defined, the Randomized Search method is used to optimize the performance of the models. This method is widely used for hyperparameter optimization. Random Search sets up random combinations of hyperparameter values to train the model and provide performance scores. This approach can significantly enhance the models’ accuracy and generalizability (Bergstra et al. 2012).

Cross-validation (CV) is a widely used technique in ML for selecting the best-performing model and evaluating its generalization ability (Géron 2019). It is a powerful tool for reducing the risk of overfitting, which occurs when a model is too complex and performs well on the training set but poorly on new data (Bates et al. 2023). By dividing the data into subsets, training the model on one subset, and testing on the others, CV allows us to estimate the model’s true error and select the best performance (Bates et al. 2023). Using multiple folds in CV also improves the accuracy of the model evaluation and reduces the variance of the estimated error (Bengio et al. 2003). This study will use the 5-fold technique to validate and compare the performance of four types of ML models (RF, XGBoost, ANN, and SVR).

2.6.7 Feature Ranking and Importance

Some methods exist to calculate feature importance and ranking for the RF, XGBoost (Gini Index), and SVR (Permutation Feature Importance) algorithms. On the other hand, Neural Networks are often perceived as black boxes, posing challenges in extracting interpretable information or feature explanations for other purposes due to their complex and nonlinear nature. The following paragraphs describe how can be calculated the importance of each method.
**Gini impurity**

It is a metric used to measure the impurity or disorder in a dataset. It quantifies how well a feature separates the data into different classes or categories (Strobl et al. 2008).

\[
\text{Gini Index} = 1 - \sum_{i=1}^{n} p_i^2
\]  

(2.6)

where \( n \) is the number of classes or categories and \( p_i \) is the probability of a data point belonging to class \( i \).

In the context of decision trees, when a tree node splits on a particular feature, the impurity of the resulting child nodes should be lower than that of the parent node. The reduction in impurity resulting from splitting on a specific feature serves as a measure of the feature's importance. Both RF and XGBoost, as ensemble learning methods built on decision trees, leverage these impurity reduction metrics to rank features based on their capacity to enhance predictive accuracy. Features that contribute to more substantial reductions in impurity are deemed more important in the decision-making process.

**Permutation importance**

Since SVR lacks built-in support for native feature importance scores, Permutation Feature Importance is employed as an alternative technique. In this method, a model, including ones that do not natively provide feature importance scores, is trained on the dataset. Predictions are then made on the dataset, with each feature’s values (columns) randomly shuffled in multiple iterations (e.g., 3, 5, 10 times). This process yields a mean importance score for each input feature and a distribution of scores across repetitions. This approach can be applied to both regression and classification tasks, with the choice of a performance metric (e.g., mean squared error for regression or accuracy for classification) serving as the basis for the importance score.
**SHapley Additive exPlanations**

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of complex ML models. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (Lundberg et al. 2017). When applied to Artificial Neural Networks or DL models, SHAP values can offer valuable insights into the black-box nature of these complex models. By understanding the contribution of each input feature to the model’s output, practitioners can better understand how the neural network makes predictions.

The computation of SHAP values involves evaluating the model’s output for all possible combinations of feature values and averaging the contributions of each feature across these combinations. This process ensures that each feature is assigned a fair share of the credit for the model’s prediction. A SHAP summary plot often visualizes the impact of each feature on a particular prediction. The color-coded bars indicate the direction and magnitude of the effect. For instance, positive values (often represented in red) suggest a feature that pushes the prediction higher, while negative values (blue) suggest a feature that pushes the prediction lower (Lundberg et al. 2017).

### 2.7 Processed-based modeling

The General Lake Model (GLM) is a lake water balance and vertical stratification model. It is an open-source, one-dimensional lake hydrodynamic model that balances fluxes of mass and energy on a daily (or subdaily) time step, and tracks state variables (such as temperature) with Lagrangian layers resolved in the vertical dimension (Hipsey et al. 2019). GLM contains complex vertical mixing routines that redistribute heat in response to prevailing conditions and external forcing.

At the core of GLM’s functionality lies the water balance equation, a fundamental element in modeling the mass dynamics of lakes. The equation accounts for the inflow, outflow, and changes in storage within the lake:
\[
\frac{dV}{dt} = P - E - Q \quad (2.7)
\]

where, \( V \) represents the volume of the lake, \( P \) denotes the inflow (precipitation), \( E \) signifies the outflow (evaporation), and \( Q \) stands for the discharge.

GLM’s sophistication extends to the consideration of thermal dynamics within lakes. The heat balance equation governs the vertical temperature distribution, crucial for understanding stratification and mixing processes:

\[
C \frac{dT}{dz} = \frac{\partial}{\partial z} \left( K \frac{\partial T}{\partial z} \right) + \Phi \quad (2.8)
\]

where, \( C \) is the heat capacity, \( T \) stands for temperature, \( z \) represents the vertical dimension, \( K \) denotes the vertical turbulent diffusivity, and \( \Phi \) symbolizes the heat sources and sinks.

GLM’s ability to perform well relies on having precise input data. Having enough data about weather, lake structure, and external influences is crucial for setting up the model accurately and obtaining trustworthy results. The model involves several parameters (like those linked to vertical mixing, wind protection, and water clarity), usually adjusted based on individual lakes if sufficient training data are accessible (Xiaowei Jia et al. 2021). For this study, the model will be run with default parameters since there isn’t daily water profile data available for calibrating the model.

## 2.8 Performance Evaluation

Model training and testing will be conducted via the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination \((R^2)\). The RMSE measures the deviation of predicted values from the actual values, and the RMSE zero value indicates the best fit. The formula for the RMSE is as follows:
\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \]  
(2.9)

where \( y_i \) and \( \hat{y}_i \) are the actual and predicted values, respectively, and \( n \) is the number of observations.

The MAE measures the absolute deviation of predicted values from the actual values and is defined as:

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|. \]  
(2.10)

The \( R^2 \) coefficient measures the proportion of variance in the dependent variable that is predictable from the independent variable(s) and is defined as:

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \]  
(2.11)

where \( \bar{y} \) is the mean of the dependent variable.

\section*{2.9 Synthesized Working Framework}

Developing a ML model involves several key steps, including defining the input variables (predictors), pre-processing the data, applying the ML algorithm with hyperparameter tuning, and evaluating the model. Figure 2.6 depicts the procedural framework that this master's thesis follows.
Figure 2.6: A summary framework for this master’s thesis.
3.1 Exploratory Data Analysis

In the analysis of LWT prediction, a correlation matrix employing the Pearson correlation coefficient quantifies the degree of linear association between various variables. The correlation matrix, shown in Figure 3.1, was examined before initiating the LWT prediction process.

As illustrated in Figure 3.1, positive correlations were observed among $\text{air\_temp}$, $\text{air\_temp7d}$, and $\text{temp\_measure}$. Specifically, the correlation analysis revealed a strong positive relationship between these variables. Additionally, $\text{doy}$ exhibited a moderate positive correlation (Pearson’s coefficient $= 0.5$) with $\text{temp\_measure}$. $\text{doy}$ controls incoming solar radiation and precipitation regimes, ultimately affecting LWT.

The moderate negative correlation between latitude and $\text{vol\_lake}$ implies that lakes at higher latitudes exhibit comparatively smaller volumes. Also, the moderate positive correlation between $\text{surf\_area\_depth}$ and longitude suggests that lakes with a higher surface area relative to their maximum depth are more prevalent at eastern positions, and the strong negative correlation between $\text{surf\_area\_depth}$ and latitude indicates that lakes with a higher surface area relative to maximum depth are predominantly found at lower latitudes.
Figure 3.1: Correlation matrix of input features to predict LWT
3.2 LWT Modeling

3.2.1 Hyperparameter tuning

An initial sensitivity analysis was conducted to ensure that the ML models were properly tuned and capable of delivering reliable predictions of LWT. This process allowed us to define the appropriate range of values for each hyperparameter, as detailed in Table 3.1. Subsequently, we applied the Randomized Search technique to identify the optimal hyperparameter combinations for each model, as summarized in Table 3.1. As a result, the RF model was configured with 200 estimators, a maximum of 5 features, a minimum sample leaf of 1, a minimum sample split of 2, and a depth of 20. The SVR model was set with a radial basis function kernel, a regularization parameter of 10, and an epsilon value of 0.3. The XGBoost model employed 600 estimators, a learning rate of 0.1, a depth of 9, a gamma value of 0.3, and a subsample of 1.0. The GPBoost model was configured with a learning rate of 0.8, maximum depth of 11, and minimum data in leaf of 1. Lastly, the DL model consisted of 2 layers with 48 neurons, a batch size of 128, 700 training epochs, a dropout rate of 0.1, and a learning rate of 0.01.

3.2.2 Cross-validation comparison

After conducting hyperparameter tuning, 5-fold cross-validation is employed to assess how statistical analysis results generalize to an independent dataset. Table 3.2 demonstrates consistent performance across all methods, with lower values for both RMSE and MAE. The $R^2$ values are consistently closer to 1 for all cases, indicating the ideal scenario. XGBoost outperformed other models, achieving the lowest RMSE and MAE values in the 5-fold cross-validation analysis.
Table 3.1: Selected hyperparameter values after a random search strategy for each of the test ML algorithms at predicting LWT

<table>
<thead>
<tr>
<th>Model</th>
<th>Hyperparameters</th>
<th>Range</th>
<th>Selected value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>Number of estimators</td>
<td>[50,100,200,300,400]</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Maximum features</td>
<td>[1, 2, 3, 4, 5, 6, 7, 8, 9, 10,11,12]</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Minimum sample leaf</td>
<td>[1,3,5]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Minimum sample split</td>
<td>[1,2,5,10]</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Depth</td>
<td>[10,20,30,50]</td>
<td>20</td>
</tr>
<tr>
<td>XGBoost</td>
<td>Number of estimators</td>
<td>[100,300,600,900,1000]</td>
<td>600</td>
</tr>
<tr>
<td></td>
<td>Learning rate</td>
<td>[0.001, 0.01, 0.1, 0.2]</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Depth</td>
<td>[3,4,5,6,7,8,9,10]</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>gamma</td>
<td>[0, 0.1, 0.2, 0.3, 0.4]</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Subsample</td>
<td>[0.6, 0.7, 0.8, 0.9, 1.0]</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>colsample bytree</td>
<td>[0.6, 0.7, 0.8, 0.9, 1.0]</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>reg alpha</td>
<td>[0, 0.1, 0.2, 0.3]</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>reg lambda</td>
<td>[0, 0.1, 0.2, 0.3]</td>
<td>0.2</td>
</tr>
<tr>
<td>GPBoost</td>
<td>Maximum depth</td>
<td>[1, 2, 3, 4, 5, 6, 7, 8, 9, 10,11,12]</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Learning rate</td>
<td>[0.001, 0.01, 0.1, 0.8,1]</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Minimum data in leaf</td>
<td>[1,3,5]</td>
<td>1</td>
</tr>
<tr>
<td>SVR</td>
<td>kernel</td>
<td>['linear', 'poly', 'rbf', 'sigmoid']</td>
<td>rbf</td>
</tr>
<tr>
<td></td>
<td>Regularization</td>
<td>[0.1,1,10]</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Epsilon</td>
<td>[0.1,0.3, 0.5]</td>
<td>0.3</td>
</tr>
<tr>
<td>DL</td>
<td>Number of layers</td>
<td>[1,2,3,4]</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Number of neurons</td>
<td>[1,11, 22,33,44,55]</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Batch size</td>
<td>[32, 64, 128, 256]</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Epochs</td>
<td>[500, 700, 1000]</td>
<td>700</td>
</tr>
<tr>
<td></td>
<td>dropout rate</td>
<td>[ 0.1, 0.2, 0.3]</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Learning rate</td>
<td>[0.2, 0.01, 0.001]</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Table 3.2: Model evaluation using 5-fold cross-validation for LWT

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (°C)</th>
<th>MAE (°C)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.89</td>
<td>0.58</td>
<td>0.98</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.56</td>
<td>0.32</td>
<td>0.99</td>
</tr>
<tr>
<td>GPBoost</td>
<td>0.95</td>
<td>0.72</td>
<td>0.97</td>
</tr>
<tr>
<td>DL</td>
<td>0.74</td>
<td>0.50</td>
<td>0.99</td>
</tr>
<tr>
<td>SVR</td>
<td>0.61</td>
<td>0.29</td>
<td>0.99</td>
</tr>
</tbody>
</table>

3.2.3 Overall testing performance comparison

Figure 3.2 (A-E) and Table 3.3 provide an overview of the test performance of the ML models applied to the unseen dataset. The performance metrics reported in the table are RMSE, MAE, and $R^2$. Among the models, RF demonstrated the lowest RMSE of 1.44°C and MAE of 1.09°C. The $R^2$ value for RF was 0.96, signifying a strong degree of explained variance. DL and SVR models also exhibited robust performance with RMSE of 1.75°C and 1.78°C, MAE of 1.25°C and 1.26°C, respectively and $R^2$ of 0.94. While GPBoost and XGBoost showed competitive results, they presented slightly higher RMSE values and marginally lower $R^2$ values than the models above.

Table 3.3: Test Model Performance for LWT

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (°C)</th>
<th>MAE (°C)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>1.44</td>
<td>1.09</td>
<td>0.96</td>
</tr>
<tr>
<td>XGBoost</td>
<td>1.89</td>
<td>1.36</td>
<td>0.93</td>
</tr>
<tr>
<td>GPBoost</td>
<td>2.13</td>
<td>1.31</td>
<td>0.92</td>
</tr>
<tr>
<td>DL</td>
<td>1.75</td>
<td>1.25</td>
<td>0.94</td>
</tr>
<tr>
<td>SVR</td>
<td>1.78</td>
<td>1.22</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Figure 3.2: Lake Water Temperature ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values.
Figure 3.2: (Continued) Lake Water Temperature ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values.
Figure 3.2: (Continued) Lake Water Temperature ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values.
Figure 3.3 shows the probability density function of the observed temperature compared to the predictions of the five ML models. Among the models, RF and XGBoost exhibit higher density, particularly in lower temperatures. Their patterns show an upward curve when the observed values display a downward curve, specifically in the temperature range of 14 to 18°C. Both models generally align with the peaks observed in the overall distribution. The DL model closely tracks the density values of the observed temperatures across the entire range. However, as the observations decrease in density, the DL model maintains a relatively constant pattern (Between 14°C to 18°C). The SVR model slightly exceeds the observed values and shares a similar issue with the DL model in the mentioned temperature interval. Nevertheless, it stabilizes before experiencing a slight decrease. Finally, the GPBoost model closely follows the observed density values, decreasing when they decrease and increasing when they rise, although slightly above the observed values.

Figure 3.3: Probability density function, observed temperature vs LWT ML models
3.2.4 Understanding the drivers

Figure 3.4a shows the feature importance of the RF model in simulating vertical LWT profiles: `air_temp7d` is the most influential feature, contributing significantly with 47.7%. Following is the `air_temp` feature, accounting for 23.2%. `doy`, which regulates incoming solar radiation and precipitation regimes, holds the third position with 20.8% importance, while `depth_meas` contributes 3.4%. These four variables collectively contribute 91.7% to the model’s predictive accuracy. Other variables each have a contribution of less than 1%. The XGBoost model’s feature importance exhibits a similar pattern to RF, as Figure 3.4b depicts. However, the trio of variables (`air_temp7d, air_temp, and doy`) accounts for an even more substantial portion, totaling 94.3%. The remaining variables collectively contribute less than 0.9%, constituting the remaining 5.7%. This highlights the importance of a few key variables in both models for predicting vertical LWT profiles.

In Figure 3.4c, the GPBoost feature importance highlights that `depth_meas` has the highest importance, contributing 31.3%. Following closely is `vol_lake`, representing 14.3%, and `air_temp7d` with 12.3%. `inflow_lake` contributes 10.8%, `doy` holds 9.9%, `w_Average` is at 5.3%, and `air_temp` is also at 5.3%. Collectively, these seven variables contribute to 89.2% of the total feature importance.

In Figure 3.4d, the SVR permutation importance indicates that `doy` holds the highest importance, contributing 32.1%. Next in line is the `air_temp7d` feature, making up 31.8%. `vol_lake` represents 9.0%, and `depth_meas` contributes 7.1%. `surf_area` holds 6.1%, while `air_temp` is at 5.7%. Together, these six variables contribute to 91.7% of the total feature importance.
Figure 3.4: Feature importance for the LWT ML models.
Figure 3.4: (Continued) Feature importance for the LWT ML models.
Assessing the importance of features in the DL model poses challenges compared to simpler models like decision trees. However, an analytical approach utilizing SHAP additive explanations enables us to examine the impact of individual features on the model’s output. This is achieved by perturbing each feature and observing resultant changes in predictions. Careful consideration is necessary for interpretations derived from this method, as it can provide valuable insights into how variations in features influence the model’s predictions for LWT in the lake.

In Figure 3.5, the y-axis denotes variable names arranged in order of importance from top to bottom, accompanied by mean SHAP values. The x-axis represents SHAP values, indicating the degree of change in log-odds. These values show the distribution of the impacts each feature has on the model output. The gradient color represents the original value for each variable, with each point on the plot corresponding to a row from the original dataset. The plot reveals that the top five influential variables are air_temp7d, doy, depth meas, vol lake, and air temp. The feature surf_area_depth appears to be associated with positive (red) and negative (blue) SHAP values. A positive SHAP value indicates a contribution towards higher predicted LWT, suggesting that lakes with a larger surface area relative to their depth may experience more significant temperature variations. Conversely, a negative SHAP value suggests a contribution towards lower predicted temperatures, indicating that lakes with a smaller surface area relative to their depth may exhibit more stable temperature conditions. When depth meas has high (red) feature values and negative SHAP values, it suggests that higher values of depth meas are associated with lower predicted LWT. This implies that deeper areas in the lake contribute to lower predicted temperatures. Conversely, when depth meas has low (blue) feature values and positive SHAP values, it indicates that lower values of depth meas are associated with higher predicted LWT. Lower air temperatures contribute to lower predicted temperatures, while higher air temperatures contribute to higher predicted temperatures. The transition from blue to red indicates a shift in this relationship, with the magnitude of the SHAP values indicating the strength of these associations.
3.2.5 Model Performance Analysis in Stratified and Non-Stratified Water Column Conditions

In the preceding section, it was discussed that the presence of a thermocline in a vertical temperature profile indicates stratification in the water column. Subject to determining whether a profile is stratified or not, the test dataset was partitioned accordingly. The results for the models under stratified conditions are detailed in Table 3.4. For RF, XGBoost, DL, and SVR, $R^2$ values ranged between 0.45 and 0.54, with RMSE between 4.12 and 4.47 °C. The GPBoost model exhibited the best performance among these models, showcasing an $R^2$ of 0.86 and an RMSE of 2.26°C.
Table 3.4: LWT Model Performance under Stratified Water Column Conditions

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (°C)</th>
<th>MAE (°C)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>4.12</td>
<td>3.29</td>
<td>0.54</td>
</tr>
<tr>
<td>XGBoost</td>
<td>4.35</td>
<td>3.35</td>
<td>0.52</td>
</tr>
<tr>
<td>GPBoost</td>
<td>2.26</td>
<td>1.52</td>
<td>0.86</td>
</tr>
<tr>
<td>DL</td>
<td>4.24</td>
<td>3.21</td>
<td>0.53</td>
</tr>
<tr>
<td>SVR</td>
<td>4.47</td>
<td>3.57</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Moving on to the results under non-stratified conditions, as illustrated in Table 3.5, the SVR model showed the poorest performance with the lowest $R^2$ value of 0.68 and the highest RMSE of 4.24°C. Conversely, the GPBoost model demonstrated the best performance, achieving the highest $R^2$ of 0.95 and the lowest RMSE of 1.6°C.

Table 3.5: LWT Model Performance under Non-Stratified Water Column Conditions

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (°C)</th>
<th>MAE (°C)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>3.31</td>
<td>2.74</td>
<td>0.78</td>
</tr>
<tr>
<td>XGBoost</td>
<td>3.46</td>
<td>2.64</td>
<td>0.79</td>
</tr>
<tr>
<td>GPBoost</td>
<td>1.6</td>
<td>1.04</td>
<td>0.95</td>
</tr>
<tr>
<td>DL</td>
<td>3.1</td>
<td>2.2</td>
<td>0.81</td>
</tr>
<tr>
<td>SVR</td>
<td>4.24</td>
<td>3.34</td>
<td>0.68</td>
</tr>
</tbody>
</table>

3.3 GLM

This study conducted uncalibrated GLM simulations on Lake Texoma for 2016. Inputs included meteorological variables (e.g., shortwave radiation, precipitation) and hydrological
variables (e.g., reservoir inflow and outflow). Parameters governing mixing efficiency and lake morphology were also considered. The simulation divided the water column into layers of 0.1 meters with a maximum depth of 32 meters. Lake Texoma was chosen due to its extensive dataset compared to other lakes, and the year 2016 was selected to cover multiple seasons.

Figure 3.6 displays the raw results for the entire simulation year, while figure 3.8 presents selected simulation results for days with available measurements, comparing GLM, DL, and observed data. GLM simulations calculated LWT only up to 7 meters and did not accurately reflect stratification patterns. During winter dates (figure 3.7a and 3.7b), the water column is mixed, as captured by the DL model. In summer (figure 3.7c), the DL model follows the stratification behavior of the observed data with slightly lower values. Finally, in fall (figure 3.7f), the GLM and DL simulations are in the same temperature range, but the DL was able to calculate temperature deeper in the column while the GLM only reached up to 7 meters. These two models have an average temperature 3°C lower than the simulated data. Finally, Figures 3.8a, 3.8b, and 3.8c display the DL modeling results for some of the above profiles, allowing us to visualize the vertical temperature profile of the lake based on the bathymetry.

Figure 3.6: GLM LWT simulations over 2016
Figure 3.7: LWT profile simulations vs observed data, year 2016
Figure 3.8: LWT simulations interpolated onto bathymetric map, Lake Texoma, year 2016
3.4 DO Modeling

3.4.1 Hyperparameter tuning

An initial sensitivity analysis was conducted to ensure that the ML models were properly tuned and capable of delivering reliable predictions of DO. Subsequently, Randomized Search technique was applied to identify the optimal hyperparameter combinations for each model, as summarized in Table 3.6. As a result, the RF model was configured with 1000 estimators, a maximum of 3 features, a minimum sample leaf of 3, a minimum sample split of 10, and a depth of 30. The SVR model was set with a radial basis function kernel, a regularization parameter of 30, and an epsilon value of 0.3. The XGBoost model employed 400 estimators, a learning rate of 0.01, a depth of 5, a gamma value of 0.2, and a subsample of 0.9. The GPBoost model was configured with a learning rate of 0.05, maximum depth of 12, and minimum data in leaf of 1. Lastly, the DL model consisted of 3 layers with 36 neurons, a batch size of 128, 700 training epochs, a dropout rate of 0.1, and a learning rate of 0.01.

3.4.2 Cross-validation comparison

After conducting hyperparameter tuning, 5-fold cross-validation is employed to assess how statistical analysis results generalize to an independent dataset. Table 3.7 demonstrates consistent performance across all methods, with lower values for both RMSE and MAE. The $R^2$ values are consistently closer to 1 for all cases. DL outperformed other models, achieving the lowest RMSE (mg/l) and MAE (mg/l) values in the 5-fold cross-validation analysis.
Table 3.6: Selected hyperparameter values after a random search strategy for each of the test ML algorithms at predicting DO

<table>
<thead>
<tr>
<th>Model</th>
<th>Hyperparameters</th>
<th>Range</th>
<th>Selected value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>Number of estimators</td>
<td>[50,100,200]</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Maximum features</td>
<td>[1, 2, 3, 4, 5, 6, 7, 8, 9, 10,11,12]</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Minimum sample leaf</td>
<td>[1,3,5]</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Minimum sample split</td>
<td>[1,5,10,15]</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Depth</td>
<td>[10,20,30,50]</td>
<td>30</td>
</tr>
<tr>
<td>XGBoost</td>
<td>Number of estimators</td>
<td>[100,200,300,400,500]</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>Learning rate</td>
<td>[0.001, 0.01, 0.1, 0.2]</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Depth</td>
<td>[3,4,5,6,7,8,9,10]</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>gamma</td>
<td>[0, 0.1, 0.2, 0.3, 0.4]</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Subsample</td>
<td>[0.7, 0.8, 0.9, 1.0]</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>colsample bytree</td>
<td>[0.6, 0.7, 0.8, 0.9, 1.0]</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>reg alpha</td>
<td>[0, 0.1, 0.2, 0.3]</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>reg lambda</td>
<td>[0, 0.1, 0.2, 0.3]</td>
<td>0.2</td>
</tr>
<tr>
<td>GPBoost</td>
<td>Maximum depth</td>
<td>[1, 2, 3, 4, 5, 6, 7, 8, 9, 10,11,12]</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Learning rate</td>
<td>[0.01, 0.05, 0.1, 1.0]</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Minimum data in leaf</td>
<td>[1,3,5]</td>
<td>1</td>
</tr>
<tr>
<td>SVR</td>
<td>kernel</td>
<td>['linear', 'poly', 'rbf', 'sigmoid']</td>
<td>rbf</td>
</tr>
<tr>
<td></td>
<td>Regularization</td>
<td>[0.1,1,10,20,30,40]</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Epsilon</td>
<td>[0.1,0.3, 0.5]</td>
<td>0.3</td>
</tr>
<tr>
<td>DL</td>
<td>Number of layers</td>
<td>[1,2,3,4]</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Number of neurons</td>
<td>[1,12,24,36,48]</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Batch size</td>
<td>[32, 64, 128, 256]</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Epochs</td>
<td>[500, 700, 1000]</td>
<td>700</td>
</tr>
<tr>
<td></td>
<td>dropout rate</td>
<td>[ 0.1, 0.2, 0.3]</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Learning rate</td>
<td>[0.2, 0.01, 0.001]</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Table 3.7: Model evaluation using 5-fold cross-validation for DO

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (mg/l)</th>
<th>MAE (mg/l)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>0.92</td>
<td>0.62</td>
<td>0.92</td>
</tr>
<tr>
<td>RF</td>
<td>0.94</td>
<td>0.58</td>
<td>0.93</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.94</td>
<td>0.61</td>
<td>0.93</td>
</tr>
<tr>
<td>GPBoost</td>
<td>0.90</td>
<td>0.61</td>
<td>0.93</td>
</tr>
<tr>
<td>DL</td>
<td>0.90</td>
<td>0.56</td>
<td>0.93</td>
</tr>
</tbody>
</table>

3.4.3 Overall testing performance comparison

When tested on unseen data, models show similar performance (See Table 3.8). $R^2$ values range from 0.81 to 0.83, and RMSE ranges from 1.18 to 1.32 mg/l. Figure 3.9 (A-E) illustrates the comparison between observed DO levels and the corresponding predictions generated by the ML models, with a density estimate of the data distribution. Discrepancies are visually represented by a vertical line of data points centered around the value 0 in the observations: the model tends to overestimate DO concentrations compared to the actual observed values.

Table 3.8: Test DO Model Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE (mg/l)</th>
<th>MAE (mg/l)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>1.24</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>RF</td>
<td>1.23</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>XGBoost</td>
<td>1.32</td>
<td>0.87</td>
<td>0.81</td>
</tr>
<tr>
<td>GPBoost</td>
<td>1.20</td>
<td>0.73</td>
<td>0.84</td>
</tr>
<tr>
<td>DL</td>
<td>1.18</td>
<td>0.77</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Figure 3.9: Dissolved oxygen ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values.
Figure 3.9: (Continued) Dissolved oxygen ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values.
Figure 3.9: (Continued) Dissolved oxygen ML model performance after hyperparameter optimization using the testing measured data (x-axis) and ML predicted (y-axis) values.
Figure 3.10 shows the probability density function of the observed DO compared to the predictions of the five ML models. A notable inconsistency emerges around 7-8 mg/L DO, where the SVR, DL, GPBoost, and RF models show simulated probability density values decreasing while the observed values continue to increase. Additionally, the presence of two peaks in the probability density function predicted by the SVR, RF, GPBoost, and DL models in this range, as opposed to the single peak observed in the data, suggests that these models capture additional complexities or noise in the data. Within this range, the XGBoost model shows rising and falling patterns at the peak of the probability density distribution but with slightly lower density values. This discrepancy may indicate that while the XGBoost model captures the overall trend of the data, it predicts a slightly lower likelihood for the specific range of DO concentrations around 7-8 mg/L.

Figure 3.10: Probability density function, observed temperature vs DO ML models
3.4.4 Understanding the drivers

Figure 3.11a shows the feature importance of the RF model in simulating vertical DO profiles. \textit{predicted\_lwt} (Predicted water temperature) is the most influential feature, contributing significantly with 22.3%. Following is the \textit{air\_temp} feature, accounting for 20.8%, while \textit{air\_temp7d} contributes 17.7%. \textit{depth\_meas} contributes 15.9 %, \textit{doy} 9.5%, \textit{wind} 3.7%, and \textit{vol\_lake} 2.4%. These variables collectively contribute to 92.2% of the model’s predictive accuracy. Other variables each have a contribution of less than 2 %.

Regarding XGBoost, figure 3.11b, the feature \textit{air\_temp} is the most influential, contributing significantly with 56.7%, while \textit{air\_temp7d} contributes 17.3 %, and \textit{depth\_meas} 5.5%. These variables contribute to 79.5% of the importance. Other variables each have a contribution of less than 5%. This highlights the importance of a few key variables in the XGBoost model for predicting vertical DO profiles.

In Figure 3.11c, the GPBoost feature importance highlights that \textit{depth\_meas} has the highest importance, contributing 35.1%. Following is \textit{predicted\_lwt}, representing 15.0%, and \textit{air\_temp7d} with 12.4%. Also, \textit{vol\_lake} contributes 6.4%, \textit{wind\_avg7} holds 5.7%, \textit{inflow\_lake} 5.6%, \textit{air\_temp} 5.1%, and \textit{doy} 4.0%. Collectively, these variables contribute to 89.4% of the total feature importance.

In Figure 3.11d, the SVR permutation importance indicates that \textit{air\_temp} holds the highest importance, contributing 24.8%. Next in line is the \textit{predicted\_lwt} feature, making up 24.1%. Then, \textit{depth\_meas} represents 13.1%, and \textit{air\_temp7d} contributes 12.0%. \textit{doy} 5.6%, \textit{vol\_lake} 5.0%, \textit{wind\_avg7} 4.4%, and \textit{prec\_cum7} 3.0% jointly contribute to 92% of the feature importance.
Figure 3.11: Feature importance for the DO ML models.
Figure 3.11: (Continued) Feature importance for the DO ML models.
Figure 3.12 displays the top five influential variables in predicting DO: \textit{air\_temp}, \textit{depth\_meas}, \textit{air\_temp7d}, \textit{doy}, and \textit{predicted\_lwt}. The figure reveals two strong patterns. Firstly, lower air temperatures are associated with higher predicted DO, while higher air temperatures correspond to lower predicted DO. Secondly, shallower depths are associated with higher predicted DO, whereas deeper depths are connected to lower predicted DO.

Figure 3.12: SHAP additive explanations for the Deep Learning model of DO.
Chapter 4

Discussion

This master’s thesis studied the efficacy of five ML techniques in predicting LWT profiles on a daily basis across 12 reservoirs in the RRB. After calibrating the models with the training data (hyperparameter tuning), the evaluation metrics (RMSE, MAE, and $R^2$) were applied to the overall unseen data in the test dataset (See Table 3.3), which are commonly used metrics to evaluate the performance of ML models in predicting variables such as LWT and DO. These metrics provide valuable insights for assessing the accuracy and reliability of models, particularly in scenarios related to climate change or environmental changes impacting fish species and ecosystem services (Vishwakarma et al., 2022). The results generally perform similarly to the tested LWT machine learning models, with $R^2$ ranging from 0.90 to 0.94 and RMSE within 1.44°C and 2.13°C. In contrast, water temperature tasks using ML have demonstrated RMSE values ranging from 0.325 to 1.69°C, with some models significantly outperforming traditional linear regression approaches (Zhu et al. 2019; Feigl et al. 2021). However, it is essential to interpret the results cautiously, considering the physics related to the modeled variable. Additionally, although all models initially exhibited satisfactory performance metrics on the entire dataset, their performance decreased when assessed under stratified versus non-stratified conditions, a point which will be analyzed further in the following paragraphs.

Ensemble learning methods such as RF and XGBoost are widely used in various fields like environmental science, medicine, and agriculture due to their high classification accuracies and effectiveness in nonlinear mapping between features and target variables (Wang et al. 2022; Jafarzadeh et al. 2021). However, these methods can encounter challenges in datasets with highly correlated features (Dorado-Guerra et al. 2022). For instance, the correlation
analysis revealed a strong correlation between LWT measurements and air temperature, along with the average air temperature of the last seven days. In the feature importance analysis for both LWT models (RF and XGBoost), the dominance of the average temperature of the last seven days, the day of the year, and the air temperature for the same day explained the 90% of the variability of the LWT modeling. While some of these features are justified due to their impact on heat exchange between the atmosphere and the lake surface (Schmid et al. 2022; Hipsey et al. 2014), their disproportionately high importance suggests a potential oversight of other crucial variables such as depth, volume (less than 1.1

While RF and XGBoost are recognized for their effectiveness in machine learning applications, concerns regarding overfitting have been raised. Overfitting can occur when these models assign excessive importance to specific features that are correlated with the target variable in the training data, potentially impacting their generalizability to new, unseen LWT profiles. Additionally, while ensemble methods like RF and XGBoost are flexible in capturing complex nonlinear relationships, they may produce models that do not always align with the expected linear or physical relationships inherent in LWT dynamics. In this context, alternative methods may offer more interpretable and reliable results, warranting careful consideration before choosing the appropriate modeling approach (Huang et al. 2022).

The GPBoost model, which combines tree-boosting and mixed effects models, proved effective for modeling LWT. While this method has not been previously applied to this specific task, previous research by Pollak et al. (2023) successfully utilized GPBoost for modeling pain, demonstrating its capability to predict with an intuitive interpretation of results. GPBoost extends linear mixed effects models by considering the response variable as the sum of a potentially nonlinear mean function and random effects. Its iterative training process, involving learning covariance parameters and the regression function, enhances adaptability and accuracy (Schmid et al. 2022). Although GPBoost did not exhibit the highest performance among all LWT models (see Table 3.3), it maintained satisfactory metrics, with an $R^2$ of 0.85 and an RMSE of 1.5°C.

Despite being less accurate than other models studied, GPBoost aligns more closely
with the physical aspects of the problem among the ensemble methods when analyzing the variables that collectively contribute to the model’s predictive accuracy (depth is the most influential driver, followed by air temperature for the same day, inflow, the average temperature of the last seven days, volume, and day of the year). The observed errors may be attributed to the model’s avoidance of overfitting and its ability to capture inherent uncertainties associated with the natural physics of the problem (Modeling LWT in lakes presents challenges due to the complex dynamics influenced by heating and cooling variations, showing a basic temporal structure with seasonal differences between summer and winter (Jordan S. Read et al. 2019; Schmid et al. 2022). Furthermore, it performed well under stratified and non-stratified datasets, as depth, the main driver in the LWT feature importance, captures the stratification, mixing processes, and water column stability (J. S. Read et al. 2013; Kling 1988; Magee et al. 2017b), distinguishing it from other ML models studied. The interpretability, adaptability, and alignment with the physical characteristics of the problem position GPBoost as a promising approach for further exploration and application in environmental studies.

While the SVR and DL models initially demonstrated satisfactory performance metrics on the whole dataset (See Table 3.3), their performance metrics decreased when assessed under stratified versus non-stratified conditions (See Tables 3.4 and 3.5). For the SVR model, the day of the year is the most influential variable despite not being a direct physical variable, which provides valuable information on various driving factors, including solar radiation variability, monthly precipitation, runoff regimes, seasonal atmospheric cycles driving winds, cloudiness, atmospheric pressures, and evaporation (Weng 2012). Competing evidence can be found in the study by Moazenzadeh et al. (2018), which focuses on predicting evaporation in northern Iran and challenges the notion that variables like the day of the year are the most influential in SVR models, as observed in this study. Additionally, the interpretability of the DL model, often referred to as ”black boxes” (Teng et al. 2022), presents a significant challenge. The SHAP method provides a means to interpret these models, even when highly complex, by offering explanations based on the contributions of individual features.
The observed relationships between influential variables and LWT align with physical expectations.

Additionally, Figure 3.3 illustrates the probability density function for all the evaluated LWT models. The GPBoost model closely aligns with the bimodal data distribution, effectively capturing temperature variations across the spectrum. In contrast, methods such as SVR, RF, and XGBoost tend to overestimate the density probability for cooler temperatures. This discrepancy suggests that these models predict a higher likelihood or frequency of occurrence for cooler temperatures than is observed in the data, indicating a potential limitation in their ability to account for factors influencing temperature variations in lake environments. On the other hand, the GPBoost model predicts cooler temperatures more accurately, resulting in a probability density function that better reflects the observed temperature distribution, which suggests a more comprehensive understanding or representation of the complex thermal dynamics within lakes, potentially attributed to the GPBoost model’s prioritization of depth as a primary component for model accuracy, as highlighted in the feature importance analysis section. Previous studies have demonstrated the significant influence of lake depth on various factors such as stratification, stability, surface heat flux differences, hypolimnion temperature, and the onset and fall overturn dates of stratification (Magee et al. 2017b). Consequently, the GPBoost model’s depth consideration enables it to outperform SVR, RF, and XGBoost, which rely on a limited set of meteorological variables (See Figure 3.4a,3.4b,3.4d) to simulate thermal dynamics in lakes. Future work will investigate whether custom loss functions can be defined for those models to better account for depth’s importance in thermal dynamics simulation.

The observed discrepancies in the performance of the GLM should not be hastily interpreted as indicative of inferior modeling capabilities. It is crucial to acknowledge that the GLM simulations were run uncalibrated, which may not accurately capture the complexities of Texoma Lake’s ecosystem. Thus, any evaluation of its performance must be approached with caution, recognizing the need for further refinement and calibration. The absence of an observed dataset for calibration poses a significant limitation, primarily because of the
discrete nature of the available data. Prior research has underscored the difficulties of cali-
branding models based on limited datasets, highlighting the potential for inaccuracies in model
predictions (Meyer-Jacob et al. 2017). Instead, the comparison demonstrates that ML tech-
niques can deliver faster and more precise results with the same input data and timeframe.
Additionally, incorporating GLM parameters and variables like thermal properties, boundary
conditions, biological and chemical factors, and ice dynamics into ML models can enhance
model performance. However, including these specialized parameters may necessitate addi-
tional data collection efforts, given their specific and sometimes complex nature (S. Liu et al.
2019; Beaulieu et al. 2020).

Regarding DO modeling, the results show a very similar performance of the tested DO
machine learning models, with $R^2$ ranging from 0.81 to 0.84 and RMSE within the range of
1.18 mg/l to 1.32 mg/l (see Table 3.8). In comparison, previous studies employing ML for
DO tasks have reported RMSE values ranging from 0.98 to 1.28 mg/L (M. Liu et al. 2022).
These models incorporated factors such as water temperature, water clarity, and chlorophyll-
a levels to make accurate predictions. In this study, the DO feature importance results
(see 3.11) show that the XGBoost model for DO exhibits an issue as it assigns excessive
importance to specific features that correlate with the target variable in the training data,
similar to the LWT model using XGBoost, raising concerns about the model’s reliability for
this specific task. In contrast, the GPBoost, RF, and SVR models predominantly rely on
LWT, air temperature, depth, day of the year, wind, and volume. These models, with their
more diverse features explaining the process, outperform XGBoost.

Also, the DO model’s tendency to overestimate DO concentrations compared to the
observed values (See Figure 3.9) is primarily evident during the summer months when DO
levels near zero are observed in the deeper layers of the lake (hypolimnion), characterized
by reduced DO and cool water. DO modeling is inherently complex, involving intricate
physical processes such as diffusion, which are influenced by temperature, salinity, pressure,
and mixing dynamics (Scully 2013). The current set of features utilized by the model may
not fully capture these nuanced dynamics, suggesting the need for further refinement and
inclusion of additional variables.

In this study, ML models incorporating data from multiple reservoirs and lakes have demonstrated their potential to provide reasonable approximations for LWT and DO despite acknowledged challenges such as overfitting and interpretability. The ability of ML models to capture intricate nonlinear relationships positions them as promising tools for initial estimations, particularly in scenarios where obtaining extensive datasets poses practical difficulties. Notably, this study achieved significant performance metrics utilizing only air temperature, wind, precipitation, and a few morphological and hydrological characteristics. ML offers a practical alternative, balancing computational efficiency and acceptable accuracy. Unlike GLM, which requires a significant amount of data for robust outcomes and calibration, ML emerges as a pragmatic approach for obtaining meaningful insights within the constraints of time and data availability. It presents the potential for successful predictions even without the diverse array of data—such as lake mixing, water clarity, and heat fluxes—typically demanded by physically-based models.

Finally, developing a predictive model for these crucial environmental variables holds immense promise in advancing our understanding of aquatic ecosystems. Such a model becomes a valuable tool, enabling the assessment of the impact of various factors on water quality and empowering informed decision-making for the sustainable management of water resources.
Chapter 5

Limitations and Future Work

When simulating LWT, process-based lake ecosystem models consider various factors influencing this variable, such as heat exchanges (short wave and long wave radiation, evaporation, and conduction), mass exchanges (water vapor, inflows, and outflows), and mechanical energy (wind stress) (Imberger et al. 1989; Prats, Jordi et al. 2019). These models also account for internal mixing and stratification processes, driven by energy exchanges and influenced by chemical, thermal gradients, and biological activity (Imberger et al. 1989; Boehrer et al. 2008; Prats, Jordi et al. 2019). However, these models need extensive data for calibration, which may be unavailable for many water bodies. ML becomes a suitable approach in such cases, especially for water and fishing planning tasks. The proposed approach uses only air temperature, wind, precipitation, and a few morphological and hydrological characteristics. Despite its good performance, the approach has limitations. First, the model does not consider all factors influencing thermal behavior, potentially leading to oversights in predictions (Chakravarthy et al. 2022). Second, ML models overlook water transparency, a crucial factor for one-dimensional models, particularly in clear waters (Prats, Jordi et al. 2019; Henderson-Sellers 1988; Heiskanen et al. 2015).

Data-driven models are commonly used in environmental modeling but face challenges due to limited training data and potential discrepancies with physical laws. The absence of continuous data per lake limits the recognition of nuanced temperature patterns, such as the slow and muted change in LWT during the fall, in contrast to the highly variable spring warming period. The temperature dynamics in lakes exhibit fluctuations due to variations in heating and cooling. These patterns have a basic temporal structure, such as seasonal differences between summer and winter, coupled with short-term responses to
prevailing weather conditions (Jordan S. Read et al. 2019). Continuous data, along with appropriate training and predictor data, could enhance the simulation of these dynamics using ML methods, including LSTMs (Long Short-Term Memory), a type of recurrent neural network. Additionally, continuous data serves not only for ML models but also for running and calibrating physics-based models, facilitating effective comparisons.

Using a series of models, such as one to predict the onset of stratification and another for LWT, could help identify if a lake with specific temperature and meteorological conditions will exhibit stratification. However, implementing this method is challenging due to the requirement for continuous data to accurately determine the onset of stratification. Obtaining such continuous data is not straightforward and typically necessitates field campaigns conducted over various dates.

Future work involves developing a Process-guided Deep Learning (PGDL) model that integrates an LSTM neural network and a process-based model (PBM), specifically the GLM. Conservation of energy is crucial in LWT predictions within PBMs, as it influences the physical validity of predicted outcomes. Evaluating reservoir LWT prediction performance includes assessing the satisfaction of the energy conservation law \( E_{TR} = E_{SR} \) and using error indices to compare measured and predicted values (Kim et al. 2023). PGDL models leverage the strengths of PBM and data-driven models to improve predictive accuracy while ensuring physical consistency (PGDL assumes that PBM adequately captures the underlying physics, and DDM can capture any remaining unknown physics) (Kim et al. 2023).

Modeling DO presents inherent challenges due to its dependence not only on meteorological and hydrological factors but also on physical processes like diffusion (Pasternak et al. 1967). This process involves DO movement from higher to lower concentrations, influenced by temperature, salinity, and pressure, playing a crucial role in oxygen transfer between the atmosphere and water. Furthermore, the physical mixing of water layers, driven by wind or turbulence, significantly influences the distribution of DO (Flaim et al. 2020). The rate at which organic material is oxidized, known as biochemical oxygen demand, is another critical factor in modeling oxygen dynamics, among other variables (Matear et al. 2000). While
this initial approach provides a starting point, further advanced techniques such as extreme learning machines and deep belief networks are necessary. These techniques have been utilized to predict DO content in various aquatic environments, including aquaculture systems and rivers (Adeniran et al. 2017; W. Liu et al. 2023). These models incorporate traditional variables and innovative approaches like feature engineering and optimization methods to enhance prediction accuracy (Nemati et al. 2015).

In conclusion, modeling lake water temperature and dissolved oxygen presents challenges, ranging from the complexities of physical processes to the limitations of available data and modeling techniques. While traditional process-based models offer valuable insights into ecosystem dynamics, they often require extensive calibration data. On the other hand, machine learning approaches show promise for enhancing predictive accuracy, particularly in scenarios where continuous data is scarce. However, integrating these approaches with physics-based models, such as the proposed Process-guided Deep Learning model, holds great potential for advancing our understanding and prediction capabilities in aquatic ecosystem modeling. Future research efforts should bridge the gap between data-driven and physics-based modeling approaches, leveraging their strengths to achieve more robust and accurate predictions for effective management and conservation strategies for freshwater ecosystems worldwide.
Chapter 6

Conclusions

This thesis presented a ML-based approach for predicting and learning the drivers of LWT and DO across a number of lakes of the Red River Basin of the south in the United States, using meteorological and hydrological input data. The main conclusions obtained are numbered here:

- GPBoost emerges as the most effective ML method in predicting LWT and DO. Its performance is attributed to incorporating more interpretable physical variables relevant to the problem, allowing the model to learn better.

- RF, XGBoost, and SVR exhibit signs of overfitting and oversimplifying the modeling problem in the context of the lake water ecosystem. Alternative methods may lead to more interpretable and reliable results that are better suited to capture those complexities.

- A comparison between the LWT machine learning models and an uncalibrated GLM (physics-based model) developed in this thesis reveals better performance for the ML algorithms. It is important to note that the uncalibrated GLM model may not accurately capture the complexities of Texoma Lake. Therefore, any discrepancies in the GLM’s performance should be interpreted cautiously. However, comparing the GLM with ML techniques highlights the potential of ML for achieving faster and more precise results with the same dataset and timeframe, requiring fewer parameters to calibrate than the physics-based model does to achieve high accuracy.

- This study demonstrates the potential of ML algorithms in predicting LWT and DO
using a limited set of variables, including air temperature, wind, precipitation, and specific morphological and hydrological factors. This positions ML as a valuable tool for research to enhance the understanding of climate change impacts on lake ecosystems.

- ML algorithms effectively balance computational efficiency with satisfactory accuracy. Moreover, they offer significant prospects for enhancing our understanding of aquatic ecosystems and enabling informed decision-making in sustainable water resource management.
Bibliography


Adrian, Rita et al. (2009). “Lakes as sentinels of climate change”. In: Limnology and Oceanography 54.6, pp. 2283–2297. DOI: 10.4319/lo.2009.54.6_part_2.2283.


76


Flerchinger, GN, Wei Xiaio, Danny Marks, TJ Sauer, and Qiang Yu (2009). “Comparison of algorithms for incoming atmospheric long-wave radiation”. In: *Water resources research* 45.3.


Isabela Suaza Sierra was born and raised in Apartadó, Colombia. She graduated from Universidad Nacional de Colombia in June 2022 as a Geological Engineer. Before graduating, she worked as a Teaching Assistant for three years in Differential Equations and Hydrogeology and completed a six-month internship at the Colombian National Environmental Licensing Authority (ANLA).

In the Fall of 2022, Isabela started her Master’s in Geological Science in the Department of Earth, Environmental, and Resource Science (DEERS) at The University of Texas at El Paso, mentored by Dr. Hernan Moreno. She joined the GeoSenSE group as a Research Assistant, focusing on developing machine learning models for predicting water temperature and oxygen levels in lakes and reservoirs, along with hydrological modeling for the Southwestern United States. She also was enrolled in the Geographic Information Science and Technology (GIST) certificate program. Isabela presented her research at various conferences, including AGU 2023, HydroML 2023, and DEERS Colloquium. She served as an Outreach Officer for SACNAS (Society for Advancement of Chicanos/Hispanics and Native Americans in Science), advocating for diversity in STEM.

She was honored with three White Awards and received the Professor Nicholas E. Pingitore Jr. Innovation Award in 2023, which is presented each year to a master’s student who demonstrates outstanding innovative ideas and commitment to success. She was selected for the prestigious NCAR internship in the summer of 2023. Isabela’s exceptional academic performance earned her the Outstanding Graduate Student in Geological Sciences recognition at the 2024 Honors Convocation upon graduating in May 2024.

Currently, Isabela is starting her Ph.D. in Geology at the University of Texas at El Paso. Her interests include water management, hydrological modeling, remote sensing, machine learning, and using environmental data for decision-making.