UNIVERSITY OF TRENTO

Doctoral School in Civil, Environmental and Mechanical Engineering

Curriculum 1. Civil and Environmental Engineering

Azadeh Yousefi

# Investigating the Crucial Factors Impacting the Accuracy of Machine Learning Models for Lake Surface Water Temperature and Ice Thickness Prediction

2023 - Doctoral thesis

UNIVERSITY OF TRENTO - Italy
Department of Civil, Environmental and Mechanical Engineering

Doctoral School in Civil, Environmental and Mechanical Engineering

Civil and Environmental Engineering

XXXV cycle 2020/2023

Doctoral Thesis - December 2023

Azadeh Yousefi

# Investigating the Crucial Factors Impacting the Accuracy of Machine Learning Models for Lake Surface Water Temperature and Ice Thickness Prediction

Supervisors

Marco Toffolon, Sebastiano Piccolroaz

Credits of the cover image .....



University of Trento
Doctoral School in Civil, Environmental and Mechanical Engineering
*http://web.unitn.it/en/dricam*
Via Mesiano 77, I-38123 Trento
Tel. +39 0461 282670 / 2611 - *dicamphd@unitn.it*

# Abstract

Climate change is having a profound impact on freshwater ecosystems, with lakes being highly sensitive to environmental changes. In recent years, the increasing popularity of machine learning techniques in this field can be attributed to their ability to analyze intricate patterns and relationships in meteorological data, lake dynamics and their response to climate change. This study aims to explore how meteorological variables impact lake surface water temperature and ice thickness.

First, we evaluate nine machine learning algorithms for lake surface water temperature prediction in synthetic lakes, comparing their performance and investigating the effects of input variables on accuracy. Our study, based on a numerical model (distinct from a real lake), indicates that considering air temperature and day of the year suffices for acceptable outcomes. Additional predictors provide minimal improvement. Furthermore, better results can be achieved by pre-processing air temperature through time averaging or incorporating past values specially for deep lakes. Despite exploring various machine learning algorithms with the same inputs, no single optimal choice emerged (although artificial neural networks exhibited slightly better results).

Second, we select a machine learning technique, artificial neural network, to model the influential factors of lake surface water temperature response on 2024 lakes worldwide, based on the CCI Lakes dataset. Our analysis reveals that, in general, the day of the year is the most relevant factor, suggesting that the mean (climatological) year is already a good approximation. Removing it from the set of predictors, air temperature, shortwave and downward longwave radiation and relative humidity gain the predominant roles and the incorporation of other meteorological variables could significantly or moderately improve the models' performance across different climatic zones.

Third, we investigate the influence of various meteorological variables on ice thickness prediction in two distinct lakes in Sweden using artificial neural network. Among the input variables, the day of the year assumes a significant role in simulating ice thickness. Additionally, shortwave radiation and specific humidity prove to be pivotal predictors. During the period of ice formation, aside from the day of the year, the negative degree days linked to negative air temperature also stand out as influential predictors.

Our findings demonstrate that machine learning techniques offer a promising avenue for studying lake dynamics and their response to environmental changes because of being flexible to change the input variables and to analyze their importance on the model leading to understand the physics behind it. Through the random regeneration of each feature, we can assess its impact on the model by

measuring the extent to which it reduces the model's performance compared to the model incorporating all variables. The choice of meteorological variables plays a critical role in model performance, emphasizing the need to select relevant input variables for optimal results.

# Contents

# Chapter 1. Introduction

## 1.1. Aims and motivation

The thermal response of lakes is a complex phenomenon, influenced by several factors, including local climate conditions and anthropogenic activities. The current warming trend of lake water temperature, as highlighted by Woolway and Merchant (2018) and Woolway et al. (2020), poses a significant threat to the biological and chemical processes in aquatic ecosystems, ultimately affecting the survival of aquatic life. In this context, it becomes crucial to understand the factors that control the thermal response of lakes to changing climate conditions, such as air temperature (AT) and solar radiation, and to quantify the expected alterations in water temperature patterns. Such knowledge is of paramount importance for effective management of aquatic resources and the preservation of freshwater ecosystems. Several studies have indicated that increasing water temperatures can have cascading effects on the food web, leading to changes in species composition and distribution, and altering the nutrient cycles in the ecosystem (Carnicer et al., 2011). Thus, quantifying the expected changes in lake water temperature can enable the identification of potential risks and inform management decisions to mitigate the impacts of climate change on freshwater ecosystems (O'Reilly et al., 2003).

Water temperature is a fundamental abiotic factor that influences the ecology of lakes, water quality, and human activities. Temperature affects the distribution, abundance, and behavior of aquatic organisms in lakes, and changes in temperature can have cascading effects on the entire ecosystem. Temperature also affects water quality by influencing the physical, chemical, and biological properties of water (Gebrekiros, 2016). Additionally, high water temperatures can promote the growth of harmful algal blooms, which can produce toxins and impact the health of humans and animals that rely on the lake for drinking water or recreation (Chapra et al., 2017). Finally, water temperature can impact various human activities in and around lakes, such as swimming, boating, fishing, and water supply. Therefore, it is essential to monitor and understand water temperature dynamics in lakes to inform decision-making regarding lake management, water quality, and human activities. By understanding the impacts of meteorological variables on lake surface water temperature (LSWT), we can develop effective strategies to protect and manage these important ecosystems.

Ice thickness is another vital indicator of seasonal dynamics, influencing aquatic life, water quality, and human activities such as spiritual rituals, competitions in ice skating, tournaments for ice fishing, and the extended utilization of winter ice roads in lake ecosystems (Knoll et al., 2019). Predicting

and understanding the dynamics of ice thickness (IT) on lakes is scientifically necessary because it can analyze the impact of climate on ice phenology, the lake's physical, chemical, and biological processes, and can also inform decisions regarding ice-related activities, such as transportation, fishing, and recreation, to ensure safety and minimize environmental impacts (Masterson, 2009; Obertegger et al., 2017).

This study also focuses on modelling and understanding the dynamics of IT in lakes and understanding the main influential factors affecting them by machine learning techniques. The relationship between these factors is intricate, non-linear, and time-varying, making it difficult to predict the changes in LSWT and IT accurately. A comprehensive understanding of the complex interaction among the factors is essential to develop accurate models for predicting LSWT and IT, which is crucial for the management of freshwater resources and aquatic ecosystems.

## 1.2. Factors affecting LSWT and IT modeling approaches

Several models have been developed to forecast LSWT and IT, which can be broadly classified into two categories: physically based models and data-driven models. Physically based models rely on fundamental equations to simulate the thermal dynamics of lakes, such as heat transfer, mixing, and stratification processes, without relying on LSWT or IT data. However, empirical relations are usually used for calibration, which require LSWT and IT measurements. In contrast, data-driven models rely on statistical and machine learning (ML) techniques to establish relationships between LSWT or IT, and various meteorological variables. These models utilize LSWT or IT data for calibration and prediction. The application of data-driven models is more straightforward than physically based models, and they can provide predictions with a lesser amount of data about the lake’s morphology. However, data-driven models do not explicitly include the underlying physical processes governing the thermal response of lakes. In contrast, physically based models provide a more comprehensive understanding of the thermal dynamics of lakes that require less amount of data in comparison to data-driven models but more kinds of input data and a higher computational cost. Therefore, choosing the appropriate modeling approach depends on the specific application and the availability of data and computational resources.

Physical models for both LSWT and IT require a variety of input data, including meteorological data, lake characteristics such as bathymetry, as well as hydrological data such as water inflow and outflow, which can influence IT and LSWT. Siegert et al. (2012) developed a physical model for IT that requires detailed meteorological and hydrological data. These models are based on fundamental principles of thermodynamics and fluid dynamics and are thus more interpretable and provide a better understanding of the underlying physical processes. Physical models for LSWT are typically based

on a set of deterministic equations describing the heat budget for a bulk volume (e.g., Ragotzkie, 1978) or for a discretized version (one-, two-, or three-dimensional, e.g., Irving et al., 2006, Hipsey et al., 2019) of the water body. These models require a comprehensive understanding of the physical processes that govern the thermal response of lakes, including heat transfer, mixing, stratification, and advection processes.

Data-driven models for predicting LSWT and IT are characterized by a wide range of stochastic methods, including both linear and non-linear regression models as well as more advanced approaches such as autoregressive, periodic autoregressive (Benyahya et al., 2007), and evolutionary polynomial regression models (Doglioni et al., 2008). These models are developed to establish relationships between the observed output and other influential factors, such as AT (Assel, 1976). These data-driven models require an extensive repository of both input factors and desired outcomes. Presently, acquiring finely detailed data in terms of spatial and temporal resolution is more convenient compared to the past. The use of ML algorithms, such as artificial neural networks (ANN), decision trees (DT), and support vector regression (SVR), has also gained popularity in recent years as a general-purpose computational tool for developing data-driven models (Mohri et al., 2018). These ML algorithms can handle large and complex datasets and are capable of capturing non-linear relationships between variables, which makes them suitable for modeling the highly non-linear dynamics of LSWT and IT. However, the performance of these models depends on the quality and quantity of the available data, as well as the accuracy of the selected input variables and the chosen parameters used to optimize the model.

The reason we use ML models is primarily driven by their exceptional flexibility. ML models can autonomously learn complex patterns and relationships from data. This adaptability allows them to handle various situations and accommodate changing information. ML models enable us to make sense of intricate data patterns and understand the importance of predictors.

## 1.3. List of symbols

The symbols used in this study are presented in Table 1.1. This table serves as a reference, facilitating readers' comprehension of the assorted terms and symbols employed throughout our research.

Table 1.1. List of symbols used in the thesis.

| | |
|---|---|
| AI | Artificial Intelligence |
| ANFIS | Adaptive Neuro-Fuzzy Inference System |
| ANN | Artificial Neural Network |
| AP | Air Pressure |
| AT | Air Temperature |
| BOM | Beginning Of the Month |
| BPNN | Backpropagation Neural Network |

| | |
|---|---|
| CDOY | Cosine of Day Of the Year |
| CWT | Continuous Wavelet Transform |
| DL | Deep Learning |
| DOY | Day Of the Year |
| DT | Decision Tree |
| DWT | Discrete Wavelet Transform |
| ESA | European Space Agency's |
| ERT | Extremely Randomized Tree |
| GA | Genetic Algorithm |
| GAM | Generalized Additive Model |
| GBT | Gradient Boosting Trees |
| GEP | Gene Expression Programming, a variant of Genetic Programming |
| GP | Genetic Programming |
| KNN | K-Nearest Neighbor |
| LSTM | Long Short-Term Memory |
| LSWT | Lake Surface Water Temperature |
| LWR | Longwave Radiation |
| MAMM | Moving Average with Min-Max |
| ML | Machine Learning |
| MLPNN | Multi-Layer Perceptron Neural Network |
| PGRNN | Physics-Guided Recurrent Neural Network |
| PSO | Particle Swarm Optimization |
| R | Rainfall |
| RF | Random Forest |
| RH | Relative Humidity |
| RMSE | Root Mean Square Error |
| RNN | Recurrent Neural Network |
| RT | Regression Tree |
| SCDOY | Sine and Cosine of Day Of the Year |
| SDOY | Sine of Day Of the Year |
| SMHI | Swedish Meteorological and Hydrological Institute |
| SVM | Support Vector Machine |
| SVR | Support Vector Regression |
| SWR | Shortwave Radiation |
| T | Temperature |
| WS | Wind Speed |
| IT | Ice thickness |
| AI | Artificial Intelligence |
| WT | Wavelet Transform |

## 1.4. Thesis outline

The following list highlights and summarizes the main contributions of this thesis and its outline. Each following chapters contain an introduction, literature review, and methodology specific to that section. This comprehensive approach ensured a thorough examination of each aspect of the research topic, including the identification of key research gaps and the development of novel methodology.

### 1.4.1. Chapter 2 – Materials and methods

Chapter 2 presents the datasets and methodology used for the accurate modeling and prediction of LSWT and IT. For LSWT models, two synthetic lakes are used to explore how lake depths impact the analysis. Besides, LSWT data for 2024 lakes from CCI datasets (derived from the satellite

observations for more than 2000 water bodies) are considered to analyze the effect of different climate regions. For IT model, we focused on two lakes in Sweden. In the methodology section, ML methodologies, such as ANN, DT, K-nearest neighbors (KNN), and SVR, to model and predict LSWT and IT are explained. These ML algorithms are capable of detecting complex nonlinear relationships between various input variables, such as AT, solar radiation, wind speed, and the output variables, LSWT and IT. The performance of the ML models is highly dependent on the selection of appropriate hyperparameters (parameters that are determined by users), which are determined by the user and affect the algorithms' learning behavior and prediction accuracy.

In this section, the optimization of hyperparameters for each ML model is explained in detail. This involves selecting the optimal values for hyperparameters. The chosen hyperparameters for each model are elaborated in Appendix A, providing transparency and reproducibility of the modeling process. Each section of the methodology is presented with a thorough elaboration of its respective materials, including preprocessing steps, and model evaluation metrics.

Additionally, this study provides a comprehensive explanation of several crucial aspects related to ML performance evaluation. Notably, the concept of feature ranking (FR) is extensively discussed, wherein each feature is assigned a rating that indicates its relevance in the ML model. The intricate details of this method, including the implementation and significance, are elucidated in this study, facilitating a comprehensive understanding of its applicability and effectiveness in ML-based prediction tasks.

To evaluate the ML performance, we used specific metrics to assess the effectiveness and accuracy of the models and explained them in detail. These metrics serve as quantitative measures to gauge the performance of ML algorithms in solving a given task. They provide insights into how well the models are able to generalize and make predictions based on the available data.

### 1.4.2. Chapter 3 – Selecting machine learning techniques and the features affecting LSWT by using a synthetic lake

Chapter 3 delves into the application of various ML techniques for LSWT prediction and modeling. The evaluation of different preprocessing techniques and ML approaches is discussed in detail, highlighting the strengths of each approach. The memory of the forcing, which refers to the influence of meteorological data from previous days on LSWT, is also examined to provide a more comprehensive understanding of the factors that affect LSWT.

To evaluate the influence of lake depth on LSWT, a synthetic case study is used. We used a physically based model named General Lake Model (GLM; Hipsey et al., 2019) to generate artificial time series

of LSWT data based on the general morphological and climatological features of Lake Caldonazzo (Italy) but considering different depths. The thermal dynamics of the synthetic lake is assumed to reproduce the main characteristics of real-world lakes, but in perfectly controlled conditions, providing a reliable testing ground for the ML models. A literature review is also conducted, summarizing the previous studies that employed ML approaches for LSWT prediction and modeling. This review serves as a foundation for the current study and provides insights into the current state-of-the-art in this field.

A large number of ML techniques are employed in this study: ANN methods including backpropagation neural network (BPNN), multi-layer perceptron neural Network (MLPNN) and long short-term memory (LSTM), DT and its derivatives, KNN, and SVR. The evaluation of the results demonstrates that ANN just slightly outperforms the other ML techniques in terms of prediction accuracy. In addition, a critical analysis of role of preprocessing is developed, considering techniques like normalization and wavelet decomposition to preprocess the input data and improve the performance of the ML models.

### 1.4.3. Chapter 4 – The influence of climate on the lakes' thermal response

In Chapter 4, the developed ML model is applied to CCI dataset (http://cci.esa.int/lakes), obtained from satellite observations consisting of 2024 lakes worldwide. This large dataset allows for a robust evaluation of the model's performance, as well as an exploration of the influence of regional climate differences on the predictors of LSWT. To achieve this, the most influential predictors in each region are identified using a feature importance analysis.

The feature importance analysis is performed to identify the most significant predictors of LSWT, considering the meteorological inputs as well as the day of the year (DOY), which has the greatest impact on LSWT values. The methods used for retrieving the features' importance are explained in detail. This involves ranking the predictors based on their contribution to the model's performance, which is measured using evaluation metrics. The identified predictors can provide insights into the drivers of LSWT variability.

The dataset used in this study consists of meteorological inputs from the ERA5 reanalysis dataset (Hersbach et al., 2020) and LSWT data from the CCI LSWT dataset. The ERA5 dataset provides global meteorological data with high spatial and temporal resolution, making it a valuable resource for LSWT prediction and modeling. The CCI dataset is also a reliable source of LSWT data.

### 1.4.4. Chapter 5 – Ice dynamics in boreal lakes

In Chapter 5, the ML model is applied to predict IT in lakes located in Sweden, northern Europe. We study how the predictors are related by analyzing their correlations. Furthermore, we incorporate the concept of memory of predictors to better understand its impact on the analysis. The feature importance is used to assess the effect of each predictor variable on the IT model and prediction.

The identified predictors can provide insights into the drivers of IT variability. Furthermore, understanding the key drivers of IT variability can help identify potential climate change impacts on ice cover and inform mitigation and adaptation strategies. The dataset used in this study includes meteorological data from the ERA5 reanalysis dataset and IT from Swedish Meteorological and Hydrological Institute (SMHI, https://www.smhi.se).

### 1.4.5. Chapter 6 – Conclusions

In this section, general conclusions are drawn based on the findings of the study. The concluding remarks were developed by synthesizing the results of the different sections and identifying common themes and patterns. The conclusions emphasized the strengths and weaknesses of ML methodologies to model and predict LSWT and IT, as these variables play a critical role in climate change impacts. The study demonstrated the effectiveness of various ML algorithms in predicting LSWT and IT and highlighted the key predictors influencing these variables in different regions.

The study also highlighted the need for reliable data sources, such as meteorological data for accurate predictions and model validation. Additionally, the importance of considering the memory of previous days of the data was emphasized, as it improved the accuracy of predictions.

### 1.4.6. Appendix

At the end of the thesis, the hyperparameters that we used for the ML models are briefly explained and the values and methods for each model are written in order to be able to redo the analysis. Moreover, the review of the papers using machine learning to simulate and forecast water temperature in rivers is written in a table. The metrics used to evaluate the performance of ML in synthetic models are included, together with the hypsographic curves defining the synthetic case studies.

# Chapter 2. Materials and methods

## 2.1. Available data

Various types of datasets were employed based on the distinct sections of the research. The data concerning lake surface water temperature (LSWT) were obtained from a physically based model and the satellite observations, and ice thickness (IT) datasets are measurement data which will be elaborated upon in the subsequent sections.

### 2.1.1. Generation of LSWT for a synthetic lake

To select the features affecting the machine learning (ML) models, we employed a synthetic example, giving us complete control over variables for comparing and evaluating diverse approaches. This constructed instance allows us to examine how lake depth affects LSWT responses, while ensuring consistent meteorological inputs, enabling us to assess the capabilities of ML models under varying circumstances.

The synthetic case study employed the General Lake Model (GLM; Hipsey et al., 2019) to generate LSWT data. While specific model intricacies are of less concern, this approach lets us concentrate on ML model performance without the intricacies of actual data. This artificial scenario enables us to delve into the significance of depth while fixing the other parameters, thereby isolating its impact. The study selected Lake Caldonazzo, a medium-sized peri-alpine lake as a reference, incorporating realistic morphology and climatic conditions.

### 2.1.2. LSWT from the CCI-lakes database

This dataset consists of processed satellite observations covering global inland water bodies from 1992 to 2020. It's a product of the European Space Agency's (ESA) Lakes Climate Change Initiative (Lakesa_cci; http://cci.esa.int/lakes) project, encompassing over 2000 water bodies. The dataset includes the Lakes Essential Climate Variable with six interconnected parameters describing the lake's physical condition, including, lake surface temperature, water level and extent, ice cover and thickness, and lake water leaving reflectance which LSWT derived from the satellite observations. LSWT data come with associated per-pixel quality levels, indicating an evaluation of the accuracy of the data and its estimated level of uncertainty. The validation process for LSWT relies on comparing satellite data with in-situ measurements, gathered through direct communication with limnologists and organizations willing to provide data for validation objectives (Carrea et al., 2022).

### 2.1.3. Ice thickness for Swedish lakes

Two lakes located in Sweden were selected as the focal points for our case studies for IT model. The data for these investigations were sourced from the database maintained by the Swedish Meteorological and Hydrological Institute (SMHI). Notably, SMHI engages in the comprehensive measurement and documentation of various hydrological parameters, including water discharges, lake levels, IT, and the temporal evolution of ice freeze-up and break-up events. The lakes we have chosen are Lake Runn (in southern part of Sweden) and Lake Gouta (in the northern part of Sweden) because both had larger amount of IT data.

### 2.1.4. Retrieval of meteorological information from ERA5

The meteorological variables were acquired from ERA5 (Hersbach et al., 2020), which is the fifth generation of the ECMWF reanalysis of global climate and weather. ECMWF introduced an extension of ERA5, encompassing an additional 21 years of global atmospheric reanalysis data, effectively extending the dataset back to 1979, making it an increasingly comprehensive tool for climate research and analysis. ERA5 is accessible at https://cds.climate.copernicus.eu/cdsapp#!/dataset/derived-near-surface-meteorological-variables. The dataset encompasses a range of meteorological parameters extracted based on the latitude and longitude of the lakes. These parameters include air temperature (K) 2 meters above the lake surface, wind speed (m/s), air pressure (Pa), rainfall (kg/m²/s), specific humidity (kg/kg), and downward longwave and shortwave radiations (W/m²). The values were originally recorded on an hourly basis, but we converted them into daily mean values.

### 2.1.5. Koppen's climate classification

The Köppen climate classification classifies climates into five primary categories, and each of these categories is further segmented based on distinctive patterns of temperature and precipitation throughout the seasons. These major divisions are designated as tropical, dry, temperate, snow and polar (Köppen, 1918). We classified the performance (explained in section 2.3) of our model based on Köppen regions to evaluate the effect of climate on LSWT (Chapter 4).

## 2.2. Machine learning algorithms

Machine learning (ML) is a subfield of artificial intelligence (AI) that involves the development of algorithms that can automatically learn patterns and relationships in data, without being explicitly programmed. ML algorithms utilize statistical and mathematical techniques to analyze data and identify patterns that can be used to make predictions or decisions. There are several types of ML

algorithms, including supervised learning, unsupervised learning, and reinforcement learning, which are applied depending on the nature of the problem and the available data. ML algorithms have been extensively applied in diverse fields, such as natural language processing, image recognition, and predictive modeling, to provide solutions to complex problems. In this work, we used supervised ML approaches because of having the observed values to evaluate the model. For the sake of ensuring the reliability and consistency of the outcomes, it is recommended to execute the model multiple times, given that the initial weights and biases are randomly assigned in each iteration. In this research, we carried out the model running procedure 20 times and then computed the mean of the results obtained. In this chapter, we review the main features of the ML algorithms that we analyzed in this study.

The ML algorithms listed below are applied to forecast LSWT as a comparative analysis in Chapter 3. In Chapter 4, where the dataset involves 2024 lakes and we aim for efficient performance with the sparse observed LSWT data, we opt for the Backpropagation Neural Network (BPNN). In Chapter 5, considering the outcomes of Chapter 3 and having two distinct case studies, the Long Short-Term Memory (LSTM) model exhibits relatively superior performance, especially for IT prediction where memory of previous days is crucial. Despite taking more time to run compared to BPNN, LSTM proves more robust in IT prediction, providing an additional reason for its selection.

### 2.2.1. Artificial Neural Network (ANN) and derived algorithms

ANN algorithms are intricate systems that consist of interconnected processors and follow organizational principles resembling the recognition of patterns, prediction, and optimization of weights and biases (Jain et al., 1996). The algorithm of the network comprises interconnected nodes, referred to as neurons, arranged in parallel layers (Figure 2.1). The initial layer, known as the input layer, receives input data, which then propagates through a variable number of intermediate layers, called hidden layers, until it reaches the final layer, known as the output layer. The output layer provides the simulation results. The links between nodes in adjacent layers transfer signals from upstream neurons to downstream neurons. Within each hidden and input layer, the neurons are provided with an additional parameter known as biases. The signal emitted by a neuron is appropriately weighted before being transmitted to the subsequent neuron. It is then summed with signals from other neurons and the bias of the same layer. This sum is then processed by an activation function, which produces the output signal that is transmitted downstream.

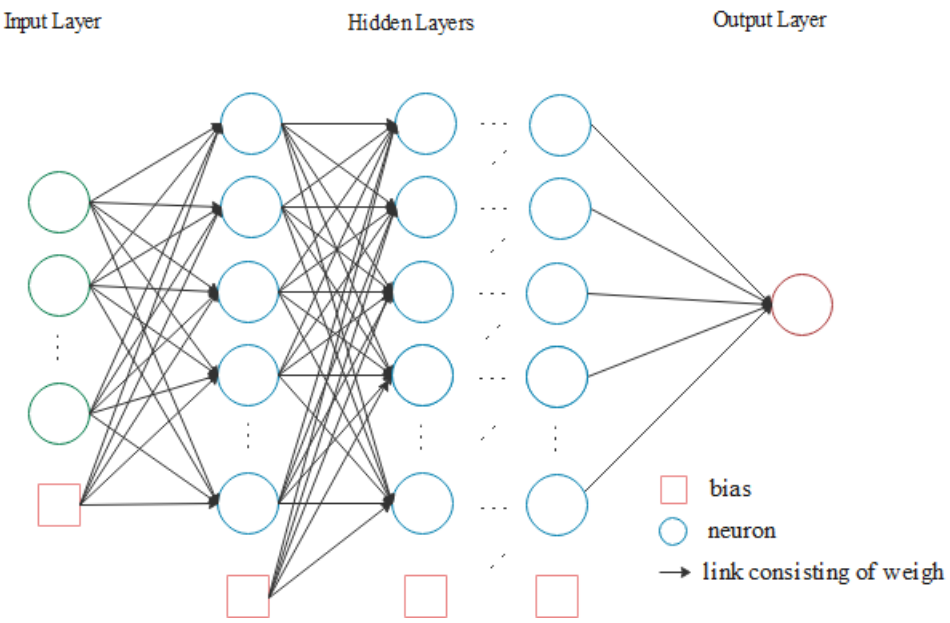



Figure 2.1. The general scheme of an Artificial Neural Network (ANN).

ANN involves determining the number of hidden layers and the number of neurons within each layer, as well as selecting the activation function type and the optimization parameters for weight and bias calibration. The optimization procedure employed defines the specific approach within the ANN framework. Typically, the optimization process commences with the assignment of random values to the weights and biases. During the initial run, the ANN generates an output that is then compared to the desired target. Subsequently, the weights and biases are updated. This iterative process continues until the discrepancy between the output and the target is minimized.

Within our analyses, we included three distinct ANN types: the multi-layer perceptron neural network (MLPNN), the backpropagation neural network (BPNN), and the Long Short-Term Memory (LSTM), which is under the category of recurrent neural networks (RNN). Additionally, we considered the adaptative neuro-fuzzy inference system (ANFIS), which comprises a single layer of neural network. Subsequent sections of this study delve into the further examination of these methodologies.

The MLPNN is an ANN architecture characterized by the presence of multiple hidden layers. In the context of this study, we employ the term "MLPNN" specifically to describe a feed-forward neural network. The notable characteristic of the MLPNN is its weight updating process, which is based on the error propagated from the input layer to the output layer in a unidirectional manner.

The BPNN approach is distinguished by its weight and bias update algorithm, which aims to minimize errors (Li et al., 2012). Initially, the value of each neuron is computed using the initial weights. Subsequently, the weights are adjusted in a backward direction, starting from the output layer and moving towards the input layer. This adjustment is guided by the understanding of how much each

node contributes to the prediction error, considering observed data (Hecht-Nielsen, 1992). Unlike in the MLPNN, where weights are updated solely based on the error and move in a forward direction (Dudek, 2019), in BPNN, the weights are updated in a backward manner from the output layer towards the input layer.

LSTM method belongs to the category of RNN, a subset of ANN that exhibits sequential connections between time steps (Sainath et al., 2015). Compared to standard RNNs, LSTM has the capability to retain information for extended durations. Within the hidden layers of LSTM, specialized units called memory blocks are employed. These memory blocks consist of memory cells responsible for storing and regulating the flow of information through three gates (Figure 2.2). These gates are the forget gate, responsible for deciding whether to retain or discard information; the input gate, which allows the addition of new information to the cell; and the output gate, which outputs the pertinent information. The hidden state, which encapsulates the neural network's memory, preserves the prior information (Gonzalez-Dominguez et al., 2014).

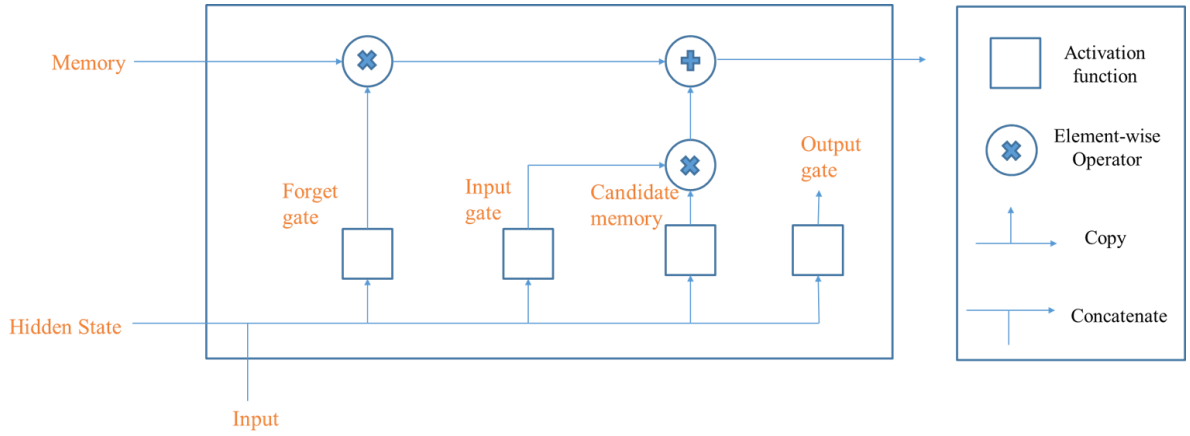



Figure 2.2. The scheme of long short-term memory (LSTM), a branch of ANN.

ANN is a versatile modeling approach that effectively captures intricate, nonlinear relationships among dependent and independent variables, as well as predictor variables. It utilizes a diverse training algorithm, allowing for comprehensive analysis. Liu and Chen (2012) highlighted the cost-effectiveness of ANN compared to the three-dimensional semi-implicit Eulerian-Lagrangian finite-element method (Zhang and Baptisa, 2008), making it suitable for water quality prediction and ecological management. Despite these advantages, the ANN is regarded as a black box model, necessitating significant computational resources and being prone to overfitting (Tu, 1996). To address overfitting concerns, dropout techniques have been proposed (Piotrowski et al., 2020; Zhu and Piotrowski, 2020), whereby nodes and their corresponding connections are randomly dropped during the training phase (Srivastava et al., 2014).

### 2.2.2. The Adaptative Neuro-Fuzzy Inference System (ANFIS)

The ANFIS is a specific subcategory of ANN that is based on the Takagi-Sugeno-type fuzzy inference system (Anikin and Zinoviev, 2015). ANFIS utilizes fuzzy logic principles to train a model set and calculate the parameters of the membership function, which represents the degree of truth in fuzzy logic (Opeyemi and Justice, 2012). There are three widely employed ANFIS methods: (i) ANFIS with grid partition method known as ANFIS_G, (ii) ANFIS with subtractive clustering (SC) referred to as ANFIS_S, and (iii) ANFIS with fuzzy c-means clustering (FCM) called ANFIS_F. The primary difference among these algorithms lies in the optimization of the fuzzy rules. ANFIS_G requires a larger number of fuzzy rules, while ANFIS_S and ANFIS_F aim to provide a model with fewer fuzzy rules. For instance, ANFIS_G cannot be used with more than six input variables. In our study, we employed ANFIS_G, taking into consideration its rules and limitations. Based on a first order Sugeno model, the fuzzy logic involves two if-then rules. The ANFIS network comprises five layers (Figure 2.3). Layer 1, known as the fuzzification layer, consists of adaptive nodes that determine the degree of fuzzy membership of the inputs. In our study, the triangular membership function yields more accurate results compared to other functions. Layer 2 involves the multiplication of connected signals ($w_i$), and the output of this node represents the firing strength of the rule. In layer 3, node N normalizes the firing strength ($\overline{w_i}$). Layer 4 computes the output of the adaptive nodes as the product of the previous layer's results and a first-order polynomial ($f_i$) according to the first order Sugeno model. Finally, in layer 5, the summation of all connected signals is calculated (Opeyemi and Justice, 2012).

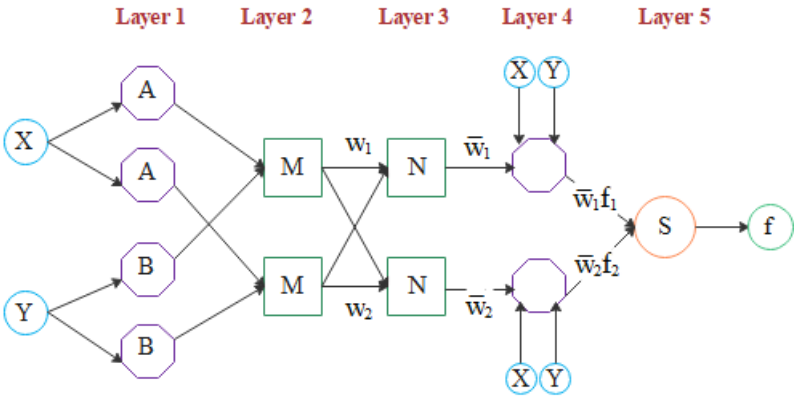



Figure 2.3. The scheme of Adaptative Neuro-Fuzzy Inference System (ANFIS), a branch of ANN that exploits fuzzy logic (adapted from Opeyemi and Justice, 2012).

### 2.2.3. Decision Tree (DT) and derived algorithms

DT are utilized in classification or regression scenarios to discern and forecast by means of partitioning, thereby enabling the identification of features and extraction of patterns. DTs are

commonly employed in prediction models and exploratory data analysis (Myles et al., 2004). As illustrated in Figure 2.4, internal nodes and root nodes traverse the tree structure to obtain accurate outputs, referred to as leaf nodes, which represent categories or patterns (Navada et al., 2011). In our investigation, we assessed three approaches: (1) the standard DT method, along with two derived algorithms, namely (2) extremely randomized tree (ERT) and (3) random forest (RF).

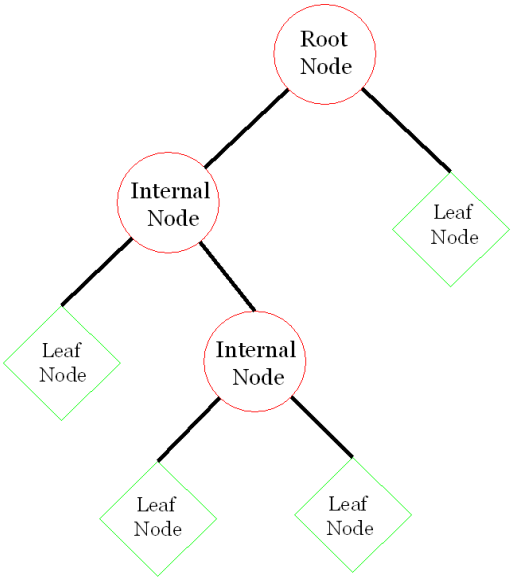



Figure 2.4. The scheme of Decision Tree (DT).

The conventional DT follows a top-down approach, starting from the root and extending towards the leaf nodes, where no further splitting occurs (Figure 2.4). The root and internal nodes are divided into branches that terminate at child nodes. In regression scenarios, the attribute or feature used for splitting aims to minimize the variance at each split by considering the weighted average variance of the child nodes (Sharma, 2020).

The ERT method adopts a randomized approach by randomly selecting features and cutting points (threshold values) for splitting (Geurts et al., 2006). It undergoes repetitive training processes, constructing a model comprising multiple trees utilizing the entire dataset (Alswaina and Elleithy, 2018; Heddam et al., 2020). ERT exhibits superiority over conventional DT due to its ability to prevent overfitting, efficient computational performance, random feature selection and cut point values, and improved prediction accuracy (Galelli and Castelletti, 2013). In the context of predicting water temperature in shallow lakes, Heddam et al. (2020) discovered that ERT demonstrates greater robustness compared to ANN and RF.

RF is a versatile approach that encompasses both classification and regression tree groups. It operates by randomly selecting features and training them using bootstrap sampling, which involves random samples with replacement (Figure 2.5). The final decision of the RF is determined by the majority consensus among the individual trees. RF offers several advantages, including accurate

prediction, the integration of favorable features, built-in performance evaluation, assessment of feature importance and relative weights based on compositional similarity (Svetnik et al., 2003). To reliably estimate the significance of features, a large number of trees should be utilized. Although the measurement of feature significance may vary in each run, the relative importance rankings of the features remain relatively stable (Liaw and Wiener, 2002).

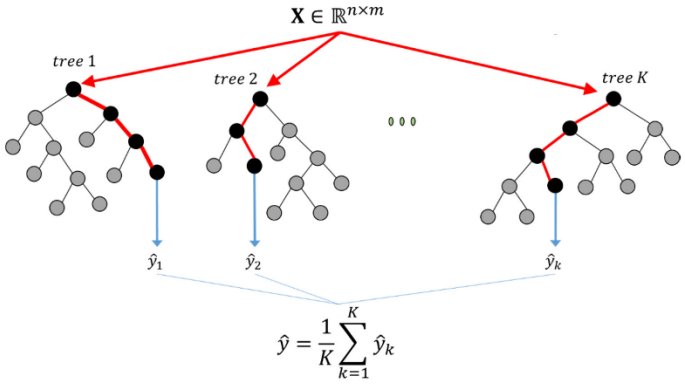



Figure 2.5. The scheme of Random Forest (RF) (Aldrich, 2020).

### 2.2.4. Support Vector Machines (SVM)

SVM is a machine learning approach used for classification and regression analysis, where the regression variant is known as support vector regression (SVR). SVM offers several advantages, including good performance with a small number of data points, low computational requirements, and robustness in the presence of outliers in the model (Steinwart and Christmann, 2008). Figure 2.6 illustrates a classifier that utilizes hyperplane division, which needs to be optimized not only for the training data but also for any other test data (Gunn, 1998). In an n-dimensional Euclidean space, a hyperplane is a flat subset with a dimension of n-1, effectively dividing the space into two separate parts (Veling et al., 2019). SVR aims to minimize the differences (offsets) between the features and the hyperplane, while SVM seeks to maximize the separation distances between the features (Quan et al., 2020). In Figure 2.6, the circles and triangles represent the features that are divided by the hyperplane with the best margin. Since SVR only relies on a subset of the training data and is independent of the input space dimensions, it is advantageous to use it for high-dimensional input spaces (Drucker et al., 1997). The hyperparameters of the kernel function, which maps the nonlinearity in regression algorithms to optimize the hyperplane's location, and the penalty can be fine-tuned using optimization algorithms such as genetic algorithm (GA) to achieve more accurate results (Quan et al., 2020).

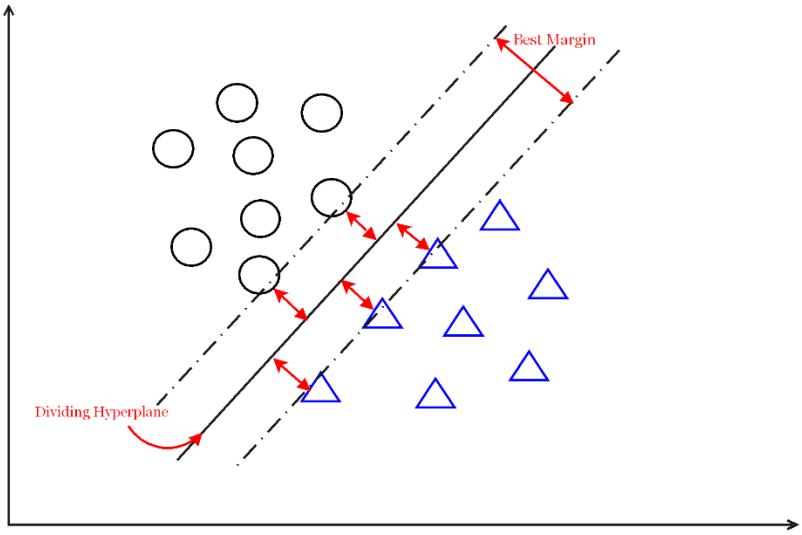



Figure 2.6. The scheme of support vector machines (SVM). The circles and triangles are the features that need to be divided by a hyperplane with the best margin.

### 2.2.5. K-Nearest Neighbor (KNN)

The KNN algorithm is employed when there is limited prior knowledge about the data distribution (Peterson, 2009). Classification in KNN involves assessing the similarity or distance between samples (Keller et al., 1985). By identifying the nearest neighbors among the training data (as shown in Figure 2.7), candidates for a particular category are evaluated based on their similarities to the categories of the K nearest neighbors (each category is represented by black circles in Figure 2.7). The category with the highest score corresponds to the category assigned to the sample (Jiang et al., 2012).

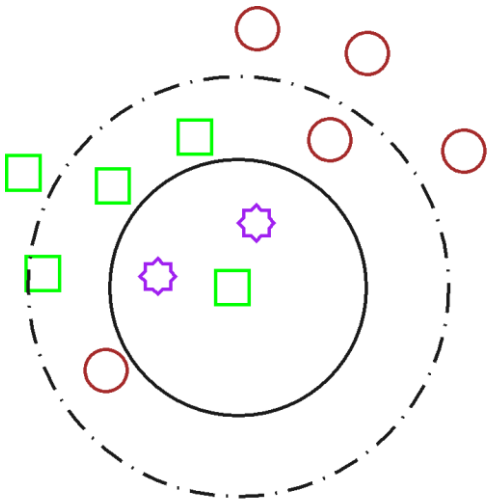


Figure 2.7. The scheme of KNN scheme. Small circles, squares and stars represent different categories.

## 2.3. Metrics used to evaluate ML performance

Our analysis of the ML performance is based on the values of the root mean square error (RMSE) between the simulated and observed target variable, and due to the specific requirements of each project, we found it necessary to utilize additional metrics. RMSE is computed for both the training and test data using the standard definition.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(sim_i - obs_i)^2} \tag{2.1}$$

Where $i$ is the time index, $N$ is the number of samples, $sim_i$ (simulated) is the variable (either LSWT or IT) modelled by ML, and $obs_i$ is the observed value (obtained from GLM simulation). Perfect fit is obtained for RMSE = 0 (°C or cm, depending on the variable).

The Nash-Sutcliff efficiency index (NSE) accounts for the variability of the observations to rescale the RMSE:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(sim_i - obs_i)^2}{\sum_{i=1}^{N}(obs_i - \overline{obs})^2} = 1 - \frac{RMSE^2}{\sigma_{obs}^2} \tag{2.2}$$

where $\overline{obs}$ is the mean of the observations. NSE = 1 indicates perfect fit, while using the mean of the observations as a predictive model would lead to NSE = 0. The information is equivalent to that obtained by RMSE if the variance of the observations, $\sigma_{obs}^2$, does not change.

Considering the distinctive seasonal patterns inherent in the LSWT series, we opted for a modified definition of the NSE index (Schaefli and Gupta, 2007; Piccolroaz et al., 2016). In this adaptation, we employed the inter-annual mean value corresponding to each day of the year, commonly known as the mean year, as a benchmark model. This departure from using the overall mean of the target series allows for a more accurate assessment of the model's performance within the context of the specific seasonality observed in the data. In order to delve into the model's efficiency in capturing these non-seasonal dynamics, it is advisable to utilize a modified version of NSE. The Nash-Sutcliffe Efficiency with inter-annual mean year.

$$NSE^* = 1 - \frac{RMSE^2}{\frac{1}{n}\sum_{i=1}^{n}(obs_i - \overline{obs}_w)^2} \tag{2.3}$$

where $\overline{obs}_w$ describes the annual variability of the mean year which is predicted LSWT by ML model using only sine and cosine of day of the year (DOY) as input (presented in Figure 2.8). This metric

provides valuable insights regarding the model's ability to account for variations that are not encompassed within the seasonal component.

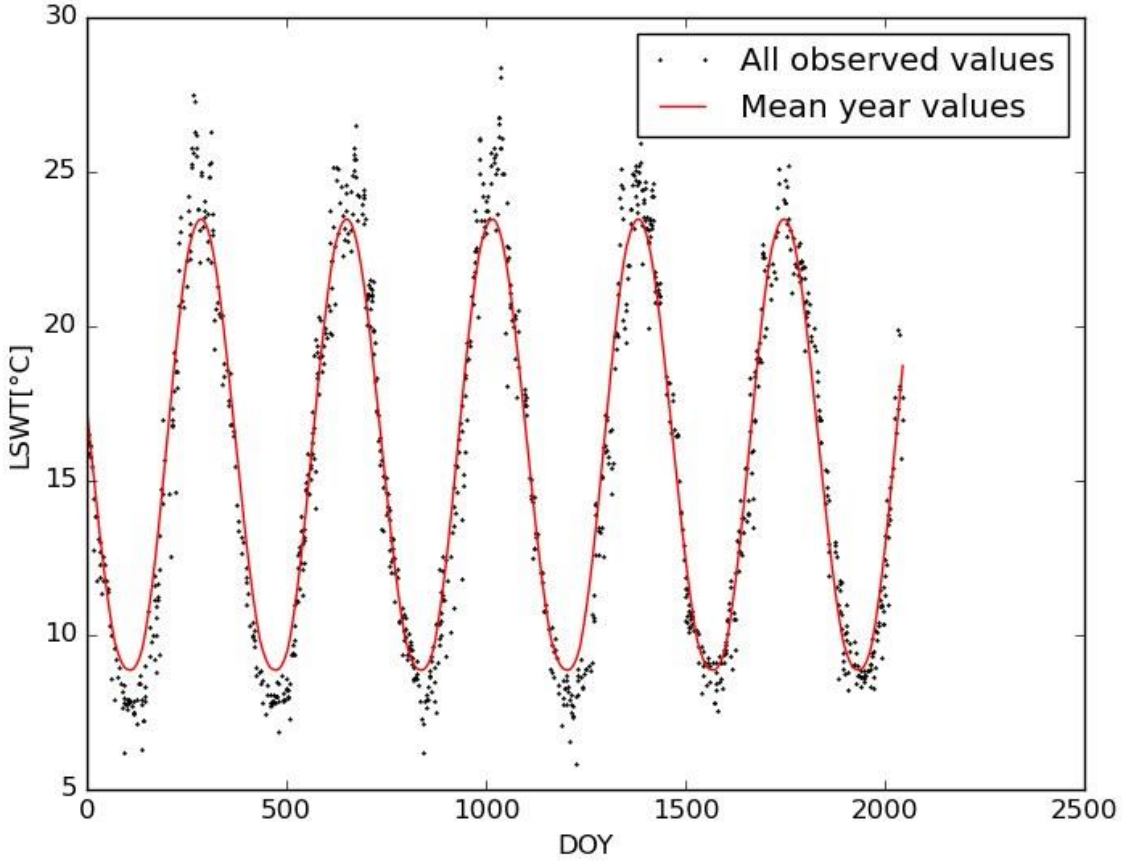



Figure 2.8. Mean year value obtained by ML model, using sine and cosine of DOY as input for Lake Garda.

Notably, an NSE* value of 1 serves as an indication of an impeccable match between the model and the data (equivalent to NSE=1), signifying a remarkable level of accuracy in capturing the underlying dynamics. Conversely, an NSE* value of 0 suggests that the model's performance is comparable to assuming the mean year of measurements as a predictor. This implies that the model fails to exhibit any superior predictive capability beyond a simplistic approach based on the inter-annual day-to-day average of the observed values. This metric is employed in Chapter 4 to gain a more profound understanding of the influence of meteorological variables on LSWT prediction, in addition to those obtained using only the Day of the Year (DOY).

## 2.4. Averaging

Lakes display a filtering tendency when subjected to external influences, owing to the heat capacity of water mass. To effectively capture this phenomenon, we introduced a temporal averaging feature during the pre-processing stage. By strategically considering various window sizes, we pinpointed

the optimal temporal averaging duration. This selection was based on its profound impact on our analysis, allowing us to capture the memory of previous days. Due to this, we explored three different averaging methods to retain the memory of meteorological variables from previous days. In all the analyses that we performed; we used a daily time step.

A backward weighted average can be defined, in general, as

$$\overline{X_t} = \frac{\sum_{i=0}^{N_{ave}} \gamma_i X_{t-i}}{\sum_{i=0}^{N_{ave}} \gamma_i} \tag{2.4}$$

where $\overline{X_t}$ represents the average of the variable at time $t$, $N_{ave}$ denotes the number of days defining the averaging window, the subscript $i$ refers to the day number (with $i = 1$ the day before $t$ and $i = N_{ave}$ the earliest and furthest day in the time window), so that $X_{t-i}$ represents the value of the variable on the $i^{\text{th}}$ day before time $t$, and $\gamma_i$ is the weight.

The first method employs a straightforward averaging technique by assuming 'uniform' weights

$$\gamma_i = 1 \tag{2.5}$$

and corresponds to the standard average.

The second method adopts weights that are inversely proportional to the number of days before the current time,

$$\gamma_i = \frac{1}{i} \tag{2.6}$$

so, we term it as 'hyperbolic' average.

The third method assigns an exponential decay to the weights,

$$\gamma_i = \exp\left(-\frac{i}{T_e}\right) \tag{2.7}$$

where $T_e$ is the *e*-folding time scale; this method will be termed 'exponential'. In order to have a simple comparison with the previous approaches, we fixed $T_e = N_{ave}/3$, so that the furthest value (the earliest day in the averaging window) has a weight $\gamma_i = \exp(-3) \cong 0.05$.

Both in the second and third method, the use of variable weights allows us to account for the historical progression of the forcing, but with greater emphasis placed on more recent values. The main difference is that the hyperbolic weights are characterized by a fixed decay rate with the temporal distance, while the use of a small weight for the furthest value in the exponential average determines

a smooth transition between values inside and outside the averaging window, whose extension ($N_{ave}$) however determines the decay rate of the weights.

We investigated the effect of using a weighted average of the forcing variables. This approach enabled us to consider the influence of past forcing events on the present system state, leading to a more comprehensive understanding of the dynamics involved. We conducted a comprehensive analysis for the two types of datasets considered in this work (synthetic lakes and lakes belonging to the CCI database) by examining various window sizes, with $N_{ave}$ ranging from 0 to 100 days, to obtain the average of all variables. The results are presented in Figures 2.9, 2.10 and 2.11, showing the RMSE obtained by the BPNN model applied to reconstruct the LSWT. We note that this analysis anticipates the results discussed in chapters 3, 4, and 5, because of the need to introduce a general framework for the use of weighted averages. More details about the ML application are provided in the respective chapters.

For the synthetic lake (see Chapter 3), as depicted in Figure 2.9, the hyperbolic averaging demonstrated superior performance (lower errors) compared to the two other methods, which are however satisfactory. Selecting the optimal window size for uniform and exponential averaging is crucial to decrese the error efficiently. Based on our findings, approximately 8 days for the shallow (5 m depth) lake and 15 days for the deep (60 m depth) lake proved to be the most suitable window size choices. A more extensive analysis on the best averaging window in the case of the uniform averaging is developed in Chapter 3 for AT. In such analysis, we did not apply all the methodologies, but only the uniform averaging, to determine the window size across various depths, obtaining a shorter window size for shallower lakes. Anyway, further analyses (not shown) suggest that the qualitative behaviour is consistent among the different approaches.

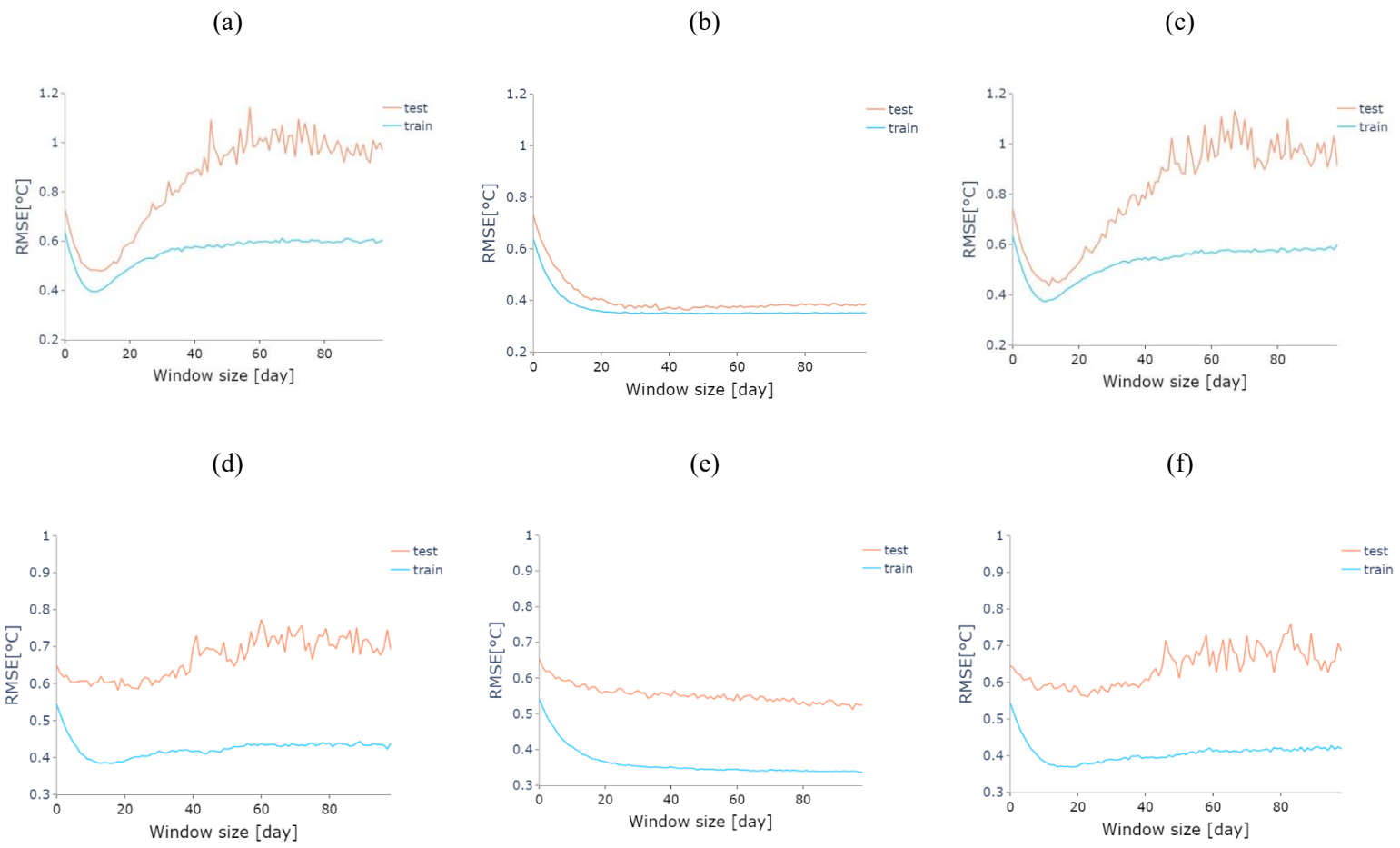
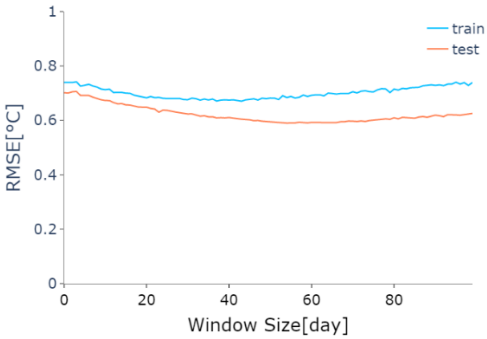


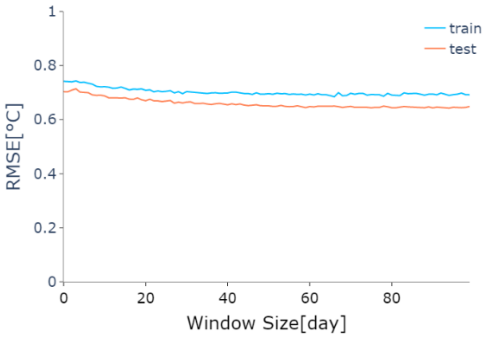

Figure 2.9. The results of using averaged meteorological data in synthetic lakes simulated by BPNN where the results refer to: (a, d) uniform averaging, (b, e) hyperbolic averaging, and (c, f) exponential averaging, for lakes with depth 5 m and 60 m, respectively. The window size is $N_{ave}$, used in equation (2.4).

Within the CCI dataset (see Chapter 4), we selected representative lakes from distinct Köppen regions: Lake Victoria (maximum depth of 83 m) for the tropical region, Dead Sea (maximum depth of 306 m) and Lake Chad (maximum depth of 2 m) as a shallow lake for the dry region, Lake Garda (maximum depth of 346 m) for the temperate region, Lake Erie (maximum depth of 64 m) for the snow region, and Lake Zhari Namco (maximum depth of 71.55 m) for the polar region. Figure 2.9 illustrates the results, indicating that uniform and exponential averaging consistently performs best based on higher performances (explained in section 2.3), even with a window size of up to 100 days. Moreover, when examining shallower lakes such as Lakes Victoria, Chad, and Erie, we observe that the minimum values for both uniform and exponential averaging align with our expectations, reaching to minimum error, especially when considering the training set. At approximately a window size of 40, we attain the lowest errors for these instances, and the optimal window size is notably smaller when compared to deeper lakes. It worths noting that Lake Zhari Namco (Figure 2.10, p, q, r) exhibits more noises compared to the other lakes, possibly due to a smaller available dataset in comparison to the other lakes. Different from Chapter 3, in the analysis of Chapter 4 we employed exponential averaging to define the best window size for each lake but, as already discussed, the results are consistent with all methods.

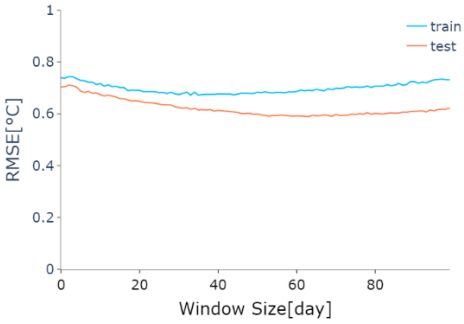
(a)



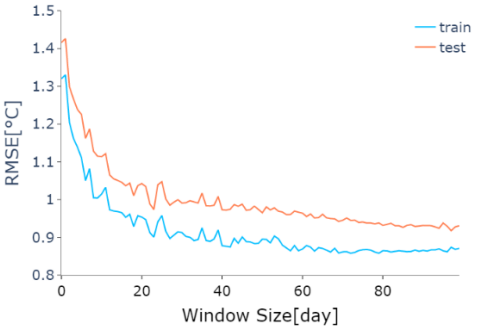

(b)



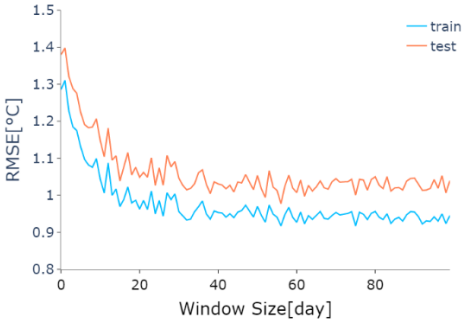

(c)



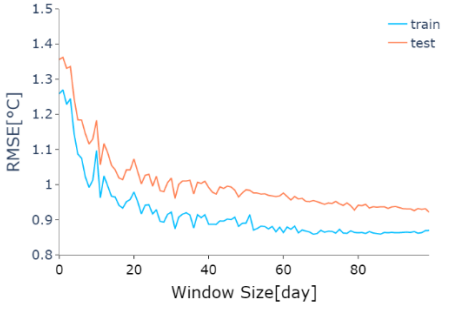

(d)



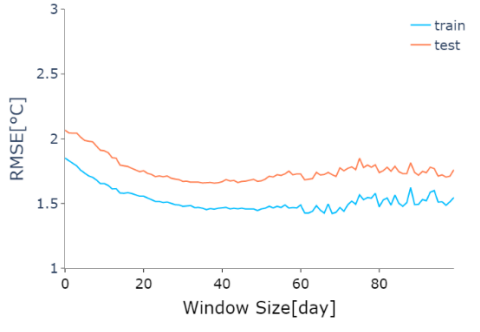

(e)



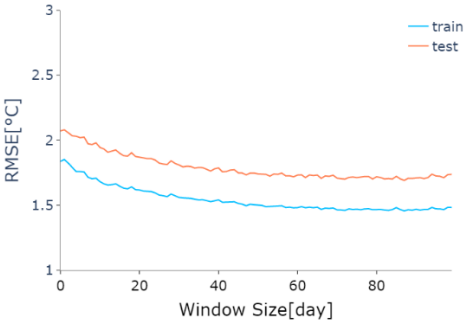

(f)



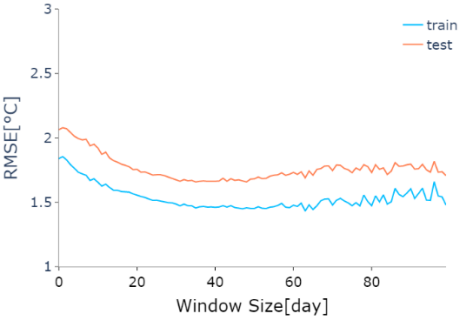

(g)




(h)




(i)




(j)




(k)




(l)

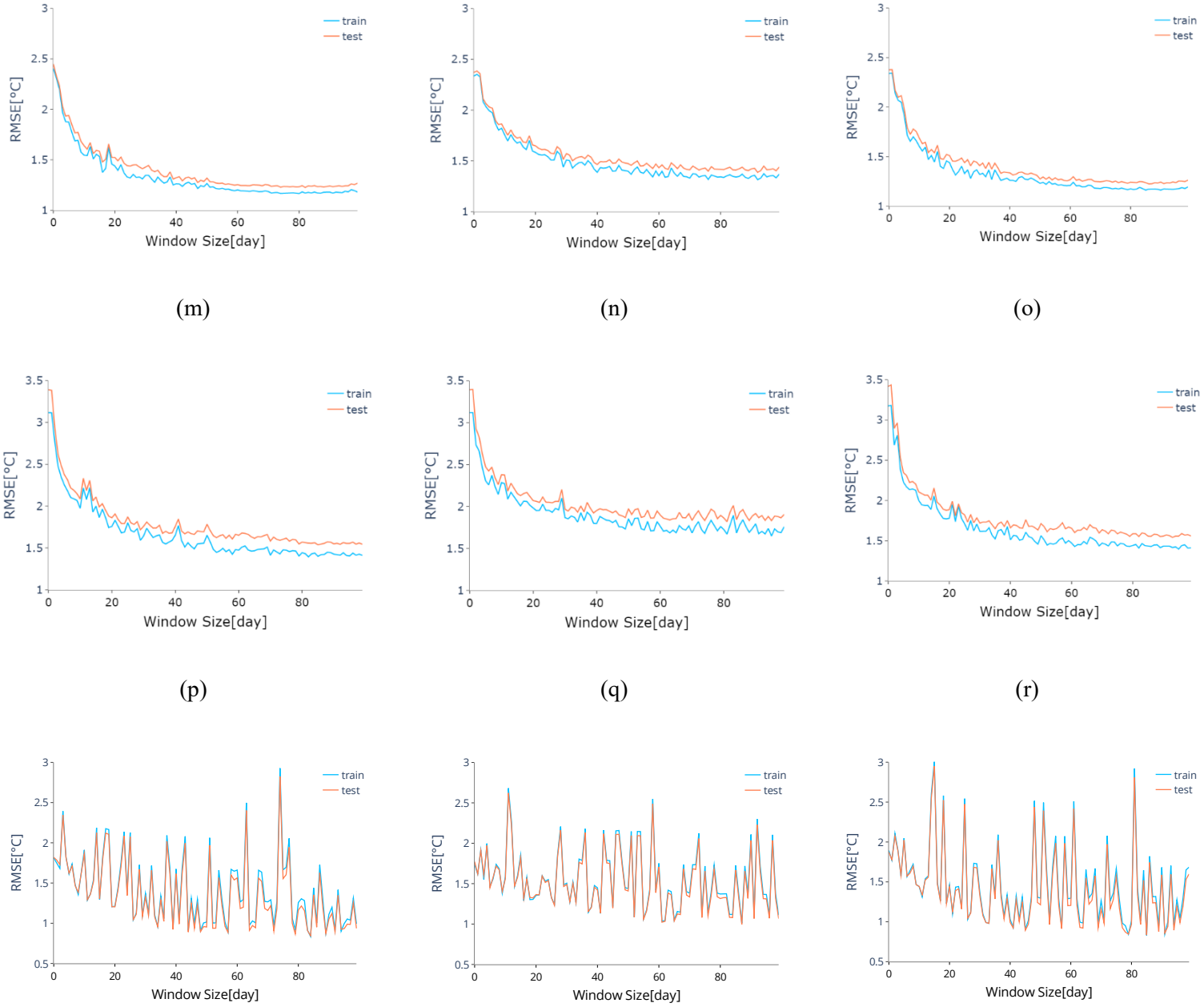



Figure 2.10. The results of using averaged meteorological data in CCI lakes simulated by BPNN; where the first column from left is according to the uniform averaging, the second column for hyperbolic averaging and third column for exponential averaging. The results refer to: (a, b and c) Lake Victoria in tropical region, (d, e and f) Dead Sea and (g, h and i) Lake Chad in dry region, (j, k and l) for Lake Garda in temperate region, (m, n and o) for Lake Erie in snow region and (p, q and r) for Lake Zhari Namco in polar region. The window size is $N_{ave}$, used in equation (2.4).

In the case of ice-covered lakes (Chapter 5), Lakes Runn and Gouta are are the objects of the analysis, and they are both deep lakes (32 m and 58 m, accordingly). As shown in Figure 2.11, the uniform and exponential averaging seem to more clearly select an optimal value of the time window. Accordingly, in Chapter 5, we implemented exponential averaging to define the optimal window size.

Generally, the results of this section suggests that there is not a best averaging method for all cases, but the uniform and exponential methods hold a notable advantage and are consequently recommended for averaging the meteorological variables.

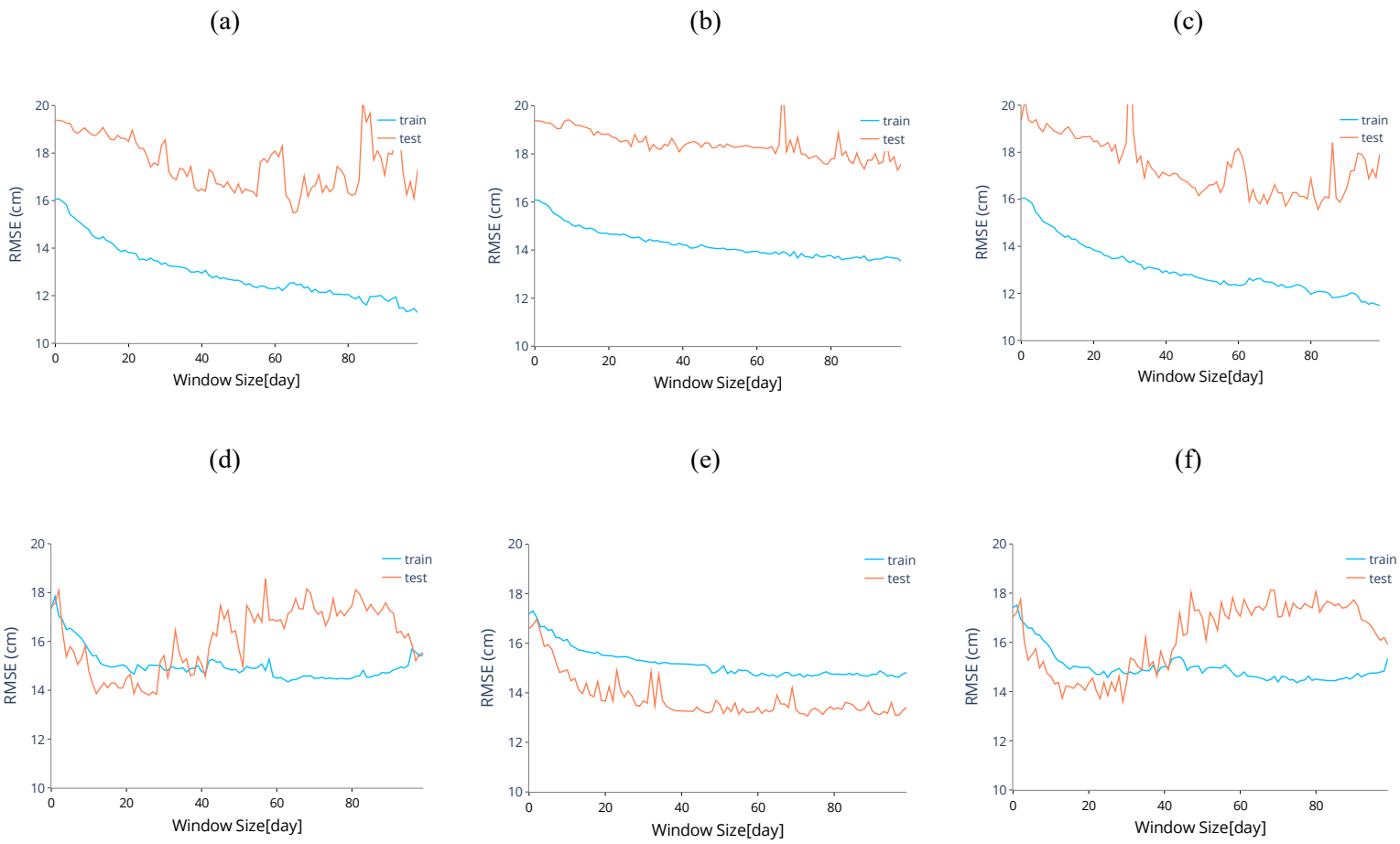



Figure 2.11. The results of using averaged meteorological data in ice covered lakes simulated by BPNN where the results refer to: (a, d) the uniform averaging, (b, e) the hyperbolic averaging, and (c, f) the exponential averaging for Lake Runn and Lake Gouta, respectively. The window size is $N_{ave}$, used in equation (2.4).

## 2.5. Training and test sets selection

Splitting the dataset into training and test sets is an essential step in ML, as it allows the model's performance to be evaluated on data that the model has not seen before. There are several ways to split the dataset, each with its own advantages and limitations. Cross-validation is a resampling method that divides the dataset into multiple partitions, creating various combinations of training and test sets to comprehensively assess model performance. Its advantage lies in maximizing the utilization of all available data for both testing and training, ensuring diversity and adequacy in evaluation. There are several options for cross-validation. In random cross-validation, data is randomly partitioned into training and test sets in each assessment iteration. Although by randomly selecting data points for the training and test sets this method helps ensure that both sets are representative of the overall dataset, random cross-validation faces challenges when dealing with

autoregressive models. In fact, in these cases the same values are used for both input and reference data, compromising the independence of training and test sets. Last block evaluation is another option, particularly useful for time-ordered data where the model is trained on the first part of the time series and tested on the last part. The drawback of last block evaluation is that it does not maximize the utilization of available data, as the last fold (block) is excluded from the model training process. This limitation becomes apparent when only one forecast per time series can be calculated. In such cases, the forecast might reflect characteristics specific to the test set rather than capturing the broader patterns present in the entire dataset (Bergmeir & Benítez, 2012).Another option is the k-fold cross validation method (Anguita et al., 2012), which is one of the successful ways that help to choose the best set for training and test in a non-random manner, thus being appropriate also for autoregressive models. This method involves partitioning the available data into k equally sized subsets or folds, where each fold is used once as the validation data while the remaining k-1 folds are used for training. This process is repeated k times, with each fold being used as the validation data once. This allows for a more accurate assessment of the model's performance, as it is evaluated on multiple test sets rather than a single test set as in the last block evaluation. In addition, the k-fold approach helps to mitigate issues such as overfitting and biased training, as it provides a more representative sample of the data for training and validation. The choice of k is typically based on the size of the available data and computational resources, with a larger k resulting in a more computationally intensive process but potentially providing a more accurate assessment of model performance (Yadav and Shukla, 2016).

Regarding the IT model (Chapter 5), we opted for the k-fold cross validation technique to determine the optimal divisions for test and training sets. Across all the models, we partitioned our dataset into five folds because this configuration yielded more robust results for both simulation (using the training set) and prediction (with the test set) compared to other fold numbers. We initially split the entire dataset into five folds, including the missed observed IT. Subsequently, we provided the model with the training set, including the days for which we had IT data, and conducted predictions for all days within both the training and test sets.

In the case of the LSWT models, we used last fold as the test set (last block approach). In the CCI dataset (Chapter 4), the distribution of the data was uneven, with more frequent data in the final fold. Consequently, the final fold is selected as the test set, while the remaining dataset was used as the training set. Additionally, we observed a more efficient model using the last block approach compared to other folds as the test set, avoiding overfitting. For the synthetic lakes (Chapter 3), as our primary focus was on assessing ML techniques, we employed a straightforward data splitting approach.

Although this method could not assess other folds as the test set, it produced satisfactory results without encountering the issue of overfitting.

## 2.6. Optimizing the hyperparameters

The learning process of a neural network involves various hyperparameters (parameters determined by the user), such as the number of hidden layers, the number of nodes in each layer, the choice of activation function, optimizer, learning rate, and others. Detailed information of hyperparameters for each ML technique is explained in Appendix A. Optimizing these hyperparameters is essential to achieve more accurate forecasts. In this study, we have selected the GA as an optimization method to approximate the best set of hyperparameters. GA is a powerful optimization technique capable of handling both discrete and continuous problems with multiple objectives. Its robustness in finding solutions makes it suitable for avoiding local optima.

GA is an adaptive algorithm originally proposed by Holland et al. (1992), inspired by evolutionary theory, where stronger genes have a higher likelihood of being preserved in subsequent generations. It begins with a set of random solutions, evaluate their fitness using a predefined function, and favor the selection of superior solutions for reproduction. New solutions are created through crossover and mutation, augmenting the population while replacing weaker solutions. This process continues until a specific termination condition is met.

The optimization's goal is to minimize the RMSE for both the training and test sets. Additionally, we provide a range for the hyperparameters to expedite and streamline the optimization process. We employed the GA method to optimize the hyperparameters of the models in Chapter 3, ensuring a fair comparison of the ML methods. In Chapter 4, optimization was initially conducted in an independent way for all lakes; then, the most frequently occurring values or methods for hyperparameters were chosen to create a model with identical hyperparameters to be used for all cases. Chapter 5 involved not only optimization for each lake but also a trial-and-error approach to attain identical model hyperparameters that are effective for both lakes. By effective, we refer to the hyperparameters that yield efficient performance for both lakes.

## 2.7. Feature selection

### 2.7.1. Feature ranking

Feature ranking (FR) falls under the category of feature selection algorithms and serves as a valuable tool in ML. While one might assume that incorporating more features (inputs) would lead to greater specificity, this is not always the case, as highlighted by Koller and Sahami (1996). Reunanen (2003)

discusses several reasons for employing feature extraction techniques to reduce the number of inputs. Firstly, measuring only a subset of data proves to be more cost-effective. Secondly, excluding unnecessary features can potentially enhance prediction accuracy. Thirdly, utilizing fewer input features allows for the development of simpler and faster predictors. Lastly, determining the importance of features aids in comprehending the challenges associated with prediction tasks. Consequently, the identification of the most important forcing has become increasingly important in the realm of ML in order to increase the efficiency of the model (Blum and Langley, 1997; Vivencio et al., 2007).

In essence, the FR algorithm assigns a rating to each feature, reflecting its significance in determining the model's performance. In our study, we employed the feature importance approach (Izetta Riera and Granitto, 2009) to rank the predictors. This approach involves evaluating how the performance (explained in section 2.3) of the calibrated ML model changes when compared to the reference case (i.e., the performance achieved during training) after shuffling (rearranging) the values of a specific feature to obtain a random sequence of that predictor (presented in Figure 2.12). The decline in model error indicates the model's dependence on that particular feature: the greater the decline, the more relevant the feature. We conducted the shuffling (rearranging) process independently for all features used as predictors, repeating it 30 times for each feature.

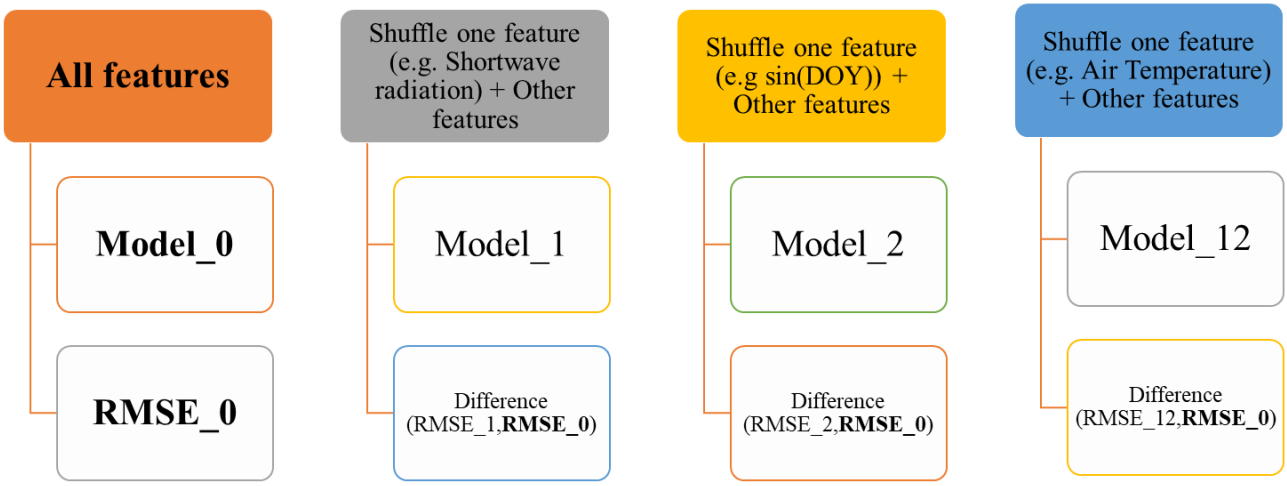



Figure 2.12. Scheme of feature ranking approach.

By ranking the features based on their relevance and importance, we can identify those that convey the most valuable information to the model. This ranking allows us to prioritize the retention of high-ranking features while discarding lower-ranking or irrelevant ones. By doing so, we can eliminate unimportant features, reduce dimensionality, and enhance the model's ability to extract meaningful patterns from the data. This approach is applied in Chapters 4 and 5. However, in Chapter 3, we approached feature selection differently, focusing on individual features as well as analyzing various

combinations. In Chapter 3 we aimed at emphasizing the discovery of the most efficient feature combinations, while the focus in the other two chapters was on identifying significant features in distinct climate regions (Chapter 4) and various simulation periods (Chapter 5).

### 2.7.2. Correlation

There are different methods for feature selection that can be categorized as ad-hoc, where inputs are chosen arbitrarily or based on domain knowledge, and analytical, where statistical measures like correlation or mutual information assess the relationship between potential inputs and outputs (Maier et al., 2010). Therefore, examining the correlation among input variables, the target variable, and the inputs themselves can provide insights into feature selection and reveal how the variables are interrelated.

To explore these relationships in our cases, we conducted Pearson and Spearman's rank-order correlation analyses. Spearman's rank-order correlation measures the strength and trend of monotonic relationships between variables (Spearman, 1910), while Pearson correlation evaluates the strength of linear correlation. Spearman's measures are suitable when dealing with variables measured on an ordinal scale, unlike Pearson's, which requires continuous variables. The choice between these measures may also depend on theoretical considerations, ease of comparison with previous studies, or specific research needs. In Chapter 3, we performed correlation analyses using both Pearson and Spearman's rank-order methods to assess various correlation approaches. Chapters 4 utilized Spearman's method to account for the non-linearity of variables. We skipped performing the correlation analysis for the ice model because we verified that the results of correlation and feature ranking align in Chapter 4, and we wanted to avoid redundant analyses.

Collinearity arises when two or more variables exhibit strong correlations, potentially impacting classifier performance by introducing additional weight to input variables or introducing noise into final outcomes (Gómez-Escalonilla et al., 2021). In the study of Ferreira et al. (2021), the pairs of predictors exhibiting a Pearson's correlation exceeding 95% were assessed for collinearity with all other predictors (multicollinearity). The predictor displaying the highest correlation with other predictors within each pair was then removed.

Permutation importance, a method used to assess variable importance, may face challenges in correctly identifying significant predictors in datasets with highly correlated variables, as highlighted by Strobl et al. (2008). The act of randomly permuting a variable disrupts its correlative links with other predictors, potentially impacting the model's ability to discern important features. This issue becomes particularly pertinent in datasets with environmental variables, characterized by strong

multicollinearity (e.g., elevation and precipitation). In such cases, permutation importance may not be the optimal choice for evaluating variable importance (Wade et al., 2023).

# Chapter 3. Selecting machine learning techniques and the features affecting LSWT by using a synthetic lake[1]

## 3.1. Introduction and literature review

Local climate conditions have a profound influence on the lake surface water temperature (LSWT), and the ongoing warming trend is posing a significant threat to the biological and chemical processes within aquatic ecosystems (Woolway and Merchant, 2018; Woolway et al., 2020; O'Reilly et al., 2003). Consequently, it is of utmost importance to gain a comprehensive understanding of the underlying factors controlling the thermal response of lakes to climate change and accurately assess the expected changes (O'Reilly et al., 2003). In pursuit of this goal, a diverse range of models has been proposed, spanning from simple correlations to intricate numerical models.

This study focuses specifically on the LSWT due to its convenient observability through in-situ measurements or remote sensing techniques. However, predicting LSWT presents notable challenges due to the complex interplay among various meteorological factors, including wind speed, air temperature (AT), radiation, heat fluxes, as well as lake bathymetry and inflows/outflows (Livingstone, 2003; Piccolroaz et al., 2015). Despite these challenges, several effective models have been developed for forecasting LSWT, which can be broadly classified into two main categories: physically based models and data-driven models. As the following paragraphs have been discussed in section 1.1, we elaborated on them here as well.

Physically based models aim to solve the fundamental equations that govern the exchange and transport of heat in lakes. These models have the potential to provide outcomes without relying solely on LSWT data, although the calibration of empirical relations using LSWT data remains an integral part of the modeling process (Ragotzkie, 1978; Hipsey et al., 2019; Irving et al., 2006).

The data-driven modeling category encompasses a wide range of stochastic methods, including linear and non-linear regression, autoregressive models, periodic autoregressive models (Benyahya et al., 2007), and evolutionary polynomial regression (Doglioni et al., 2008). These models establish

[1] The contents of this chapter have been derived from: Yousefi, A. and Toffolon, M., 2022. Critical factors for the use of machine learning to predict lake surface water temperature. *Journal of Hydrology*, *606*, 127418. https://doi.org/10.1016/j.jhydrol.2021.127418

relationships between LSWT and influential factors, with AT being a predominant factor, in order to forecast the thermal dynamics of water. Additionally, this category includes the application of machine learning (ML) algorithms, which have gained significant popularity as versatile computational models (Mohri et al., 2018). In recent years, there has been a notable increase in endeavors to predict LSWT using supervised learning approaches based on available data, particularly since 2019. Table 3.1 provides an overview of previous studies on LSWT prediction using ML in lakes, highlighting the various ML algorithms employed and the features of the lakes investigated.

Among the ML algorithms, artificial neural network (ANN) has emerged as the most widely used and successful approach for LSWT prediction (Sharma et al., 2008; Liu and Chen et al., 2012; Sener et al., 2012; Samadianfard et al., 2016; Read et al., 2019; Saber et al., 2020; Heddam et al., 2020; Jia et al., 2019 and 2021; Quan et al., 2020; Zhu et al., 2020a; Table 3.1).

Appendix B includes an examination of the relevant literature on ML-based water temperature prediction in rivers (Table B.1), as there are similarities between this problem and the prediction of water temperature in lakes. It is worth noting that lakes, being lentic water bodies, are less influenced by the advection of water from upstream regions, which is often a significant factor in rivers. However, the thermal behavior of lakes is complicated by the development of thermal stratification throughout the year.

Hybrid models, which lie at the intersection of physically based and data-driven models, encompass various approaches. According to Karpatne et al. (2017), hybrid models can be implemented in different ways. One approach involves utilizing data from physically based models as input for data-driven models, often used for downscaling climate variables. Additionally, the outputs of physically based models can be employed to control the training phase of data-driven models by providing initial estimations of target outputs (Read et al., 2019). Another approach is the integration of data-driven methods within the framework of physically based models, aiming to improve parameter estimates or replicate complex processes with greater accuracy.

In their study, Read et al. (2019) employed traditional physically based models to estimate weights and biases for pre-training ML models, as well as to calibrate the model using physical laws during the pre-training and training phases. They conducted a comparison between their hybrid model, which combined ML and physically based models, and the two methods used individually. The results demonstrated that ML outperformed the physically based model when the number of training data exceeded a specific threshold, which varied depending on the dataset. Conversely, the performance

of the process-based model did not exhibit further improvement beyond a certain number of training data, highlighting the notion that ML approaches require and benefit from a larger quantity of data.

Furthermore, Read et al. (2019) conducted an analysis using different datasets grouped based on seasonal or climatic scales. Their findings emphasized that ML yielded more accurate results when the data exhibited similar patterns. The study confirmed that, in their case studies, the hybrid model emerged as the most successful approach compared to the individual ML and process-based methods.

A distinct form of hybrid model known as air2water has been developed for predicting LSWT (Piccolroaz et al., 2013, 2018, 2020, 2021; Toffolon et al., 2014; Piccolroaz, 2016). This model employs simplified equations derived from physical laws, condensing intricate details into a reduced set of parameters that are calibrated against measurements. Notably, air2water only requires AT as input, rendering it as user-friendly as other data-driven models while surpassing the performance of many such models (Heddam et al., 2020). Its simplicity and effectiveness also make it suitable for comparative evaluations of ML algorithms.

This research addresses gaps in existing studies by investigating the use of ML methods with optimization to predict water temperature, specifically concentrating on LSWT. Notably, the studies outlined in Table 3.1 did not encompass such a comprehensive exploration of methods as well as implementing an optimization. Previous studies, like Read et al. (2019), have compared hybrid versus ML models, but they did not explore optimizing the integration of physical information, such as considering the memory of previous days of predictors. While pre-processing methods including denoising and rescaling have been commonly used (Saber et al., 2020; Zhu et al., 2020a), there is a lack of comprehensive comparisons, especially for LSWT models, including an effective assessment of different preprocessing methods like wavelet transformers. None of the studies have considered the effect of lake depth, which is a physical quantity that is known to strongly influence the thermal response of lakes (e.g., Toffolon et al., 2014). Therefore, our study aims to fill these research gaps through a systematic analysis of predictor selection, pre-processing methods, and the performance of various ML algorithms.

Table 3.1. A list of recent papers dealing with the use of ML tools to predict LSWT.

| **Reference** | **ML model** | **Optimization / pre-processing** | **The best method(s)** | **Investigated lake or region** | **Lake depth (m)** | **Lake area ($km^2$)** |
|---|---|---|---|---|---|---|
| Heddam et al. (2020) | ERT, Multivariate adaptive regression splines (MARS), M5 model tree (M5Tree), RF, MLPNN | Sigmoid and identity activation function for MLPNN | air2water; ERT among ML methods | 25 lowland lakes in Poland | 1.3 – 15.5 | 1.545 – 114.88 |
| Jia et al. (2019) | PGRNN, RNN | Pre-training using physical simulations - | PGRNN | Minnesota and Wisconsin | max 25.3 | 40 |

| Reference | ML model | Optimization / pre-processing | The best method(s) | Investigated lake or region | Lake depth (m) | Lake area ($km^2$) |
|---|---|---|---|---|---|---|
| | | (density-depth relationship - energy conservation) - hyperbolic tangent transfer function | | | | |
| Jia et al. (2021) | PGRNN | Pre-training using physical simulations - constrain of ML (energy balance) - hyperbolic tangent transfer function | PGRNN | Minnesota and Wisconsin | max 25.3 | 40 |
| Liu and Chen (2012) | ANN | Hyperbolic tangent sigmoid transfer function | 3D hydrodynamic model | Yuan-Yang | max 4.5 | 37 |
| Quan et al. (2020) | SVR, ANN, GA-SVR, and optimized parameters of improved support vector machine (M-GASVR) | GA | M-GASVR | Longyangxia | different depths from 35 to 55 | 131,142 |
| Read et al. (2019) | Process-guided deep learning (PGDL) hybrid modelling, DL model, and a process-based (PB) model | Pre-training using physical Simulations - constrain of ML (energy balance) - hyperbolic tangent transfer function | PGDL | 68 Canadian lakes | Maximum depth from 5.2 to 54 | From 0.07 to 196.65 |
| Saber et al. (2020) | ANN | WT | ANN | Mead in Nevada and Arizona | Min 140, Max 162 | 640 |
| Samadianfard et al. (2016) | GEP, ANFIS, ANN | GP - sigmoid activation function | GEP | Yuan-Yang in Taiwan | max 4.5 | 37 |
| Sener at al. (2012) | ANN | Sigmoid activation function | ANN | Eğirdir, Turkey | 16 | 482 |
| Sharma et al. (2008) | Multiple regression, regression tree, ANN, Bayesian multiple regression | - | multiple regression | 2348 lakes in Minnesota and Wisconsin | - | - |
| Zhu et al. (2020a) | MLPNN, WT and MLPNN integrated model (WT-MLPNN), non-linear regression model (S-curve), air2water | WT | air2water, WT_MLPNN, and MLPNN | 8 lowland lakes in Poland | from 1.4 to 15.5 | from 2.15 to 98.51 |
| Zhu et al. (2020b) | The non-linear regression model (S-curve), The linear regression | - | S-curve | Morskie Oko | 29.7 | 0.33 |

## 3.2. Characterization of the physical problem

Before delving into the performance evaluation of ML methods in simulating LSWT, it is valuable to comprehend the underlying physics of the investigated process. The temperature of water in lakes is influenced by a multitude of meteorological and limnological variables (Magee and Wu, 2017). Gaining a deeper understanding of the heat budget provides valuable insights into the factors that exert influence.

One well-known phenomenon is the stratification of lakes, where distinct layers with different temperatures (and densities) form. As stratification intensifies, these layers behave as separate regions. This stratification is particularly pronounced in temperate lakes during summer, with a well-mixed surface layer known as the epilimnion, which is separated from the deep waters (hypolimnion) by a layer where the largest temperature difference is localized, known as the thermocline (Boehrer and Schultze, 2008). In the well-mixed surface layer, the average temperature across the volume exhibits a strong correlation with LSWT. The LSWT, in turn, depends on the net heat flux exchanged through the lake surface. Several factors control this net heat flux, which also have implications for the prediction of LSWT. These factors include shortwave solar radiation, longwave radiation influenced by AT and cloud cover, latent heat influenced by AT, water temperature, relative humidity, and wind speed, sensible heat influenced by AT, water temperature and wind speed, and the heat contribution of rainfall (Schmid et al., 2014).

### 3.2.1. Minimal model of the thermal response of a lake

In this section, we revisit the heat budget of lakes by examining a simplified one-dimensional (1D) representation,

$$\frac{\partial}{\partial t}(\rho c_p S T) + \frac{\partial}{\partial z}(S\phi + I) = 0 \tag{3.1}$$

where $t$ is time [s], $z$ is the vertical coordinate [m], $T(z,t)$ is the water temperature [K], $\rho(T)$ is water density [kg m$^{-3}$], $c_p(T)$ is the specific heat capacity [J kg$^{-1}$ K$^{-1}$], $S(z)$ is the horizontal surface area [m$^2$], $\phi(z,t)$ is the heat flux [W m$^{-2}$], which includes diffusion, convection and penetrating shortwave solar radiation, and $I(z,t)$ represents the contribution [W] of lateral inflows and outflows. Equation (3.1) is solved with proper boundary conditions at the lake bottom (heat exchanged with sediments) and at the surface (heat exchanged at the air-water interface).

We can consider a bulk version of equation (3.1) to understand the behavior of LSWT, assuming that it is correlated with the average temperature in the surface layer. By integrating equation (3.1) over

the thermally reactive volume $V_r$ [m$^3$] at the surface, the volume-averaged temperature $\bar{T}(t)$ follows the equation

$$\frac{d}{dt}(\rho c_p V_r \bar{T}) = S_0 \phi_{net} + S_D \phi_D + \sum_{j=1}^{N_I} \rho c_p Q_{I,j} T_{I,j} - \sum_{j=1}^{N_O} \rho c_p Q_{O,j} T_{O,j} \tag{3.2}$$

where $S_0$ is the surface area and $\phi_{net}$ is the net heat flux [W m$^{-2}$] exchanged with the atmosphere (assumed positive towards the lake), $S_D$ is the area at the bottom of the surface layer and $\phi_D$ is the heat flux exchanged with deeper waters (positive if entering the surface layer), and the last two summations include the effects of inflows and outflows. $Q_{I,j}$ and $T_{I,j}$ are, respectively, the discharge and the temperature associated with the $j$-th inflow; $Q_{O,j}$ and $T_{O,j}$ are the equivalent variables for the $j$-th outflow. The net heat flux $\phi_{net}$ is the combination of multiple factors:

$$\phi_{net} = H_{SW} + H_{LW_{in}} - H_{LW_{out}} \pm H_L \pm H_C + H_P \tag{3.3}$$

where $H_{SW}$is the fraction of the shortwave solar radiation penetrating into the lake and absorbed in the surface volume, depending on water transparency; $H_{LW_{in}}$ is the net longwave radiation emitted from the atmosphere to the lake, depending on AT and cloud cover; $H_{LW_{out}}$ is the longwave radiation emitted from the water, depending on LSWT; $H_L$ is the latent heat flux considering evaporation and condensation, which depends on AT, LSWT, relative humidity and wind speed; $H_C$ is the sensible heat flux because of convection, depending on AT, LSWT and wind speed; $H_P$ is the heat flux associated with rainfall. The deep-water heat flux $\phi_D$ can be affected by subsurface fluxes and by the geothermal heat flux at the sediment-water interface.

The role of the depth, and in particular of the thickness $h = V_r/S_0$ of the thermally reactive surface layer, can be highlighted considering a simplistic (only for illustration purposes) version of equation (3.2):

$$\alpha \frac{dT}{dt} = \beta \sin(\omega t) \tag{3.4}$$

where $\alpha$ includes the heat capacity of the surface layer (linearly dependent on $h$), and the sinusoidal function on the right-hand side mimics the temporal variability of the net heat flux. In this simplified formulation, we refer only to single harmonics of a Fourier decomposition of the heat flux because the linearity in $T$ (a rough approximation of the actual behaviour) allows to consider them separately. Hence, $\beta$ is the amplitude of the sinusoidal variability associated with a given frequency $\omega$: for

instance, two dominant frequencies of the heat flux are related to the variability of solar radiation at daily and annual periods. The solution of equation (3.4) is straightforward:

$$T = \frac{-\beta}{\alpha\omega}\cos(\omega t) + c \tag{3.5}$$

with $c$ an integration constant depending on the initial conditions. It is immediately evident that the amplitude of the water temperature oscillations at a given frequency is inversely dependent on $\alpha$ (for a similar analysis, see Toffolon et al., 2014). Thus, a deeper surface layer (larger $h$, larger thermal inertia, longer memory) will reduce the difference between the maximum and minimum temperature, for instance, at a daily time scale. Of course, larger $h$ are more likely to occur for deep lakes. Moreover, we can see that the amplitude is also inversely dependent on the frequency of the forcing, so that higher frequencies are more damped than lower frequencies, suggesting the 'filter' analogy introduced in the main text.

The depth of a lake is a critical parameter to consider when modeling its temperature. Shallow lakes tend to have uniform water temperature throughout the entire water column, whereas deep lakes exhibit a significant temperature difference between the surface and the hypolimnion during the summer (Calamita et al., 2021). Consequently, the behavior of lakes varies depending on their depth, with deep lakes being particularly influenced by stratification dynamics (Piccolroaz et al., 2015). The significance of depth can also be highlighted through minimal physical models, as demonstrated in Toffolon et al. (2014). It is worth noting that lakes can be described as "filters," where the input signal (such as air temperature and solar radiation) undergoes physical processing, resulting in a smoothed output signal, such as lake surface water temperature (Calamita et al., 2021; Piccolroaz et al., 2015).

## 3.3. Materials and methods

### 3.3.1. Case study

Our primary objective is to elucidate the critical factors necessary for the accurate application of a ML model in predicting LSWT. In order to achieve this, we opted to create a synthetic example instead of utilizing real lake data. By doing so, we obtained complete control over all the variables involved, enabling us to compare and analyze the performance of various approaches. Furthermore, this synthetic example allows us to investigate the impact of lake morphology, particularly depth, on LSWT response while maintaining consistent meteorological forcing. Consequently, we can assess the ability of ML models to simulate thermal dynamics under diverse conditions.

The synthetic case study utilized the widely used General Lake Model (GLM) to generate a 1D representation (Hipsey et al., 2019). The specific details of the model are not significant for our analysis, as any other model could have been employed. It is important to acknowledge that all the findings and conclusions drawn from our study are based on numerical simulations rather than real-world cases. However, it is well-established that these simulations accurately capture the primary dynamics of lakes. The advantage of employing a fully controlled synthetic case is that it allows us to focus solely on the performance of the ML model, without introducing additional uncertainties associated with real data, such as missing or erroneous values. Additionally, it enables us to compare lakes with different depths while keeping all other factors constant, thereby isolating the significance of depth in our analysis. This approach could potentially be extended to explore other physical variables. Nevertheless, it is important to note that the artificial case lacks the complexity of the real world, and only the factors included in the lake model are evaluated.

For the general characteristics of the synthetic case study, we selected a medium-sized peri-alpine lake in the temperate region as a reference (see Table 3.2). To ensure realistic values for lake morphology and climatic conditions, we utilized the hypsometric curve of Lake Caldonazzo (northern Italy) and extracted corresponding meteorological data from ERA5 for the specific latitude and longitude (Hersbach et al., 2020). It is important to emphasize that a real lake was solely used as a reference to design the synthetic case study, and we do not aim to simulate actual dynamics. The choice of an artificial lake enabled us to work with a comprehensive and coherent dataset, facilitating the examination of the effects of various forcing factors at different depths. Specifically, we investigated a shallow case (maximum depth of 5 m), a deep case (maximum depth of 60 m), and two intermediate depths (20 m and 40 m). To achieve these scenarios, we modified the hypsometric curve of Lake Caldonazzo (with a maximum depth of 49 m) by stretching the vertical coordinate while maintaining the surface area to preserve the interface with the atmosphere. The inflow to the lake was artificially adjusted to compensate for evaporation and maintain relatively stable water levels from year to year, minimizing the thermal impact of inflows. An illustration of the seasonal evolution of thermal stratification in the shallow and deep lakes can be found in Figure 3.1. It suffices to mention here that the shallow lake is consistently well-mixed, while the deeper cases clearly stratify during summer.

The water temperature variations along the water column in the synthetic lakes (shallow and deep) are depicted in Figure 3.1. The temporal evolution is captured through numerical simulations using the one-dimensional lake model, GLM. Specific calibration was not conducted as the case studies refer to synthetic lakes. The model adopted default parameters from the calibrated model of a real lake, Caldonazzo, with the sole adjustment being the modification of depth. The figure displays the

patterns over two consecutive years, showcasing the distinctive seasonal variations and highlighting the contrasting characteristics between the shallow and deep lake scenarios. The hypsographic curves of the synthetic lakes can be found in Appendix D.

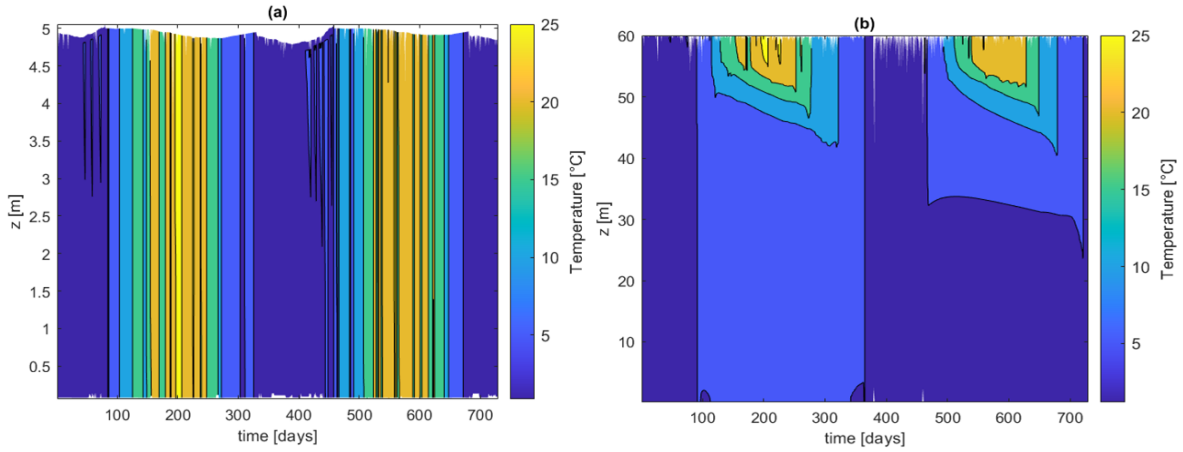



Figure 3.1. Evolution of water temperature in two lakes with the same hypsometry, but rescaled with different depths (a: maximum depth 5 m; b: maximum depth 60 m), as a function of depth and time (from 01/01/2015 to 31/12/2016), obtained from the physically based model GLM.

The ML application employed data from the GLM to obtain the LSWT as the target variable, while meteorological data from ERA5 were utilized as predictors. To ensure consistency, all data sets were standardized to a daily time step. Table 3.2 presents the main characteristics of the data set, including its duration and other pertinent details.

Table 3.2. Features of synthetic case studies.

| | Shallow lake | Deep lake |
|---|---|---|
| Lake's maximum depth (m) | 5 | 60 |
| Lake's average depth (m) | 2.5 | 30.0 |
| Lake's area ($km^2$) | 5.2 | 5.2 |
| Lake's volume ($10^6$ $m^3$) | 13 | 157 |
| Hydraulic retention time (years) | 4.6 | 52 |
| Date range for training set | 01/01/1979 – 31/12/2014 | |
| Date range for test set | 01/01/2015 – 31/12/2018 | |
| Latitude / Longitude | 46.01° N, 11.24° E | |
| Elevation (m a.s.l.) | 450 | |
| Range* of WS (m/s) | 0.46 – 7.8 | |
| Range* of SWR ($W/m^2$) | 4.05 – 342.5 | |
| Range* of AT (°C) | -17.2 – 27.1 | |

* For daily averaged values

### 3.3.2. Setup of ML models

The configuration of a ML model involves several stages, including the selection of input variables (predictors), their pre-processing, and the application of the ML algorithm. Each component of the methodology, as depicted in Figure 3.2, is examined in detail in the subsequent paragraphs.

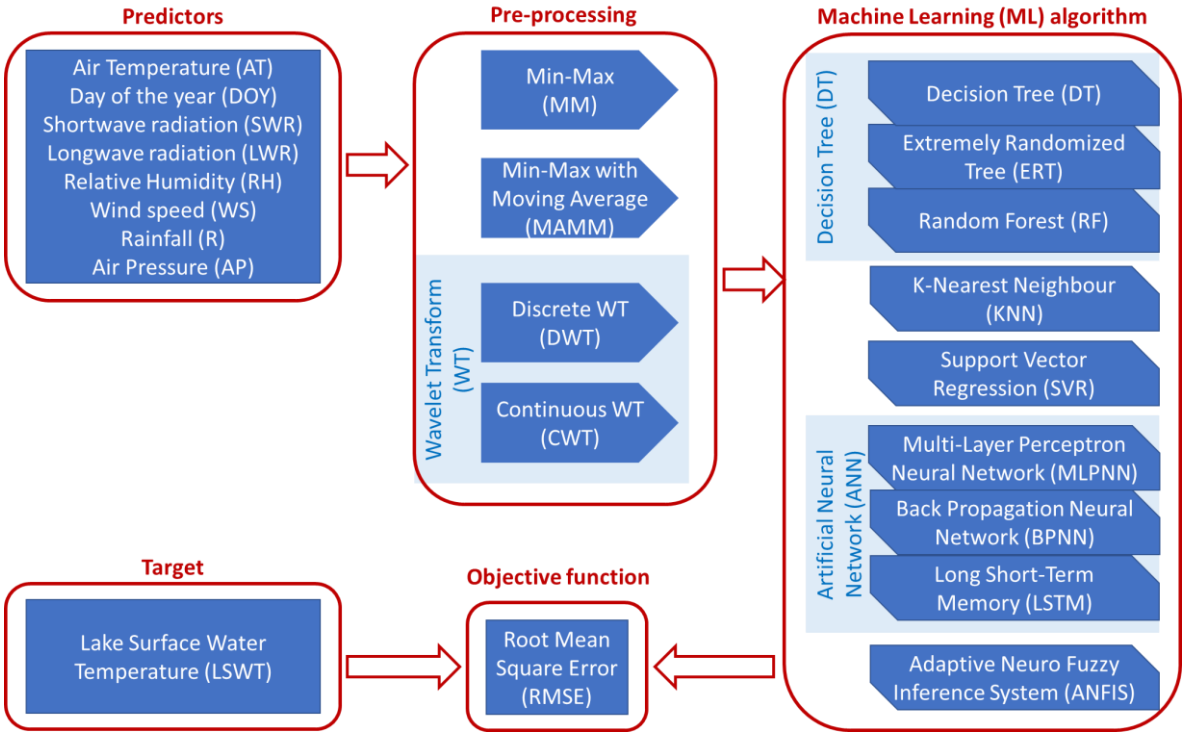



Figure 3.2. Workflow summarizing the steps of the comparative analysis of the performance of the different ML methods.

The initial stage involves selecting the predictors, which are the variables utilized as input for the ML algorithm. In line with our objective of predicting LSWT, a comprehensive range of predictors has been considered, taking into account the heat flux components outlined earlier. These include AT, shortwave radiation (SWR), downward longwave radiation (LWR), relative humidity (RH), wind speed (WS), rainfall (R), and air pressure (AP). Additionally, the day of the year (DOY) was incorporated as a feature. We opted not to incorporate the inflows to avoid introducing features that are characterized by a broad range of variability, which would also require modifying the water budget, and limit the analysis to the usual meteorological forcing. Both individual predictors and their combinations were evaluated. The impact of different predictor choices is examined in section 3.4.1.

Regarding the use of the DOY as a predictor, caution is advised due to its discontinuous behavior at the year's change (from day 365 to day 1), which does not align with the continuous variation of LSWT. To address this issue, we explored the utilization of the sine and cosine functions of DOY (SCDOY), denoted as sin(DOY⁄n_DOY) and cos(DOY⁄n_DOY) respectively, where n_DOY represents the number of days in a year. These transformed variables, collectively referred to as

SCDOY for simplicity, contain an equivalent amount of information as DOY but exhibit a continuous pattern.

The subsequent step involves pre-processing the input data to ensure their comparability by normalizing their range of variation. Various pre-processing strategies were examined, including popular techniques such as Min-Max scaling, Robust scaling, and Standard scaling. In our preliminary tests, Min-Max scaling yielded the most favorable outcomes for our specific problem. Consequently, we adopted this method for all subsequent analyses.

LSWT exhibits a filtering behavior in response to external forcing due to the water mass's heat capacity, as discussed in section 3.2.1. Consequently, in the pre-processing step, we incorporated the option to temporally average certain predictors, particularly AT. The hyperbolic averaging results slightly better than uniform averaging which the detailed information is explained in section 2.4. Although, as the matter of simplicity, we consider the unform average for AT. Alternatively, we also considered the inclusion of AT from previous days as a distinct strategy to achieve a similar effect. This resulted in an increased number of input variables for the ML model. Detailed discussions on the different pre-processing strategies can be found in section 3.4.2.

As part of the standard procedure, the available data were divided into two subsets, commonly referred to as the training and test data sets. The optimization of ML algorithm parameters was performed using both the training and test data sets, which accounted for the entire data set. Although other proportions (e.g., 50%, 60%, and 70%) were tested, the results remained largely consistent due to the ample availability of data, reducing the risk of overparameterization. Subsequently, the ML model's performance was evaluated using the test data set to ensure that the calibrated parameters were not influenced by overfitting or underfitting and could be generalized to other scenarios. The chronological separation (first 80%, last 20%) of the training and test data sets was implemented to preserve the temporal patterns of the variables within each data set. Random splitting was also examined, but the results of both data sets exhibited minimal variations, making it challenging to identify the optimal parameters.

The third step involves selecting the appropriate ML algorithm, which serves as the core of the approach and encompasses various structural parameters known as hyperparameters. These hyperparameters are determined prior to training the data (e.g., Probst et al., 2019). The distinction between hyperparameters and parameters is customary and revolves around the fact that the former define the characteristics of the algorithm itself (e.g., the number of layers and neurons in an ANN), while the latter pertain to the coefficients in the equations utilized by the algorithm (e.g., the values of weights and biases in an ANN). Both the hyperparameters and parameters are chosen and

optimized for individual case studies and combinations of input variables. The optimal values for the hyperparameters are identified using optimization algorithms, whereas the calibration of parameters depends on the specific ML approach employed.

The selection of hyperparameters can be accomplished through various optimization methods, including Grid Search, Random Search, Bayesian methods, and particle swarm optimization (PSO). In this study, we specifically employed the genetic algorithm (GA) for all the ML approaches examined. GA is an evolutionary optimization method that yields satisfactory results while maintaining reasonable computational demands (Di Francescomarino et al., 2018).

A range of ML algorithms have been utilized for predicting water temperature in lakes, as indicated in Table 3.1. These include linear and non-linear regression models (e.g., logistic regression), ANN, decision trees (DT), support vector regression (SVR), K-nearest neighbor algorithm (KNN), random forest (RF), and gradient boosting machines (GBM). While linear and logistic regression models can provide satisfactory outcomes despite their simplicity, more advanced ML algorithms have demonstrated greater suitability for water temperature prediction in broader contexts (Sharma et al., 2008; Sener et al., 2012).

In this study, we conducted experiments with nine distinct ML approaches, aiming to encompass a wide range of options proposed in the literature. Among these approaches, four are variations of ANN with specific characteristics. The multilayer perceptron neural network (MLPNN) and backpropagation neural network (BPNN) are closely related, differing primarily in their parameter calibration strategies. Long short-term memory (LSTM) is a deep learning technique that preserves temporal memory within the system. The adaptive neuro-fuzzy inference system (ANFIS) combines a neural network structure with fuzzy logic. Additionally, three approaches fall within the DT framework: the standard DT, RF, and extremely randomized tree (ERT). The remaining two approaches are SVR and KNN. Section 2.2 provides a summary and explanation of the main features of these ML algorithms, while Appendix A contains details regarding the hyperparameters. The comparative evaluation of the nine ML approaches is presented in section 3.4.3.

To analyze the different pre-processing strategies and predictors in sections 3.4.1 and 3.4.2, we selected the BPNN as a representative branch of ANN. BPNN is commonly used for capturing nonlinear relationships between dependent and independent variables, and it offers various training methods (Tu, 1996).

As the fourth and final step of the workflow, we assessed the performance of the ML approaches. We employed the RMSE of the LSWT as the objective function, where lower values indicate a better fit. While alternative metrics such as mean absolute error, correlation coefficient, and Nash-Sutcliffe

efficiency index can also be used, they did not yield significant differences in the ranking of the best ML approaches. As our study focuses on a synthetic case study, we considered the output of GLM as the observed truth for the analysis.

To enhance the robustness of the ML results and mitigate the possibility of inconsistent outcomes due to suboptimal initial parameter guesses or other numerical artifacts, we performed each algorithm 20 times (similar to the approach by Piotrowski et al., 2021, who conducted 50 runs). The difference in the runs, quantified using the standard deviation of the simulated LSWT, can be found in Appendix C. Unless specified otherwise, the performance metrics presented in the analysis are averaged across the runs. More detailed information is provided in Appendix C. Additionally, to gauge the method's robustness and the variation across different runs, we computed the standard deviation of the 20 LSWT estimates for each day of the simulated period. These values were then averaged over the entire record to obtain the average value, $\sigma_R$.

## 3.4. Results

In this section, we delve into the examination of the influential inputs, effective pre-processing methods, and the performance of ML approaches. We also analyze the impact of the lake's depth and incorporate the lake's physical characteristics in the statistical analysis.

### 3.4.1. Identification of predictors

Gaining a deep understanding of the underlying physics of thermal dynamics is essential for achieving accurate predictions, particularly when using approaches that do not rely on physical laws. ML algorithms, being versatile tools, heavily rely on the careful selection of predictors for successful application. As highlighted in section 3.3.2, there is a wide range of potential predictors for predicting LSWT. Previous analyses (Table 3.3) have considered various predictors depending on the specific variable being simulated, such as daily averaged temperature, monthly or annual averages or maximums. These predictors include AT, DOY, Gregorian calendar, lake surface area, maximum and average water depth, altitude, WS, indicators for frozen and snowing conditions, R, percentage of cloud cover, inflow and outflow rates, geopotential height, LWR, SWR, pH, total dissolved solids, RH, Secchi depth, and specific variables for seasonal evaluations, such as the maximum number of daylight hours in June, mean July AT, mean annual AT, mean July precipitation, July SWR, and July cloud cover (Sharma et al., 2008). Many of these predictors have already been discussed when exploring the factors influencing heat flux at the lake surface. It is worth noting that for each predictor, careful consideration must be given to the temporal and spatial scales of integration, as these choices can significantly impact the results (Piccolroaz et al., 2016; Toffolon et al., 2020).

Table 3.3. ML algorithms used to predict LSWT in the literature, with the indication of the predictors used.

| Machine Learning Algorithm | Author (s) | Best ML model | Predicted Quantity | Influential Predictor(s) | Water body |
|---|---|---|---|---|---|
| ANN | Sharma et al. (2008) | ANN | Annual maximum near-surface lake water temperature, and not daily LSWT | Up to 17 predictors were used in different combinations, including mean AT, mean July AT, DOY and year | Lake |
| | Sener at al. (2012) | ANN | Daily water temperature | AT, RH, AP, and depth | Lake |
| | Samadianfard et al. (2016) | GEP | Hourly water temperature in different depths | AT, SWR, AP, RH, R, WS, soil T | Lake |
| | Jia et al. (2019) | PGRNN | Daily water temperature profile | 10 predictors, including AT, SWR, LWR, WS, RH, frozen and snowing indicators | Lake |
| | Read et al. (2019) | Process-Guided Deep Learning (PGDL) | Daily water temperature profile | AT, SWR, LWR, WS, RH and precipitation (rain or snow) | Lake |
| | Saber et al. (2020) | ANN | Daily and 6-hour water temperature profile | AT, WS, RH and AP | Lake |
| | Jia et al. (2021) | PGRNN | Daily water temperature profile | Meteorological variables AT, SWR, LWR, WS, RH, R, snow in addition to depth and DOY | Lake |
| | Zhu et al. (2020a) | WT-MLPNN | Daily LSWT | AT | Lake |
| | Quan et al. (2020) | Optimized parameters of improved support vector machine (M-GASVR) | Daily surface water temperature | Surface T, AP, near-surface air specific humidity, near-surface full wind speed, surface precipitation rate, and water depth | Reservoir |
| SVM | Heddam et al. (2020) | ERT | Daily LSWT | AT, the Gregorian calendar (year, month and day) number | Lake |
| DT | Zhu et al. (2020b) | Non-linear regression model | Monthly and yearly average LSWT | AT | Lake |

Statistical analysis has frequently been employed to identify the most influential predictors for LSWT. For example, Zhu et al. (2020b) examined both linear and non-linear relationships between AT and LSWT. They found that non-linear S-shaped functions yielded favorable correlations between AT and LSWT. Mohseni et al. (1998) also conducted research on AT and LSWT, concluding that non-linear S-shaped functions were effective in capturing their relationship. In a similar vein, Yang et al. (2020) utilized various methods such as trend analysis, contribution analysis, and Pearson correlation

analysis to identify the most influential predictors for LSWT. Through their investigation, they highlighted the significance of AT and SWR as influential predictors for LSWT.

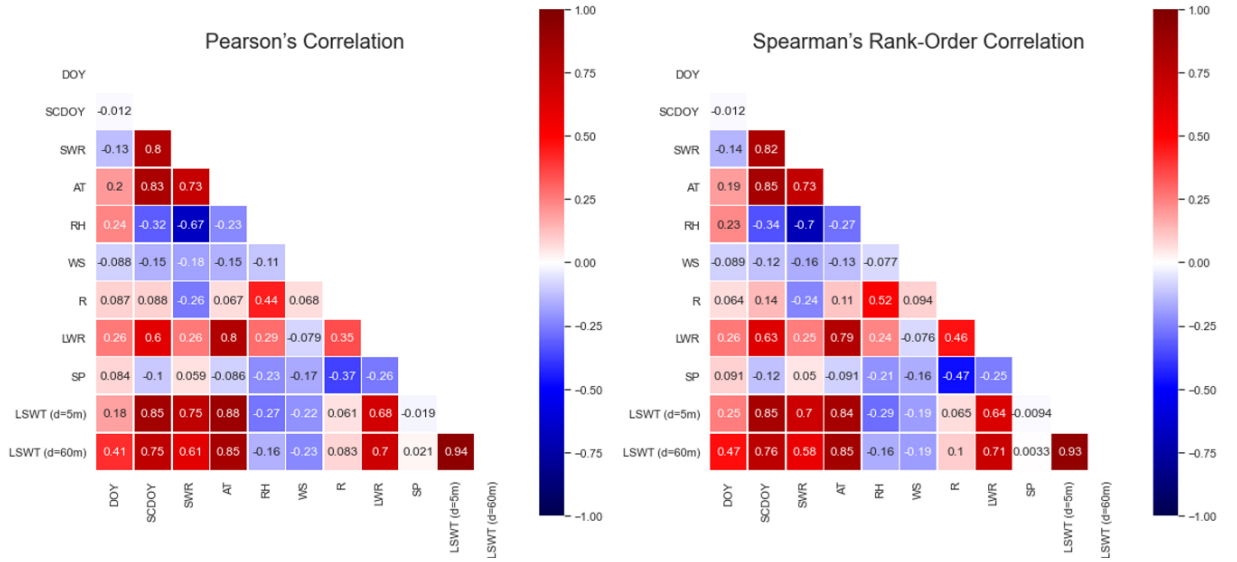



Figure 3.3. Correlation between variables: (a) Pearson's (linear) correlation; (b) Spearman's Rank-Order (non-linear) correlation.

It is crucial to consider not only the correlations between predictors and LSWT, but also the intercorrelations among the predictors themselves. To explore these relationships in our synthetic shallow and deep test cases, we conducted Pearson and Spearman's rank-order correlation analyses explained in section 2.7.2. It is worth mentioning that in our specific cases, it is unnecessary to eliminate features due to collinearity since the Pearson's correlations do not exceed 95% (explained in section 2.7.2).

Our findings indicate that AT is the most significant predictor for LSWT, exhibiting a strong linear correlation. Additionally, the variable SCDOY (sin(DOY/n_DOY) and cos(DOY/n_DOY)) emerges as a highly influential predictor. The linear version of time (DOY) demonstrates lower correlation compared to SCDOY, particularly in the shallow case. Alongside SCDOY and AT, LWR and SWR exhibit similar trends compared to other predictors. Notable correlations include R with RH and LWR, albeit with low scores with respect to LSWT. It is worth noting that the influence of R might be relatively minor due to the simplified representation of rainfall's impact on inflows in our physically based lake model, which may not fully capture real-world scenarios.

To assess the impact of predictor selection on the performance of ML in reproducing LSWT, we examined various combinations of input variables based on the literature review summarized in Table 3.3. Specifically, we focused on the BPNN algorithm, one of the widely employed ML methods. The

outcomes of this analysis, highlighting the influence of each predictor and their combinations on model performance, are presented in Table 3.4.

Table 3.4. RMSE of LSWT obtained by means of BPNN for the shallow (depth = 5 m) and deep (depth = 60 m) lake, depending on the predictors used and separating the performances of the train and test data sets. The last column is the error relative to the reference case (defined by the predictors AT + SCDOY, highlighted in bold). The negative values correspond to worsening the results. All RMSE values are averaged over 20 independent runs.

| **Predictor** | **Lake depth (m)** | **Train RMSE (°C)** | **Test RMSE (°C)** | **Difference with respect to the test RMSE (°C) for the case AT+SCDOY** |
|---|---|---|---|---|
| Air Temperature (AT) | *5* | *3.044* | *3.105* | *-1.645* |
| | 60 | 3.629 | 3.911 | -2.619 |
| Day Of the year (DOY) | *5* | *1.856* | *2.125* | *-0.666* |
| | 60 | 1.596 | 1.746 | -0.454 |
| Sine and cosine of the Day Of the year (SCDOY) | *5* | *1.781* | *2.070* | *-0.610* |
| | 60 | 1.434 | 1.579 | -0.286 |
| Shortwave radiation (SWR) | *5* | *5.575* | *5.556* | *-4.096* |
| | 60 | 6.182 | 6.383 | -5.090 |
| Downward longwave radiation (LWR) | *5* | *5.810* | *5.904* | *-4.444* |
| | 60 | 5.300 | 5.285 | -3.993 |
| Relative Humidity (RH) | *5* | *7.928* | *8.141* | *-6.681* |
| | 60 | 7.604 | 7.884 | -6.591 |
| Wind speed (WS) | *5* | *8.030* | *8.452* | *-6.992* |
| | 60 | 7.520 | 7.799 | -6.506 |
| Rainfall (R) | *5* | *8.227* | *8.594* | *-7.134* |
| | 60 | 7.703 | 7.946 | -6.654 |
| Air Pressure (AP) | *5* | *7.520* | *7.884* | *-6.424* |
| | 60 | 7.112 | 7.385 | -6.093 |
| AT + SCDOY | ***5*** | ***1.312*** | ***1.460*** | ***0** (reference case)* |
| | **60** | **1.160** | **1.292** | **0** (reference case) |
| AT + SCDOY + WS | *5* | *1.313* | *1.452* | *0.008* |
| | 60 | 1.151 | 1.294 | -0.002 |
| AT + SCDOY + AP | *5* | *1.308* | *1.455* | *0.005* |
| | 60 | 1.164 | 1.284 | 0.008 |
| AT + SCDOY + SWR | *5* | *1.381* | *1.534* | *-0.074* |
| | 60 | 1.204 | 1.328 | -0.036 |
| AT + SCDOY + LWR | *5* | *1.287* | *1.422* | *0.038* |
| | 60 | 1.153 | 1.282 | 0.010 |
| AT + SCDOY + SWR + LWR | *5* | *1.375* | *1.525* | *-0.066* |
| | 60 | 1.189 | 1.322 | -0.030 |
| AT + SCDOY + SWR + LWR + WS | *5* | *1.369* | *1.509* | *-0.049* |
| | 60 | 1.186 | 1.328 | -0.036 |
| AT + SCDOY + SWR+ LWR+ WS+ RH | *5* | *1.371* | *1.497* | *-0.037* |

| Predictor | Lake depth (m) | Train RMSE (°C) | Test RMSE (°C) | Difference with respect to the test RMSE (°C) for the case AT+SCDOY |
|---|---|---|---|---|
| | 60 | 1.185 | 1.309 | -0.017 |
| AT + SCDOY + SWR + LWR + WS+ RH + R | *5* | *1.358* | *1.484* | *-0.024* |
| | 60 | 1.175 | 1.307 | -0.015 |
| AT + SCDOY + SWR + LWR + WS + RH + R+ AP | *5* | *1.363* | *1.496* | *-0.037* |
| | 60 | 1.176 | 1.301 | -0.009 |

Analyzing individual predictors in isolation, we observed that SCDOY yielded the lowest RMSE. In the deep lake scenario, the training set exhibited an RMSE of 1.43°C, while the test set had an RMSE of 1.58°C. Correspondingly, the shallow case yielded values of approximately 1.78°C and 2.07°C, respectively. These results suggest that an average year effectively captures the intra-annual variability of LSWT. AT ranked second among the predictors; however, the errors significantly increased (3.04°C - 3.11°C for training and test sets in the shallow case, and 3.63°C - 3.91°C in the deep case). The inclusion of other predictors did not yield satisfactory results.

Nevertheless, combining different variables led to notable improvements. Specifically, considering AT and SCDOY together resulted in favorable performance (1.31°C - 1.46°C for the shallow case and 1.16°C - 1.29°C for the deep case), and we established this combination as the reference for subsequent analyses. Further addition of variables to the predictor set yielded only marginal enhancements at the cost of increased information demand by the model (as indicated by the difference in RMSE shown in the last column of Table 3.4). However, we acknowledge that in climatic conditions distinct from the temperate case examined in this study, or in real lake scenarios, other variables may also hold importance.

We provide a more detailed illustration of the impact of considering AT and DOY or SCDOY on the model by examining Figure 3.4. This figure displays the simulated LSWT (using BPNN once again) for both shallow and deep lake cases in individual runs. To enhance clarity, the plots are limited to a two-year period. Due to its lower thermal inertia (heat capacity per unit of surface area), the shallow lake exhibits greater LSWT fluctuations in response to meteorological forcing, as compared to the deep lake (Toffolon et al., 2014, 2020).

The plot clearly demonstrates that when AT is considered alone, the predicted range of LSWT variability is excessively large. On the other hand, relying solely on DOY provides a smoother representation, similar to the mean year, but fails to capture interannual variability. The use of SCDOY, with its gradual transition at the end of each year, proves to be more appropriate. The bottom row of Figure 3.4 highlights the benefits of combining SCDOY and AT to predict LSWT, leading to improved accuracy and capturing both intra- and interannual variability.

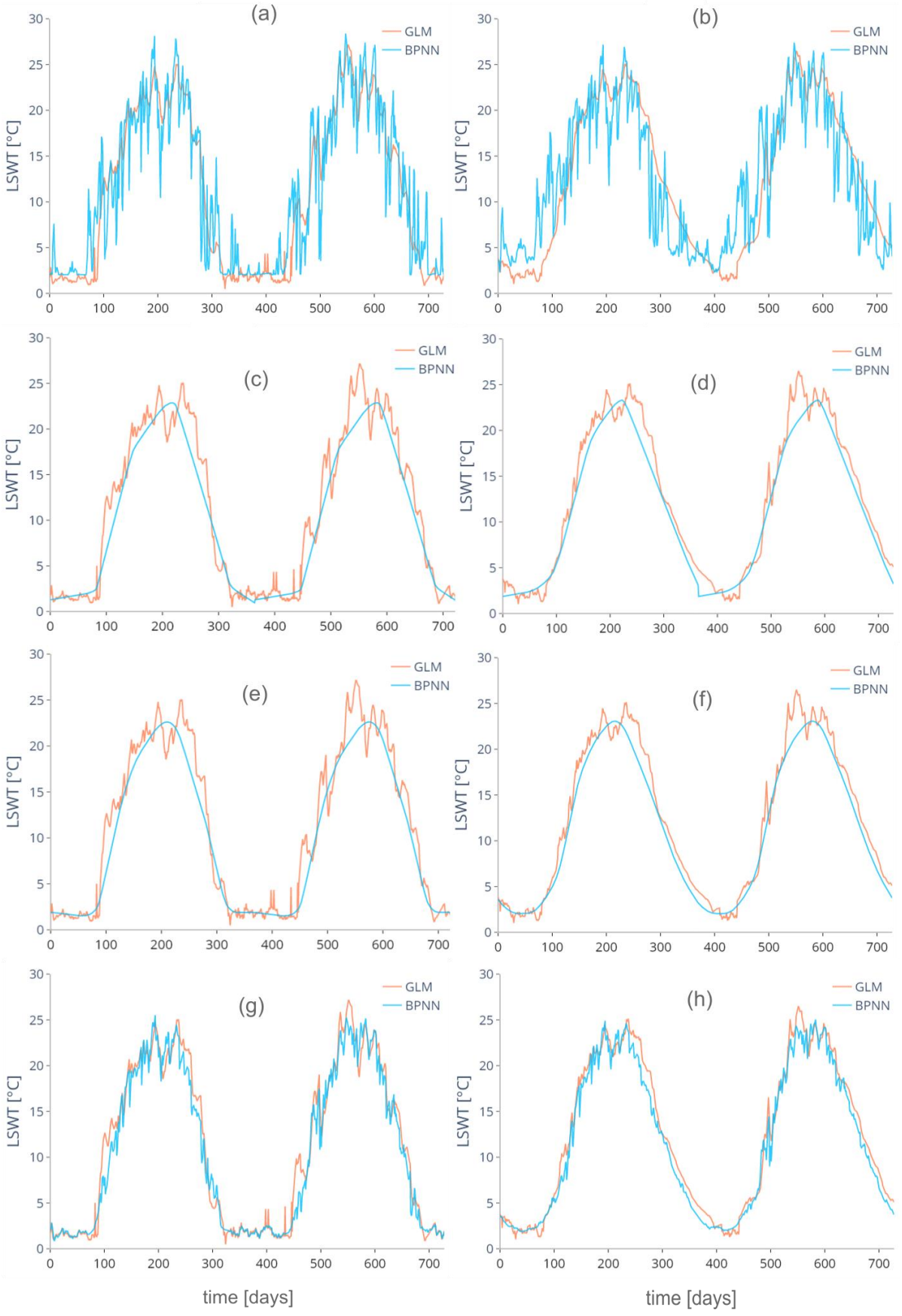



Figure 3.4. Effect of using AT (a,b), DOY (c,d), SCDOY (e,f), and the combinations of AT and SCDOY (g,h) on the LSWT modelled by the BPNN for a shallow (5 m, left column) and a deep (60 m, right column) lake.

### 3.4.2. The importance of pre-processing

Given that the input values exhibit different units and ranges, it is necessary to apply pre-processing methods such as standardization and normalization to eliminate anomalous values and ensure data is on a comparable scale. Sener et al. (2012) utilized the Min-Max transform, scaling input data (RH, AT, AP, and water depth) from 0 to 1, enabling comparability. More advanced pre-processing techniques include moving average with Min-Max (MAMM) and wavelet transform (WT), which not only remove noise from the data but also smooth it. This section analyzes the impact of pre-processing methods and their influence on ML performance. It should be mentioned that in section 2.4, we explained the averaging of all variables and here we just focus on averaging of AT which is the most influential forcing in this study, and we avoid repeating the averaging of all variables here.

We have observed that lakes act as filters, attenuating high-frequency inputs, with this effect becoming more pronounced in deep lakes. This process is dependent on the lake's thermal inertia and should be incorporated in pre-processing if the ML algorithm lacks the ability to capture system memory. When considering AT as a predictor, a general linear relationship between AT and LSWT cannot be established. Instead, a hysteresis effect exists between the two variables (Toffolon et al., 2014). Linear relationships between daily-averaged values may be valid for very shallow lakes but become problematic for deep lakes, where the LSWT of a given day heavily relies on the previous day's value. If the ML algorithm does not adequately account for the system's previous state, pre-processing becomes crucial for a more physically meaningful representation of the lake system.

To simulate the filtering behavior of a lake, a simple pre-processing approach involving data smoothing and denoising was employed using Moving Average (right-aligned window, considering previous days' AT) combined with MAMM. Different window sizes were evaluated to minimize the RMSE. The results demonstrate that the RMSE exhibits a minimum value at an optimal window size, with distinct patterns for shallow and deep lakes (Figure 3.5e,f). In the shallow case, the RMSE initially decreases rapidly and then increases at a similar rate for window sizes longer than the optimal value (Figure 3.5e). In the deep case, the error for the test dataset is less variable compared to the training dataset, although it still exhibits a minimum within the range of window sizes similar to that of the training dataset (Figure 3.5f). Consequently, the optimal window size was selected based on the minimum average of training and test errors, resulting in 8 days for the shallow case and 13 days for the deep case, confirming the influence of lake depth in amplifying the filtering effect on LSWT.

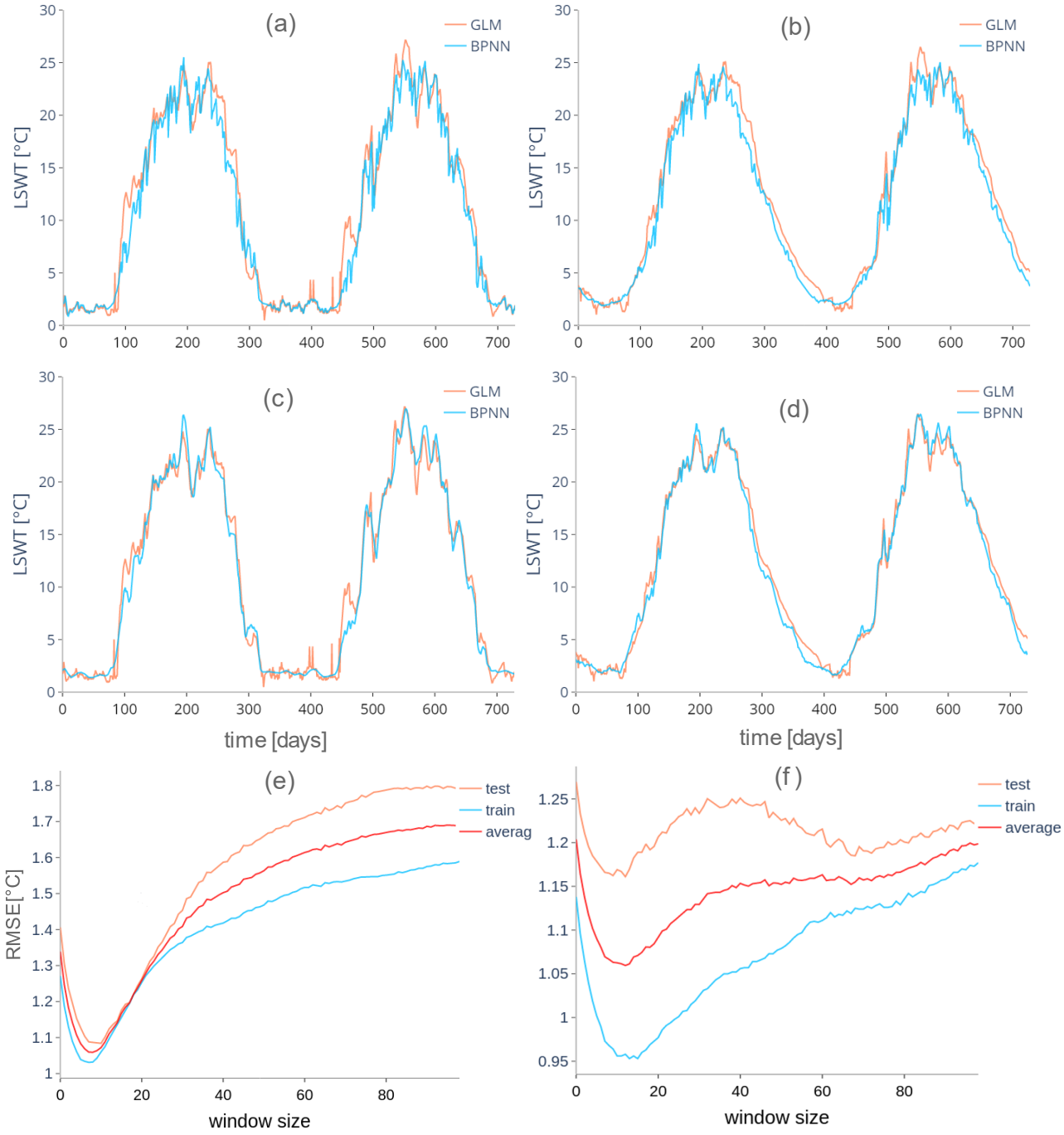



Figure 3.5. Effect of using a moving average for pre-processing the AT data (in addition to SCDOY of the simulated day) to predict LSWT by BPNN in a shallow (left column) and a deep (right column) lake. (a,b) Results in the standard BPNN application, with AT input only from the same day as the predicted LSWT. (c,d) Results with the optimal window size of the moving average. (e-f) Variation of the RMSE as a function of the window size (all RMSE values are averaged over 20 independent runs).

To evaluate the optimal window size based on depth, further analyses are necessary. However, initial insights can be gained by considering two intermediate depths, namely 20 m and 40 m. By comparing the RMSE values across the four cases, a threshold depth appears to influence the qualitative behavior (illustrated in Figure 3.6). While a distinct minimum (representing the optimal window size) is

observed for the shallow lake (5 m), the behavior of the deeper cases shows more similarity. It is worth noting that the 5 m lake does not exhibit stratification (Figure 3.6a), and the heat capacity remains constant throughout all seasons. In contrast, as the depth increases, stratification occurs, and the thermally active layer varies seasonally. In our case study, the thermocline is approximately around 10 m (Figure 3.6b), suggesting that any deeper lake, such as one with a depth of 20 m, would likely exhibit qualitatively similar behavior. We speculate that the variability in heat capacity throughout the year could lead to less pronounced minima in the RMSE curves since the optimal window size may need to change with the season.

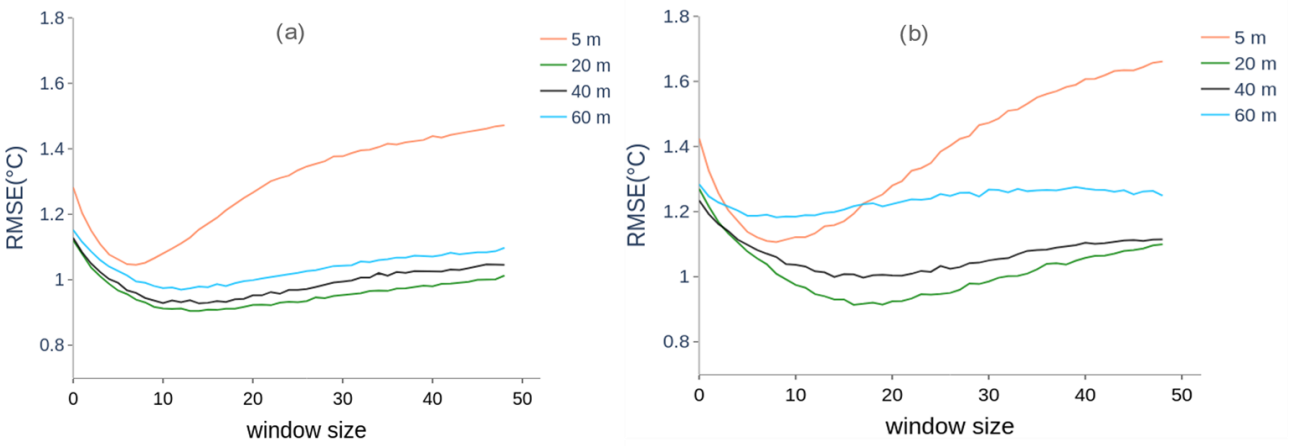



Figure 3.6. The effect of the depth on the window size of the moving average, using AT+SCDOY (ML method: BPNN) for the training (a) and test (b) dataset in different depths (5, 20, 40 and 60 m). All RMSE values are averaged over 20 independent runs.

A more advanced technique for denoising and smoothing the input signal is the WT, which overcomes limitations of the Fourier transform in determining specific frequencies and scales. WT involves fitting a mother wavelet within specified periods and can be categorized as continuous WT (CWT) or discrete WT (DWT), where the data is decomposed into levels. In this study, we utilized the wavelet transform library in Python to determine the optimal mother wavelet among 106 types for DWT and 21 types for CWT. DWT decomposes the raw AT data into coefficients that measure the intensity of signal variations, ultimately resulting in smoothed data. The level of decomposition for the input data was determined as ∫[log_10 (N)], with N representing the length of the time series and the int function rounding the values to the nearest integer.

Applying the different pre-processing methods to the AT data, with BPNN as the ML approach, the results are summarized in Table 3.5. It is observed that the Moving Average with MAMM method outperforms other methods for both shallow and deep lakes, while WT yields the least favorable results. It should be noted that the choice of the most effective mother wavelets may vary depending on the optimization approach employed. Within the Min-Max framework, smoothing the AT data shows a significant improvement compared to the standard approach, with RMSE decreasing from

approximately 1.31-1.46°C (Min-Max) to 1.06-1.14°C (MAMM) in the shallow case, and from 1.16-1.29°C (Min-Max) to 0.98-1.19°C (MAMM) in the deep case.

Table 3.5. Comparison of the performances with different pre-processing methods for a shallow and deep lake (ML method: BPNN). The last row shows the performance of the model that uses also AT values from previous days. All RMSE values are averaged over 20 independent runs.

| **Pre-processing method** | | **Shallow lake (depth = 5 m)** | **Deep lake (depth = 60 m)** |
|---|---|---|---|
| Min-Max | Train RMSE (°C) | 1.312 | 1.160 |
| | Test RMSE (°C) | 1.460 | 1.292 |
| Moving average + Min-Max | Train RMSE (°C) | 1.061 | 0.983 |
| | Test RMSE (°C) | 1.136 | 1.189 |
| | Best window size | 8 | 13 |
| DWT | Train RMSE (°C) | 1.754 | 1.365 |
| | Test RMSE (°C) | 1.968 | 1.538 |
| | Best Wavelet | 'coif8' | 'coif16' |
| CWT | Train RMSE (°C) | 1.801 | 1.421 |
| | Test RMSE (°C) | 2.085 | 1.558 |
| | Best Wavelet | 'cgau4' | 'fbsp' |
| AT from previous days + Min-Max | Training RMSE (°C) | 0.959 | 0.887 |
| | Test RMSE (°C) | 0.943 | 1.136 |
| | Best number of previous days | 15 | 20 |

### 3.4.3. Retaining the history of the forcing

In conventional ML approaches (excluding LSTM), such as the BPNN method considered as our reference, the system's historical information is not explicitly retained. Therefore, pre-processing the input data can be beneficial, as demonstrated in the previous section. However, utilizing a simple moving average over an optimal period to capture the effect of previous AT values is a crude approach, as it assigns equal weight to all days within the averaging window, disregarding the potentially stronger impact of more recent days. Various methods can address this limitation, such as assigning different weights based on the temporal distance from the simulated day. In this section, we explore an alternative strategy that explicitly incorporates past inputs into the ML model.

We continue to employ BPNN as the reference ML algorithm, with SCDOY and AT as the most relevant input data. However, while the use of SCDOY remains unchanged, the AT record is duplicated multiple times with a variable shift of N days: AT synchronized with LSWT, AT from the previous day, AT from two days prior, and so on. Consequently, the number of input nodes in the BPNN increases by N-1. By explicitly including the previous forcing conditions (assuming AT as a

proxy) in the ML algorithm, the model weighs them based on their actual relevance to LSWT. The performance of the BPNN algorithm significantly improves (Figure 3.7) as we add an increasing number of previous days, until the improvement reaches a saturation point, beyond which additional information becomes irrelevant. Optimal performance is achieved by considering approximately 15 days for the shallow lake (Figure 3.7a) and about 20 days for the deep lake (Figure 3.7b), confirming that the system's memory is relatively long and slightly increases with depth. The last row in Table 3.5 illustrates that incorporating AT values from previous days leads to considerable improvement (0.96-0.94°C for the shallow lake and 0.88-1.14°C for the deep lakes). While both the moving average and the utilization of single values from previous days as additional inputs show similarities, we included both approaches in Table 3.5. However, it is evident that the latter approach is not a pre-processing method like those discussed in Section 3.4.2.

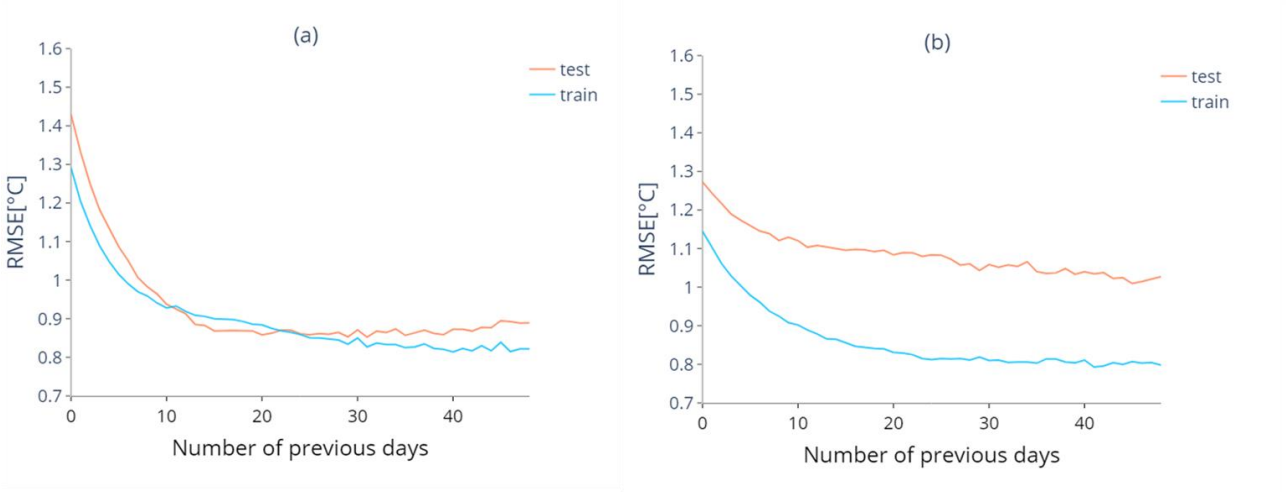



Figure 3.7. Variability of the RMSE using AT from a variable number of previous days in the shallow (a, 5 m) and deep (b, 60 m) lake. All RMSE values are averaged over 20 independent runs.

Figure 3.8 provides further evidence that incorporating AT values from previous days as additional inputs yields greater improvements in the model's performance compared to applying a moving average to the AT data. It is noteworthy that both methods outperform the approach of considering all input variables together on the same day as LSWT. The combination of SCDOY and the historical AT values proves to be the most effective input configuration for accurately simulating LSWT in lakes of this nature.

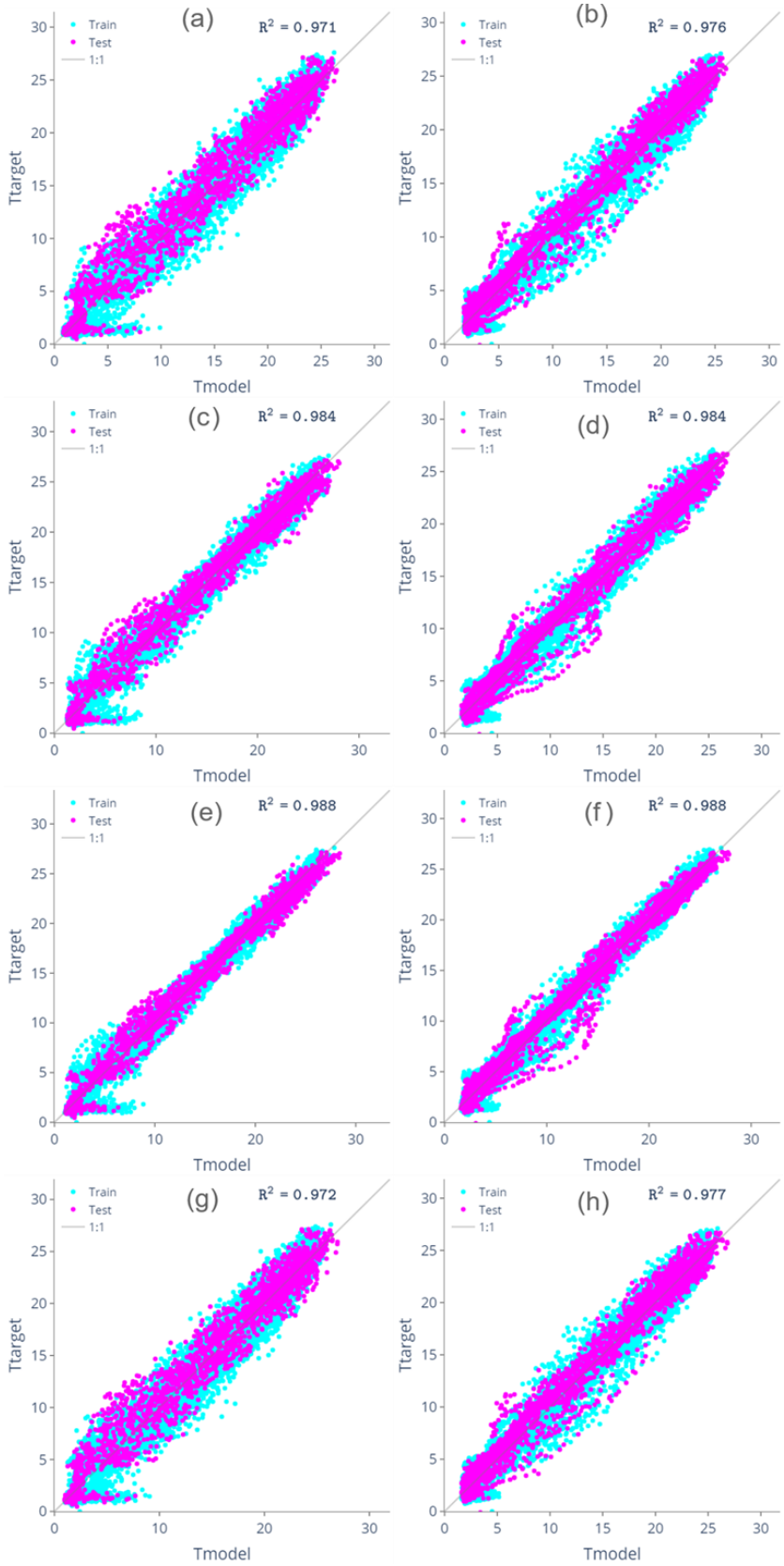



Figure 3.8. Scatter plots of observed and predicted LSWT using AT + SCDOY as predictors for the shallow lake (left column, subplots a, c, e, g) and the deep lake (right column, subplots b, d, f, h) for different ways to include AT in the inputs: (a, b) reference case (using AT and LSWT of the same day); (c, d) moving average of AT with windows of 8 (shallow) and 13 (deep) days; (e, f) ATs from the previous 15 and 20 days for shallow and deep lakes, respectively, as additional inputs; (g, h) considering all meteorological predictors of the same day as LSWT.

Figure 3.9 presents an alternative representation of Figure 3.8, reinforcing the observation that integrating Air Temperature (AT) values, both averaged and from preceding days, yields superior results compared to the utilization of AT+SCDOY and all variables. This underscores the significance of considering the specific contributions of AT values in enhancing model performance.

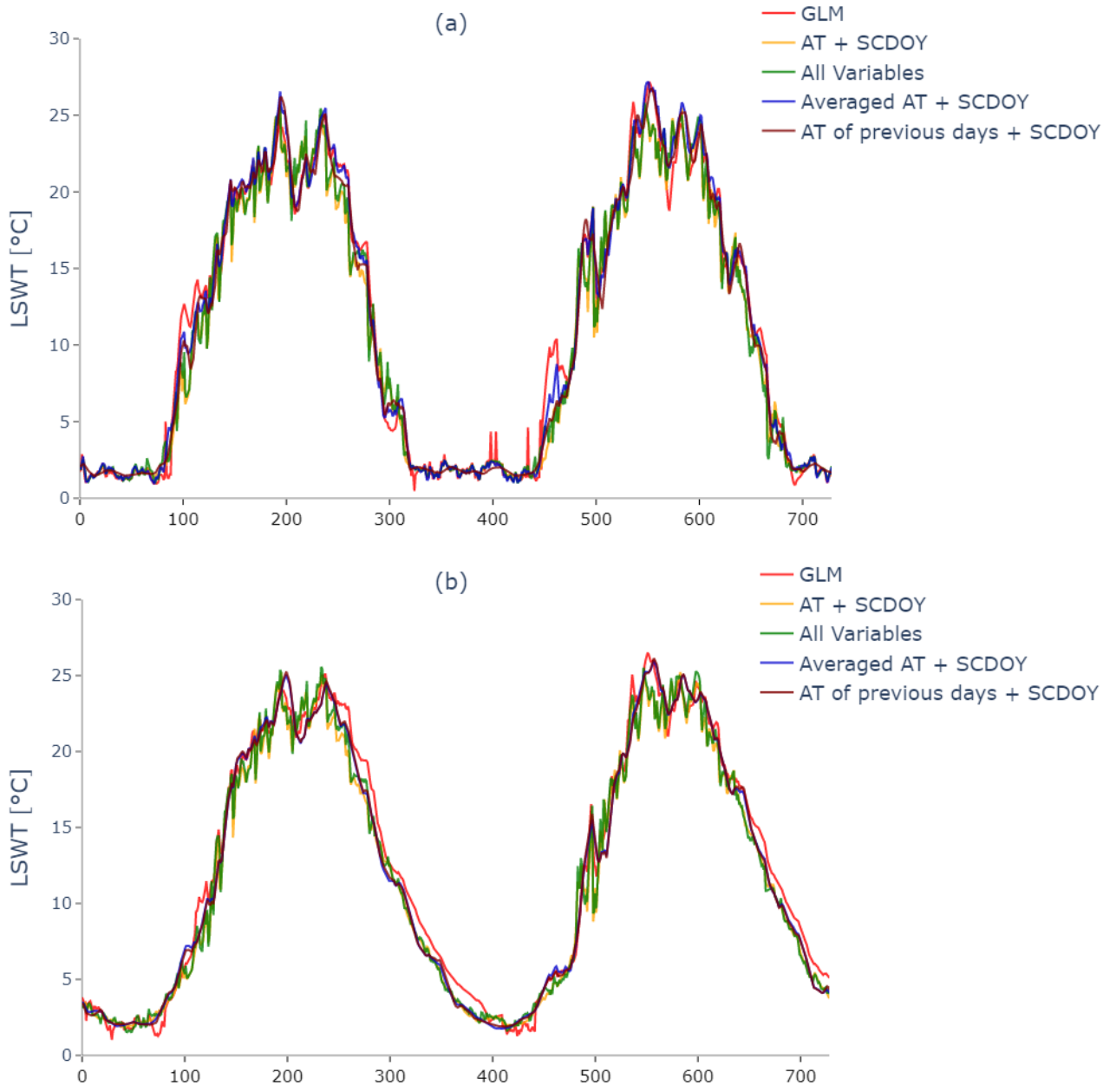



Figure 3.9. Plots of observed and predicted LSWT using as predictors AT + SCDOY, all variables, averaged AT +SCDOY and AT of previous days + SCDOY for the shallow lake (a) and the deep lake (b) . All RMSE values are averaged over 20 independent runs.

### 3.4.4. The choice of the ML algorithm

Based on our comprehensive literature review (summarized in Table 3.1), the ANN emerges as the most widely employed ML approach. However, numerous other methods have also been proposed in the field. In this study, we aim to delve into the specifics of various ML algorithms and subsequently compare their performances in synthetic case studies, encompassing both shallow and deep lakes.

For our comparative analysis, we selected nine commonly utilized ML algorithms, taking into account their applicability to stream temperature prediction as well. These algorithms include MLPNN, BPNN, LSTM, ANFIS, DT, RF, ERT, KNN, and SVR (Table 3.6). Detailed information regarding these methods can be found in section 2.2. To ensure a fair and consistent comparison, we established a standardized procedure for the analysis. This involved utilizing the same predictors (AT and SCDOY), optimizing the hyperparameters using the GA optimization technique, and applying Min-Max scaling as the pre-processing method for the input data without any additional smoothing. When evaluating the performances, we focused on the RMSE for both the training and test data sets separately. It is important to note that a low error for the training set and a high error for the test set indicate overfitting, where the model fails to generalize effectively to new data sets.

Table 3.6. Comparison of the performances among different ML algorithms: RMSE of train and test data set using AT and SCDOY as input (with MM for pre-processing). A negative difference indicates the worsening of the performances for the test with respect to the train data set. All RMSE values are averaged over 20 independent runs.

| **Method** | **Phase** | **RMSE (°C)** | |
|---|---|---|---|
| | | **Depth = 5 m** | **Depth = 60 m** |
| Decision tree (DT) | *Train* | *1.230* | *1.428* |
| | Test | 1.626 | 1.668 |
| | (difference) | (-0.395) | (-0.240) |
| Random forest (RF) | *Train* | *1.159* | *1.117* |
| | Test | 1.540 | 1.362 |
| | (difference) | (-0.381) | (-0.246) |
| Extremely randomized tree (ERT) | *Train* | *1.295* | *1.141* |
| | Test | 1.479 | 1.292 |
| | (difference) | (-0.184) | (-0.150) |
| K-Nearest neighbuor (KNN) | *Train* | *1.243* | *1.093* |
| | Test | 1.525 | 1.311 |
| | (difference) | (-0.283) | (-0.218) |
| Support vector regression (SVR) | *Train* | *1.369* | *1.171* |
| | Test | 1.497 | 1.302 |
| | (difference) | (-0.128) | (-0.131) |
| Multi-layer perceptron neural network (MLPNN) | *Train* | *1.319* | *1.212* |
| | Test | 1.467 | 1.378 |

| | | | |
|---|---|---|---|
| | (difference) | (-0.148) | (-0.166) |
| Long short-term memory (LSTM) | *Train* | *1.353* | *1.286* |
| | Test | 1.490 | 1.264 |
| | (difference) | (-0.137) | (0.022) |
| Backpropagation neural network (BPNN) | *Train* | *1.312* | *1.160* |
| | Test | 1.460 | 1.292 |
| | (difference) | (-0.148) | (-0.133) |
| Adaptive neuro fuzzy inference system (ANFIS) | *Train* | *2.911* | *3.743* |
| | Test | 2.848 | 3.456 |
| | (difference) | (0.063) | (0.287) |

As shown in Table 3.6, the performances of most ML algorithms were comparable for the case of the shallow lake: RMSE in the range 1.16-1.37°C for the training data set, with the exceptions of ANFIS providing the worst results (2.91°C), and in the range 1.46–1.63°C for the test data set (2.85°C for ANFIS). In the case of the deep lake, the RMSE for the training data set ranges from 1.46 to 1.63°C for most ML algorithms, with ANFIS exhibiting the poorest performance (2.91°C). Similarly, for the test data set, the RMSE ranges from 1.26 to 1.67°C, with ANFIS again producing unacceptable results (3.46°C). Notably, RF, which initially performed well during training, experiences a significant deterioration in performance for the test set, with an increase in RMSE of 0.38°C for the shallow lake and 0.25°C for the deep lake. On the other hand, LSTM, which preserves information from previous states, performs effectively for the deep lake in both the training and test data sets, with only a slight difference of 0.02°C between the two. This observation aligns with the notion that lakes retain the memory of previous conditions, indicating that satisfactory results can be achieved without extensive pre-processing of the data involving AT smoothing or the inclusion of previous AT values.

To gain further insights into the performance of ML methods, it is beneficial to consider the characteristics of specific algorithms. Let's examine the performances of two similar methods: ERT and RF. ERT is a modified version of DT that selects model parameters randomly and trains the model repeatedly using the data. In contrast, RF determines node splits based on optimal decisions, while ERT randomly splits nodes. Moreover, RF divides inputs into subsamples, allowing replacement and resulting in repeated data, whereas ERT avoids data repetition by not employing replacement during input selection (Geurts et al., 2006). The results presented in Table 3.6 indicate that ERT exhibits greater robustness, as the decline in performance from training to test is less pronounced compared to RF.

The findings suggest that LSTM, MLPNN, and BPNN are successful methods for predicting LSWT. These methods demonstrate the smallest difference in RMSE between the training and test data sets

among the various ML algorithms (Table 3.6). Notably, these methods belong to the domain of deep learning (DL), which implies that DL is capable of extrapolating favorable performances to the test data set more effectively than other ML approaches. DL models consist of multiple hidden layers, which introduce more degrees of freedom in weight and bias combinations (the parameters calibrated during training), leading to more accurate results. Conversely, ANFIS, which leverages the structure of ANN but incorporates fuzzy logic in its hidden layer (further details in section 2.2), yields less accurate predictions compared to the other ML methods examined in this study.

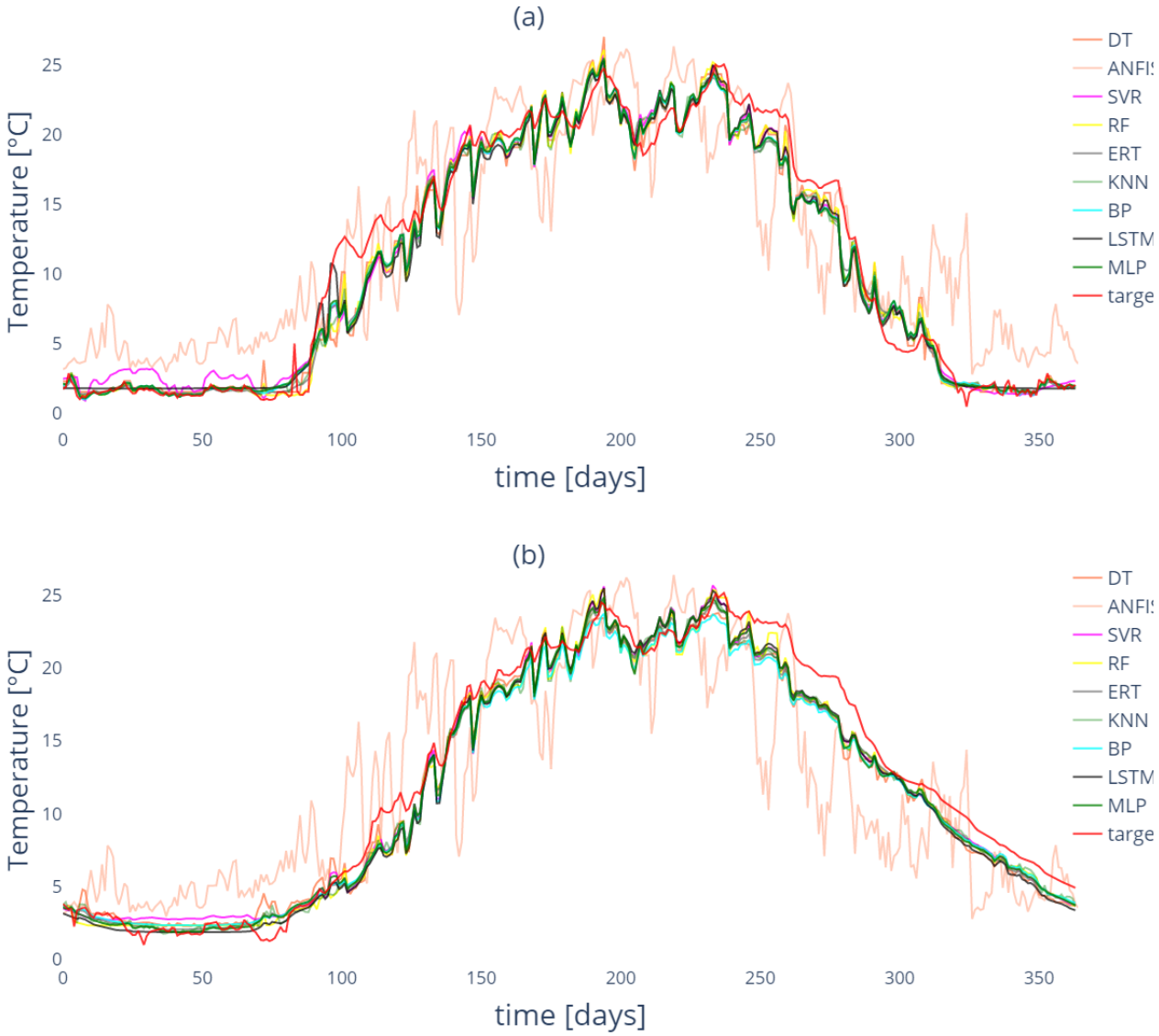



Figure 3.10. The evolution of LSWT in a single year (2015) as predicted by different ML approaches using AT+ SCDOY for the shallow (a) and deep (b) lake; the target is LSWT simulated by the physically based model GLM.

The behavior of shallow and deep lakes can be distinguished by the intensity of fluctuations in LSWT. The shallow lake exhibits significant variability on a short time scale, whereas the deep lake experiences smaller fluctuations (as depicted in Figure 3.10) (Ding and Mao, 2021). However, the

damped variability observed in the deep lake is confined to periods when stratification does not occur. During summer, LSWT exhibits increased fluctuations in the deep lake due to the smaller surface volume's response to heat flux. Replicating such behavior poses a challenge for ML methods. DL algorithms demonstrate superior ability to track the target (LSWT simulated by GLM model) compared to other ML methods. On the other hand, the poorest performing method, ANFIS, yields less accurate results.

### 3.4.5. The benchmark of a physically based data-driven model

To establish a benchmark for evaluating the performance of ML models, we compared the results with those obtained from the physically based, data-driven model called air2water (Toffolon et al., 2014). The input data for the 6-parameter version of air2water consists of AT and DOY. This model incorporates a differential equation that considers the time derivative of LSWT, allowing for the retention of the lake's thermal inertia. By calibrating the parameters of air2water, we achieved values of 1.098°C and 1.078°C (for training and test data, respectively) in the case of the shallow lake, and 0.946°C and 1.021°C in the deep lake. Thus, air2water exhibits performance comparable to the best ML models (Table 3.6), and only slightly lags behind the modified BPNN with previous days' AT values (Table 3.5).

It is noteworthy that air2water retains the memory of previous conditions through the time derivatives of LSWT, similar to stochastic autoregressive algorithms. Additionally, air2water incorporates a parameter that adjusts the thermal inertia of the system based on a simplified representation of seasonal stratification (Toffolon et al., 2014). LSTM, on the other hand, follows a similar approach by employing a complex structure of ANN with internal loops. In contrast, the inclusion of previous days' AT values, as demonstrated in our modified BPNN application (Figure 3.8g, h), explicitly incorporates the historical forcing data rather than the state of the physical system (such as the previous day's LSWT) (Austin and Colman, 2007).

### 3.4.6. Discussion

ML has proven to be a successful approach for predicting LSWT, and it offers advantages over physically based models by not requiring a wide range of data on lake morphology, meteorological forcing, and heat fluxes. However, the selection of appropriate predictors plays a crucial role in the success of ML models. Our analysis of artificial case studies reveals that the most influential inputs for forecasting LSWT are DOY or alternatively, SCDOY, and AT. While other predictor combinations may slightly improve accuracy, they require additional data. The significance of AT is expected, as it impacts various components of the heat flux and serves as a proxy for meteorological conditions.

More interestingly, DOY (or SCDOY) allows for the reproduction of the climatological mean year, capturing the annual cycle of LSWT that repeats approximately every year (Toffolon et al., 2020).

In addition to its climatological significance, DOY contributes to the inclusion of seasonal variability in stratification, which is particularly important in lakes deep enough to experience periods of non-mixing. The intra-annual variability range between minimum and maximum LSWT is substantial in temperate lakes, with inter-annual variability often making a minor contribution to the overall RMSE (Piccolroaz et al., 2016). The morphological and hydrological features of lakes are known to affect their thermal response. Although we focused on two depth cases (shallow: 5 m, and deep: 60 m) and did not conduct a comprehensive comparative analysis across multiple lakes, our findings suggest that the smoother behavior observed in deep lakes, attributed to their larger heat capacity, may result in lower RMSE values compared to shallow lakes. However, it should be noted that the dynamics of deeper lakes are generally more complex due to the nonlinear effect of stratification on LSWT.

Optimization and pre-processing methods are well-known techniques to enhance the performance of ML models (Isik et al., 2012). In our study, we observed that applying a moving average to the input AT effectively smoothed the high-frequency response of LSWT, mimicking the physical filtering effect of the larger water mass in a lake. Another approach to mitigate LSWT oscillations involves reducing the influence of day-to-day AT variability by incorporating AT values from previous days.

Overall, ML models demonstrate their utility in LSWT prediction, and careful selection of predictors, consideration of seasonal variability, and appropriate pre-processing techniques can further improve their performance (Isik et al., 2012; Toffolon et al., 2020; Piccolroaz et al., 2016).

In our modeling exercise, we did not identify a single ML approach that consistently outperformed the others. While a group of methods showed reasonably good performance (varying depending on the shallow/deep case), only ANFIS exhibited unacceptable errors. Our findings align with the results of our literature review, which also lacked clear indications of a dominant ML approach for LSWT predictions (Table 3.3). The most commonly used model in the literature is and its variations, which have become somewhat of a standard for LSWT predictions. However, it is worth noting that traditional methods sometimes outperformed advanced algorithms. For example, Sharma et al. (2008) demonstrated that multiple regression, which considers linear relationships, can be more effective than ANN. Similarly, Quan et al. (2020) found that SVR with GA outperformed ANN in reservoir applications. Heddam et al. (2020) reported that ERT was more robust than ANN and RF for LSWT prediction in shallow lakes, while also requiring less computation time than ANN. Interestingly, according to their study, a physically based data-driven model called air2water outperformed all the ML methods, confirming the outcome of our analysis. Liu and Chen (2012) compared the accuracy

of ANN with a 3D hybrid model in the shallow Yuan-Yang Lake (maximum depth 4.5 m) and found that ANN was less accurate in forecasting water temperature below the surface compared to LSWT. Similarly, Samadifard et al. (2016) investigated ANN, gene expression programming (GEP), and ANFIS in the Yuan-Yang Lake and, consistent with our results, demonstrated that ANN generally outperformed ANFIS. It should be noted that achieving high performance with ANN requires careful optimization method selection, as well as an appropriate number of hidden layers and neurons. Additionally, different branches of ANN, such as MLPNN, RNN, and BPNN, are available.

The incorporation of physical constraints can help ML models achieve more accurate predictions. In our analysis, we focused on pre-processing AT data by including information from previous days to simulate the combined effect of historical forcing and system memory. Other researchers have directly included constraints within their models. For instance, Jia et al. (2019) added two physical constraints to their LSTM model for simulating temperature profiles in Lake Mendota: the requirement that density increases with depth and the conservation of energy. Furthermore, Jia et al. (2021) pretrained their LSTM model using a physically based model in cases of data deficiency to improve accuracy, effectively creating a hybrid model that combines ML with physics-based principles.

While ML methods generally perform better than linear regression models, physically based data-driven models like air2water tend to outperform most ML methods we investigated. These models retain some physical principles while offering flexibility through specific parameter calibration for different case studies. This raises the question of why ML approaches are used. We believe that the flexibility of ML represents a significant advantage when the underlying physics of the system is not sufficiently understood or when simplified models are not available for a specific class of problems. This is particularly relevant in global-scale analyses that consider different types of lakes using the same approach. For example, Piccolroaz et al. (2020) demonstrated that air2water performs well for temperate lakes but shows diminished performance for tropical and equatorial lakes, where other factors may become as relevant as Air AT. Thus, conducting a global analysis of LSWT in lakes using ML methods deserves further attention.

## 3.5. Conclusions

Predicting the water temperature of freshwater bodies, including LSWT, can be achieved using various methods, with ML algorithms gaining popularity as an effective approach. In this study, we conducted a comprehensive review of common ML approaches for LSWT prediction and aimed to understand the factors influencing their performance. Specifically, we investigated the impact of different predictors used as input data, the potential improvement gained by considering the history

of forcing and system memory in thermal response (such as smoothing AT input or incorporating previous days' values), and the performance of various ML approaches. Our comparative analysis was based on a literature review and the application of nine ML methods to two deliberately designed case studies: a shallow lake (maximum depth 5 m) and a deep lake (maximum depth 60 m). In these case studies, LSWT was artificially generated using a physically based one-dimensional model (GLM). By utilizing synthetic cases, we were able to control all relevant variables and compare the shallow and deep cases without external factors affecting the thermal response. However, it is important to note that these results are derived from synthetic cases and may not fully represent real lakes.

The outcomes of our analysis, using BPNN as the reference ML approach, revealed that a combination of AT and DOY - particularly with a continuous representation of DOY using SCDOY - provided the minimum necessary information for obtaining reasonable results. Further inclusion of additional variables did not lead to significant improvements. Additionally, we demonstrated that smoothing the AT input signal, such as employing a moving average, notably enhanced the predictive performance. Another effective approach, albeit more complex, involved incorporating previous days' AT values as inputs to the ML model to capture the history of forcing conditions. This finding underscores the importance of incorporating memory effects in ML approaches that do not inherently account for them. While we only tested LSTM as an option, it proved beneficial. Finally, the comparison of nine ML algorithms did not reveal substantial differences in performance, although LSTM exhibited slightly greater robustness compared to other approaches, while ANFIS consistently produced unsatisfactory results. Consequently, our analysis suggests that the success of ML applications for LSWT prediction primarily relies on identifying appropriate predictors, while the choice of the best ML algorithm remains subjective and lacks a definitive solution.

# Chapter 4. The influence of climate on the lakes' thermal response

## 4.1. Introduction

Lake surface water temperature (LSWT) affects the period and severity of lake overturn, which often results in changes the quality of water (Ptak and Nowak, 2016; Missaghi et al., 2017; Gray et al., 2019). LSWT is increasing due to global warming (about 0.3 °C per decade) (Schneider and Hook, 2010; O'Reilly et al., 2012; Ptak et al., 2020). For instance, Dokulil et al. (2021) showed the increase of +0.58 °C decade$^{-1}$ and +0.42 °C decade$^{-1}$ for the annual maximum LSWT and air temperature, respectively, in the same period.

Predictions of LSWT has driven studies measuring the trend of global warming and its effect on the LSWT, also taking into account the impacts of other meteorological variables such as shortwave radiation. To develop LSWT models for predictive purposes, two primary types of models are introduced, those that are physically based and those that are data-driven. Physically based models, including fundamental equations of the lakes' heat budget, demand not only the meteorological variables but also data such as inflow, outflow and lake's physical features (Piotrowski and Napiorkowski, 2018). Data-driven models utilize observed LSWT and meteorological variables without requiring knowledge of the physics of lakes. Data-driven models consist of machine learning (ML) methods, in which thanks to available observed output(s), we can focus on supervised branch. ML approaches are more flexible than physically based models due to removing or changing the predictors for further analysis.

Different ML approaches are utilized to predict LSWT such as multiple regression model, based on linear relationships, (Sharma et al., 2008; Zhu et al., 2020b), random forest (Heddam et al., 2020), support vector machine (SVM) (Quan et al., 2020), artificial neural network (ANN) (Sharma et al., 2008; Sener et al., 2012; Read et al., 2019; Zhu et al., 2020a). The deep learning methods (ANN with more than one hidden layer) are successful popular approaches utilized in many studies. Although, in Sharma et al. (2008) study, multiple regression model outperformed ANN. Moreover, Quan et al. (2020) showed that using genetic algorithm (GA) as the optimization for SVM makes this method more accurate than ANN.

As input(s) of ML to predict LSWT some variables are used such as air temperature (AT), wind speed, shortwave and downward longwave radiations, relative (or specific) humidity, rainfall, water depth and air pressure, among which AT is one of the most influential indicators (Sharma et al. 2008; Sener et al. 2012; Read et al. 2019; Heddam et al. 2020; Zhu et al. 2020b).

While previous studies, mentioned in this section, employed meteorological variables to predict LSWT using ML for specific cases (with the exceptions of Sharma et al., 2008 and Read et al., 2019), it remains uncertain whether the identified influential variables hold equal significance for all other lakes. None of these studies have considered the relevance of meteorological variables across diverse climate regions. This study aims to fill this gap by examining the impact of meteorological variables on LSWT in 2024 lakes worldwide. Utilizing Backpropagation Neural Network (BPNN) as an ML approach, we seek to understand how the importance of meteorological variables in LSWT prediction is influenced by different climate regions. Essentially, this study extends the findings of Chapter 3 by generalizing them with optimized models. We also investigated the accuracy of BPNN as a ML approach to check whether it would be useful for prediction in different cases, having low number of valid data with reasonable quality by remote sensing.

## 4.2. Methodology

The implementation of the model entails data preparation and the utilization of ML model for simulation. Recognizing the fact that lakes retain a memory of past conditions and forcing events, we took into account weighted average of the forcing variables over a specific time span preceding the day of simulation (explained in section 2.4). We identified the optimal time span for each lake, during which the model achieves the lowest RMSE. This strategy was implemented to address the challenge posed by the memory effects of lakes, enabling us to improve the predictive capabilities of our model.

The sequential progression of the methodology is depicted in Figure 4.1 through a visual representation in the form of a flowchart and subsequently explained in the following description.

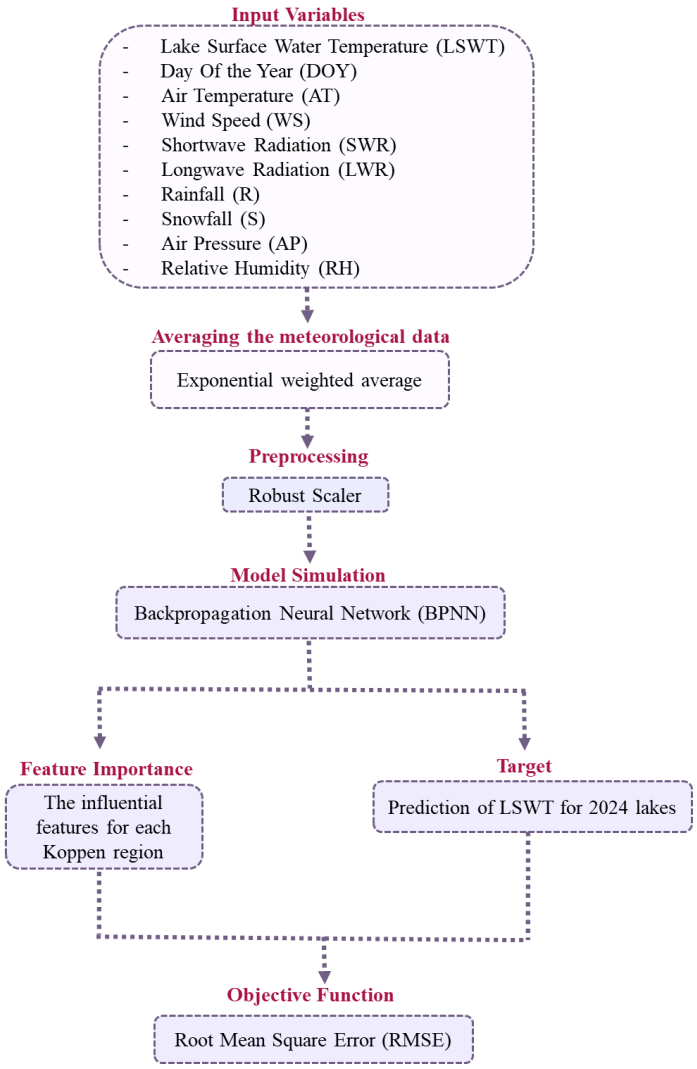



Figure 4.1. Framework.

### 4.2.1. Case Study

This study investigates 2024 lakes, the descriptions of meteorological variables in all lakes can be found in Table 4.1. The meteorological data and observed LSWT of the lakes are taken from ERA5 (2017b), which is provided by the Climate Change Service (Hersbach, 2020), and ESA-CCI dataset providing long-term global LSWT data (http://cci.esa.int/lakes), respectively. Based on the Köppen climate classification (explained in section 2.1.5), lakes are categorized into five distinct regions, each representing different climatic conditions: tropical, dry, temperate, snow and polar. The respective ranges for each variable in each region are detailed in Table 4.1.

Table 4.1. Ranges of (daily mean) meteorological variables and LSWT.

| Regions / Parameters | Tropical | Dry | Temperate | Snow | Polar |
|---|---|---|---|---|---|
| AT (°C) | 7.23 – 40.32 | -42.28 – 42.17 | -28.05 – 40.59 | -64.69 – 36.27 | -63.31 – 29.60 |
| WS ($ms^{-1}$) | 0.40 – 17.13 | 0.54 – 19.39 | 0.33 – 19.89 | 0.33 – 20.08 | 0.21 – 23.24 |
| SWR ($Wm^{-2}$) | 8.75– 375.74 | 4.42 – 425.92 | 0.14 – 435.50 | 0 – 399.04 | 0 – 426.15 |
| LWR ($Wm^{-2}$) | 245.65 – 473.15 | 84.47 – 481.42 | 126.99 – 468.70 | 73.88 – 444.64 | 66.81 – 426.05 |
| AP (Pa) | 71083.60 – 103598 | 54204.20 – 105421 | 60866.1 – 104827 | 53143.1 – 105968 | 48887.5 – 106025 |
| RH ($kgkg^{-1}$) | 8430.54 – 646156 | 3.12 – 628073 | 52.63 – 581016 | 0.01 – 420107 | 0.02 – 275082 |
| R ($kgm^{-2}s^{-1}$)* | 0 – 40.81e-4 | 0 – 23.24e-4 | 0 – 38.28e-4 | 0 – 15.24e-4 | 0 – 16.70e-4 |
| S ($kgm^{-2}s^{-1}$)* | 0 – 0 | 0 – 8.02e-4 | 0 – 14.60e-4 | 0 – 16.13e-4 | 0 – 16.73e-4 |
| LSWT (°C) | 13.95 – 37.64 | 0 – 40.89 | 0 – 36.77 | 0 – 36.82 | 0 – 24.62 |

*The unit is equal to mm/s

### 4.2.2. Data acquisition

The input data of ML consists of meteorological variables, Day of the year (DOY). Observed LSWT (°C) that is lake surface skin temperature used as the target for incorporation into the objective function during model training. The meteorological retrieved from ERA5, explained in section 2.1.4, variables include air temperature (°C) 2 m above the lake surface, wind speed ($ms^{-1}$), air pressure (Pa), rainfall and snowfall ($kgm^{-2}s^{-1}$), specific humidity ($kgkg^{-1}$) converted to relative humidity ($kgkg^{-1}$), surface downward longwave radiations and shortwave radiations ($Wm^{-2}$). The detailed information of these variables is available at https://cds.climate.copernicus.eu/cdsapp#!/dataset/derived-near-surface-meteorological-variables.

As shown in Chapter 3 (Figure 3.4), DOY jumps from 365 to 1 when the year changes, and the discontinuity does not correspond to the continuous change in LSWT. Accordingly, we applied the sine and cosine of DOY (SCDOY), i.e., sin(DOY/nDOY), abbreviated as SDOY, and cos(DOY/nDOY), abbreviated as CDOY, where nDOY is the number of days in a year. By representing the DOY using the sine and cosine functions, we can capture the same information while facilitating a smoother transition between the end of each year and the beginning of the following year (Yousefi and Toffolon, 2022).

### 4.2.3. Preprocessing

In ML applications, it is customary to normalize and standardize data, particularly when their range of variation spans different scales (refer to Table 4.1). Data scaling allows for the evaluation of various features expressed in different units and enhances the overall results (Fahle et al., 2020). Several scaling methods exist, and after testing Min-Max, Standard, and Robust scalers (Varoquaux et al., 2015), we determined that the Min-Max scaler yields better outcomes in our cases (also shown for the synthetic lake in Chapter 3).

The process of splitting the dataset into training and test sets is a fundamental step in ML, as it enables the evaluation of the model's performance on unseen data. There are multiple ways to split the dataset, each with its own advantages and limitations. Based on trial and error and the need to have available measured LSWT (not daily values) in the last fold for all cases, we chose the $5^{th}$ fold out of 5 folds as the test set and other folds as the training set for all lakes for sake of simplicity.

### 4.2.4. Machine learning model

ANNs are intricate systems designed to imitate pattern recognition and prediction. They comprise interconnected processors and layers of neurons that process input data through hidden layers to produce simulation results. The weights and biases of the neurons are continuously optimized to minimize discrepancies between the output and the desired target. Within this framework, the backpropagation neural network (BPNN), a prominent ANN architecture, assumes a crucial role by updating the weights backward, in contrast to the forward weight updates of feedforward ANN. This method is explained in detail in section 2.2.1.

Achieving optimal performance with ML models necessitates the meticulous selection of relevant hyperparameters. These user-defined parameters exert significant influence over the algorithms' learning dynamics and predictive accuracy. Hence, we experimented with different hyperparameters and opted for the most effective ones that proved suitable for all 2024 lakes (represented in Appendix A.2). Accordingly, we opted for one set of hyperparameters for all the lakes, while the model's parameters were tuned for each individual lake.

To ensure the robustness and stability of the results, it is advisable to run the model multiple times, where the initial values of weights and biases are randomly selected for each run. In this study, we conducted the model running process 20 times and subsequently calculated the average of the obtained results. This approach of averaging the outputs from multiple runs enhances the reliability of the outcomes (Yousefi and Toffolon, 2022).

This study assessed the model's performance through the examination of two primary performance metrics: the root means square error (RMSE) and the modified Nash-Sutcliffe efficiency index (NSE*), introduced in section 2.3. Both indicators were evaluated for both the training and test sets, as detailed in previous studies (Nevitt and Hancoc, 2000; McCuen et al., 2006). The RMSE and NSE* are obtained from average on 20 predicted LSWT.

## 4.3. Results

This section of the thesis represents the comprehensive analysis of features correlations, the performance of the BPNN model in predicting LSWT for a dataset encompassing 2024 lakes. Additionally, we performed the feature ranking for each Köppen region to determine the most influential factors.

### 4.3.1. Optimal average window size

Taking into account the impact of preceding forcing events on the current state of the system allowed us to gain insight into the underlying dynamics and improve ML predictions. Different from the analysis in Chapter 3, here we used the exponential weighted average of each variable (as explained in section 2.4) and determined the best window size based on BPNN model performance in a way to have the lowest RMSE of the model. Through the examination of individual time spans for each lake over a period of 90 days, and subsequent calculation of the corresponding RMSE, we identified the time span that corresponds to the minimum RMSE for each lake. This approach enabled us to assess the significance of temporal dependencies and forcing memories of previous days.

Figure 4.2 illustrates the optimal time window for calculating the average across all lakes in each region. Overall, the average window sizes seem to exhibit a relatively uniform pattern across all regions, with the temperate region displaying the largest window size values.

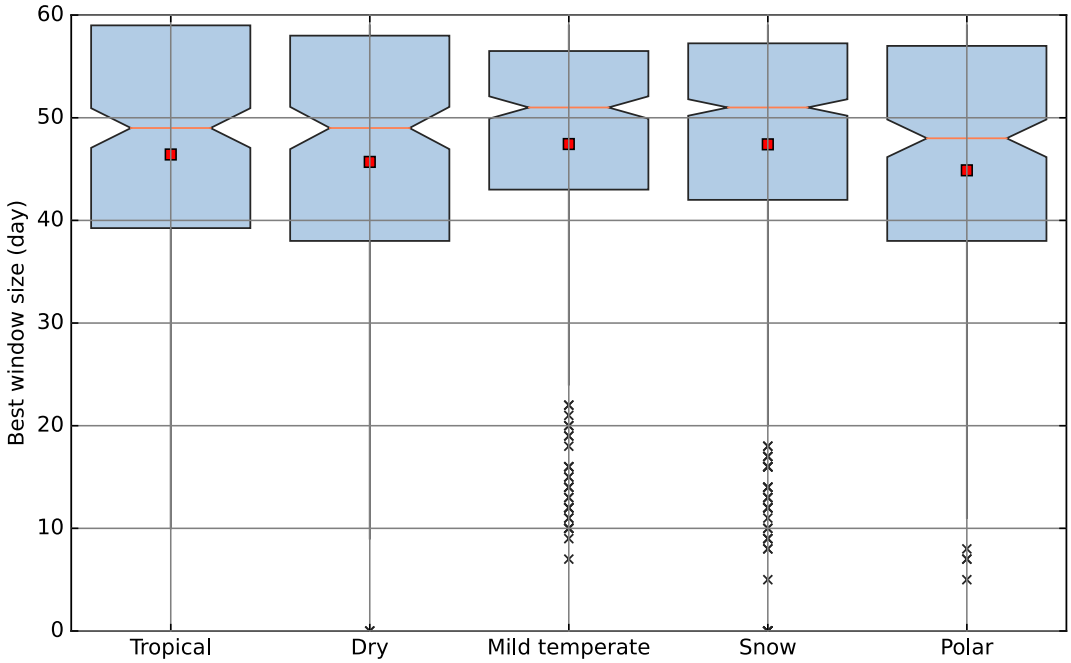



Figure 4.2. Boxplot of best window size for averaging for all meteorological variables in different Köppen regions, showing the 25, 50 and 75 percentiles. The mean value is presented as red square.

When determining the optimal window size, we observed fluctuations in averaging feature results (RMSE) across various window sizes. To mitigate these fluctuations to get the best window size, we opted to apply smoothing techniques. For instance, in the case of Victoria (as discussed in section 2.4), we utilized smoothed values (as illustrated in Figure 4.3) to improve our ability to detect the minimum, which, in this instance, occurred at approximately 40 days.

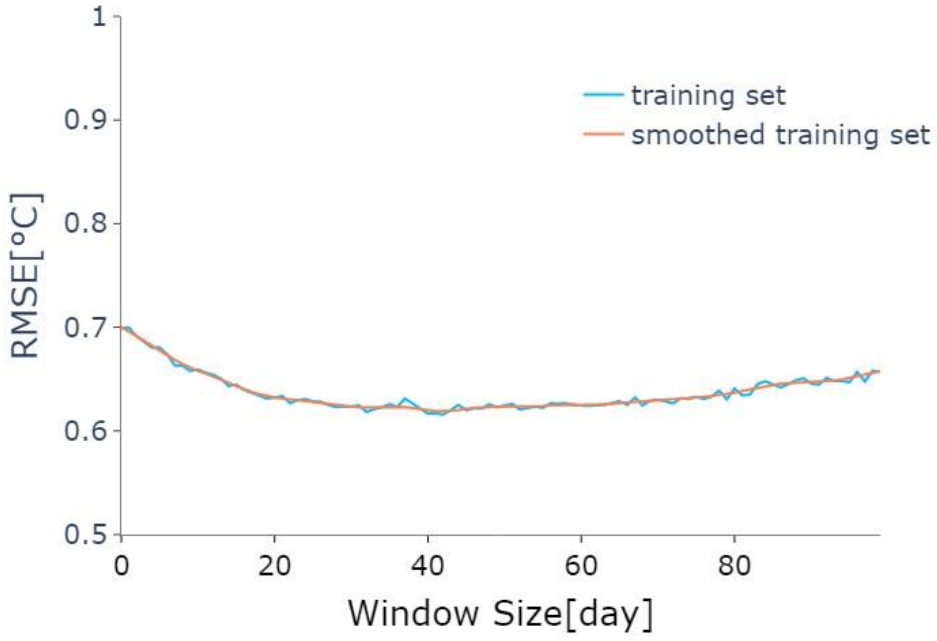



Figure 4.3. The actual and smoothed RMSE (°C) over different window sizes for Lake Victoria.

### 4.3.2. Cross correlation among variables

Gaining a deep comprehension of the principles governing thermal dynamics is essential for achieving precise forecasts, particularly when employing non-physical methodologies such as ML. Accordingly, for each lake, we chose and utilized Spearman's rank-order correlation to assess the interrelationships between variables, and then averaged for each Köppen region. The method quantifies the magnitude and direction of a nonlinear relationship between two variables, as proposed by Spearman (1910) explained in section 2.7.2. It is crucial to acknowledge that there exist correlations not only between the predictors and LSWT but also among the predictors themselves. This analysis also aids in verifying the accuracy of the feature ranking presented in section 4.3.4.

Given that the LSWT data is not available on a daily basis while the features are, we utilized linear regression to fill in the missing values of LSWT. Accordingly, we constructed a linear regression model incorporating meteorological variables alongside SCDOY as input. The objective function was set as the RMSE using observed LSWT. Subsequently, we utilized the model to predict the missing LSWT values. Therefore, the correlations were computed using the daily values. It should be noted that these correlations are based on average meteorological variables which we used in other analysis as well.

Figure 4.4 illustrates that AT, LWR, SWR, RH, and SCDOY exhibit the highest correlations not only with LSWT but also among themselves, across all regions. SDOY, CDOY, AT, WS, SWR, LWR, AP, RH, R and S are higher correlated with LSWT in temperate, temperate, snow, polar, snow, temperate

and tropical, temperate, tropical, polar and dry and temperate regions, respectively. In general, higher correlations can be observed in temperate region.

In the analysis, we focus on the magnitude of the correlation value, as even a negative value of high value indicates a strong correlation. For instance, there exists a negative correlation between AT and SCDOY, with AT exhibiting a stronger correlation with CDOY than with SDOY. It's also noteworthy that AT exhibits a high correlation with SWR, LWR, and RH which means AT includes some same information of these variables.

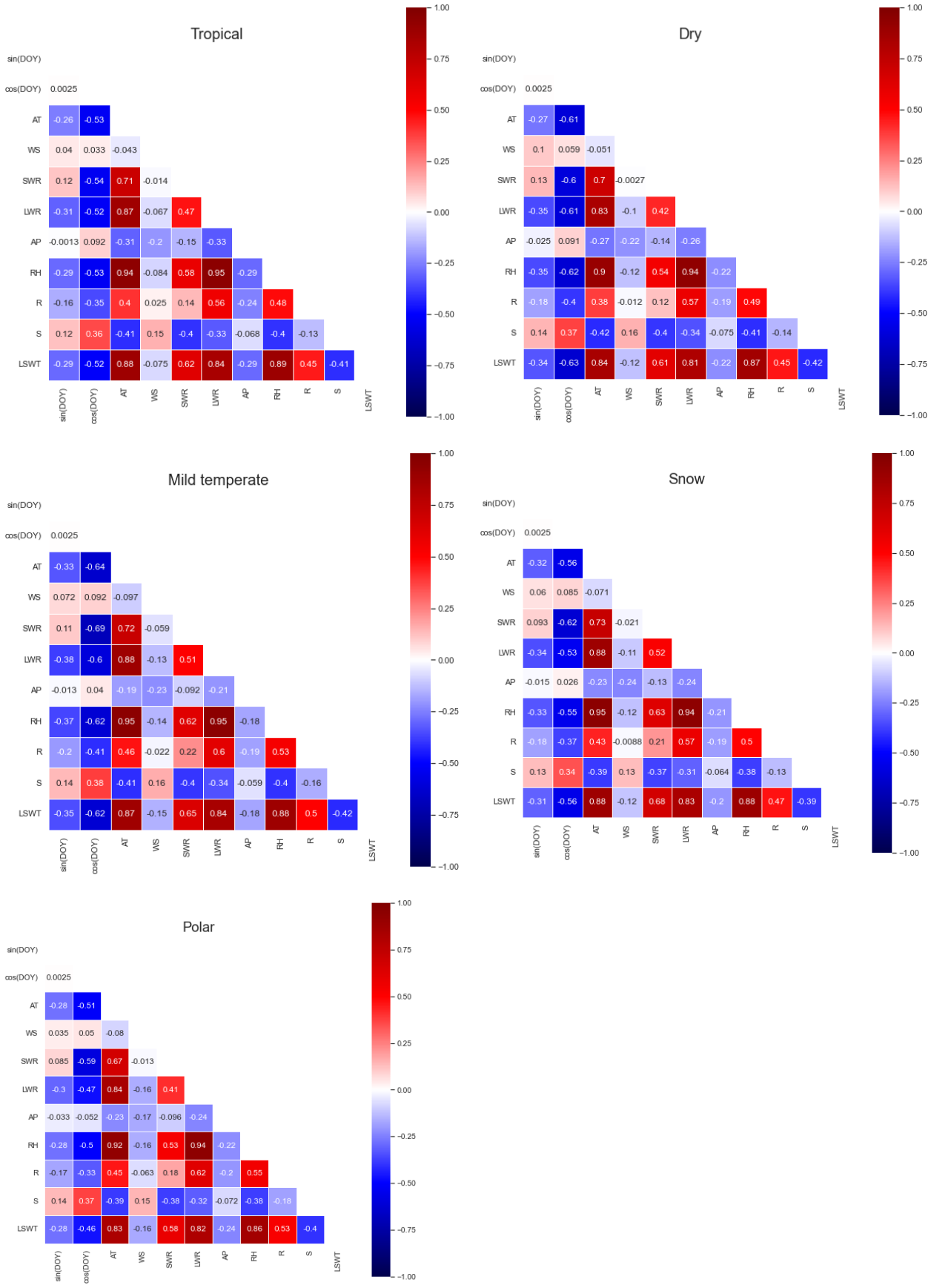



Figure 4.4. Heatmap of Spearman correlation for Köppen region.

### 4.3.3. Model evaluation

In this section, we present the results obtained from the BPNN model, considering four distinct scenarios. These scenarios involve the inclusion of meteorological variables (namely AT, SWR, LWR, WS, R, S, AP, and RH) without any averaging, both with and without SCDOY, as well as the averaged meteorological variables with and without SCDOY. These results are in Table 4.2. and depicted in Figures 4.5 to 4.12.

As shown in Table 4.2, the RMSE is lower in the training set compared to the test set. The NSE* is higher (indicating better performance) in the training set, meaning that it captures better the interannual variability. The model utilizing averaged variables outperforms the actual values in both scenarios: without SCDOY and with the inclusion of SCDOY, which further enhances the model's performance compared to the case without SCDOY. The improvement for averaged variables plus SCDOY scenario is attributed to the fact that the averaged meteorological variables such as SWR, LWR, and AT variables can effectively capture the patterns and exhibit a strong correlation with LSWT, as elaborated in section 4.3.2.

4.2. The median of RMSE (°C) and NSE* of the models in each Köppen region for both training and test sets.

| Scenario | Metrics | Tropical | | Dry | | Temperate | | Snow | | Polar | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Training | Test | Training | Test | Training | Test | Training | Test | Training | Test |
| Actual variables without SCDOY | RMSE(°C) | 2.96 | 3.03 | 2.99 | 3.03 | 3.07 | 3.13 | 3.02 | 3.08 | 2.92 | 3.07 |
| | NSE* | -0.98 | -1.07 | -0.85 | -0.87 | -0.98 | -1.07 | -0.95 | -0.98 | -0.81 | -0.88 |
| Actual variables with SCDOY | RMSE(°C) | 1.72 | 1.83 | 1.73 | 1.83 | 1.77 | 1.84 | 1.75 | 1.83 | 1.72 | 1.84 |
| | NSE* | 0.20 | 0.15 | 0.21 | 0.18 | 0.22 | 0.19 | 0.21 | 0.18 | 0.15 | 0.11 |
| Averaged variables without SCDOY | RMSE(°C) | 2.03 | 2.14 | 1.93 | 2.07 | 2.17 | 2.24 | 2.08 | 2.19 | 2.05 | 2.14 |
| | NSE* | -0.06 | -0.18 | -0.03 | -0.12 | -0.08 | -0.16 | -0.05 | -0.16 | -0.10 | -0.16 |
| Averaged variables with SCDOY | RMSE(°C) | 1.46 | 1.61 | 1.46 | 1.63 | 1.54 | 1.71 | 1.51 | 1.68 | 1.48 | 1.66 |
| | NSE* | 0.42 | 0.32 | 0.42 | 0.28 | 0.41 | 0.30 | 0.41 | 0.29 | 0.39 | 0.26 |

The plots, depicted in Figures 4.5 to 4.12, visually illustrate the RMSE and NSE* outcomes for each Köppen region across both the training and test datasets. Figures 4.5 and 4.6 display the results for actual variables without SCDOY; Figures 4.7 and 4.8 correspond to actual variables with SCDOY; Figures 4.9 and 4.10 are for averaged variables without SCDOY; and Figures 4.11 and 4.12 showcase the results of averaged variables with SCDOY. In the boxplots found within each of these figures (labeled as c and d), the averaged values are highlighted in red lines.

Overall, it is important to emphasize that the predictive errors are more pronounced in the Temperate region when contrasted with other regions. These figures reinforce the findings presented in Table 4.2, where the training results outperform the test set results, and the utilization of averaged values consistently yields superior outcomes compared to using the actual values. This suggests that incorporating the memory of previous days can enhance the results, particularly for variables like LSWT, which heavily depends on historical data from preceding days. Utilizing SCDOY results in

the NSE* exceeding zero, due to the contribution of SCDOY towards capturing the mean annual pattern, and using other features improve the prediction accuracy.

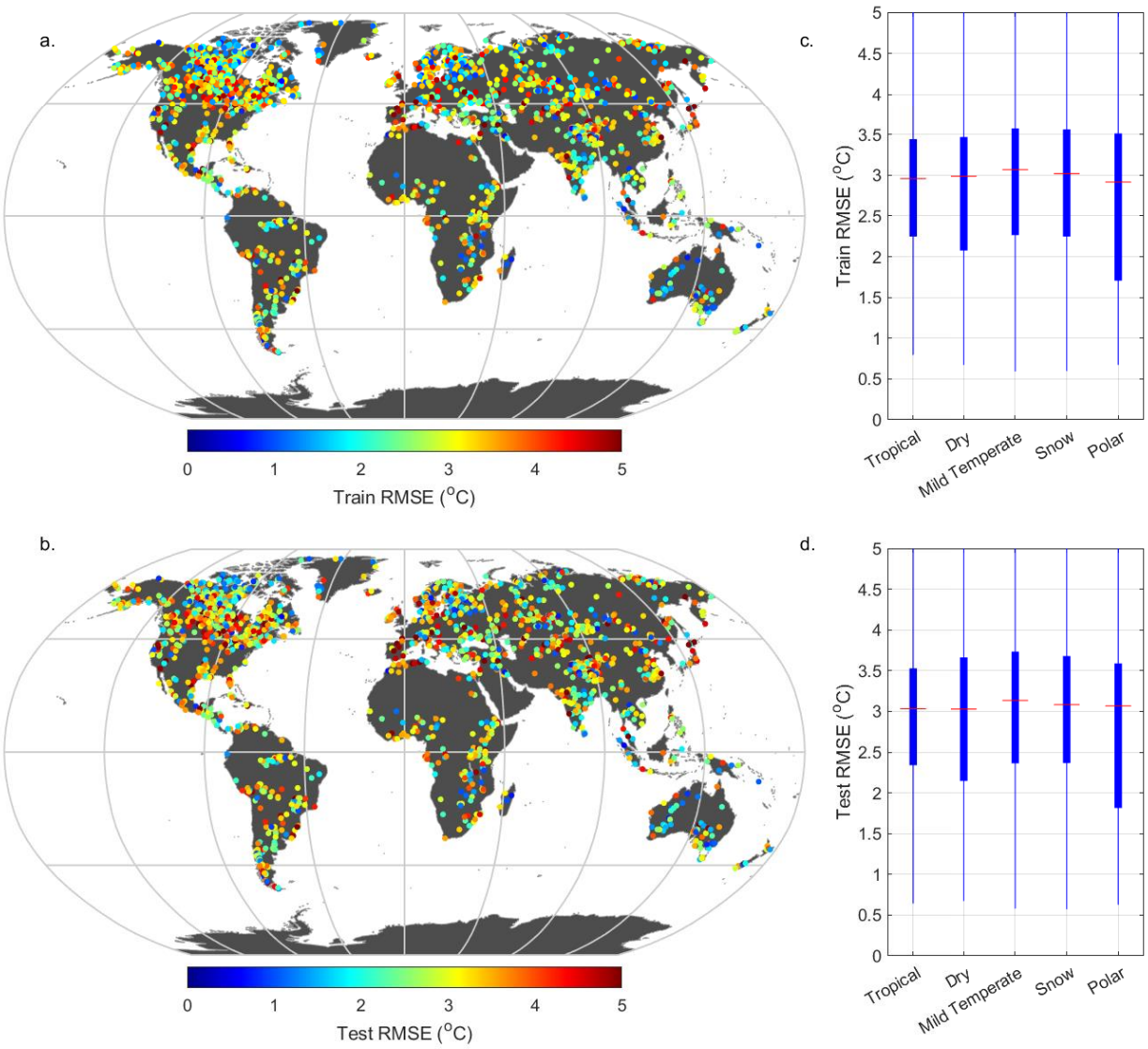



Figure 4.5. RMSE of both training and test sets for all Köppen regions, considering the actual values without SCDOY.

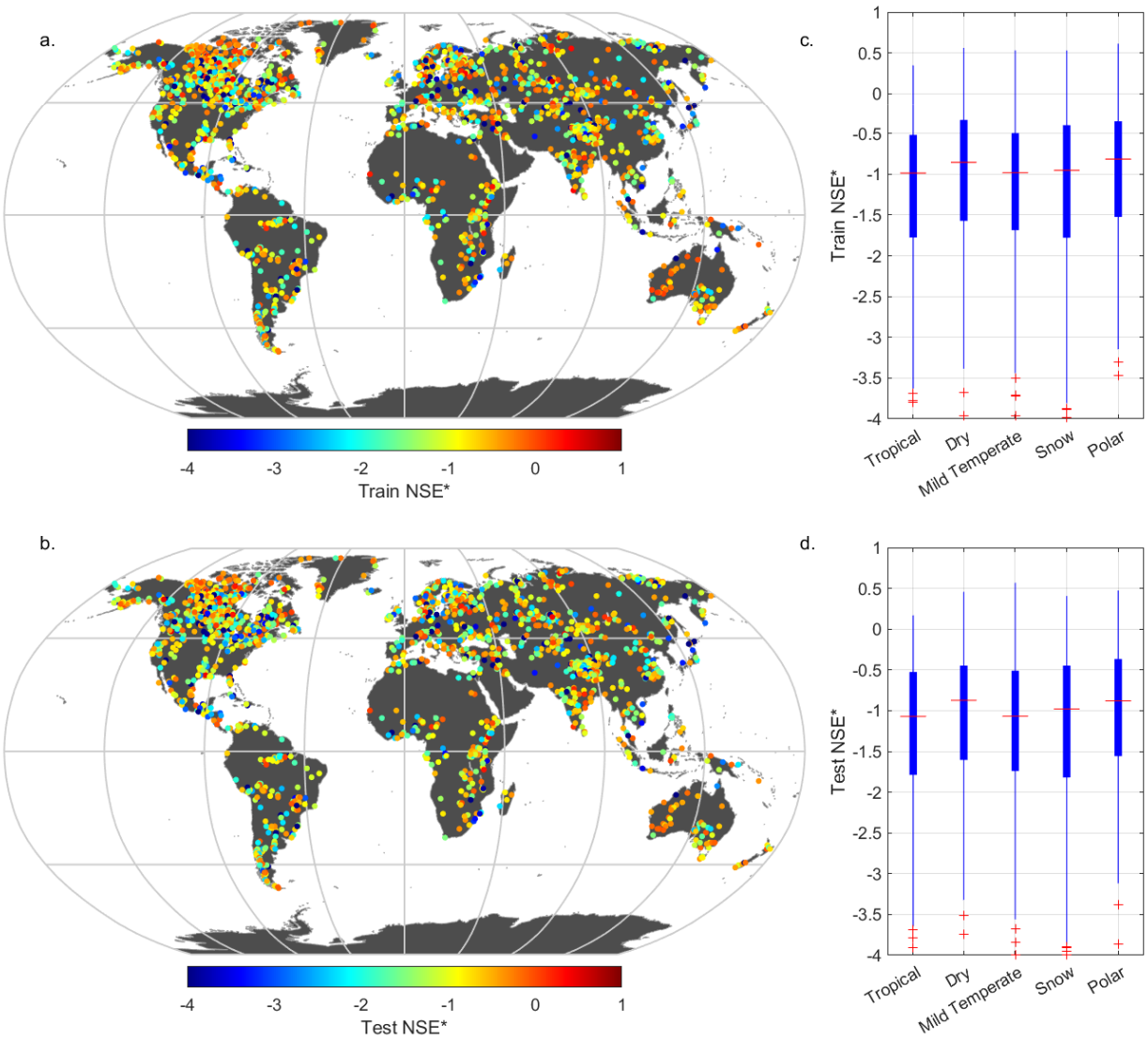



Figure 4.6. NSE* of both training and test sets for all Köppen regions, considering the actual values without SCDOY.

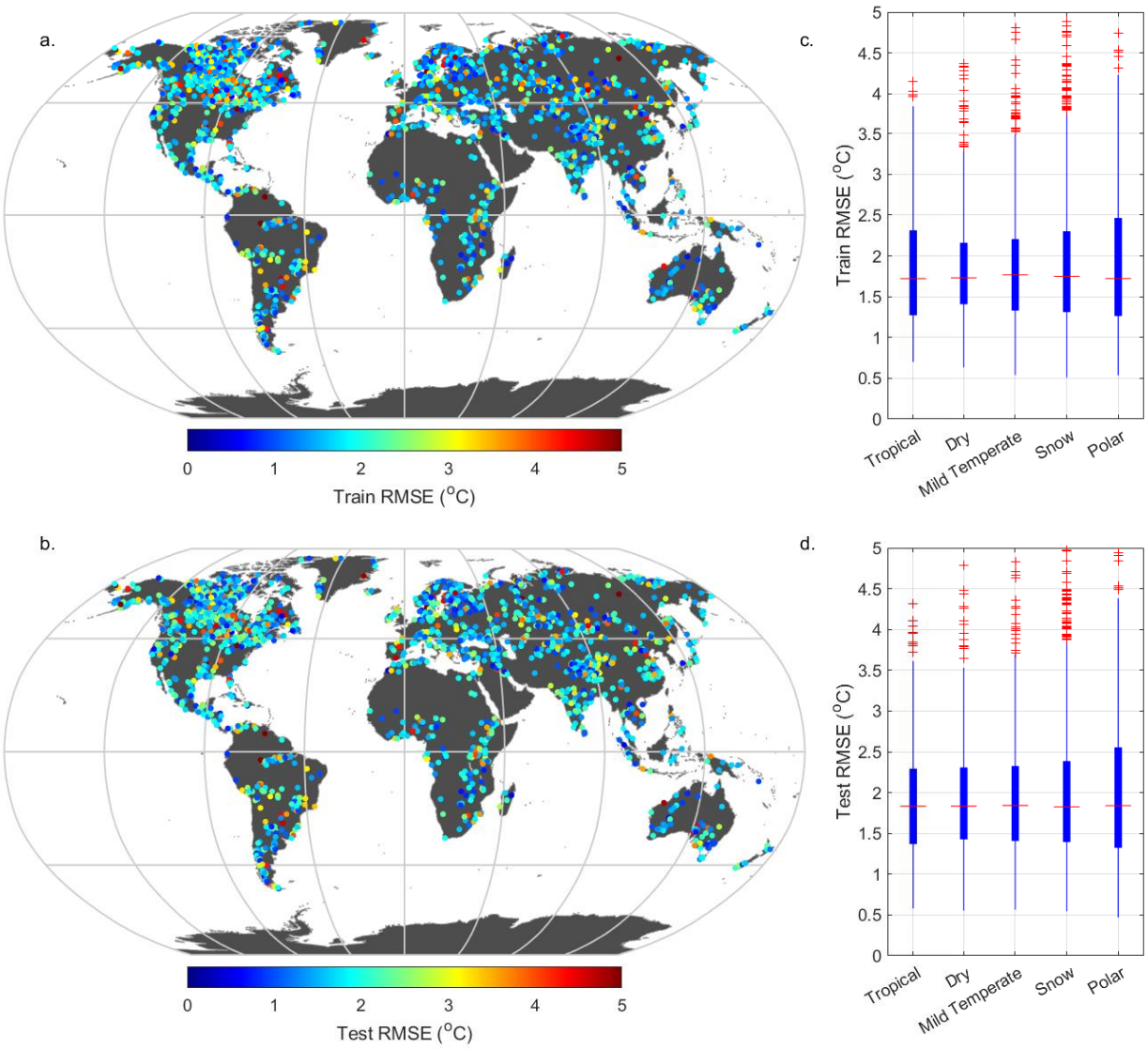



Figure 4.7. RMSE of both training and test sets for all Köppen regions, considering the actual values with SCDOY.

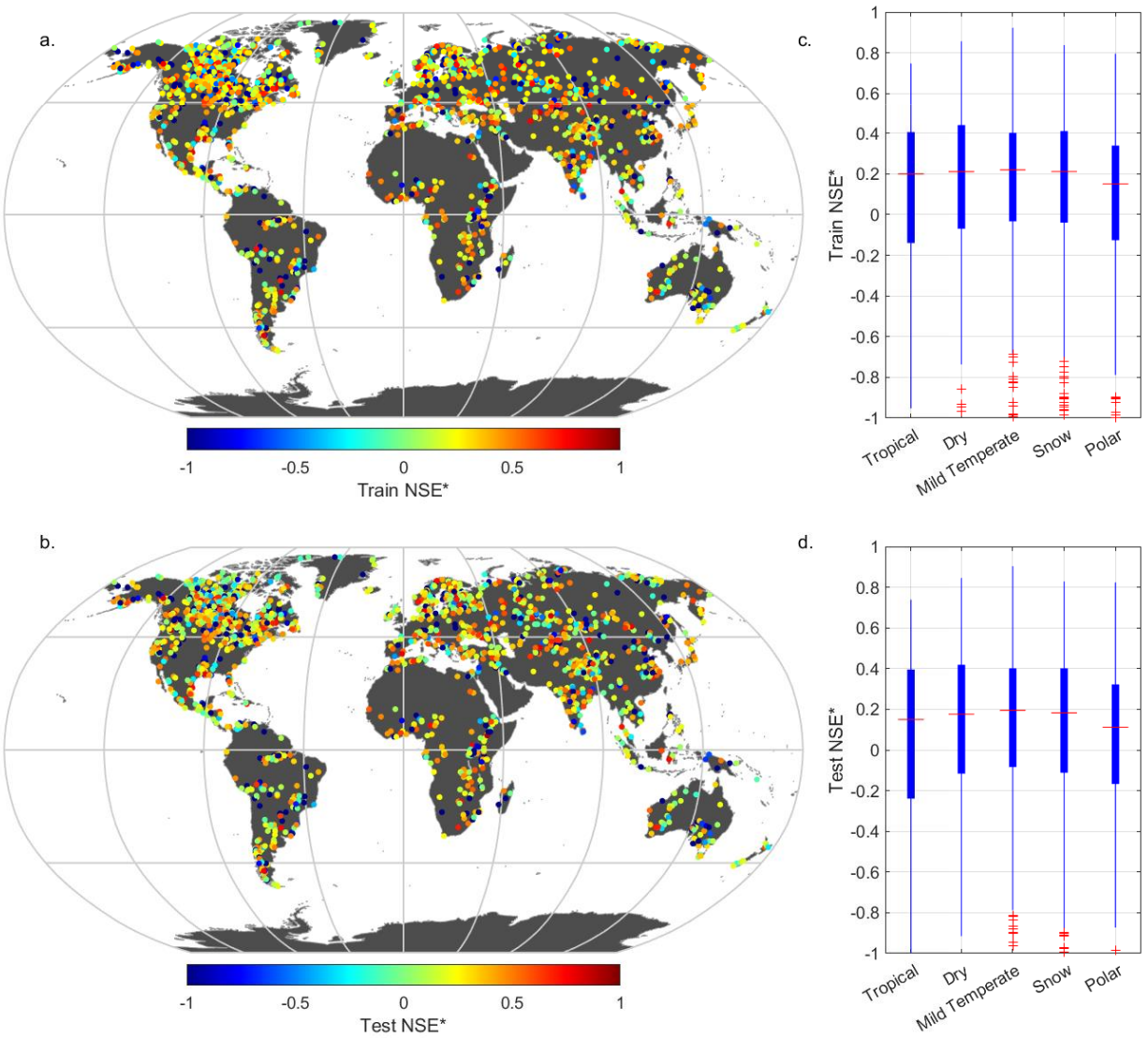



Figure 4.8. NSE* of both training and test sets for all Köppen regions, considering the actual values with SCDOY.

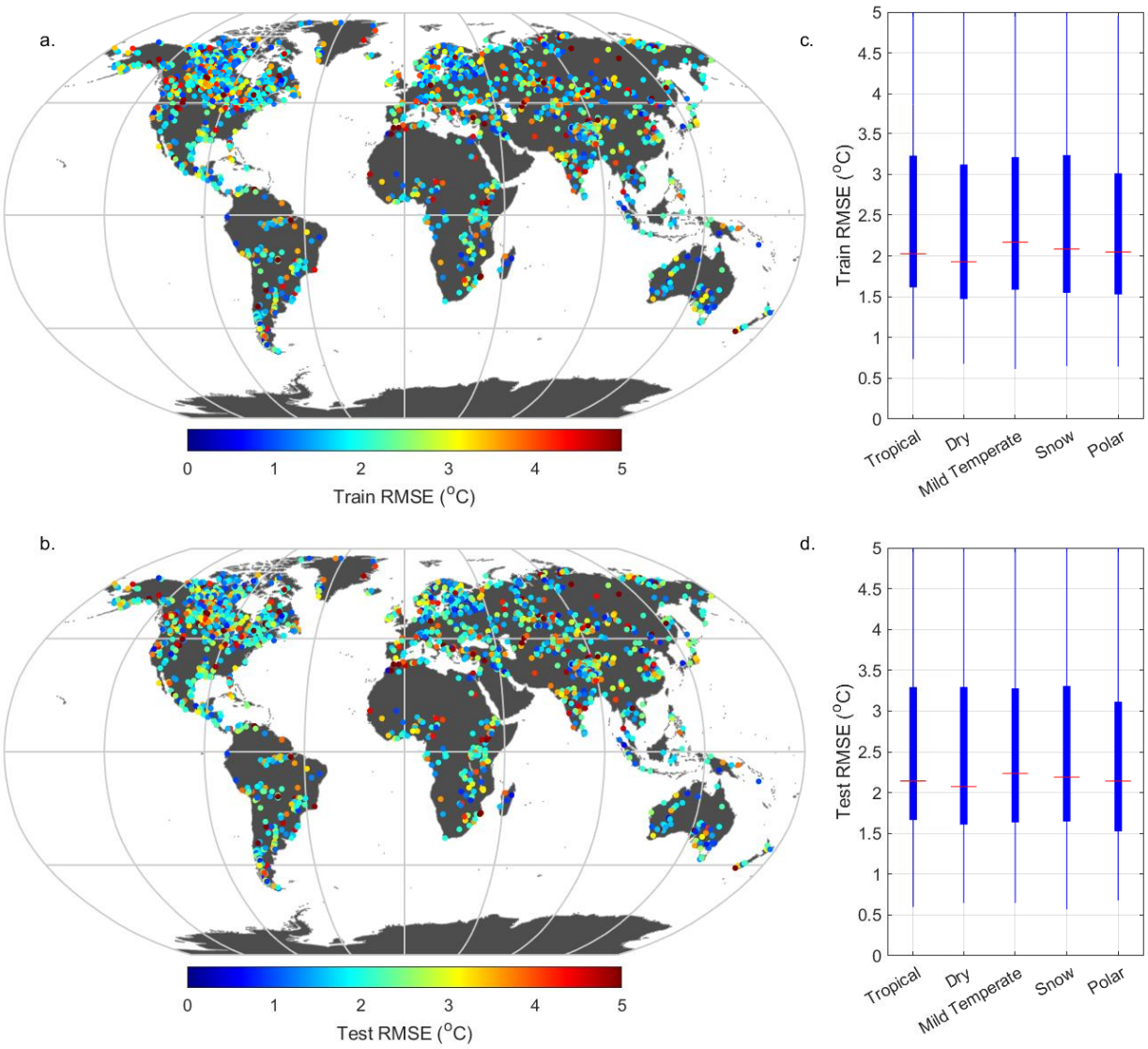



Figure 4.9. RMSE of both training and test sets for all Köppen regions, considering the averaged values without SCDOY.

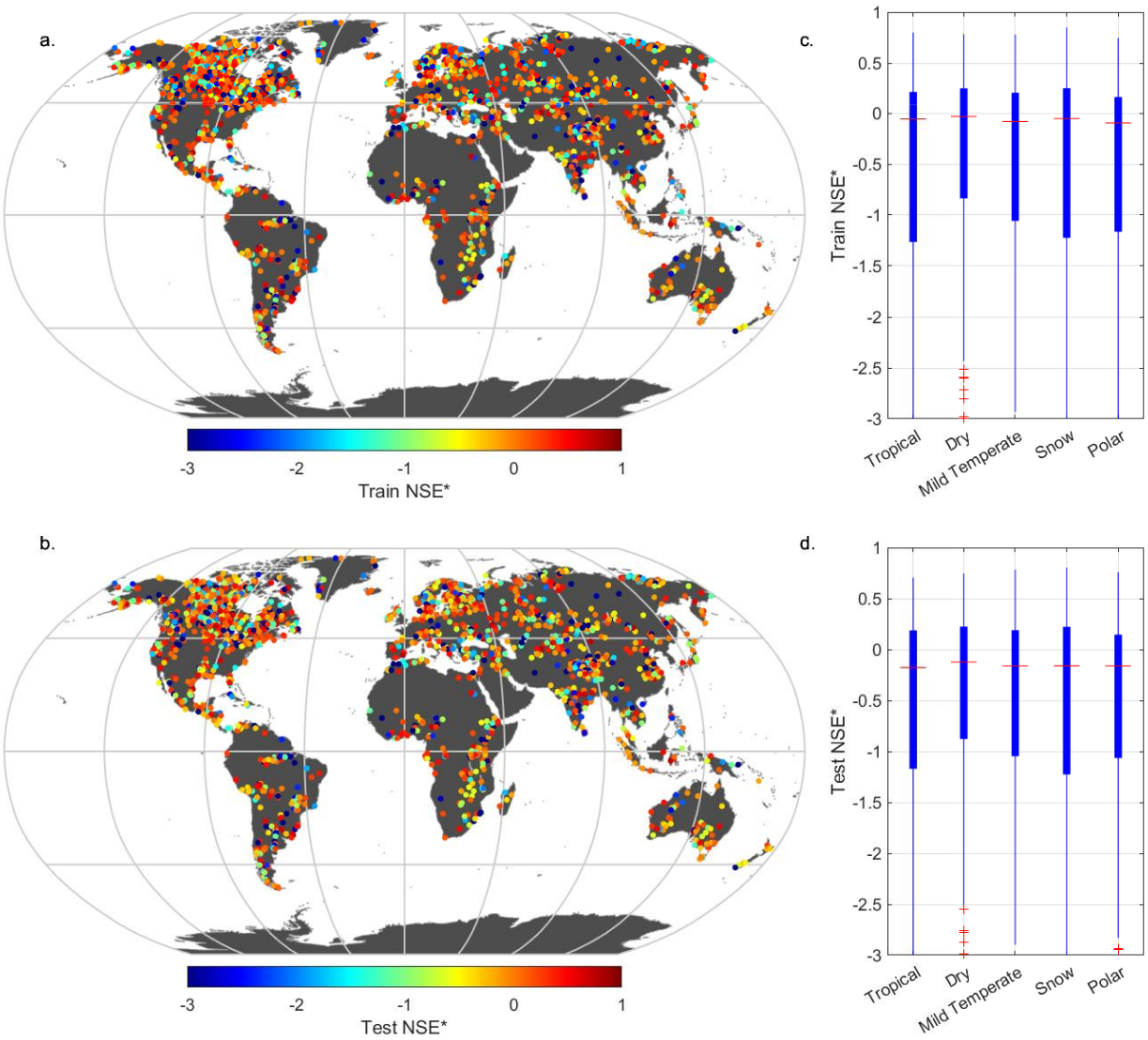



Figure 4.10. NSE* of both training and test sets for all Köppen regions, considering the averaged values without SCDOY.

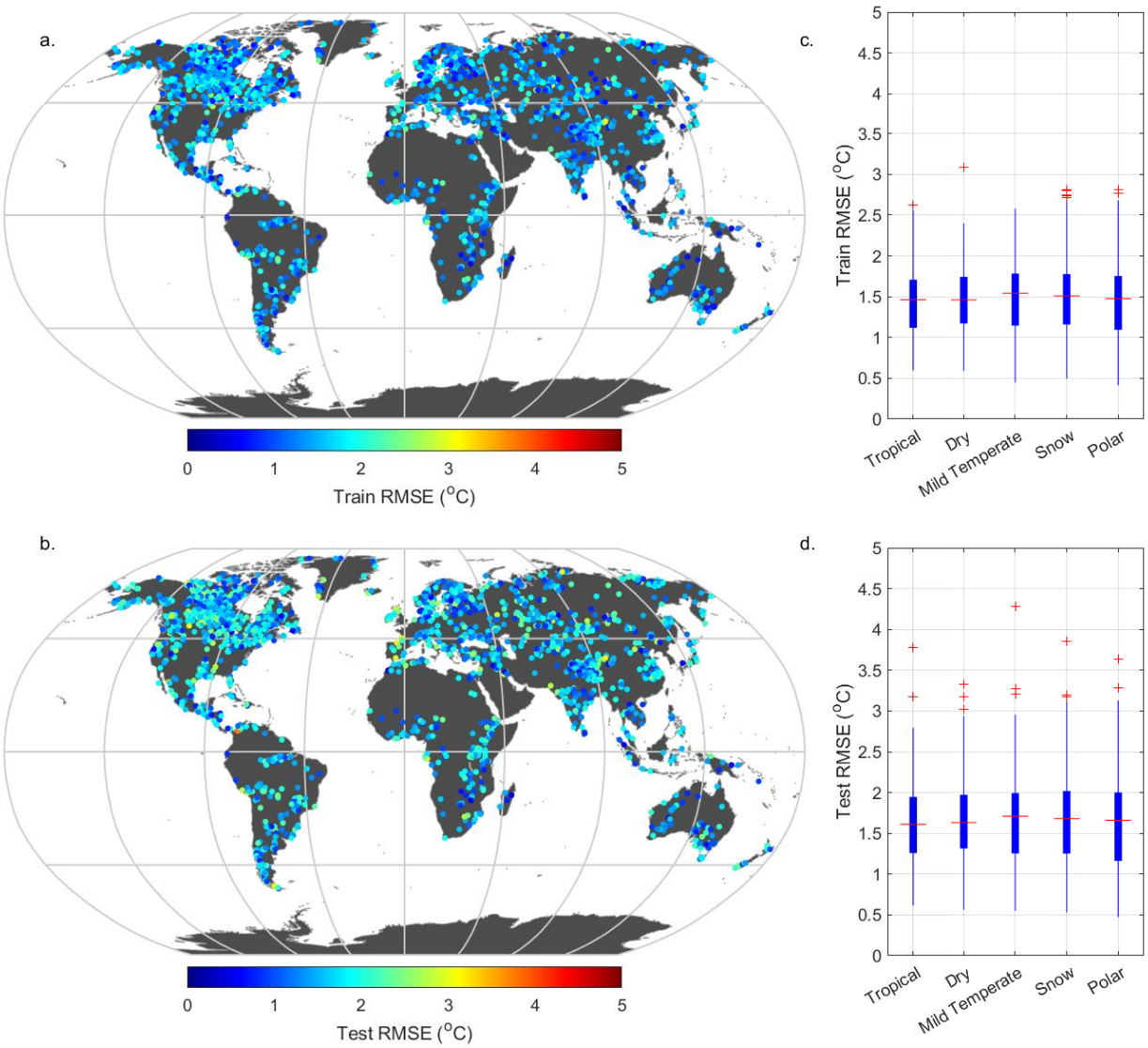



Figure 4.11. RMSE of both training and test sets for all Köppen regions, considering the averaged values with SCDOY.

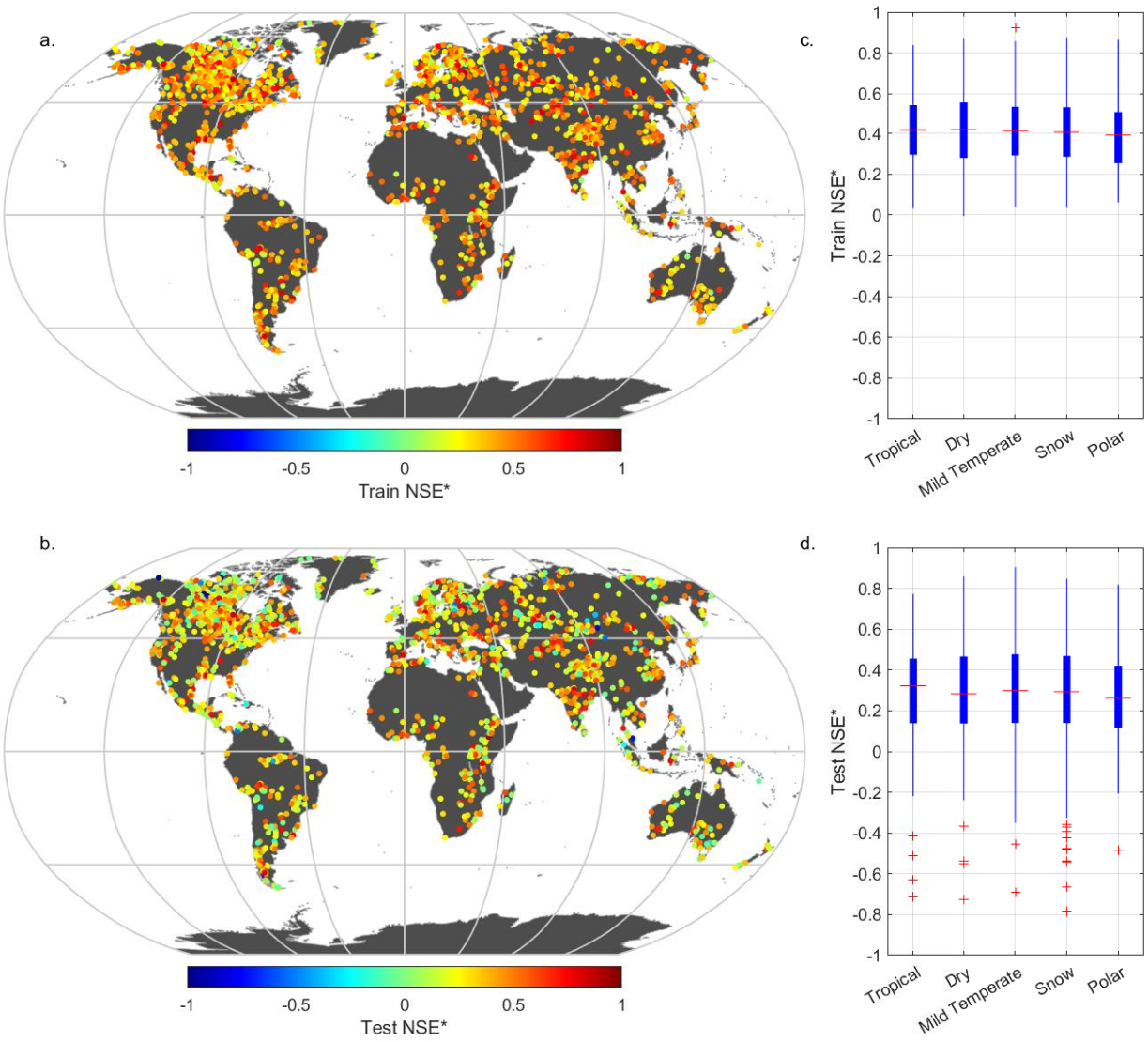



Figure 4.12. NSE* of both training and test sets for all Köppen regions, considering the averaged values with SCDOY.

Due to the impracticality of presenting detailed results for all lakes in the dataset, we have opted to evaluate a single representative lake from each Köppen region, as already anticipated in section 2.4. Accordingly, lakes Victoria, Dead Sea, Garda, Erie, and Zhari Namco have been selected to represent the tropical, dry, temperate, snow, and polar regions, respectively, based on the availability of ample data, thereby facilitating enhanced performance of the model (presented in Table 4.3). Since using the averaged variables improves the results, we considered the models with weighted average with and without SCDOY. As we discussed earlier, including SCDOY slightly improves our results.

The bias metric considers whether the model consistently predicts less (negative value) or more (positive value) than the observed values, and in these cases, it does tend to predict values on the lower side. The use of SCDOY helps achieve higher NSE* by capturing interannual variability. We anticipate that, in general, an NSE* greater than zero is obtained when incorporating SCDOY while

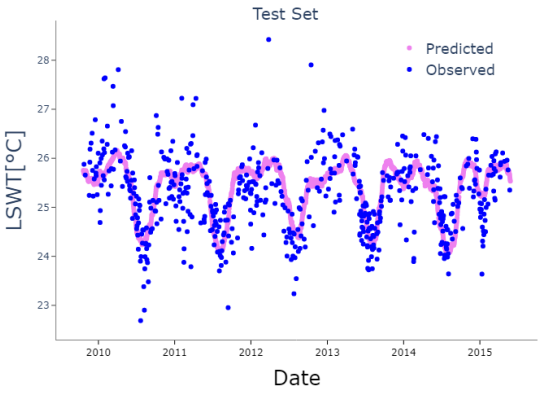
for some lakes, like theDead Sea, including additional variables (other than SCDOY) might make the predictions worse (NSE* less than zero).

Table 4.3. RMSE and NSE* for lakes representative of each Köppen region.

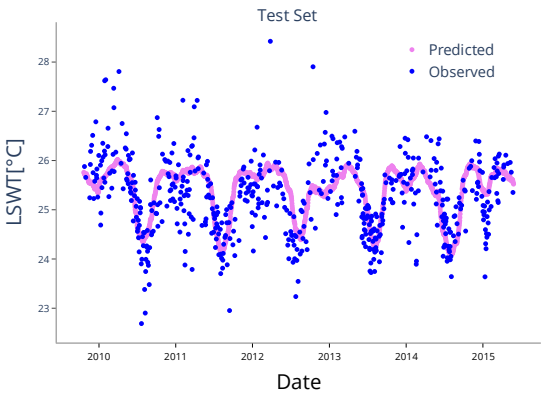
| Köppen region | Sce-nario | RMSE (°C) | | NSE* | | Bias (°C) | |
|---|---|---|---|---|---|---|---|
| | | Training | Test | Training | Test | Train-ing | Test |
| Tropical (Lake Victoria) | With SCDOY | 0.67 | 0.58 | 0.12 | 0.08 | -1.4 E-4 | 1.6E-2 |
| | Without SCDOY | 0.72 | 0.62 | -0.03 | -0.02 | 1.4 E-3 | 3.7 E-2 |
| Dry (Dead Sea) | With SCDOY | 0.99 | 0.95 | -0.08 | -0.08 | 4.9 E-4 | 1.2 E-1 |
| | Without SCDOY | 1.44 | 1.52 | -1.27 | -1.75 | 6.9 E-5 | 1.5 E-1 |
| Temperate (Lake Garda) | With SCDOY | 1.19 | 1.24 | 0.18 | 0.17 | 3.3 E-5 | 8 E-2 |
| | Without SCDOY | 2.45 | 2.50 | -2.44 | -2.40 | -3 E-4 | 2.8 E-1 |
| Snow (Lake Erie) | With SCDOY | 1.32 | 1.43 | 0.62 | 0.58 | 1 E-3 | -1.2 E-1 |
| | Without SCDOY | 1.51 | 1.66 | 0.50 | 0.43 | 1.5 E-5 | 1.3 E-1 |
| Polar (Lake Zhari Namco) | With SCDOY | 0.90 | 0.87 | 0.76 | 0.76 | - 3.1 E-4 | 3.68 E-2 |
| | Without SCDOY | 1.27 | 1.18 | 0.52 | 0.56 | -1.7 E-4 | 1.47E-1 |

The predictions are visually displayed in Figure 4.13. On the right side, the figures for cases where SCDOY is considered along with the averaged variables are shown, while on the left side, SCDOY is not included. Focusing on Lake Victoria, as shown in Figure 4.13 (a) and (b), the prediction may not appear highly accurate visually. However, it is noteworthy that the RMSE values for this lake displayed in Table 4.3 signify superior predictive performance. This occurs due to the relatively minor fluctuations in LSWT within Lake Victoria, resulting in lower RMSE values associated with reduced errors. Conversely, the LSWT exhibits wider fluctuations in other lakes. Therefore, the higher RMSE values for Lakes Garda, Erie, and Zhari Namco should not be interpreted as indicative of poor predictions. For Lake Zhari Namco, presented in Figure 4.13(i) and (j), due to the lack of zero data during the winter period, accurate prediction for those days is challenging. Consequently, it is not factored into the computation of errors.

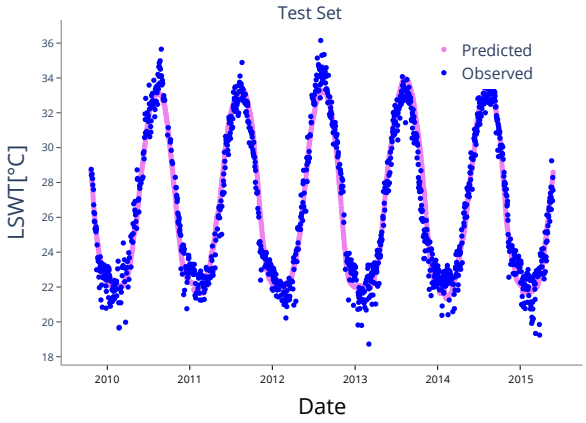
(a)


Test Set
Predicted
Observed
LSWT[°C]
Date

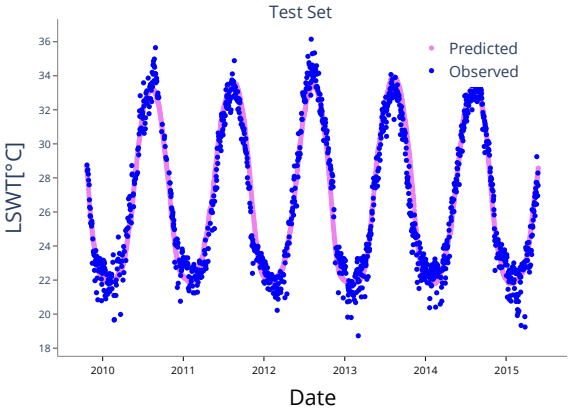
(b)


Test Set
Predicted
Observed
LSWT[°C]
Date

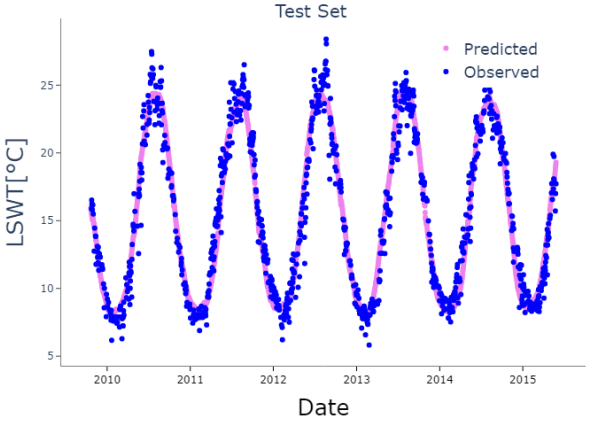
(c)


Test Set
Predicted
Observed
LSWT[°C]
Date

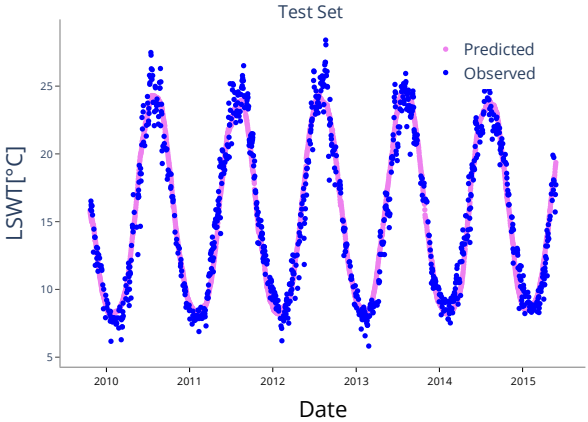
(d)


Test Set
Predicted
Observed
LSWT[°C]
Date

(e)


Test Set
Predicted
Observed
LSWT[°C]
Date

(f)


Test Set
Predicted
Observed
LSWT[°C]
Date

(g)


Test Set
Predicted
Observed
LSWT[°C]
Date

(h)


Test Set
Predicted
Observed
LSWT[°C]
Date

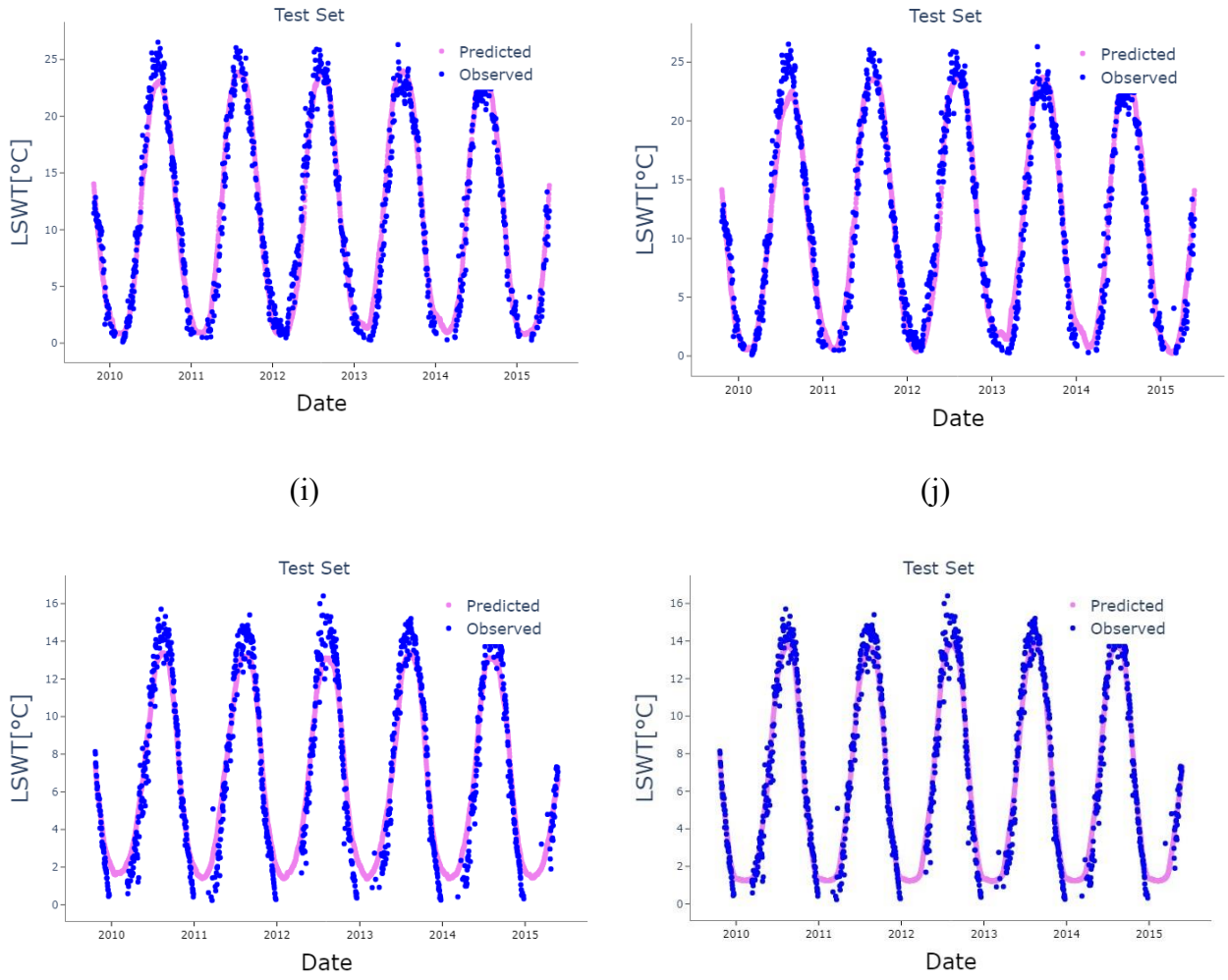



Figure 4.13. The simulation and prediction of LSWT for lakes Victoria (a, b), Dead Sea (c,d), Garda (e,f), Erie (g, h), Zhari Namco (i,j). The cases with SCDOY are on the right side and without using them in the left side.

### 4.3.4. Feature ranking

In order to assess the significance of predictors in model performance, we employed the FR algorithms introduced in section 2.7 to assign ratings to each individual feature. Additionally, we aim to identify predictors that exert a negative impact on the model, as illustrated by the case of the Dead Sea in Table 4.3. Despite the inclusion of SCDOY, resulting in a negative NSE*, this outcome indicates the use of a non-relevant predictor(s). Given our objective of creating a single, generalized model for all lakes, we needed to incorporate all predictors that could be significant in various regions.

This analysis consisted of subjecting a specific feature to 30 rounds of random shuffling (rearranging, see more details in Chapter 2), resulting in the generation of multiple random sequences for that predictor, which then lose their physically arguable influence on the target variable (LSWT). By comparing the model performance obtained by using the shuffled inputs with the reference case (all inputs are physically consistent), we were able to quantify the level of dependence on each feature. Notably, a higher degree of degradation observed in the model performance indicates a higher level

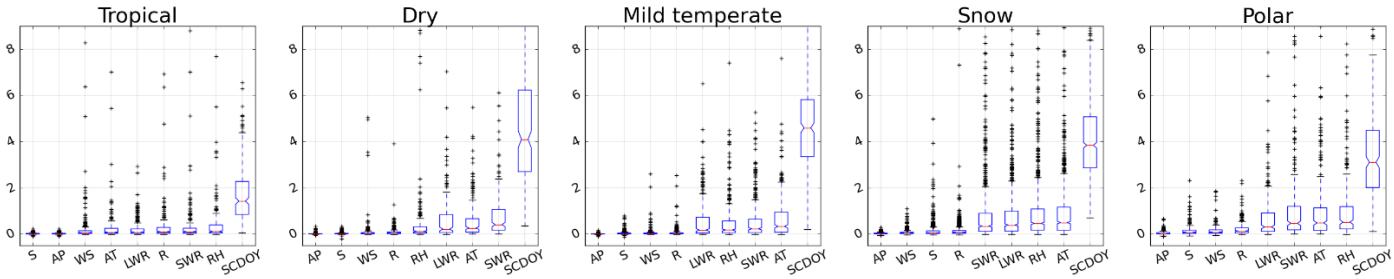
of relevance associated with the particular feature in each individual lake. Subsequently, we computed the average of the degradations for each Köppen region.

Figure 4.14 presents a visual representation of the FR for each variable, sorted by the median of errors, illustrating their respective significance within various Köppen regions. The y-axis value indicates the difference between the RMSE of the normal case (without shuffling) and the RMSE of the shuffled cases, divided by the RMSE of the normal case.

Part (a) of the figure corresponds to the scenario where SCDOY is considered, which is the primary predictor in this case. In the tropical region, apart from S, all other variables play a crucial role. This observation is in line with what we observed in Figure 4.13 (a,b), where the changes in LSWT do not precisely follow the interannual variability.

Part (b) of the figure represents cases where SCDOY is not considered. Generally, SWR, LWR, AT and RH emerge as the key predictors. This observation aligns with the correlations highlighted in section 4.3.2. We note that, in the temperate region, AT is the most relevant additional predictor when using SCDOY, as often assumed (Piccolroaz et al., 2021), but its effect is not dominant over the other factors. These findings underscore that the expected influence of these variables on the model's performance and their ability to capture the complex dynamics within each specific Köppen region is not trivial.

(a)




(b)

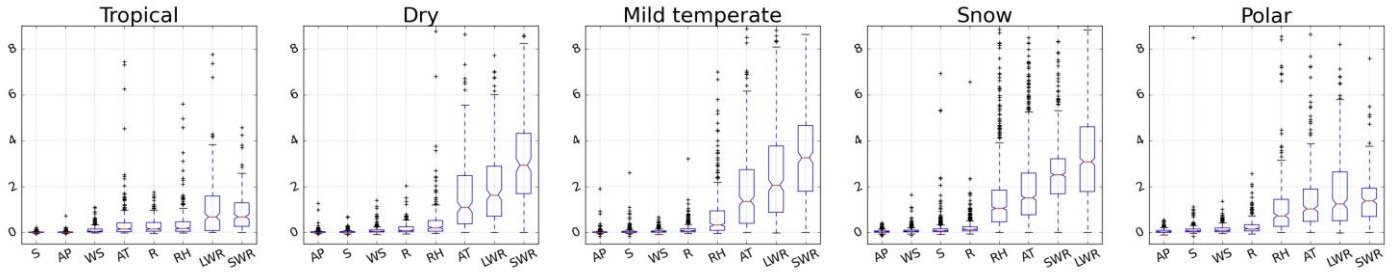



Figure 4.14. Feature ranking for each Köppen region: boxplots of the increase of RMSE (°C) obtained by shuffling the single variable indicated on the horizontal axis with considering SCDOY (a) and without it (b).

### 4.3.5. Discussion

The ML technique, BPNN, demonstrated effective simulation and prediction of LSWT, incorporating SCDOY and meteorological variables as inputs. Our study revealed that employing the average of meteorological variables, as opposed to using values from a single day, significantly enhances the model's performance. Furthermore, the inclusion of SCDOY further improves model accuracy by capturing the interannual variability in LSWT.

This chapter serves as an extension of Chapter 3, encompassing the analysis of 2024 lakes worldwide. Building upon the previous chapter's exploration of nine ML algorithms, where we found that ANN outperformed other ML methods, we chose BPNN because its commendable performance and efficiency across diverse scenarios. Unlike that analysis, which accounted for different combinations of preprocessing methods, in this chapter we directly utilize the average of all meteorological data as input for the models. This approach aims to create a standardized model applicable to all lakes globally.

In particular, in Chapter 3 we demonstrated the effectiveness of considering the ATs of previous days as additional inputs, but the computational load discouraged its usage for all lakes. Here, we opt for simplicity and use the average of all meteorological data as input features for the models, ensuring consistency across diverse regions and scenarios. Unlike our previous synthetic lake case studies, where we efficiently identified the optimal input combination (SCDOY + AT), this work retains all features in the model.

An additional focus of this study is the exploration of the effect of the geographical location on model performance. While Chapter 3 delved into the effect of varying the depth in synthetic case studies, this chapter expands its scope by considering the impact of geographical location on the selection of the feature importance. Importantly, this work offers the advantage of generalizability on a global scale, presenting the opportunity for extensive lake predictions and comparisons. The ability to utilize the forcing from ERA5 across different geographical areas adds significant value to our work.

This chapter presents a unique challenge due to the sparse data, especially when compared to Chapter 3, where daily values of LSWT were available. Despite having averaged daily meteorological variables in both studies, the scarcity of observed LSWT data in this chapter prevents us from effectively training the model while retaining the information and pattern of all days. However, by utilizing the averaged meteorological variables, we manage to preserve the memory of past conditions. This approach helps to address the limitations posed by missing observed LSWT values.

As previously mentioned, the models were executed with all features as inputs, with the expectation that including additional features in addition to SCDOY would result in NSE* values greater than zero due to its ability to capture interannual variability (noting that NSE* is zero when using only SCDOY). However, this expectation did not hold in all cases; for instance, in the Dead Sea, negative values of NSE* were obtained even when using other features in addition to SCDOY. This discrepancy suggests that the addition of certain variables had a detrimental impact on the model, reducing its performance. Consequently, while utilizing averaged meteorological variables along with SCDOY proves to be the optimal input combination, the presence of variables with negative effects necessitates their removal. Identifying these influential variables is achieved through FR analysis for each lake, and the results are aggregated within each Köppen region to discern the meteorological variables' effectiveness in different climate regions.

The FR analysis indicates that when SCDOY is used as an input, other meteorological variables have less impact compared to SCDOY itself, which indeed provides already a good description of the mean seasonal variability. In both cases, with or without SCDOY, SWR, LWR, AT, and RH emerge as the primary meteorological variables, a finding consistent with correlation analysis, irrespective of ML

model consideration. Conversely, WS, S, and AP exhibit the lowest or negative impact on the model. Notably, R demonstrates a relatively higher impact in tropical regions. In the tropical region, aside from S and AP, which have negligible positive effects, all other meteorological variables significantly influence the model. Interestingly, SWR in temperate region shows a relatively high impact, larger than AT when SCDOY is not utilized. In conclusion, we recommend excluding AP and S from the models for simulating and predicting LSWT, and the inclusion of WS is also deemed unnecessary.

## 4.4. Conclusions

In our pursuit of modeling and predicting LSWT, we exploited the power of ML, specifically BPNN. This study not only demonstrates the efficacy of BPNN for LSWT prediction but also emphasizes the model predictive capabilities for generalizing to 2024 lakes. Recognizing the significance of the thermal inertial of lakes to correctly simulate LSWT, we introduced an approach, employing a weighted average of influential meteorological variables over a defined time window. This method, carefully optimized for model performance, aimed to capture the cumulative impact of past conditions, thereby enhancing the modeling process.

Our examination delved into the intricate relationships between predictors and LSWT within specific Köppen climate regions. This analysis illuminated how LSWT responds to diverse meteorological factors across varying climates, revealing strong connections between key predictors and LSWT variations. The inclusion of SCDOY further improved outcomes by providing a reliable description of the mean seasonal variability, and letting the other factors be exploited to capture interannual variability.

We use the flexibility inherent in ML models, aiming to conduct a detailed analysis of crucial predictors specific to each Köppen climate region. This approach allowed us to unravel and understand the distinctive influences of meteorological variables on LSWT within diverse climatic contexts. Variables such as DOY (expressed through SCDOY), AT, SWR, LWR, and RH emerged as influential contributors to LSWT dynamics, with SCDOY standing out as a primary driving factor.

Our research highlights the adaptability of ML models in comprehending and predicting LSWT behaviors. This study provides insight into how environmental factors shape LSWT, contributing to a broader understanding of climate impacts. The importance of our work lies in advancing the capabilities of ML models for insightful predictions in the realm of lake temperature dynamics, generalized to other lakes.

# Chapter 5. Ice dynamics in boreal lakes

## 5.1. Introduction

Climate change is currently regarded as one of the most significant threats to lake ecosystems worldwide (Woolway et al, 2020; Heino et al., 2021), and one of the possible effects is the modification of ice phenology (Hewitt et al., 2018). In fact, winter ice-on lakes is being depleted at an unprecedented rate as a result of global warming (Sharma et al., 2019; Jenny et al., 2020). Changes in the ice cover are also important for the impact on the regional economy, as it has been recognized for the Great Lakes already several decades ago (Niimi, 1982; Magnuson et al., 1998; Hayhoe et al., 2010).

The Northern Hemisphere's lakes, distributed across various regions, are experiencing accelerated loss of lake ice due to large-scale climatic changes. This rapid ice loss has the potential to significantly impact the crucial ecosystem services provided by lake ice (Imrit et al., 2022). As shown by Sharma et al. (2020), over a 150-year period, between 1846 and 1995, ice-on occurred 5.8 days later and ice-off happened 6.5 days earlier per century. However, when considering the most recent 25 years (1995-2019), the rate of change increased significantly, with ice-on and ice-off changing at rates of 72 days/century and -32 days/century, respectively. Consequently, the duration of winter ice cover has decreased, and some lakes are experiencing more winters with minimal or no ice cover. Woolway et al. (2020) presented that higher latitudes are expected to undergo a more significant alteration in the duration of safe ice, while densely populated lower-latitude regions will experience the highest percentage change.

Lake ice regimes are largely defined by latitude, altitude and depth of the lake (Walsh et al. 1998). Ice cover is cyclical in the boreal zone, tundra, and mountainous areas in temperate climates (Kirillin et al., 2012), although it is permanent in very high elevations and high polar latitudes. Ice cover partially disconnects the lake water from the atmosphere and sunlight, and affects biological and physical processes in the lake (Leppäranta, 2014) stimulating anoxia and menacing threats such as fish kills (Barica and Mathias 1979; Shuter et al. 2012). The impact of climate change on the dynamics of dissolved oxygen (DO) in the bottom layer of alpine lakes beneath the ice cover will be determined by whether the duration of the ice cover or the characteristics of the inverse stratification dominate (Perga et al., 2024). The ice cover also prevents the transfer of momentum from the wind to water, damping turbulent mixing and reducing the lake circulation. Except for lakes largely affected by geothermal heat fluxes, the vertical heat transfer is reduced when the surface water reaches the freezing point. However, the heat flux in ice covered lakes establishes again in spring as the result of

the increased downward heat transfer to the lake caused by solar radiation (Leppäranta, 2014; Bouffard et al., 2019; Volkov et al., 2019). This largely depends on the optical properties of the ice cover, and in lakes with highly transparent ice cover (such as lakes in the Tibetan Plateau) the large amount of solar radiation penetrating under the ice promotes full mixing and warming, thus favorable conditions for aquatic life (Kirillin et al., 2021). Due to increased solar radiation received during the months when ice melts, the sensitivity of additional warming to the timing of ice loss is quantitatively greater compared to the month when ice forms (Li et al., 2012).

The ice cover on lakes consists of white ice, black ice, water, slush and snow layers that affect the ice growths as well (Marshall, 1965). Generally, two types are dominant: black ice, known as congelation ice, and white ice, known as snow-ice (Gow, 1986). White ice possesses approximately half the load-bearing strength of black ice (Barrette, 2011). Furthermore, white ice significantly restricts the penetration of photosynthetically active radiation through the ice layer, in contrast to clear black ice, which typically has minimal impact on light penetration, similar to that of lake water (Lei, 2009; Weyhenmeyer et al., 2022). The formation of white ice often occurs through a process involving the accumulation of snow on pre-existing ice, followed by melting and subsequent refreezing (Ashton, 2011). Additionally, white ice can form when rainwater falls on a layer of snow, creating slush that later freezes and transforms into white ice. Another mechanism for white ice formation is the exertion of pressure from a heavy snow load, which can force lake water to rise through cracks in the ice matrix and freeze, resulting in the formation of white ice (Brown and Duguay, 2010; Weyhenmeyer et al., 2022). Usually, IT is measured as the sum of black and white ice. As measurements are performed at point and are sparse in many cases, they may not be entirely representative of the whole lake (Assel, 1976; Adams and Roulet, 1984).

The models that simulate and predict IT are categorized into two main groups: physically based models and statistical models. Machine learning (ML) is a branch of statistical models. We chose ML models since the main advantage of ML models over physically based models is that no assumptions are introduced in the description of the process, thus allowing for a very flexible selection of the predictors.

There is an old tradition in the use of statistical models for predicting IT. As one of the first statistical models, Assel (1976) employed regression equations to establish a connection between freezing degree-days and IT across 24 sites; the predictors he studied were freezing degree-days and thawing degree-days for several winters with weekly measurements until March. Building on Assel's work, Assel et al. (2004) explored regression models to forecast the onset of ice cover on lake, specifically focusing on the beginning of the month (BOM). The first model is the climatological model, which

provides a long-term average for the BOM ice cover. Another study by Baker et al. (1976) predicted the freeze-up and ice formation on the Great lakes by four methods: freezing degree day totals, departure from normal air temperature, 30-day temperature outlooks and Lisitzin-Rodhe-Biello equation. In their study, freezing degree day totals outperformed the other three techniques.

In recent years, IT has been predicted using advanced statistical methods, notably the application of ML techniques. Despite the limited number of studies that employ ML for IT prediction, noteworthy contributions in this domain include the work of Zaier et al. (2010), who utilized an artificial neural network (ANN), a ML method, to estimate IT during the initial phase of winter ice growth across various lakes in Canada. Their investigation explored key predictors, encompassing daily snow depth, rainfall, mean air temperature, and total solar radiation. In addition, Watson et al. (2021) undertook a study aimed at forecasting the occurrence of ice cover on Lake George, New York. Employing ML classifiers, they leveraged a comprehensive range of predictor variables, comprising ice coverage date, air temperature, wind speed, relative humidity, precipitation, snow, cloud cover, and surface pressure. Despite the constraints imposed by the available data, the ML classifiers demonstrated sufficient performance in predicting complete ice coverage. These research endeavors held the promise of delivering valuable insights into the intricacies of predicting and understanding ice cover dynamics.

In addition to IT predictions, several studies have explored the application of ML for classification purposes in ice fields on lakes. As an example, a study by Xie et al. (2020) demonstrates the potential of ML techniques for efficiently studying lake ice phenology worldwide. They used a convolutional neural network (CNN) to classify lakes based on freezing patterns, achieving 91% accuracy for annual freezing lakes and 100% for non-freezing lakes. They employed support vector regression (SVR), as a ML algorithm, to extract freeze-up start and break-up end of lake ice phenology from microwave data, with strong correlation coefficients ($R^2$) of 0.8928 and 0.8899, respectively. Moreover, Wu et al. (2021) assessed four ML classifiers (multinomial logistic regression, MLR; support vector machine, SVM; random forest, RF; gradient boosting trees, GBT) for accurately mapping the presence of lake ice, lake water, and cloud cover during ice break-up and freeze-up. Samples from 17 diverse lakes across Europe and North America were used for training and validation. RF and GBT achieved accuracies above 98%, providing visually accurate representations, and exhibited strong consistency across different ice seasons and demonstrated effective spatial transferability among the 17 lakes.

This research addresses crucial aspects regarding the integration of meteorological variables in the context of an IT model. Specifically, our focus is on optimizing the use of meteorological variables to derive meaningful insights, employing ML, an aspect often understated in previous studies (Zaier et al., 2010; Watson et al., 2021). While some studies, including Zaier et al. (2010), have incorporated ML for IT prediction, none have systematically identified and ranked the influential factors for IT within the model. Our study strives to fill these gaps through a comprehensive analysis. The subsequent sections detail the analyzed empirical cases, the methods employed for IT forecasting, and the feature ranking process.

## 5.2. Materials and methods

### 5.2.1. Study sites

Two Swedish lakes were chosen as case studies from the database of the Swedish Meteorological and Hydrological Institute (SMHI). SMHI mainly measures and gathers data on water discharge to and from lakes, water level, IT, freeze-up and break-up of ice. At SMHI, hydrological data are calculated, analyzed and kept in several databases for predictions, research, and warnings. According to the fact that ML demands large amount of data, the two lakes with the longest records were selected: Lake Runn and Lake Gouta. We stress the fact that these are somehow exceptional cases because having extensive records of ice thickness is uncommon, whereas ice-on and ice-off dates are more frequently documented. Moreover, the two lakes are located in different regions, thus allowing for detecting a possibly different response to atmospheric forcing. The information about these lakes is reported in Table 5.1.

The SMHI data set provides information on various parameters such as black IT, total IT, snow layer thickness, slush, snow ice formed from slush. We have chosen to use the total IT data because it provides an integrated measurement of the total ice compared to other measures.

Table 5.1. Available information about the two case studies.

| Lake's name | Runn | Gouta |
|---|---|---|
| Maximum depth (m) | 32 | 58 |
| Average depth (m) | 8.3 | 17.2 |
| Area ($km^2$) | 63.5 | 31.6 |
| Volume ($10^6$ $m^3$) | 530.61 | 543.52 |
| Date range | 05/01/1980 – 13/04/2012 | 10/01/1980 – 19/05/2012 |
| Latitude, Longitude | 60.53° N, 15.67° E | 65.67° N, 15.38° E |
| Elevation (m a.s.l.) | 107 | 438.6 |
| SMHI measurements | | |
| Maximum IT (cm) | 42 | 56 |
| Number of available measurements | 755 | 790 |
| Data from ERA5 reanalysis * | | |
| Range of AT (°C) | -25.84 – 25.03 | -36.80 – 20.14 |
| Range of SWR (W $m^{-2}$) | 0.82 – 341.26 | 0.10 – 345.66 |
| Range of LWR (W $m^{-2}$) | 148.93 – 391.99 | 130.99 – 385.00 |
| Range of WS (m $s^{-1}$) | 0.63 – 8.42 | 0.52 – 8.91 |
| Range of AP (Pa) | 92901 – 103149 | 86891 – 95653 |
| Range of R (kg $m^{-2}$ $s^{-1}$) | 0 – $6.44\times10^{-4}$ | 0 – $3.79\times10^{-4}$ |
| Range of S (kg $m^{-2}$ $s^{-1}$) | 0 – $4.22\times10^{-4}$ | 0 – $6.12\times10^{-4}$ |
| Range of SH (kg $kg^{-1}$) | $2.47\times10^{-4}$ – $134.32\times10^{-4}$ | $9\times10^{-4}$ – $124.87\times10^{-4}$ |

* All ranges are expressed for daily averaged values and refer only to the cold season. The period considered for the analysis is from 03/01/1980 to 01/01/2020.

Owing to the satisfactory resolution offered by the ERA5 reanalysis dataset (0.5° × 0.5°; Hersbach et al., 2020), the fifth generation of ECMWF reanalysis of global climate and weather, the meteorological variables were obtained from 1979. The following variables were extracted based on latitude and longitude of the studied lakes, at 2 m above the lake surface, for the entire period of availability of IT measurements: air temperature (AT), wind speed (WS), air pressure (AP), rainfall (R), snowfall (S), specific humidity (SH), downward longwave radiations (LWR) and shortwave radiations and (SWR). Specifically, we downloaded these data from the bias-corrected reconstruction of near-surface meteorological variables product. Daily averages were computed from the higher-frequency reanalysis dataset. These are the input variables for the ML model developed to predict IT, and we will use them as 'features' or 'predictors' in the following discussion.

The day of the year (DOY) was used as a predictor, as well. Since DOY changes are not continuous, we considered the sin(DOY/nDOY) and cos(DOY/nDOY), where nDOY is the number of days in a year (see also the discussion in Chapter 3). The sine and cosine of the DOY (SCDOY) provides the

same content of information as DOY but allows for a smoother transition between the end of each year and the following one (Yousefi and Toffolon, 2022).

### 5.2.2. Data preparation

Three sequential steps are needed to implement the model: preparing the inputs, preprocessing them, and using the ML model for simulation. Figure 5.1 shows all these phases, which are described in the following sections. We compared two ML approaches, the long short-term memory (LSTM) and backpropagation neural network (BPNN). Subsequently, we chose LSTM for the analysis of feature ranking (explained in section 2.7) because of having more robust results than BPNN. The FR results are presented for the ice-on period and also ice freezing period to evaluate the importance of features on ice formation as well. We escaped the ice melting period since we have fewer data in this period.

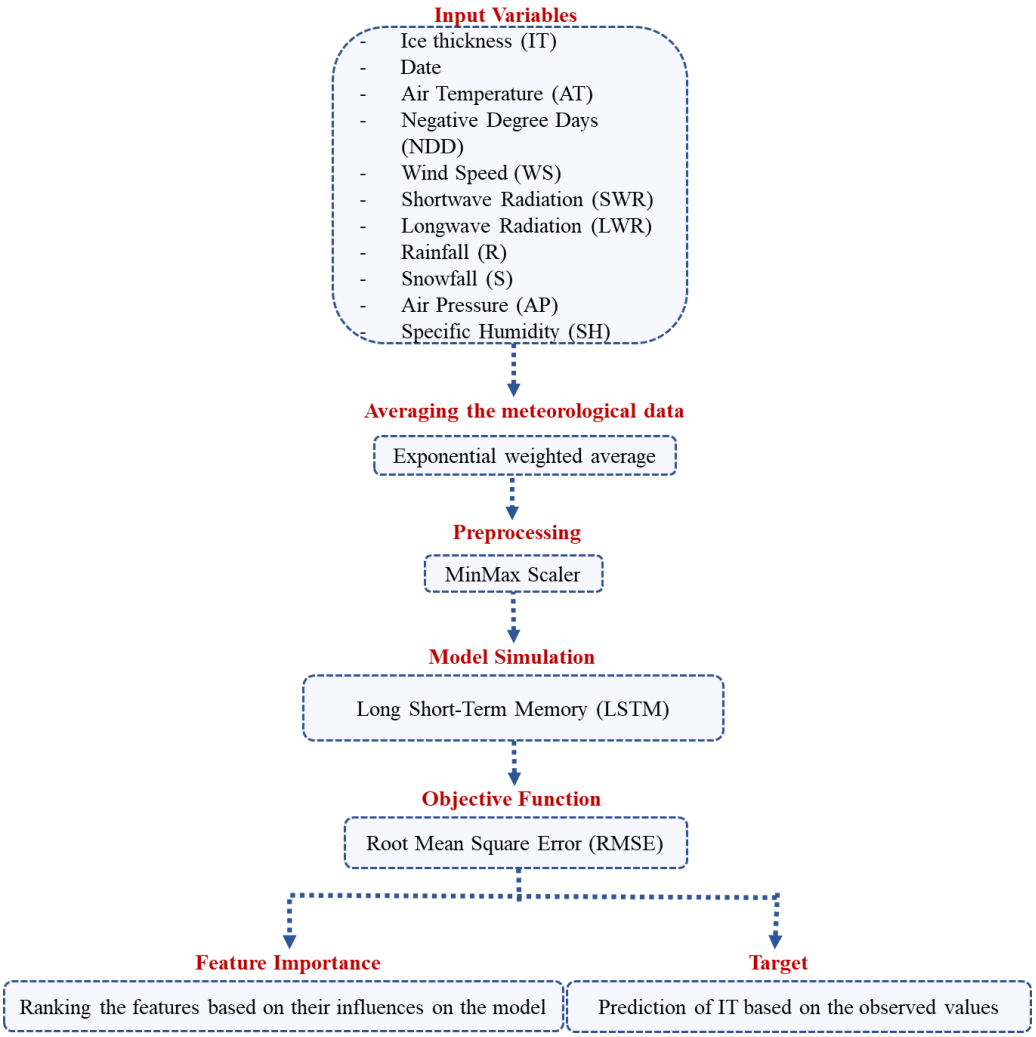



Figure 5.1. The modelling framework.

The model is expected to provide daily values of IT as a function of daily input of the meteorological variables. The available measurements of IT data are sparsely distributed in time, with an interval

typically longer than one day. A peculiarity of the IT problem is that ice is present only in a period, and for the rest of the year there is no ice-on the lake surface. The duration of the ice cover period is unknown, as well as the ice-on and ice-off dates. Therefore, there is the problem of defining how the information about the absence of ice can be properly provided to the model. After several tests, we decided to add null values (IT = 0) in the ice-off period with the same frequency of the added zeros as, on average, are available for the IT data in the ice-on period. Without assuming the ice-on and ice-off dates, we considered one month preceding the first IT values and one month after the final available IT value for the entire time series. In this way, the model receives balanced information about the two periods. In fact, on the one hand, using a much lower frequency of zeros in the ice-free period would not give enough data for the model to learn about the absence of ice during summer correctly, thus having no clues about how to predict the ice-on and ice-off dates. On the other hand, a higher frequency of zeros compared to the mean frequency of IT observations (e.g., adding them for all days in the warm season) would bias the ML tuning towards the summer months thus hampering the proper calibration of the IT in the ice-covered period. The start and the end dates for the addition of the zero values were defined, for each year, as one month after the last recorded date with ice and one month before the first recorded date with ice (identified based on the entire historical dataset), respectively.

The input data for the ML model were chosen within the set of the available meteorological variables. As lakes typically retain a memory of the previous conditions, which is the result of the history of the forcing, we developed two different strategies: (i) applying the ML model in a way to consider the memory in the model such as using LSTM algorithm and (ii) weighting the forcing over a period before the simulated day. Concerning the 2$^{nd}$ option, instead of considering the feature in the current day, we used the weighted average over a number of previous days that is determined by the best model performance (see details in section 2.4 and Chapter 4).In fact, we used the weighted average of each input variable X computed using the best time window (the one producing the lowest RMSE) with exponentially decaying weights (section 2.4). Therefore, given the deep depth of the lakes that we are studying and the absence of minimum RMSE below days of window size for lake Runn, we opted for a value that would yield a noticeable reduction in RMSE. Accordingly, 40 days is considered as the window size.

Beside considering the weighted average of each meteorological variable, we tested different ways of considering the history of AT. Several authors (Assel, 1974; Baker et al., 1976; Assel et al., 2004; Magee and Wu, 2017; Imrit et al., 2022; Weyhenmeyer et al., 2022) assumed that AT is important for the formation and evolution of IT: not directly as the instantaneous value, however, but as a cumulative effect. Therefore, we also considered the negative degree days (NDD; e.g., Bilello, 1964):

for each day, the negative value of AT is cumulated since the first negative value of AT (less than 0°C) of each winter. The NDD is considered in the analysis until it is negative. This is equivalent to considering the average but starting from a fixed date and with a constant weight. Baker et al. (1976) claim that NDD is influential in the freeze-up phase. Assel (2004) also uses the total number of NDD to look for resembling historical events and related ice cover.

### 5.2.3. Preprocessing

Since the variation range of the data covers different scales (see Table 5.1), we normalize the data. There are several scaling methods; after testing Min-Max, Standard, and Robust scalers (Varoquaux et al., 2015), we found that the Min-Max scaler works better in our case. Then, we split the data with k-fold cross validation methods, explained in section 2.5.

### 5.2.4. ML model

We compared BPNN approach and LSTM to model and predict the IT. We found that LSTM is a superior method (proven in section 5.3.2), thus we used it to get the feature importance to determine the significance and impact of the predictors on modelling IT.

We analyzed the model's performance by referring to the root mean square error (RMSE) and the standard Nash-Sutcliff efficiency index (NSE) for training and test sets (Nash and Sutcliffe, 1970; Nevitt and Hancoc, 2000; McCuen et al., 2006), as explained in section 2.3. Perfect fit is obtained for RMSE = 0 and NSE = 1; note that a value NSE = 0 would result from using the observation mean as prediction model. Since the datasets are fixed, the two indices are equivalent in terms of selecting the best model, and in the following discussion we will only refer to RMSE. The RMSE and predicted values are obtained based on the average of 20 runs of predicted IT. In this application, different from the analysis performed in Chapter 4, we did not use the modified NSE* because of the difficulty to define a meaningful mean (climatological) year.

To exclude any potential influence of zero values during the warm period, we only calculated the RMSE for the cold period for BPNN while for LSTM is the whole period. Additionally, the peaks of IT vary from year to year, and they may be difficult to capture. To address this challenge, we defined an objective function based on winter period for BPNN:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \tag{5.1}$$

where $RMSE$ is the root mean square error that we considered to evaluate the global model performance which is for the whole year (including summer), $N$ is the number of data used for model, $x_i$ and $\hat{x}_i$ are the observed IT and its average in the whole period.

## 5.3. Results and discussion

In this section, we present and analyze the performance of BPNN and LSTM, for predicting IT in the two selected lakes, Runn and Gouta. In addition, we explore how different ways of utilizing meteorological data affect the accuracy of the prediction.

In the following sections, the characterization of the two lakes, the comparison of performances of BPNN and LSTM, the methodology to choose the best part of the data set as the testing and training sets, and the importance of the training features, are presented and discussed.

### 5.3.1. Characterization of the two lakes

Using an approach that is not based on physical rules, it is especially important to comprehend how the physics underlying ice formation and melting should be included to make reliable predictions. On the other hand, understanding the physical processes would aid in determining which factors should be incorporated into the computations. ML algorithms are general-purpose tools, and the choice of the group of predictors has a significant impact on how well they can perform. The predictors that we have chosen are AT, SWR, LWR, WS, R, S, AP, SH, sine and cosine of DOY (two variables that we indicate as SDOY and CDOY for simplicity).

As already discussed, IT is influenced by the meteorological forcing not only of the current day but also of the previous days. Accordingly, we considered the weighted average of the input data for all variables.

Figure 5.2 summarizes the seasonal dynamics of IT in the two lakes. The maximum IT is larger in Lake Gouta (in the North, at a higher elevation, see Table 5.1) than in Lake Runn of around 10 cm. Also, the ice duration is longer for Lake Gouta on average, which means that ice forms earlier and lasts for longer than in Lake Runn. On the other hand, Lake Runn shows more variability of IT in the different years.

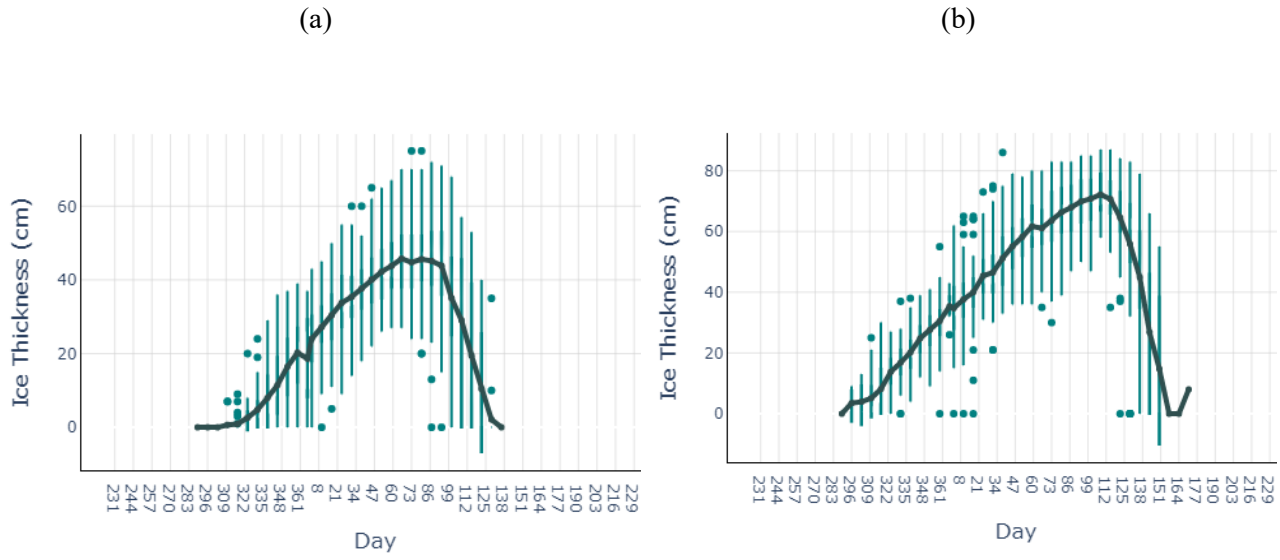


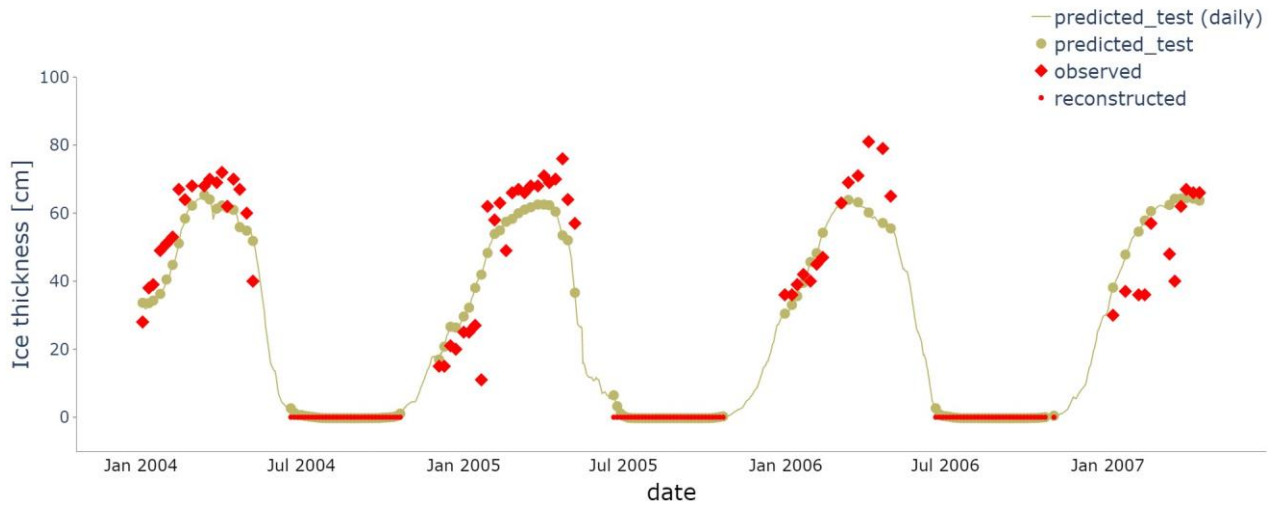

Figure 5.2. Box plots of IT considering the whole time series, with the average (climatological year) shown with a line, for Lake Runn (a) and Gouta (b). The boxes represent the variability of the data (25th-75th percentiles) in a time window of 7 days, the outliers are indicated with dots.

### 5.3.2. The reference model depending on all features

The ML model can be improved through the incorporation of the information about the previous state of the system. As it was explained in section 5.2.2, we considered LSTM method (Liu et al., 2022) in comparison to BPNN. The inputs of the models are the weighted averaged meteorological variables (AT, SWR, LWR, WS, R, S, AP, SH), NDD and SCDOY. The hyperparameters used for the two models are specified in Appendix A.3.

As indicated in Figure 5.3 for Lake Gouta, the RMSE values for the BPNN base model are 12.23 cm for the training dataset and 13.41 cm for the test dataset. The LSTM model, which incorporates memory from previous days, delivers more promising outcomes. The training RMSE achieves 11.74 cm, not far from the BPNN model, while the test RMSE settles at 14.22 cm, with a more significant improvement. Based on considerations of both computational cost and robustness, the LSTM model has been selected as the preferred option for our subsequent analyses. To clarify what we mean by robustness, LSTM is more consistent in terms of the output in each iteration. For example, for 20 random initiations of the ML model, we got the average difference of 2.60 cm and 2.43 cm between the runs for BPNN for training and test sets, respectively, while these values are 1.10 cm and 1.10 cm for LSTM method. Accordingly, we utilize LSTM algorithm for our analysis in the following sections.

(a)




(b)

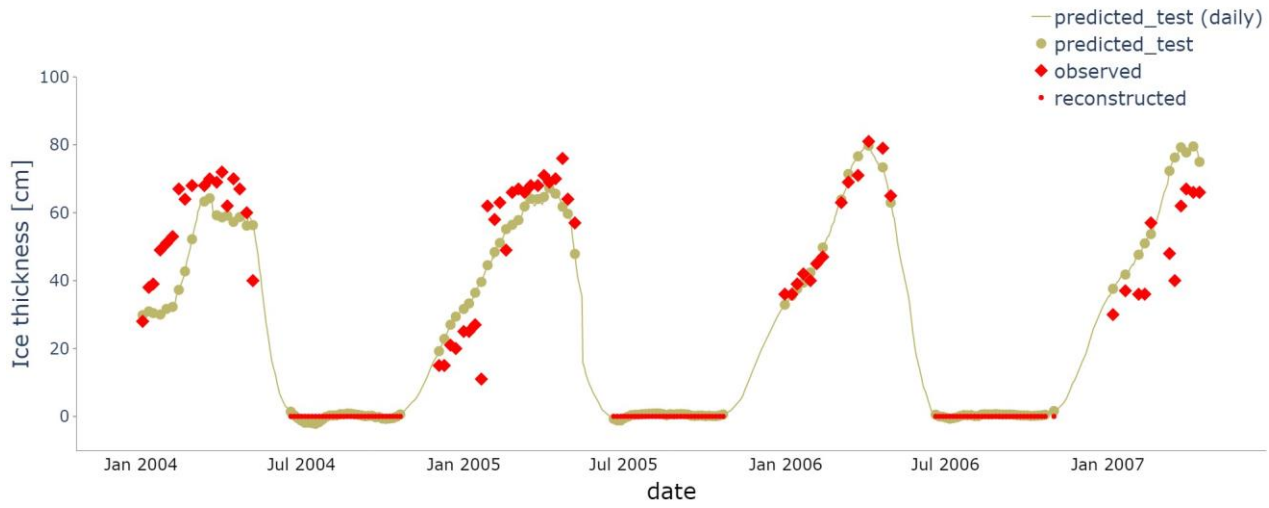



Figure 5.3. An example of the modeled IT by BPNN (a) and LSTM (b) in three years for Lake Gouta.

Following the K fold selection of Training and test sets presented in section 2.5, we divided the data into five folds, and each time, one of the folds was considered as the test set, and the remaining folds were used as training sets. We repeated this process for each fold, resulting in a total of five model performance evaluations for both the training and test sets modeled by LSTM (Figure 5.4).

Through careful selection of suitable folds for the training and test sets, the model achieves favorable performance on both datasets. For instance, in both lakes, the 1$^{st}$ fold exhibited reasonable performance when used as a training set but performed poorly as a test set. As our aim is to develop a model that can be effectively applied to the entire time series, we selected the 4$^{th}$ fold for both lakes,

as they represent the folds that are performing best in the combination of training and test (Figure 5.4).

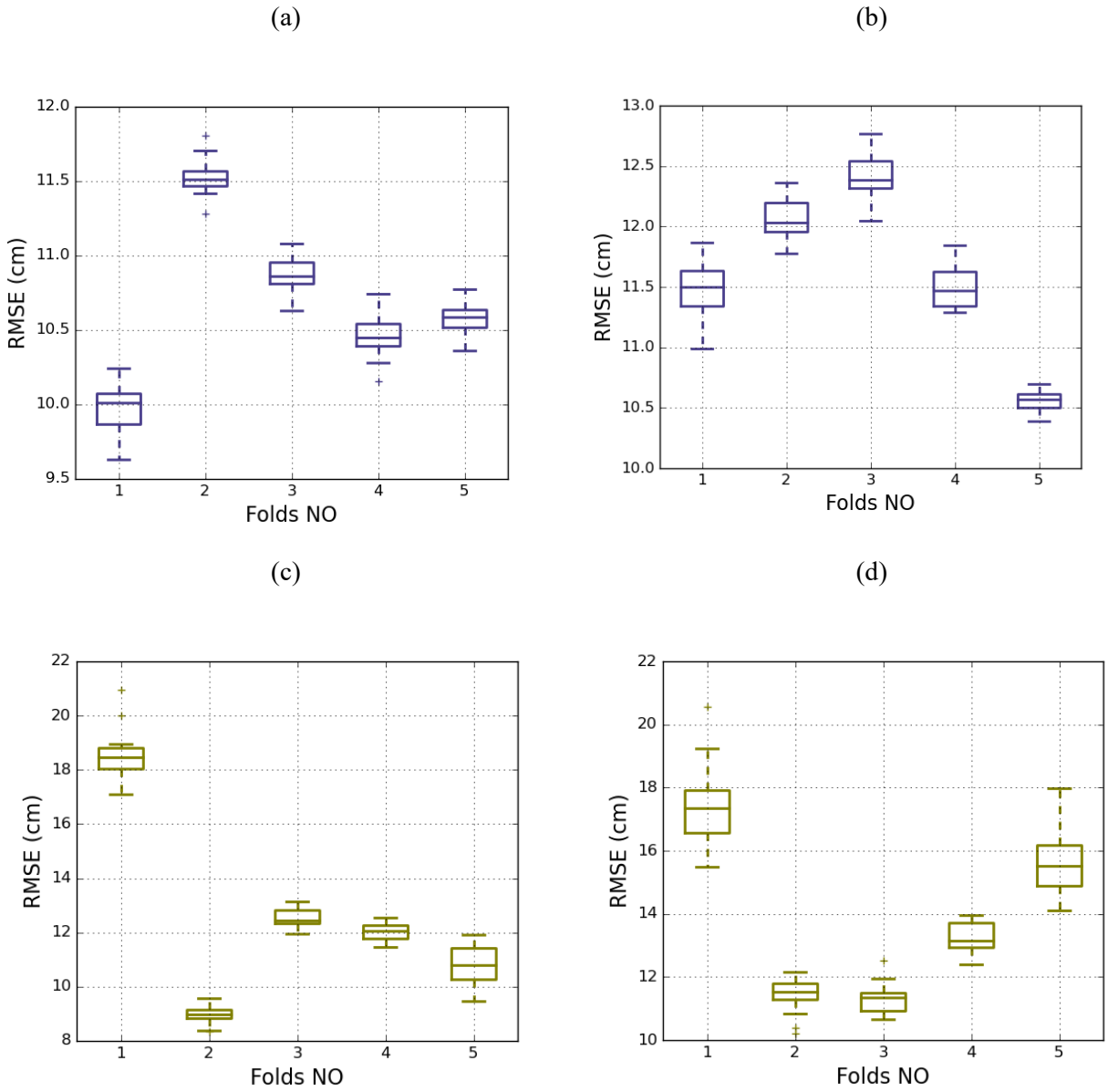
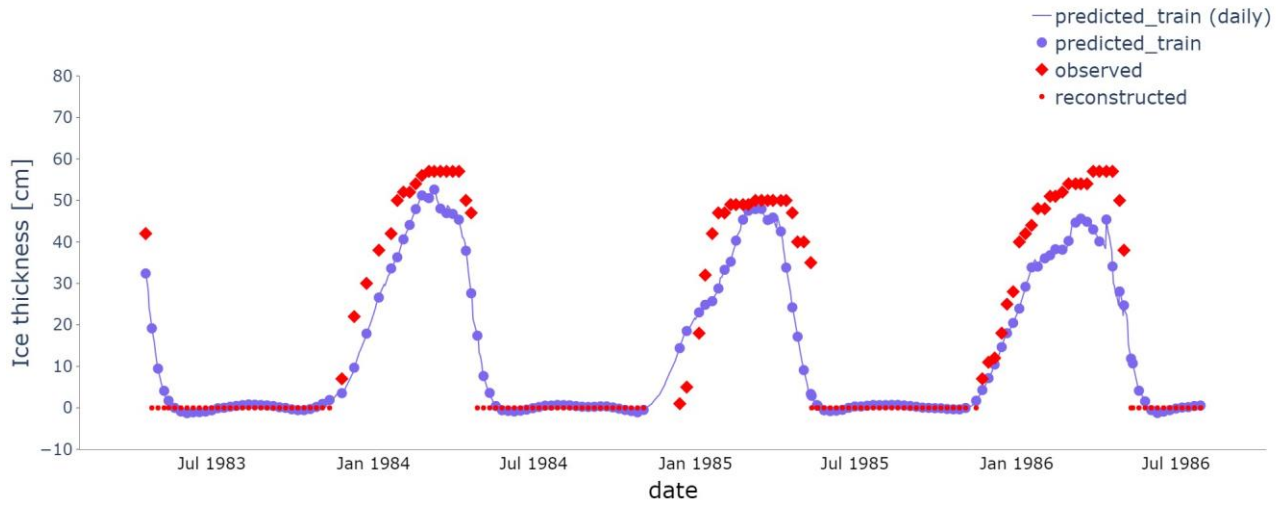



Figure 5.4. RMSE for the training set in Lake Runn (a) and Gouta (b) for the 5 folds using LSTM model. The box plots represent the distribution of the RMSE in 20 different iterations of the model. The same in (c) and (d) for the test set.

To demonstrate the interannual variability, the IT series for both lakes, including training set and test set values, have been depicted in Figure 5.5 for about three years. The RMSE values of the training and test sets were found to be 11.17 cm and 11.72 cm, respectively. The values for Lake Gouta are 11.74 and 12.45 cm for training and test sets in the LSTM model.

(a)




(b)

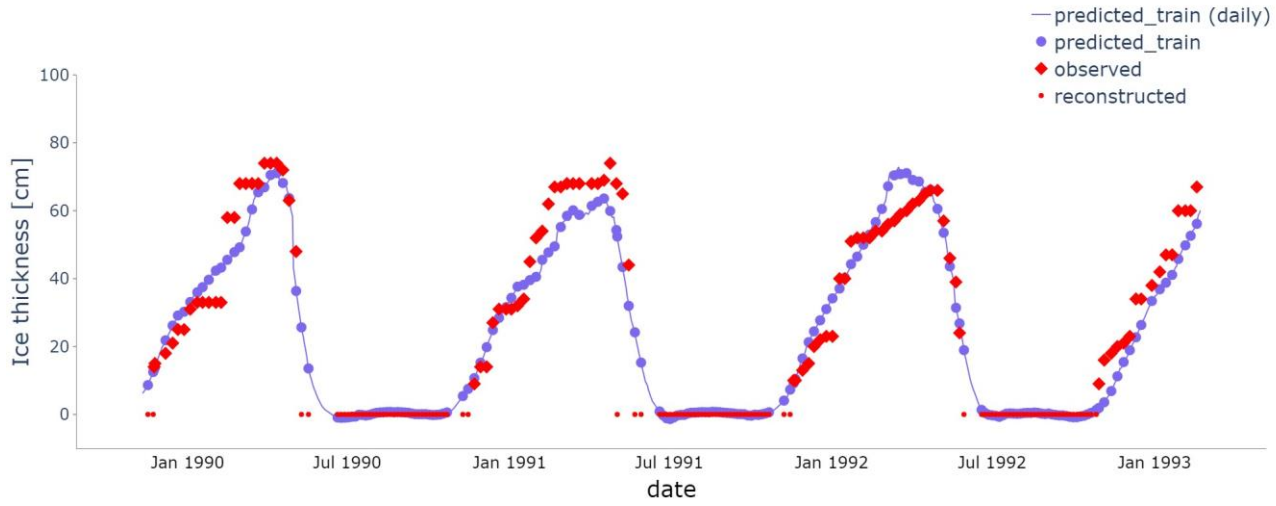



Figure 5.5. An example of the modeled IT compared with observations using LSTM model, for lake Runn (a) and Lake Gouta (b).

The parity diagrams obtained with the LSTM model for the entire series for both lakes are presented in Figure 5.6. The findings show that the better predictions for ice-on period than summer period (observed values equal to zero). The utilization of SDOY significantly impacts the model, particularly in thicker ice values that are simulated and predicted without considerable variation from year to year. Consequently, the errors are higher in thicker IT values.

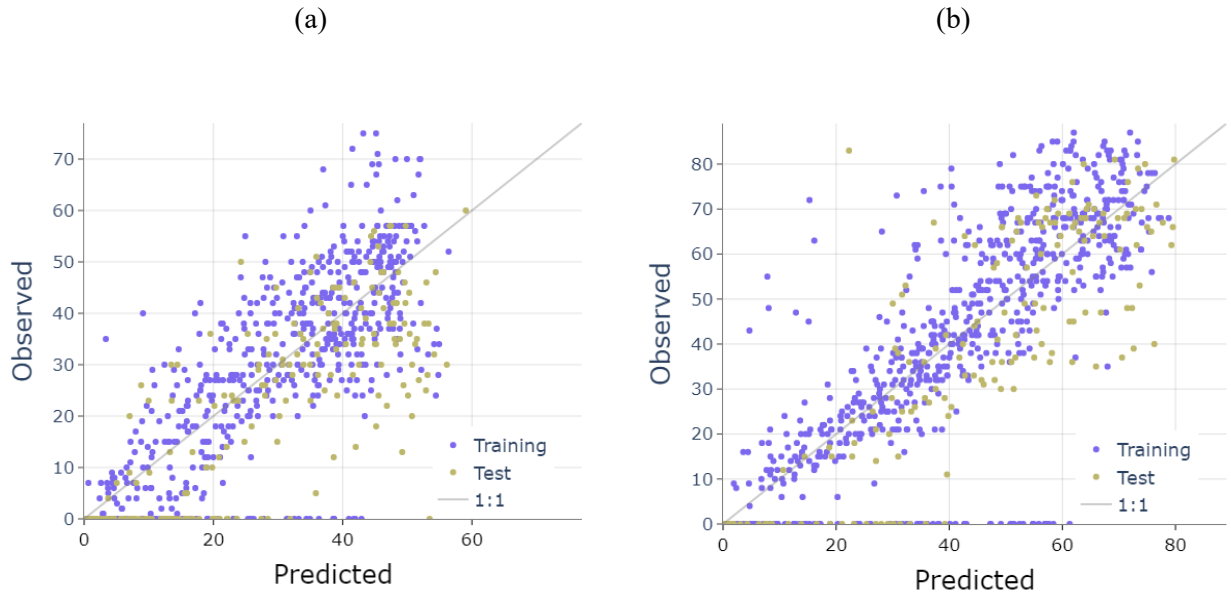



Figure 5.6. Scatter plots between modeled and observed values of IT for Lake Runn (a) and Lake Gouta (b), obtained by LSTM model.

### 5.3.3. Feature extraction

In order to select the most relevant information that allows for accurate predictions, here we analyze the main role that is played by the different predictors, and then analyze their effect based on the model's performances for the ice-on and ice formation period. We choose not to show the ice melting period due to deficiency of the data.

The assessment of the feature importance is presented only for the LSTM model. It is performed by analyzing the difference of the RMSE between the model using all features and a similar model for which the values of a specific feature are randomly shuffled, so that its physical effect is lost. In other words, the significance of a variable is quantified as the discrepancy in RMSE between the reference model and a model forced by random variations in the variable.

As illustrated in Figure 5.7, SDOY emerges as the primary predictor of IT. The y axis value is the difference of normal case (without shuffling) RMSE and the shuffled cases RMSE divided by the normal case RMSE. While SDOY is important, the CDOY impacts low because the time vector expressed by SDOY expresses the progression of time in the winter period more than CDOY in this period (presented in Figure 5.8 as an example of Lake Gouta). Other notable predictors include SWR for Lake Runn and SH for Lake Gouta. Elevated humidity levels correspond to an anticipated increase in IT (Israelachvili, 2011). Furthermore, S, AT and NDD exert significant influence over IT formation.

AT is expected to be a critical variable because it strongly affects the rate of heat transfer between the atmosphere and ice. Warmer AT causes the ice to melt, while colder AT leads to ice formation. The

NDD plays an even more crucial role in IT prediction because it represents the cumulative effect of cold temperatures on ice formation and can provide a better indication of the overall coldness of a season than AT alone.

SWR is another significant variable that influences IT prediction. It refers to the amount of solar radiation that reaches the ice surface and is absorbed or reflected. The absorbed radiation heats up the ice, leading to melting, while a larger albedo (the reflected radiation) contributes to ice formation.

When snow accumulates on top of ice, it creates an insulating layer that prevents the underlying ice from freezing as much as it would if it were exposed to the cold air directly. As a result, the IT is directly related to the amount of snowfall (S) that has accumulated on top of it throughout the winter. This quantity is implicitly considered in the weighted average of the daily snowfall.

When it comes to the specific period of ice formation, SDOY retains its primary role, followed by NDD. Owing to the limited availability of data for ice melting (approximately 100 data points), we exercise caution and opt not to present those particular results.

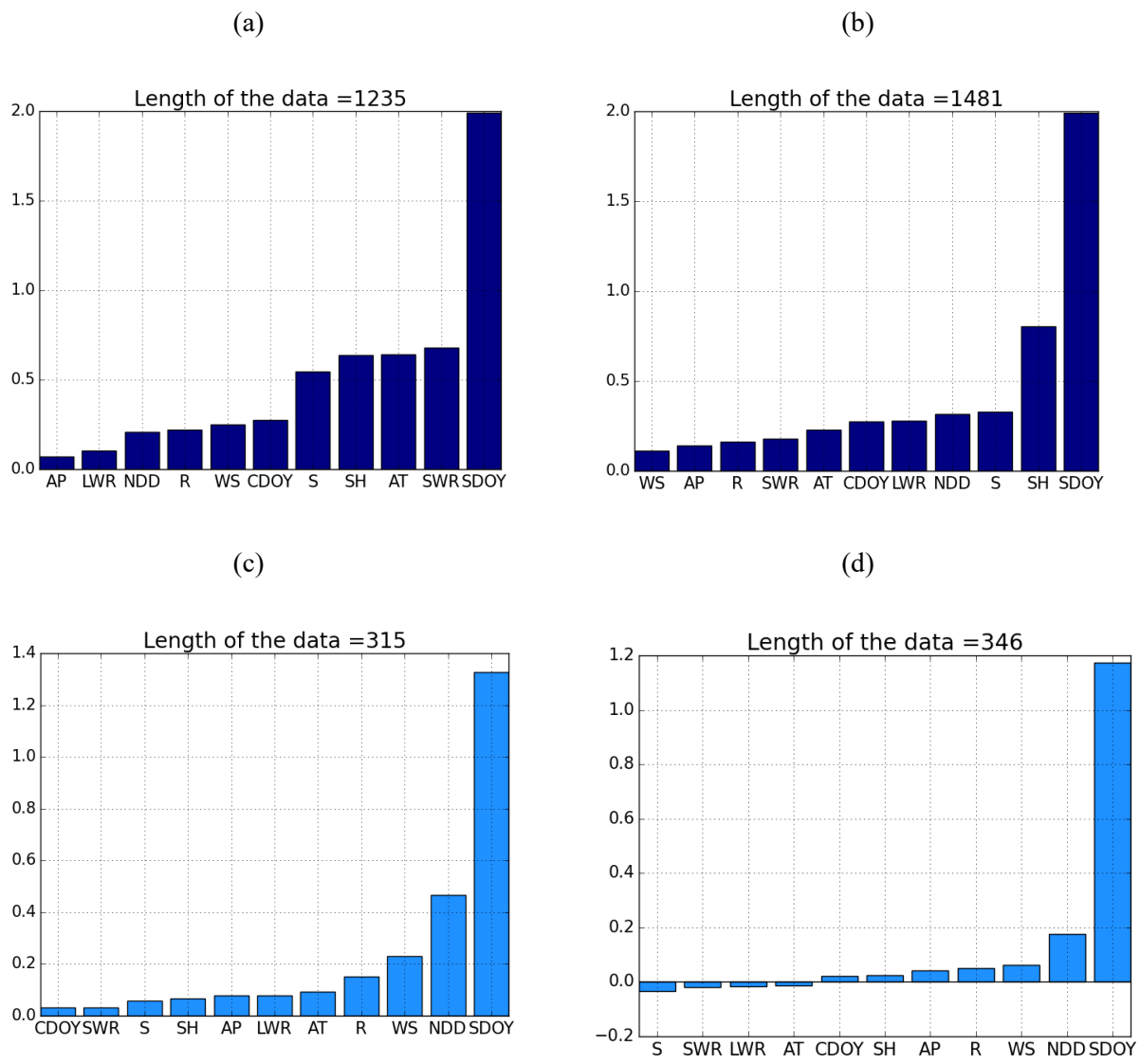



Figure 5.7. Feature importance for Lake Runn (left column) and Lake Gouta (right column) considering the whole period (a, b), the ice forming period (c, d) using LSTM model.

We separated SDOY and CDOY in this context because their impact on IT prediction varies, unlike in Chapter 4 where they were considered together. However, for LSWT prediction, both SDOY and CDOY play significant roles (see Figure 5.8).

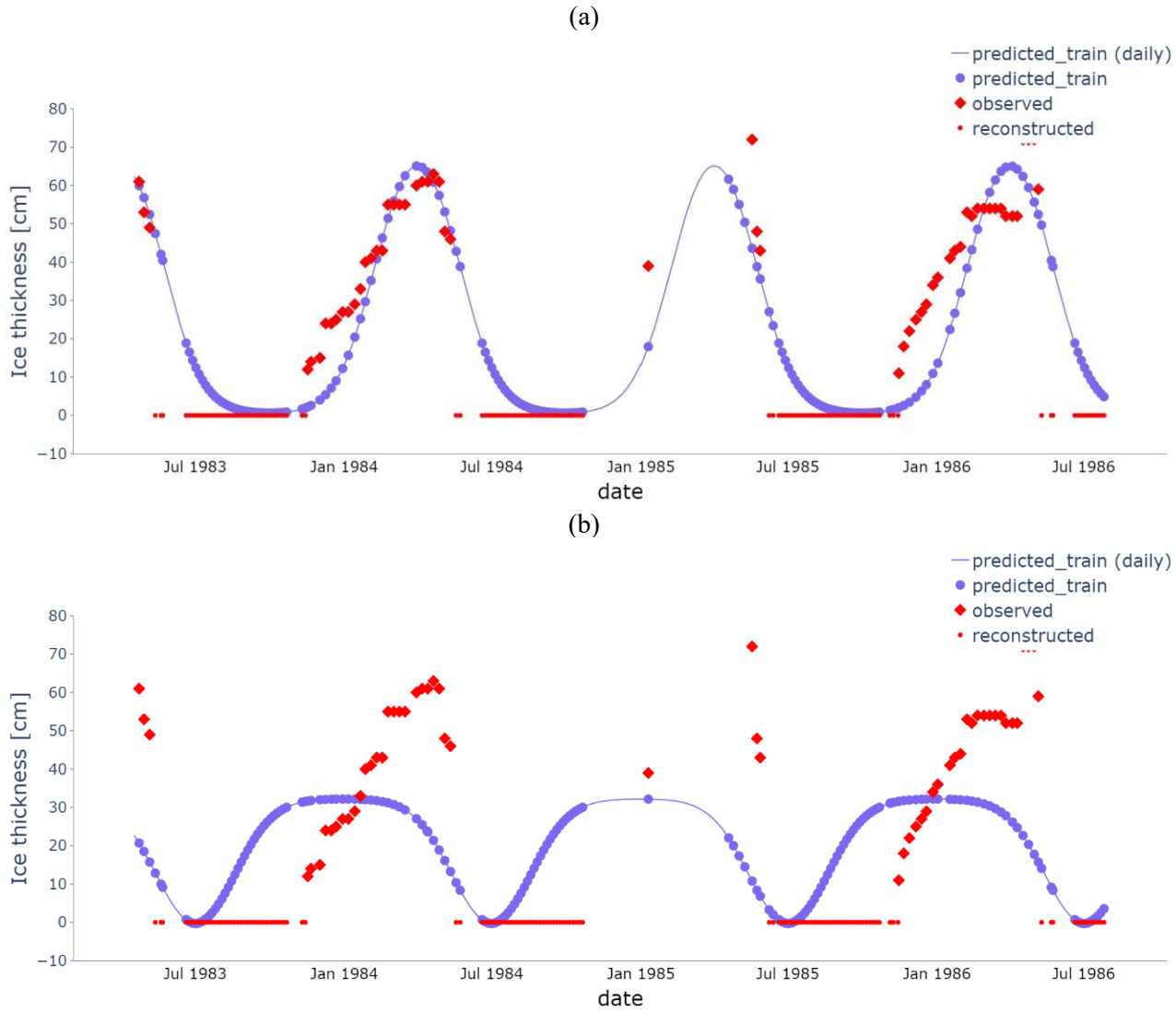



Figure 5.8. The LSTM model using solely SDOY (a) and CDOY (b) for Lake Gouta as an example.

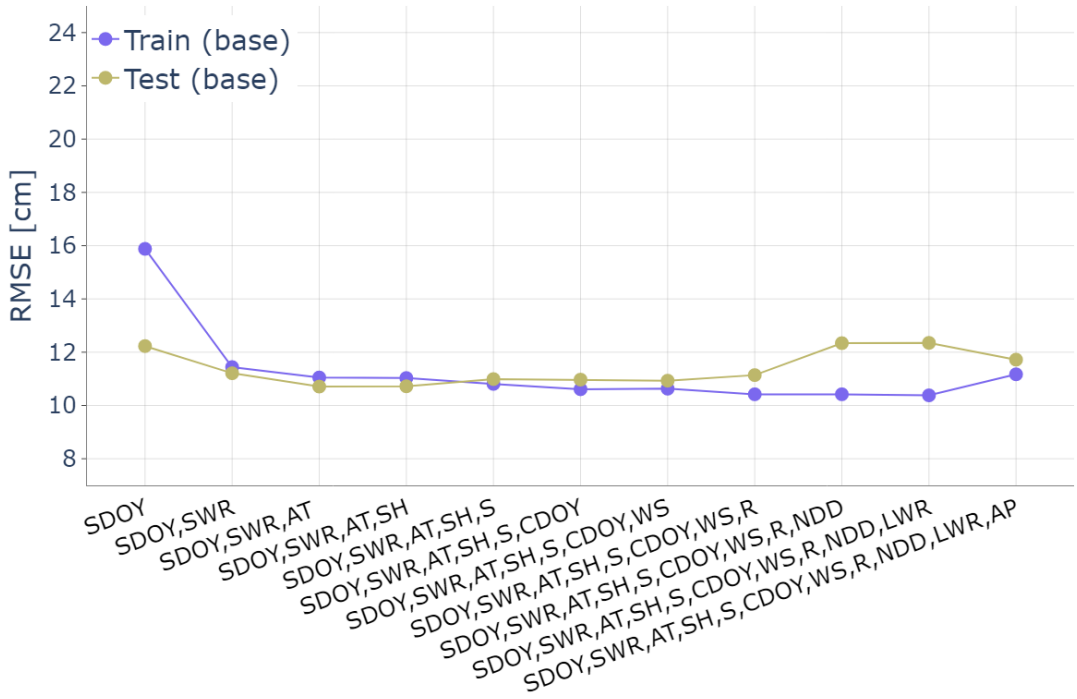
### 5.3.4. Model simplification based on features' selection

The accurate prediction of IT is a complex task that requires the consideration of various essential variables. SDOY has been shown to be a critical predictor of IT in the previous section. However, our analysis reveals that removing unnecessary variables may change the predictive model's performance slightly (Figure 5.9).

As a first analysis, we conducted an in-depth analysis of the critical variables that impact the predictive capability of the LSTM models of the two lakes. The features are added one by one with separate models, following the previously determined feature ranking: based on our findings, for both lakes, SDOY plays the most crucial role (Figure 5.7), followed by SWR as the 2$^{nd}$ most important predictor for Lake Runn, and SH for Lake Gouta.

However, considering SDOY is characterized by a noticeable decline in performance with the model that considers also SWR (Lake Runn, Figure 5.9a) or SH (Lake Gouta, Figure 5.9b). It's important to highlight that including additional predictors can deteriorate test results and potentially introduce the issue of overparameterization.

(a)




(b)

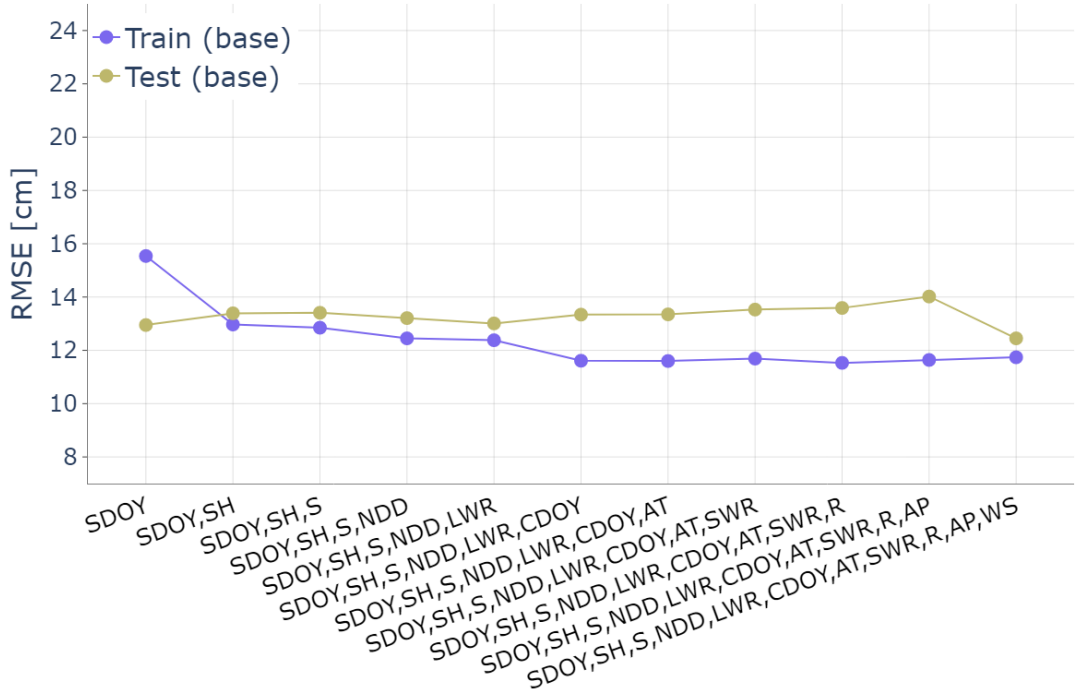
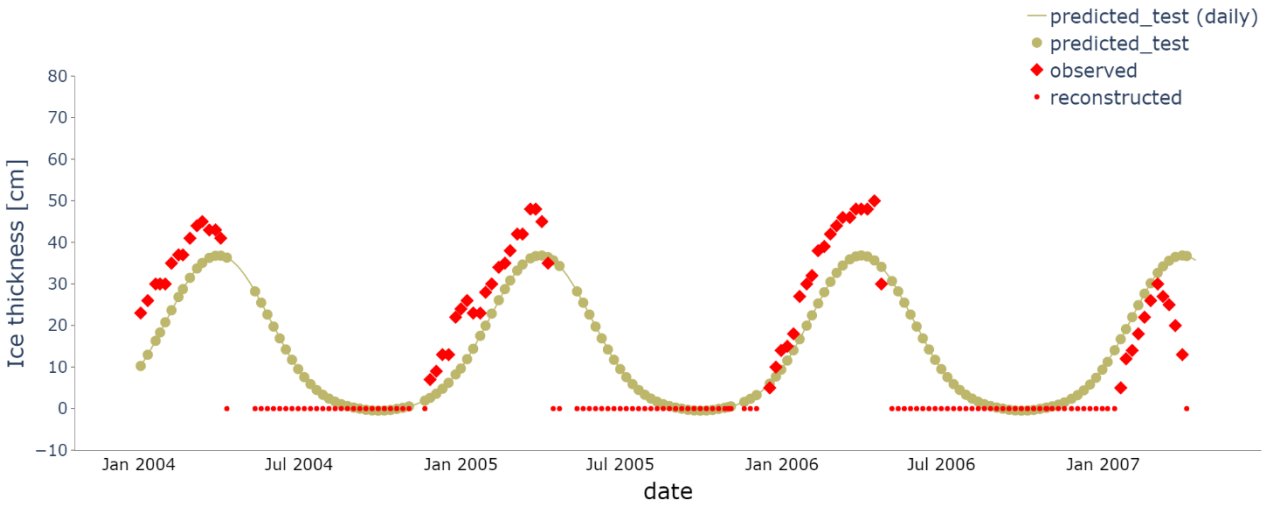



Figure 5.9. RMSE of different versions of the IT model using LSTM model obtained by different predictors selected based on the feature importance analysis for (a) Lake Runn and (b) Lake Gouta. The x-axis indicates the variable(s) added in the model.

A more detailed investigation indicates that the inclusion of SDOY as the only feature results in a reasonable prediction of IT (Figure 5.10). These results underscore the importance of this predictor to improve the models' performance.

(a)




(b)

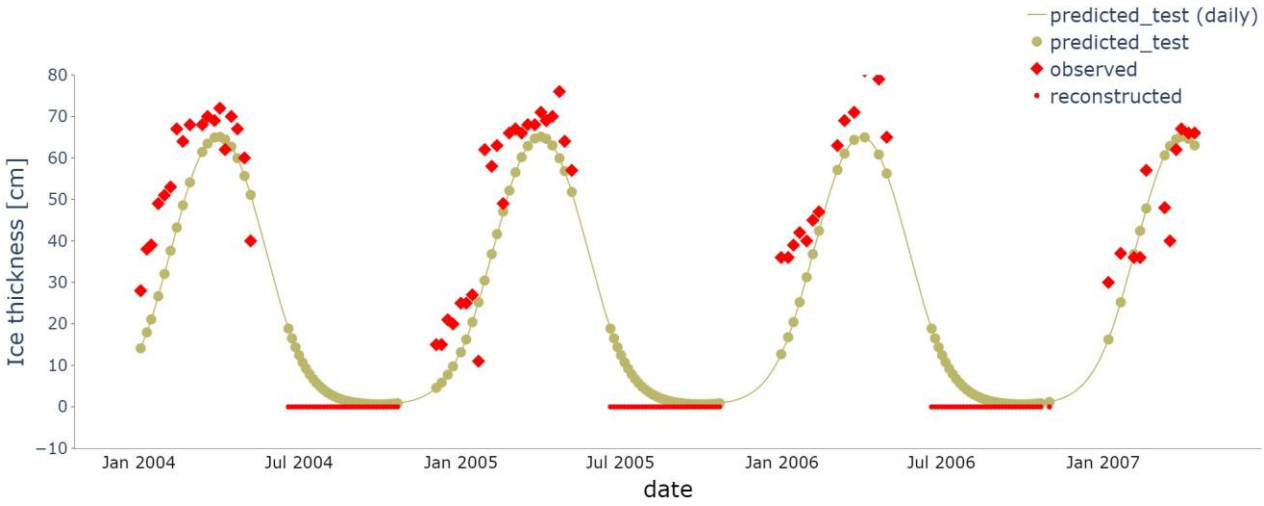



Figure 5.10. Examples of the IT model using LSTM with only SDOY as a predictor for (a) Lake Runn and (b) Lake Gouta (training set because of showing the test set in Figure 5.8).

## 5.4. Conclusions

Our study investigated the impact of meteorological variables on IT prediction in two lakes in Sweden, Runn (in southern part) and Gouta (in northern part), using ANN as an example of ML technique. Our analysis revealed that SDOY emerged as critical predictors for both lakes, emphasizing the importance of this predictor in accurately modeling IT. SWR in Lake Runn and SH

in Lake Gouta are the other influential variables. Although NDD in both lakes for ice formation is one of the main predictors, the AT itself is less influential, which emphasizes the importance of how to apply the meteorological variables. This confirms the results that have been achieved in previous studies.

Furthermore, we compared BPNN and LSTM. The second method, which considers the memory, proved to be more robust than BPNN. Accordingly, we used LSTM for IT modeling. Incorporating memory into the algorithm to represent the physical process is another solution with respect to averaging meteorological variables, which was shown to be a promising alternative approach in Chapter 4.

These findings have significant practical implications in various fields, including winter recreation, transportation, and safety. Accurately predicting IT in lakes can help ensure the safety of those engaging in recreational activities on the ice or relying on it for transportation. Moreover, predicting IT can assist in optimizing resource allocation for ice management and monitoring in lakes.

# Chapter 6. Conclusions

The prediction of lake surface water temperature (LSWT) and ice thickness (IT) in freshwater bodies is important for various reasons, including ecosystem health, resource allocation, and environmental management, as they affect the physical, chemical, and biological processes occurring in lakes and rivers. Furthermore, accurate predictions of IT can provide support in the managing and monitoring the safety of those engaging in recreational activities on the ice or relying on it for transportation.

Machine learning (ML) algorithms are becoming increasingly popular for the prediction of environmental conditions in freshwater bodies due to their ability to handle complex non-linear relationships between input and output variables. In this study, we have reviewed and compared various ML approaches to predict LSWT and IT and the effect of the main factors affecting their performances.

In the beginning of the thesis, we provided an in-depth discussion of the ML techniques utilized, along with the metrics employed for result evaluation. Furthermore, we detailed the methods employed to partition our data into training and test sets for model simulation and assessment. Additionally, we explored various methods for calculating averaged values to incorporate the memory of past data. This technique can help reduce noise in the data and better capture the underlying trend in the temperature data. A comprehensive examination was also conducted to rank the inputs of the model based on their respective impact, with a detailed explanation of the method.

In this research, we explored various ML techniques using synthetic lake scenarios, providing us with a more controlled model. We also investigated diverse preprocessing methods to enhance prediction accuracy. Subsequently, building upon the insights gained in this initial phase, we extended our model to forecast LSWT in a broader context, encompassing 2024 lakes. We then assessed the significance of different predictors in shaping the model's performance. Lastly, we extended our analysis to include the prediction of IT using ML models, and we ranked the features according to their impact on the model's predictions. Below, we offer a concise summary of each aspect of our research work.

In the first part of our study, we reviewed the most common ML approaches and attempted to disentangle the effect of the main factors affecting their performances. Our analysis involved applying nine ML methods on two specifically designed synthetic case studies: a shallow lake with a depth of 5 m and a deep lake with a depth of 60 m. The use of a synthetic case study allowed us to control all the variables involved in the problem and make a comparison between the shallow and deep cases without any other factor altering the thermal response.

Our results demonstrated that a combination of air temperature (AT) and day of the year (DOY) is the minimum amount of information required to obtain reasonable results, and adding further variables does not allow for significant improvements. Specifically, we found that adopting a continuous description of DOY through the use of trigonometric functions (SCDOY) provided the best results, likely due to its ability to capture the mean seasonal variability.

In addition to the analysis of predictors for accurate prediction of LSWT, our study also investigated the impact of smoothing the input signal for AT using a moving average and also the AT of previous days on the performance of ML model. We found that the use of these two methods led to a noticeable improvement in the performance of the ML models used in our study. Incorporating the AT data from prior days expands the input dataset, resulting in increased computational expenses. Accordingly, smoothed AT is preferable.

Furthermore, we explored the inclusion of previous days' AT values as inputs for ML models to retain the history of the forcing conditions. This can be particularly beneficial for ML approaches that cannot intrinsically consider historical data. The inclusion of previous days' AT values as inputs can enable the ML models to better capture the effects of longer-term trends and fluctuations in temperature, providing a more comprehensive understanding of the drivers of LSWT dynamics.

In the analysis of ML models, Artificial Neural Network (ANN) models demonstrated a slight superiority over other ML methods, possibly attributed to the extensive (daily) dataset available. To ensure a fair comparison, the inputs were the same, and hyperparameters of the models were optimized. Within the ANN algorithms, the Long Short-Term Memory (LSTM) algorithm excelled, particularly in scenario involving deep lake where the memory of the previous state of the system plays a crucial role. Consequently, we employed ANN for investigating both LSWT and IT in the second and third parts of our study. Due to the longer runtime of the library used for LSTM, we utilized the Backpropagation Neural Network (BPNN) for predicting LSWT at the global scale (2024 lakes), which also yielded reasonable results. For the two IT case studies, we opted for LSTM and compare it with BPNN.

Overall, this part of our study highlights the importance of careful selection and processing of input variables for ML models to accurately predict LSWT. Smoothing the input signal for AT using a moving average and including historical data as inputs can significantly improve the performance of ML models.

In the second part of our study, we aimed to expand upon the first part of the study by investigating a larger dataset of 2024 lakes from various regions around the world. Our objective was to examine the influence of meteorological variables on LSWT using the BPNN as our chosen ML approach.

While there have been previous studies highlighting the impact of meteorological variables on lake water temperature, they were often limited to specific case studies, making it uncertain whether the identified influential variables hold true for other lakes as well. Thus, we sought to test the influence of eight meteorological variables retrieved from the ERA5 dataset, along with the DOY and different combinations of these variables.

As a preliminary screening, we computed the correlations among LSWT and the meteorological variables, and as well as the correlation between all predictor variables. Considering that the availability of LSWT data is not on a daily basis, while the meteorological features are, we utilized linear regression to fill in the missing values of LSWT. Subsequently, we computed correlations using the daily values. Our analysis revealed that AT, downward longwave radiation (LWR), shortwave radiation (SWR), relative humidity (RH), and sine and cosine of DOY (SCDOY) exhibited the highest correlations not only with LSWT but also among themselves, across all regions.

Then, we used BPNN to simulate LSWT as a function of the meteorological variables. Among the analyzed variables, including SCDOY, AT, SWR, LWR and RH, we observed their substantial importance as the most influential features on simulation and prediction of LSWT within different Köppen regions. Notably, SCDOY demonstrated significant relevance in all regions, further emphasizing the fact that the mean seasonal variability dominates the model's performance, whereas capturing the interannual variability remains a challenge.

When evaluating the performance of the BPNN model, we found that utilizing averaged meteorological data significantly boosts performance compared to using the actual values. In this respect, the averaged values of AT, SWR, LWR, and RH exhibit strong correlations with SCDOY, confirming that the inclusion of mean seasonal variability significantly enhances the results. Nevertheless, considering the anomalies produced by prolonged alterations with respect to the climatological year, possibly caused by climate change, are expected to be important to capture the interannual dynamics of LSWT.

In the third segment of our research, we explored different approaches to incorporate memory of data from previous time steps into our models, including the utilization of averaged meteorological variables alongside SCDOY. We also conducted a comparative analysis between two ML methods: the first being BPNN, and the second being the LSTM method, designed to account for both long and short-term memory aspects within the model. Both models yielded satisfactory results, but LSTM exhibited greater robustness across multiple runs, which led us to favor its use for the subsequent phases of this study's analysis.

Then, we aimed to investigate the impact of meteorological variables on IT prediction in two lakes in Sweden using LSTM model. The two lakes we chose, Runn in the southern part and Gouta in the northern part, allow us to observe the effect of meteorological variables on IT prediction in different environments. Our analysis revealed that DOY (and in particular the sine of it, SDOY) emerged as a critical predictor for both lakes, but only when accompanied by another variable (SWR for Lake Runn and SH for Lake Gouta), emphasizing their importance in accurately modeling IT.

Negative degree days (NDD) in both lakes were found to be other influential variables for ice formation. NDD is a measure of the temperature below 0°C, which plays a crucial role in the ice formation process. Besides, our study also found that the AT itself was not influential as NDD in ice formation in both lakes. This result suggests that the impact of AT on IT prediction is indirect and can be mediated through other variables such as NDD.

The careful selection of pertinent meteorological variables and the adoption of a more resilient ML approach, LSTM, can significantly enhance the quality of results, contributing to a more robust and accurate outcome in the analysis. Accurate predictions of IT can assist in optimizing resource allocation for ice management and monitoring in lakes, thereby ensuring the safety of people engaging in recreational activities on the ice or relying on it for transportation.

In summary, our research underscores the critical role of meticulous meteorological variable selection for precise predictions of IT and LSWT in various lake scenarios. The DOY emerges as a pivotal predictor across all scenarios. Incorporating memory aspects in predictions proves beneficial, achieved through the implementation of averaging methods, ultimately leading to improved and more accurate results.

Another crucial aspect is the selection of suitable ML methods for lake simulation and prediction. We explored the use of BPNN to develop a model for generalizing predictions across 2024 lakes in different climatic regions. Additionally, our findings suggest that ANN, particularly LSTM, exhibit enhanced robustness and slightly superior performance compared to other ML models. The inclusion of memory considerations in the LSTM model, influential for time series datasets, leads to improved results, especially in IT scenarios. In summary, ML proves highly valuable for feature selection and the analysis of various scenarios, given its inherent flexibility.

# Appendix
# A. Hyperparameters

## A.1. Hyperparameters used for the machine learning models of synthetic lakes

For selecting the features affecting lake surface water temperature (LSWT) by using a synthetic lakes, eight of the ML methods were implemented in the Python environment using the libraries 'Keras' for multilayer perceptron neural network (MLPNN) and long short-term memory (LSTM) and 'Sklearn' for the other ML methods. Only for one method (ANFIS) we referred to a MATLAB library.

The hyperparameters were chosen by genetic algorithm (GA) optimization. A more detailed explanation is available for 'Keras' in Chollet (2015) and for 'Sklearn' in Cournapeau (2007), where we also took most of the following description. In this section, we restrict the analysis to the hyperparameters that were calibrated and provide a brief description of the meaning and a table with adopted values. For the other hyperparameters that are not mentioned here, we used the default values suggested in the libraries.

Since GA choice of the best hyperparameters is approximate, the adopted hyperparameters may not be the optimal values in absolute terms. Different hyperparameters can be obtained by varying the population size and number of iterations in GA. The objective function is the absolute value of the difference of root mean square error (RMSE) between the training and test data sets, plus the training RMSE (°C). With respect to this objective function, not only the chosen hyperparameters make the model have lower RMSE but also the difference between the results of training and test data set would be the minimum.

### A.1.1. Decision Tree (DT)

The hyperparameters for DT (optimized in Table A.1) are:

**Criterion= [ 'MSE', 'friedman_MSE', 'MAE', 'Poisson']**

A measure of the quality of a split into two branches:

- 'MSE': mean squared error
- 'friedman_MSE': mean squared error with Friedman's improvement score for potential splits,
- 'MAE': mean absolute error

- ‘poisson’: reduction in Poisson deviance to find splits.

**Max_depth**

The maximum depth of the tree. The ‘none’ is imposed, nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split (the minimum required number of samples for splitting an internal node) samples, where min_samples_split is a hyperparameter to be defined.

**Max_features= [int, float, ‘auto’, ‘sqrt’, ‘log2’]**

The number of features to consider each time to make the split decision.

- int: max_features at each split
- float: max_features is a fraction at each split; max_features * n_features (the number of features seen during fit) is the value to consider as features.
- ‘auto’: max_features is n_features.
- ‘SQRT’: max_features is the square root of n_features.
- ‘log2’: max_features is logarithm of n_features.
- None: max_features is n_features.

Table A.1. Optimized values of DT hyperparameters.

| Lake’s depth (m) | Criterion | max_depth | max_features |
|---|---|---|---|
| 5 | 'MSE' | 8 | 'auto' |
| 60 | 'friedman_MSE' | 7 | 'log2' |

### A.1.2. Random Forest (RF)

The hyperparameters for RF (optimized in Table A.2) are:

**N_estimators**

The number of trees in the forest.

**Criterion= [’MSE’, ‘MAE’]**

The function to measure the quality of a split, where ‘MSE’ and ‘MAE’ are explained in previous section.

**Max_depth**

The maximum depth of the tree. ‘None’ means that the nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

Table A.2. Optimized values of RF hyperparameters.

| Lake's depth (m) | n_estimators | Criterion | max_depth |
|---|---|---|---|
| 5 | 42 | 'MAE' | 9 |
| 60 | 26 | 'MSE' | 7 |

### A.1.3. Extremely Randomized Tree (ERT)

The explanation of the hyperparameters of ERT is similar to RF. The optimized values are reported in Table A.3.

Table A.3. Optimized values of ERT hyperparameters.

| Lake's depth (m) | n_estimators | Criterion | max_depth |
|---|---|---|---|
| 5 | 85 | 'MAE' | 7 |
| 60 | 181 | 'MSE' | 7 |

### A.1.4. K-nearest neighbour (KNN)

The hyperparameters for KNN (optimized in Table A.4) are:

**N_neighbors**

Number of neighbours required for each sample.

**Weights= [Uniform', 'Distance', Callable]**

Weight function used in prediction. Possible values:

- 'Uniform': all points in each neighbourhood are weighed equally.
- 'Distance': weight points by the inverse of their distance. In this case, closer neighbours of a query point will have a greater influence than neighbours which are further away.
- Callable: a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.

**Algorithm= ['auto', 'Ball_Tree', 'KD_Tree', 'Brute']**

Algorithm used to compute the nearest neighbours:

- 'Ball_tree': recursively divides the data into nodes defined by a centroid C and radius r, such that each point in the node lies within the hyper-sphere defined by r and C.
- 'KD_Tree': a binary tree structure which recursively partitions the parameter space along the data axes, dividing it into nested orthotropic regions into which data points are filed.

- ‘Brute’: the brute-force computation of distances between all pairs of points in the dataset.
- ‘auto’: decide the most appropriate algorithm based on the values passed to fit method.

**Leaf_size**

Leaf size passed to BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem.

**P**

Power parameter for the Minkowski metric.

- P = 1: Manhattan_distance,
- P = 2: Euclidean_distance,
- Arbitrary P: Minkowski_distance.

Table A.4. Optimized values of KNN hyperparameters.

| Lake’s depth (m) | N_neighbors | Weights | Algorithm | Leaf_size | P |
|---|---|---|---|---|---|
| 5 | 23 | 'uniform' | 'KD_Tree' | 22 | 3 |
| 60 | 20 | 'uniform' | 'KD_Tree' | 22 | 3 |

### A.1.5. Support Vector Regression (SVR)

The hyperparameters for SVR (optimized in Table A.5) are:

**Kernel= ['Linear', 'Poly', 'RBF', 'Sigmoid', 'Precomputed']**

The choice of the kernel type to be used in the algorithm. The first two kernels are linear and polynomial, and the other kernels are:

- 'RBF': radial basis function (RBF) kernel.
- 'Sigmoid': based on the hyperbolic tangent,

$$K(X,Y) = \tanh(\alpha X^T Y + c) \quad \text{(A.1)}$$

where $\alpha$ is slope, and $C$ is the constant term; X and Y are inputs of the model and observed outputs, respectively.

- 'Precomputed': used to pre-compute the kernel matrix from data matrices; matrix should be a square array of size number of samples.

**Degree**

Degree of the polynomial kernel function (‘Poly’). Ignored by all other kernels.

**Gamma= [‘Scale’, ‘auto’]**

Kernel coefficient for 'RBF', 'Poly' and 'Sigmoid':

$$'scale' = \frac{1}{n_{features} \times X.var()} \quad \text{(A.2)}$$

Where '.var' is the variance, or

$$'auto' = \frac{1}{n_{features}} \quad \text{(A.3)}$$

**C**

Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive.

**Epsilon**

Epsilon in the epsilon-SVR model. It specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value.

**Cache_size**

Specify the size of the kernel cache.

Table A.5. Optimized values of SVR hyperparameters.

| Lake's depth (m) | Kernel | Degree | Gamma | C | Epsilon | cache_size |
|---|---|---|---|---|---|---|
| 5 | 'Poly' | 5 | 'auto' | 93.16 | 0.0555 | 162.9 |
| 60 | 'RBF' | 1 | 'auto' | 158.4 | 0.0555 | 161.2 |

### A.1.6. Multilayer Perceptron Neural Network (MLPNN)

The meaning of the number of layers and neurons is explained in section 2.2 of the main text (see the section for ANN). The hyperparameters for MLPNN (optimized in Table A.6) are:

**Number of hidden layers**

The number of neurons in the hidden layer which is also used for LSTM and BPNN in the following methods.

**Activation function= ['RELU", 'Sigmoid', 'Softmax', 'Softplus', 'Softsign', 'Tanh', 'SELU', 'ELU', 'Exponential']**

- 'RELU": linear unit activation function
- 'Softmax': converts a real vector to a vector of categorical probabilities.
- 'Softplus':

$$softplus(x) = \log(\exp(x) + 1) \quad \text{(A.4)}$$

- ‘Softsign’:

$$softsign(x) = x/(|(x)| + 1) \quad \text{(A.5)}$$

- ‘SELU’: Scaled Exponential Linear Unit
- ’ELU’: Exponential Linear Unit.

**Optimization= [‘SGD’, ‘RMSprop’, ‘Adam’, ‘Adadelta’, ‘Adagrad’, ‘Adamax’, ‘Nadam’, ‘Ftrl’]**

- ‘SGD’: Stochastic gradient descent (with momentum) optimizer.
- ‘RMSprop’: maintains a moving (discounted) average of the square of gradients and divides the gradient by the root of this average.
- ‘Adam’: a stochastic gradient descent method that is based on adaptive estimation of first order and second-order moments, proposed by Kingma and Ba (2014).
- ‘Adadelta’: a stochastic gradient descent method that is based on adaptive learning rate per dimension to address the continual decay of learning rates throughout training and the need for a manually selected global learning rate.
- ‘Adagrad’: with parameter-specific learning rates, which are adapted relative to how frequently a parameter gets updated during training. The more updates a parameter receives, the smaller the updates.
- ‘Adamax’: a variant of Adam based on the infinity norm. Adamax is sometimes superior to adam, especially in models with embeddings.
- ‘Nadam’: much like Adam is essentially RMSprop with momentum, Nadam is Adam with Nesterov momentum.
- ‘Ftrl’: obtained from McMahan et al. (2013)

**Learning rate**

The parameter that controls how much to change the model in response to the estimated error each time the model weights are updated.

**Batch size**

The number of samples to work through before updating the internal model parameters.

**Number of epochs**

The number of times that the whole set of patterns is presented to the network affected by other hyperparameters such as number of training data, number of hidden layers and number of neurones (Rafiq et al., 2011).

Table A.6. Optimized values of MLPNN hyperparameters.

| Lake's depth (m) | Number of hidden layers | Number of neurons | Activation function | Dropout | Optimization | Learning rate | Batch size | Number of epochs |
|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 9 | 'RELU"' | 0 | SGD | 0.08465 | 43 | 188 |
| 60 | 15 | 14 | 'RELU"' | 0.007955 | SGD | 0.10555 | 43 | 161 |

### A.1.7. Long Short-Term Memory (LSTM)

The hyperparameters are similar to MLPNN section. The optimized values are reported in Table A.7.

Table A.7. Optimized values of LSTM hyperparameters.

| Lake's depth (m) | Number of hidden layers | Number of neurons | Activation function | Dropout | Optimization | Learning rate | Batch size | Number of epochs |
|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 8 | 'Tanh' | 0.02787 | Adam | 0.06281 | 177 | 149 |
| 60 | 1 | 8 | 'Tanh' | 0 | Adam | 0.06281 | 177 | 149 |

### A.1.8. Backpropagation Neural Network (BPNN)

The hyperparameters for BPNN (optimized in Table A.8) are:

**Activation= ['Identity', 'Logistic', 'Tanh', 'RELU"]**

Activation function for the hidden layer(s):

- 'Identity': useful to implement linear bottleneck, returns the following equation:

$$f(x) = x \tag{A.6}$$

- 'Logistic': the logistic sigmoid function, returns the following equation:

$$f(x) = 1/(1 + \exp(-x)) \tag{A.7}$$

- 'Tanh': the hyperbolic tan function, returns the following equation:

$$f(x) = \tanh(x) \tag{A.8}$$

- ‘RELU’’: the rectified linear unit function, returns the following equation:

$$f(x) = max(0, x) \quad \text{(A.9)}$$

**Solver (Optimizer)=** [‘LBFGS’, ‘SGD’, ‘Adam’]

The solver for weight optimization:

- ‘LBFGS’: an optimizer in the family of quasi-Newton methods.
- ‘SGD’: stochastic gradient descent.
- ‘Adam’: explained in MLPNN.

**Batch_size**

Size of minibatches, which the gradient is calculated across the entire batch before updating weights, for stochastic optimizers. If the solver is ‘LBFGS’, the classifier will not use minibatch. When set to “auto”, batch_size is equal to minimum of 200 and n_samples.

**learning_rate= [‘Constant’, ‘Invscaling’, ‘Adaptive’]**

Learning rate schedule for weight updates:

- ‘Constant’: a constant learning rate given by ‘learning_rate_init’.
- ‘Invscaling’: gradually decreases the learning rate at each time step ‘t’ using an inverse scaling exponent of ‘power_t’.
- ‘Adaptive’: keeps the learning rate constant to ‘learning_rate_init’ as long as training loss keeps decreasing. Each time two consecutive epochs fail to decrease training loss or fail to increase validation score; the current learning rate is divided by 5.

**Learning_rate_init**

The initial learning rate used. It controls the step-size in updating the weights.

**Max_iter**

Maximum number of iterations. The solver iterates until convergence (determined by ‘tol’) or this number of iterations. For stochastic solvers (‘SGD, ‘Adam’), note that this determines the number of epochs (how many times each data point will be used), not the number of gradient steps.

Table A.8. Optimized values of BPNN hyperparameters

| Lake’s depth (m) | Number of hidden layers | Number of neurons | Activation function | Optimization | Learning rate | Learning rate initial | Maximum iteration |
|---|---|---|---|---|---|---|---|
| 5 | 19 | 20 | 'RELU’' | 'LBFGS' | 'Adaptive' | 0.00431 | 211 |
| 60 | 15 | 19 | 'RELU’' | 'LBFGS' | 'Adaptive' | 0.00471 | 196 |

### A.1.9. Adaptive Network-based Fuzzy Inference System (ANFIS)

The features considered in this method are taken from the MATLAB toolbox for ANFIS. Based on our available data, the best hyperparameters are shown in Table A.9.

**Input Membership function**

The function based on the fuzzy logic principles trains the model set to calculate the parameters of membership function, which represents the degree of truth in fuzzy logic (Opeyemi and Justice, 2012). The following functions are introduced:

- Gaussian (Gaussmf):

$$f(x) = e^{\frac{-(x-c)^2}{2\sigma^2}} \tag{A.10}$$

where $\sigma$ is the standard deviation, $c$ is the mean.

- Generalized bell (Gbellmf):

$$f(x) = \frac{1}{1 + |\frac{x-c}{a}|^{2b}} \tag{A.11}$$

where $a$, $b$, and $c$ are parameters.

- Sigmoid (Sigmf):

$$f(x) = \frac{1}{1 + e^{-a_k(x-c_k)}} \tag{A.12}$$

where $a_k$ and $c_k$ parameters.

- Triangular membership function (Trimf):

$$f(x) = \{ \begin{matrix} 0, x \le a \\ \frac{x-a}{b-a}, a \le x \le b \\ \frac{c-x}{c-b}, b \le x \le c \\ 0, c \le x \end{matrix} \} \tag{A.13}$$

where $a$, $b$, and $c$ are parameters.

**Output Membership function**

The output type of the membership function should be chosen in one of these two ways:

- Linear
- Constant

**Optimization method**

Method used to train the model. Two optimization methods are used to reduce the error:

- Backpropagation: for all parameters.
- Hybrid method: consisting of backpropagation for the parameters associated with the input membership functions, and least squares estimation for the parameters associated with the output membership functions.

**Epochs**

A measure of the number of times that the model is trained to update the weights.

Table A.9. Optimized values of BP hyperparameters

| Lake's depth (m) | Input MF | Output MF | Optimization | Epochs |
|---|---|---|---|---|
| 5 | Trimf | Linear | Hybrid | 100 |
| 60 | Trimf | Linear | hybrid | 100 |

## A.2. Hyperparameters used for the machine learning models of CCI lakes

To evaluate the influence of climate on the lakes' thermal response, BPNN as the ML methods were implemented in the Python environment using the library of 'Sklearn'. To generalize the model for all 2024 lakes by trial and error and based on the results obtained in section A.1.8, we utilized the hyperparameters in Table A.10 to the model. In order to avoid repeating the description if these hyperparameters, for more information read section A.1.8.

Table A.10. Selected values of BPNN hyperparameters

| Number of hidden layers | Number of neurons | Activation function | Optimization | Learning rate | Learning rate initial | Maximum iteration |
|---|---|---|---|---|---|---|
| Number of features -1 | 2 | 'Tanh' | 'LBFGS' | 'Invscaling' | 0.001 | 10000 |

## A.3. Hyperparameters used for the machine learning models of boreal lakes

To evaluate ice dynamics in boral lakes, we compared BPNN and LSTM. The hyperparameters used in these two models are as shown in Tables A.11 and A.12. To avoid the repetition of the

hyperparameters' explanation, read sections A.1.7 and A.1.8. To simplify the models to be used, we unified the choice of hyperparameters for the two lakes based on trial and errors.

Table A.11. Selected values of BPNN hyperparameters

| Number of hidden layers | Number of neurons | Activation function | Optimization | Learning rate | Learning rate initial | Maximum iteration |
|---|---|---|---|---|---|---|
| 2 | Number of features -1 | 'Tanh' | 'LBFGS' | 'Invscaling' | 0.001 | 5000 |

Table A.12. Selected values of LSTM hyperparameters

| Number of hidden layers | Number of neurons | Activation function | Dropout | Optimization | Learning rate | Batch size | Number of epochs |
|---|---|---|---|---|---|---|---|
| 1 | Number of features -1 | 'Tanh' | 0 | Adam | 0.0001 | 250 | 350 |

# B. Literature review of ML models used to predict water temperature in rivers

Table B.1. ML algorithms used to predict stream temperature in rivers, with the indication of the most important predictors.

| ML Algorithm | Author (s) | The best ML method | ML models | Predictor(s) |
|---|---|---|---|---|
| Linear and/or Non-linear Regression | Mohseni et al. (1998) | Non-linear regression model | Non-linear regression model | Weekly AT |
| | Mohseni and Stefan (1999) | Linear regression | Linear regression | AT |
| | Rivers-Moore et al. (2010) | Multiple linear regression model | Linear regression, Non-linear regression model (Exponential model), Multiple linear regression model | AT, streamflow, rainfall, relative humidity (mean daily AT, minimum daily AT and RH are influential) |
| | Li et al. (2014) | Varying coefficient regression model | Varying coefficient regression model | Maximum AT, time varying coefficient model (DOY) |
| Logistic Regression | Caissie et al. (2001) | Regression model in weekly basis, Stochastic for maximum AT | Regression model (A second order Markov process), stochastic approach using Fourier and Sine function | AT |
| | Grbic' et al. (2013) | Gaussian process regression model | Gaussian Process Regression - Linear regression, logistic and stochastic models | AT, streamflow |
| | Laanaya et al. (2017) | GAM | The generalized additive model (GAM), an extension of the generalized linear model; logistic model; linear regression; residual regression model (long-term annual and the short-term) | AT, average streamflow |
| Artificial Neural Network (ANN) | Chenard and Caissie (2008) | ANN | ANN | Minimum, maximum and mean AT, DOY, water level |
| | Sahoo et al. (2009) | ANN | Empirical model (BPNN), a statistical model (multiple regression analysis, MRA), chaotic non-linear dynamic algorithms (CNDA) | SWR, AT |
| | Wenxian et al. (2010) | PSO- BPNN | BPNN | Monthly water T |
| | DeWeber and Wagner (2014) | ANN | ANN | Daily averaged AT, prior 7 days mean AT, network catchment area, predicted mean daily water T. (Forest land cover at riparian and catchment extents impact negatively) |
| | Hadzima-Nyarko et al. (2014) | ANN | ANN, linear regression model and stochastic model | Daily average AT |
| | Napiorkowski et al. (2014) | ANN | ANN | Daily average AT |
| | Piotrowski et al. (2014) | ANN | ANN | Daily average AT, daily maximum AT, daily streamflow, daily average water T, declination of the sun |

| ML Algorithm | Author (s) | The best ML method | ML models | Predictor(s) |
|---|---|---|---|---|
| | Piotrowski et al. (2015) | ANN | MLPNN, ANFIS | Mean, maximum and minimum daily AT, streamflow, SWR. |
| | Rabi et al. (2015) | MLPNN | Linear regression, stochastic modelling (non-linear regression), MLPNN | Mean daily AT |
| | Piotrowski et al. (2016) | ANN | ANFIS, MLPNN | SWR, streamflow, minimum daily AT, mean daily AT, maximum daily AT, sum of the daily averaged AT measured 2 to 6 days before the day. |
| | Temizyurek and Dadaser-Celik (2018) | ANN | ANN | AT, WS, RH, previous water T |
| | Zhu et al. (2019) | Gaussian Process Regression (GPR), Bootstrap Aggregated Decision Trees (BA-DT) | ANN, GPR, BA-DT | AT |
| | Graf et al. (2019) | WT-ANN | WT-ANN, linear and non-linear regression, ANN | Daily water T, daily AT |
| | Trinh et al. (2019) | ANN | Regression models (linear, non-linear and stochastic regression), ANN | Daily maximum AT |
| | Zhu and Heddam (2019) | Optimally Pruned Extreme Learning Machine (OPELM) | OPELM, Radial Basis Functions Neural Networks (RBFNN) | AT, streamflow, DOY |
| | Zhu et al. (2019) | MLPNN | MLPNN, ANFIS | AT, streamflow, DOY |
| | Zhu et al. (2019) | WT-MLPNN and WT-ANFIS | MLPNN and ANFIS with and without WT | Daily AT, DOY |
| | Zhu et al. (2019) | Extreme Learning Machine (ELM) model | ELM - MLPNN and simple multiple linear regression (MLR) | AT, streamflow, DOY |
| | Zhu et al. (2019) | air2stream model outperformed the three machine learning methods | MLPNN, Gaussian process regression (GPR), DT | AT, streamflow, DOY |
| | Zhu et al. (2019) | MLPNN | MLPNN, ANFIS | AT, streamflow, DOY |
| | Piotrowski et al. (2020) | ANN | ANN | AT, streamflow, declination of the sun |
| | Qiu et al. (2020) | BPNN | BPNN, radial basis function neural network, wavelet neural network, general regression neural network, Elman neural network | AT, streamflow, DOY |
| | Radulescu (2020) | MLPNN | MLPNN | AT |
| | Zhu et al. (2020) | WT_MLPNN (air2water outperformed all as a hybrid model) | MLPNN, MLPNN integrated model (WT_MLPNN), non-linear regression model (S-curve), air2water | Daily LSWT and AT |
| Decision Tree (single conjunctive rule learner, decision table, M5 model tree, and REPTree) | Goyal et al. (2012) | M5 model tree | Tree algorithm (single conjunctive rule learner, decision table, M5 model tree, and REPTree) | Specific humidity, geopotential height, meridional (north-south direction) WS, zonal (west-east direction) WS, AT |
| Support Vector Machines (SVM) | Rehana (2019) | SVR | Multiple Linear Regression Model (MLRM), SVR | AT, streamflow |
| k-nearest Neighbour algorithm (KNN) | St-Hilaire et al. (2012) | KNN | KNN | Water T from the two previous days and an indicator of seasonality (DOY), daily AT and daily streamflow |

| ML Algorithm | Author (s) | The best ML method | ML models | Predictor(s) |
|---|---|---|---|---|
| Random Forest (RF) | Holthuijzen (2017) | Gradient Boosting Machines (GBM) | GBM, RF, Spatial Statistical Network (SSN), Generalized Additive Models (GAM), Linear regression | SWR, summer streamflow, maximum weekly maximum AT, summer mean AT |
| | Lu and Ma (2020) | Extreme Gradient Boosting (XGBoost), RF | RF | 1875 data |

# C. Metrics used to evaluate ML performance

## C.1. Standard metrics

Our analysis of the ML performance is based on the values of the root mean square error (RMSE), which is computed for both the training and test data using the standard definition

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(sim_i - obs_i)^2} \tag{C.1}$$

where $i$ is the time index, $N$ is the number of samples, $sim_i$ (simulated) is the LSWT modelled by ML, and $obs_i$ is the observed value (obtained from GLM simulation). Perfect fit is obtained for RMSE = 0.

However, different metrics are also used. In this section, we review the most common ones. First, the Nash-Sutcliff Efficiency index (NSE) is equivalent to RMSE if the variance of the observations, $\sigma_{obs}^2$, does not change:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(sim_i - obs_i)^2}{\sum_{i=1}^{N}\left(obs_i - \overline{obs}\right)^2} = 1 - \frac{RMSE^2}{\sigma_{obs}^2} \tag{C.2}$$

where $\overline{obs}$ is the mean of the observations. NSE = 1 indicates perfect fit, while suing the mean of the observations as a predictive model would lead to NSE = 0.

Other metrics that are used to compute the performance of models are the mean absolute error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|sim_i - obs_i| \tag{C.3}$$

which considers the absolute value of the error instead of the square as in equation (C.1), and the mean squared error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(sim_i - obs_i)^2 = RMSE^2 \tag{C.4}$$

which is simply the square of RMSE.

Another metric is the Index of Agreement presented by Willmott (1981),

$$IA = 1 - \frac{\sum_{i=1}^{N}(sim_i - obs_i)^2}{\sum_{i=1}^{N}\left(\left|sim_i - \overline{obs}\right| + \left|obs_i - \overline{obs}\right|\right)^2} \tag{C.5}$$

which is the ratio of MSE and the potential error. $IA$ is sensitive to extreme values because of the squared differences. The values of $IA$ range between 0 and 1, where 1 represents the perfect prediction and 0 the worst.

The coefficient of determination, or R-squared ($R^2$), yields the same information as NSE. Usually, it is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\widehat{obs}_i - obs_i\right)^2}{\sum_{i=1}^{N}\left(obs_i - \overline{obs}\right)^2} \quad \text{(C6)}$$

where the only difference is the interpretation of the term $\widehat{obs}_i$, which in this case is the prediction from a statistical model based on the observations, and not from the results of a simulation model as for NSE. We do not report the $R^2$ values as we prefer to indicate them as NSE.

Finally, we use the mean error (ME), or bias, to compute the average difference between simulated and observed values:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(sim_i - obs_i) = \overline{sim} - \overline{obs} \quad \text{(C7)}$$

where $\overline{sim}$ is the average of the simulated LSWT.

## C.2. Robustness of multiple runs

The results of the ML models depend on the initial guess of their parameters, which are then suitably calibrated by means of algorithms as GA. In order to test the possible variability of the results due to the random choice of the initial parameters, we decided to run all ML models $N_r$ times, and we assumed $N_r = 20$.

Here we introduce two concepts: the robustness and the accuracy of the ML model. The robustness shows how much the results of the $N$ different simulations are close to each other. In order to have an estimate of the robustness, at each time step $i$ we compute the standard deviation among the $N_r$ runs:

$$\sigma_i = \frac{1}{N_r}\sum_{j=1}^{N_r}\left(sim_{i,j} - \mu_{sim,i}\right)^2 \quad \text{(C.8)}$$

where $j$ is the index of the run, and $\mu_{sim,i}$ is the mean of the simulated LSWT. Then, we compute the average over the whole record (Equation C.8) to obtain the index for the train and test sets separately.

$$\sigma_R = \frac{1}{N}\sum_{i=1}^{N}\sigma_i \quad \text{(C.9)}$$

The accuracy represents how much the prediction is close to the observed values. We already defined the RMSE in equation (C.1) as a measure that can be used for a single run. Here, we define two metrics for the set of $N_r$ runs. The first one is the mean of the RMSE of the single runs,

$$RMSE_A = \frac{1}{N_r}\sum_{j=1}^{N_r} RMSE_j \quad \text{(C.10)}$$

and $RMSE_B$ is the best RMSE of 20 runs; the second one is the RMSE of the average of the $N_r$ runs,

$$RMSE_{AS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{sim}_i - obs_i)^2} \quad \text{(C.11)}$$

where $\overline{sim}_i$ is the mean simulated value at the time step $i$,

$$\overline{sim}_i = \frac{1}{N_r}\sum_{j=1}^{N_r} sim_{i,j} \quad \text{(C.12)}$$

## C.3. Differences between individual runs

For each case study, an individual run of the ML model provides a simulation that is not identical to the other runs even if the hyperparameters are the same, because the number of the parameters is so large that their calibration in the training phase produces results that depend on the initial (random) guess. Therefore, we introduced a measure (equation C.7) for the robustness of the single prediction.

It is interesting to analyze a case where two runs have a very similar RMSE, but they differ significantly from each other. In Figure C.1, the simulated values in two runs (number 1 and 17) are shown: the mean difference between the two simulations is 0.308°C, although the RMSE is almost identical (1.221°C and 1.238°C, respectively).

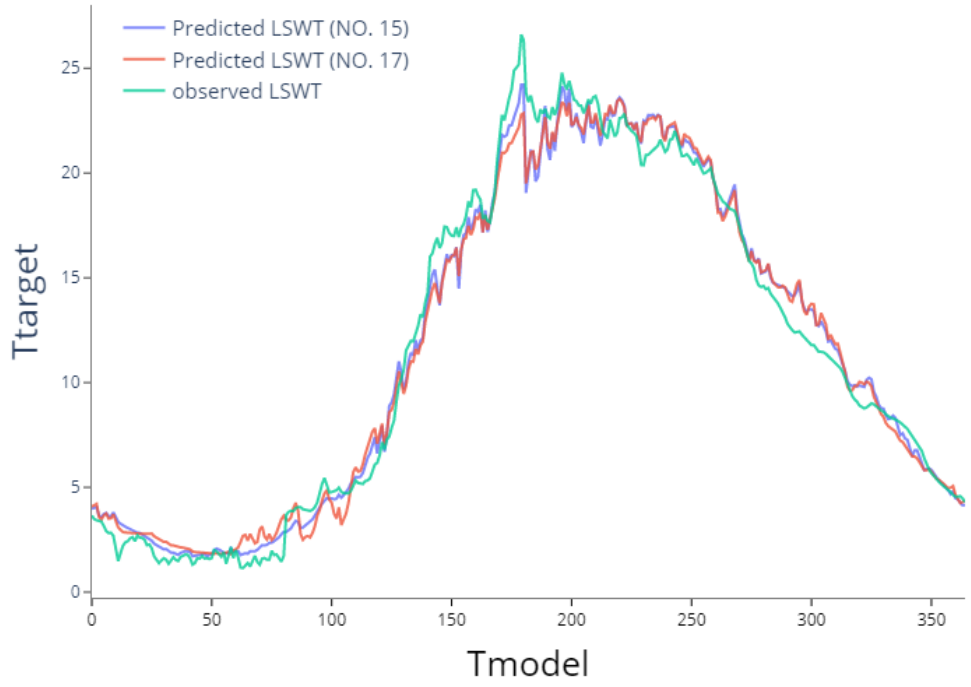



Figure C.1. Simulated LSWT from two individual runs compared with observations. The two simulations have similar RMSE, although they differ from each other.

## C.4. Analysis of the performances

Tables C.1, C.2, C.3 and C.4 report the list of all the metrics defined above to measure the performance of ML models, considering different inputs, pre-processing strategies, and ML algorithms.

Table C.1. Metrics of the BPNN model for the shallow (depth = 5 m) lake, depending on the predictors used and separating the performances of the training and test data sets.

| **Depth = 5 m** | $RMSE_A$ (°C) | | MSE (°C$^2$) | | MAE (°C) | | NSE (-) | | IA (-) | | ME (°C) | | $\sigma_R$ (°C) | | $RMSE_B$ (°C) | | $RMSE_{AS}$ (°C) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable(s) | train | test | train | test | train | test | train | test | train | test | train | test | train | test | train | test | train | test |
| AT | 3.044 | 3.105 | 9.266 | 9.644 | 2.127 | 2.178 | 0.864 | 0.869 | >0.999 | >0.999 | 0.000 | 0.239 | 0.137 | 0.183 | 3.007 | 3.059 | 3.021 | 3.076 |
| DOY | 1.856 | 2.125 | 3.482 | 4.539 | 1.326 | 1.524 | 0.949 | 0.938 | >0.999 | >0.999 | 0.000 | -0.738 | 0.257 | 0.257 | 1.782 | 2.065 | 1.796 | 2.069 |
| SCDOY | 1.781 | 2.070 | 3.172 | 4.286 | 1.262 | 1.470 | 0.953 | 0.942 | >0.999 | >0.999 | 0.000 | -0.738 | 0.013 | 0.013 | 1.778 | 2.055 | 1.777 | 2.067 |
| AT+SCDOY | 1.312 | 1.460 | 1.720 | 2.131 | 0.921 | 1.013 | 0.975 | 0.971 | >0.999 | >0.999 | 0.000 | -0.354 | 0.028 | 0.029 | 1.299 | 1.432 | 1.301 | 1.450 |
| SWR | 5.575 | 5.556 | 31.082 | 30.865 | 4.376 | 4.355 | 0.543 | 0.581 | >0.999 | >0.999 | 0.000 | -0.311 | 0.017 | 0.021 | 5.572 | 5.548 | 5.574 | 5.554 |
| LWR | 5.810 | 5.904 | 33.757 | 34.857 | 4.628 | 4.771 | 0.504 | 0.527 | >0.999 | >0.999 | 0.000 | 0.128 | 0.079 | 0.099 | 5.792 | 5.885 | 5.803 | 5.896 |
| RH | 7.928 | 8.141 | 62.847 | 66.272 | 7.119 | 7.238 | 0.076 | 0.101 | >0.999 | >0.999 | 0.000 | -0.693 | 0.040 | 0.034 | 7.922 | 8.133 | 7.925 | 8.139 |
| WS | 8.030 | 8.452 | 64.479 | 71.432 | 7.228 | 7.560 | 0.052 | 0.031 | >0.999 | >0.999 | 0.000 | -0.765 | 0.070 | 0.071 | 8.011 | 8.431 | 8.025 | 8.448 |
| R | 8.227 | 8.594 | 67.677 | 73.849 | 7.507 | 7.754 | 0.005 | -0.002 | >0.999 | >0.999 | 0.003 | -0.716 | 0.002 | 0.003 | 8.224 | 8.590 | 8.226 | 8.593 |
| AP | 7.520 | 7.884 | 56.579 | 62.182 | 6.533 | 6.822 | 0.168 | 0.156 | >0.999 | >0.999 | 0.001 | -0.492 | 0.642 | 0.618 | 7.472 | 7.840 | 7.479 | 7.846 |
| AT + SCDOY + WS | 1.313 | 1.452 | 1.723 | 2.108 | 0.921 | 1.006 | 0.975 | 0.971 | >0.999 | >0.999 | 0.000 | -0.351 | 0.035 | 0.036 | 1.300 | 1.439 | 1.299 | 1.439 |
| AT + SCDOY + AP | 1.308 | 1.455 | 1.712 | 2.117 | 0.919 | 1.010 | 0.975 | 0.971 | >0.999 | >0.999 | 0.000 | -0.344 | 0.040 | 0.042 | 1.298 | 1.432 | 1.293 | 1.440 |
| AT + SCDOY + SWR | 1.381 | 1.534 | 1.906 | 2.354 | 0.963 | 1.063 | 0.972 | 0.968 | >0.999 | >0.999 | 0.000 | -0.400 | 0.034 | 0.035 | 1.369 | 1.514 | 1.368 | 1.523 |
| AT + SCDOY + LWR | 1.287 | 1.422 | 1.657 | 2.021 | 0.901 | 0.987 | 0.976 | 0.973 | >0.999 | >0.999 | 0.000 | -0.360 | 0.041 | 0.043 | 1.270 | 1.399 | 1.271 | 1.406 |
| AT + SCDOY + SWR + LWR | 1.375 | 1.525 | 1.892 | 2.327 | 0.964 | 1.060 | 0.972 | 0.968 | >0.999 | >0.999 | 0.000 | -0.370 | 0.048 | 0.049 | 1.357 | 1.501 | 1.358 | 1.509 |
| AT + SCDOY + SWR + LWR + WS | 1.369 | 1.509 | 1.874 | 2.277 | 0.956 | 1.047 | 0.972 | 0.969 | >0.999 | >0.999 | 0.000 | -0.361 | 0.046 | 0.049 | 1.352 | 1.481 | 1.352 | 1.493 |
| AT + SCDOY + SWR + LWR + WS + RH | 1.371 | 1.497 | 1.879 | 2.240 | 0.960 | 1.038 | 0.972 | 0.970 | >0.999 | >0.999 | 0.000 | -0.326 | 0.050 | 0.055 | 1.351 | 1.475 | 1.353 | 1.478 |
| AT + SCDOY + SWR + LWR + WS + RH + R | 1.358 | 1.484 | 1.844 | 2.201 | 0.951 | 1.028 | 0.973 | 0.970 | >0.999 | >0.999 | 0.000 | -0.302 | 0.054 | 0.059 | 1.339 | 1.461 | 1.338 | 1.464 |
| AT + SCDOY + SWR + LWR + WS + RH + R +AP | 1.363 | 1.496 | 1.858 | 2.240 | 0.956 | 1.039 | 0.973 | 0.970 | >0.999 | >0.999 | 0.000 | -0.285 | 0.065 | 0.071 | 1.340 | 1.461 | 1.339 | 1.473 |

Table C.2. Metrics of the BPNN model for the deep (depth = 60 m) lake, depending on the predictors used and separating the performances of the training and test data sets.

| **Depth = 60 m** | $RMSE_A$ (°C) | | MSE (°C$^2$) | | MAE (°C) | | NSE (-) | | IA (-) | | ME (°C) | | $\sigma_R$ (°C) | | $RMSE_B$ (°C) | | $RMSE_{AS}$ (°C) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable(s) | train | test | train | test | train | test | train | test | train | test | train | test | train | test | train | test | train | test |
| AT | 3.629 | 3.911 | 13.187 | 15.303 | 2.837 | 2.989 | 0.779 | 0.757 | >0.999 | >0.999 | 0.000 | 0.154 | 0.258 | 0.266 | 3.587 | 3.881 | 3.596 | 3.878 |
| DOY | 1.596 | 1.746 | 2.654 | 3.185 | 1.186 | 1.353 | 0.956 | 0.950 | >0.999 | >0.999 | -0.001 | -0.743 | 0.482 | 0.482 | 1.436 | 1.579 | 1.474 | 1.644 |
| SCDOY | 1.434 | 1.579 | 2.056 | 2.492 | 1.041 | 1.223 | 0.966 | 0.960 | >0.999 | >0.999 | 0.000 | -0.742 | 0.012 | 0.012 | 1.431 | 1.571 | 1.430 | 1.575 |
| AT+SCDOY | 1.160 | 1.292 | 1.345 | 1.670 | 0.868 | 1.007 | 0.978 | 0.974 | >0.999 | >0.999 | 0.000 | -0.477 | 0.027 | 0.028 | 1.147 | 1.271 | 1.148 | 1.281 |
| SWR | 6.182 | 6.383 | 38.222 | 40.740 | 5.259 | 5.410 | 0.361 | 0.354 | >0.999 | >0.999 | 0.000 | -0.401 | 0.005 | 0.006 | 6.181 | 6.380 | 6.182 | 6.382 |
| LWR | 5.300 | 5.285 | 28.086 | 27.932 | 4.236 | 4.243 | 0.530 | 0.557 | >0.999 | >0.999 | 0.000 | 0.081 | 0.068 | 0.084 | 5.289 | 5.273 | 5.293 | 5.277 |
| RH | 7.604 | 7.884 | 57.815 | 62.154 | 6.828 | 6.973 | 0.033 | 0.015 | >0.999 | >0.999 | 0.000 | -0.802 | 0.081 | 0.096 | 7.593 | 7.782 | 7.598 | 7.878 |
| WS | 7.520 | 7.799 | 56.544 | 60.818 | 6.709 | 6.883 | 0.054 | 0.036 | >0.999 | >0.999 | 0.000 | -0.767 | 0.030 | 0.030 | 7.515 | 7.791 | 7.518 | 7.797 |
| R | 7.703 | 7.946 | 59.333 | 63.142 | 6.958 | 7.060 | 0.008 | -0.001 | >0.999 | >0.999 | 0.015 | -0.705 | 0.037 | 0.042 | 7.699 | 7.940 | 7.700 | 7.944 |
| AP | 7.112 | 7.385 | 50.582 | 54.544 | 6.200 | 6.435 | 0.154 | 0.135 | >0.999 | >0.999 | 0.000 | -0.515 | 0.056 | 0.050 | 7.102 | 7.373 | 7.108 | 7.382 |
| AT + SCDOY + WS | 1.151 | 1.294 | 1.326 | 1.675 | 0.862 | 1.007 | 0.978 | 0.973 | >0.999 | >0.999 | 0.000 | -0.478 | 0.037 | 0.041 | 1.137 | 1.278 | 1.135 | 1.278 |
| AT + SCDOY + AP | 1.164 | 1.284 | 1.354 | 1.650 | 0.871 | 1.004 | 0.977 | 0.974 | >0.999 | >0.999 | -0.001 | -0.469 | 0.040 | 0.041 | 1.152 | 1.267 | 1.146 | 1.268 |
| AT + SCDOY + SWR | 1.204 | 1.328 | 1.449 | 1.764 | 0.895 | 1.034 | 0.976 | 0.972 | >0.999 | >0.999 | 0.000 | -0.511 | 0.037 | 0.039 | 1.188 | 1.311 | 1.188 | 1.313 |
| AT + SCDOY + LWR | 1.153 | 1.282 | 1.329 | 1.644 | 0.859 | 1.003 | 0.978 | 0.974 | >0.999 | >0.999 | 0.000 | -0.487 | 0.037 | 0.039 | 1.141 | 1.266 | 1.137 | 1.267 |
| AT + SCDOY + SWR + LWR | 1.189 | 1.322 | 1.413 | 1.747 | 0.888 | 1.029 | 0.976 | 0.972 | >0.999 | >0.999 | 0.000 | -0.485 | 0.039 | 0.040 | 1.172 | 1.304 | 1.172 | 1.307 |
| AT + SCDOY + SWR + LWR + WS | 1.186 | 1.328 | 1.406 | 1.764 | 0.886 | 1.031 | 0.977 | 0.972 | >0.999 | >0.999 | 0.000 | -0.486 | 0.049 | 0.050 | 1.165 | 1.304 | 1.165 | 1.309 |
| AT + SCDOY + SWR + LWR + WS + RH | 1.185 | 1.309 | 1.405 | 1.714 | 0.887 | 1.013 | 0.977 | 0.973 | >0.999 | >0.999 | 0.000 | -0.432 | 0.054 | 0.059 | 1.165 | 1.290 | 1.162 | 1.286 |
| AT + SCDOY + SWR + LWR + WS + RH + R | 1.175 | 1.307 | 1.382 | 1.708 | 0.878 | 1.012 | 0.977 | 0.973 | >0.999 | >0.999 | 0.000 | -0.419 | 0.060 | 0.070 | 1.151 | 1.262 | 1.149 | 1.280 |
| AT + SCDOY + SWR + LWR + WS + RH + R +AP | 1.176 | 1.301 | 1.382 | 1.692 | 0.880 | 1.008 | 0.977 | 0.973 | >0.999 | >0.999 | 0.001 | -0.411 | 0.055 | 0.059 | 1.152 | 1.290 | 1.152 | 1.278 |

Table C.3. Comparison of the metrics obtained with different pre-processing methods and considering the input of AT from previous days, for the shallow and deep lake (ML method: BPNN).

| **Depth = 5 m** | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $RMSE_A$ (°C) | | MSE (°C$^2$) | | MAE (°C) | | NSE (-) | | IA (-) | | ME (°C) | | $\sigma_R$ (°C) | | $RMSE_B$ (°C) | | $RMSE_{AS}$ (°C) | |
| set | train | test | train | test | train | test | train | test | train | test | train | test | train | test | train | test | train | test |
| Min-Max | 1.312 | 1.460 | 1.720 | 2.131 | 0.921 | 1.013 | 0.975 | 0.971 | >0.999 | >0.999 | 0.000 | -0.354 | 0.028 | 0.029 | 1.299 | 1.432 | 1.301 | 1.450 |
| Moving Average + Min-Max | 1.061 | 1.136 | 1.126 | 1.292 | 0.743 | 0.802 | 0.983 | 0.982 | >0.999 | >0.999 | 0.000 | -0.097 | 0.026 | 0.032 | 1.047 | 1.108 | 1.049 | 1.123 |
| AT from previous days | 0.959 | 0.943 | 0.920 | 0.890 | 0.652 | 0.668 | 0.986 | 0.988 | >0.999 | >0.999 | 0.001 | 0.030 | 0.056 | 0.059 | 0.937 | 0.898 | 0.930 | 0.912 |
| DWT | 1.754 | 1.968 | 3.077 | 3.871 | 1.260 | 1.418 | 0.955 | 0.948 | >0.999 | >0.999 | -0.454 | 0.024 | 0.039 | 1.747 | 1.948 | 1.747 | 1.958 | -0.454 |
| CWT | 1.801 | 2.085 | 3.245 | 4.346 | 1.282 | 1.488 | 0.952 | 0.941 | >0.999 | >0.999 | -0.740 | 0.025 | 0.025 | 1.796 | 2.073 | 1.795 | 2.079 | -0.740 |
| **Depth = 60 m** | | | | | | | | | | | | | | | | | | |
| Min-Max | 1.160 | 1.292 | 1.345 | 1.670 | 0.868 | 1.007 | 0.978 | 0.974 | >0.999 | >0.999 | 0.000 | -0.477 | 0.027 | 0.028 | 1.147 | 1.271 | 1.148 | 1.281 |
| Moving Average + Min-Max | 0.983 | 1.189 | 0.967 | 1.414 | 0.741 | 0.876 | 0.984 | 0.978 | >0.999 | >0.999 | 0.000 | -0.202 | 0.031 | 0.036 | 0.973 | 1.153 | 0.968 | 1.174 |
| AT from previous days | 0.887 | 1.136 | 0.788 | 1.291 | 0.664 | 0.811 | 0.987 | 0.980 | >0.999 | >0.999 | 0.000 | -0.088 | 0.067 | 0.072 | 0.869 | 1.103 | 0.849 | 1.104 |
| DWT | 1.365 | 1.538 | 1.863 | 2.364 | 1.011 | 1.160 | 0.969 | 0.963 | >0.999 | >0.999 | 0.000 | -0.333 | 0.024 | 0.041 | 1.356 | 1.514 | 1.356 | 1.524 |
| CWT | 1.421 | 1.558 | 2.020 | 2.429 | 1.034 | 1.212 | 0.966 | 0.962 | >0.999 | >0.999 | 0.000 | -0.742 | 0.023 | 0.023 | 1.415 | 1.552 | 1.413 | 1.551 |

Table C.4. Comparison of the metrics obtained using different ML algorithms using AT and SCDOY as input (pre-processing: MM).

| **Depth = 5 m** | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $RMSE_A$ (°C) | | MSE (°C$^2$) | | MAE (°C) | | NSE (-) | | IA (-) | | ME (°C) | | $\sigma_R$ (°C) | | $RMSE_B$ (°C) | | $RMSE_{AS}$ (°C) | |
| set | train | test | train | test | train | test | train | test | train | set | train | test | train | test | train | test | train | test |
| DT | 1.230 | 1.626 | 1.514 | 2.642 | 0.859 | 1.090 | 0.978 | 0.964 | >0.999 | >0.999 | 0.000 | -0.340 | 0.000 | 0.002 | 1.230 | 1.623 | 1.230 | 1.625 |
| RF | 1.159 | 1.540 | 1.344 | 2.373 | 0.769 | 1.025 | 0.980 | 0.968 | >0.999 | >0.999 | -0.033 | -0.383 | 0.010 | 0.013 | 1.157 | 1.528 | 1.155 | 1.536 |
| ERT | 1.295 | 1.479 | 1.677 | 2.188 | 0.883 | 1.006 | 0.975 | 0.970 | >0.999 | >0.999 | -0.033 | -0.424 | 0.011 | 0.013 | 1.288 | 1.460 | 1.291 | 1.475 |
| KNN | 1.243 | 1.525 | 1.544 | 2.327 | 0.860 | 1.047 | 0.977 | 0.968 | >0.999 | >0.999 | -0.006 | -0.413 | 0.000 | 0.000 | 1.243 | 1.525 | 1.243 | 1.525 |
| SVR | 1.369 | 1.497 | 1.873 | 2.240 | 1.052 | 1.121 | 0.973 | 0.970 | >0.999 | >0.999 | 0.107 | -0.259 | 0.000 | 0.000 | 1.369 | 1.497 | 1.369 | 1.497 |
| MLPNN | 1.319 | 1.467 | 1.739 | 2.153 | 0.926 | 1.022 | 0.974 | 0.971 | >0.999 | >0.999 | -0.025 | -0.383 | 0.043 | 0.042 | 1.299 | 1.446 | 1.302 | 1.453 |
| LSTM | 1.353 | 1.490 | 1.831 | 2.222 | 0.983 | 1.074 | 0.973 | 0.970 | >0.999 | >0.999 | -0.024 | -0.349 | 0.085 | 0.092 | 1.322 | 1.424 | 1.321 | 1.460 |
| BPNN | 1.312 | 1.460 | 1.720 | 2.131 | 0.921 | 1.013 | 0.975 | 0.971 | >0.999 | >0.999 | 0.000 | -0.354 | 0.028 | 0.029 | 1.299 | 1.432 | 1.301 | 1.450 |
| **Depth = 60 m** | | | | | | | | | | | | | | | | | | |
| DT | 1.428 | 1.668 | 2.056 | 2.811 | 1.044 | 1.241 | 0.966 | 0.955 | >0.999 | >0.999 | 0.000 | -0.472 | 0.750 | 0.916 | 1.264 | 1.447 | 1.143 | 1.377 |
| RF | 1.117 | 1.362 | 1.247 | 1.856 | 0.847 | 1.056 | 0.979 | 0.971 | >0.999 | >0.999 | -0.001 | -0.472 | 0.007 | 0.009 | 1.114 | 1.351 | 1.114 | 1.359 |
| ERT | 1.141 | 1.292 | 1.302 | 1.668 | 0.853 | 1.011 | 0.978 | 0.974 | >0.999 | >0.999 | 0.000 | -0.522 | 0.003 | 0.004 | 1.136 | 1.285 | 1.140 | 1.290 |
| KNN | 1.093 | 1.311 | 1.196 | 1.719 | 0.808 | 1.026 | 0.980 | 0.973 | >0.999 | >0.999 | -0.009 | -0.510 | 0.000 | 0.000 | 1.093 | 1.311 | 1.093 | 1.311 |
| SVR | 1.171 | 1.302 | 1.371 | 1.695 | 0.914 | 1.025 | 0.977 | 0.973 | >0.999 | >0.999 | 0.107 | -0.355 | 0.000 | 0.000 | 1.171 | 1.302 | 1.171 | 1.302 |
| MLPNN | 1.212 | 1.378 | 1.469 | 1.901 | 0.918 | 1.085 | 0.975 | 0.970 | >0.999 | >0.999 | -0.083 | -0.603 | 0.033 | 0.034 | 1.177 | 1.296 | 1.198 | 1.366 |
| LSTM | 1.286 | 1.264 | 1.656 | 1.598 | 0.986 | 0.962 | 0.972 | 0.975 | >0.999 | >0.999 | 0.405 | -0.054 | 0.019 | 0.023 | 1.233 | 1.223 | 1.279 | 1.255 |
| BPNN | 1.160 | 1.292 | 1.345 | 1.670 | 0.868 | 1.007 | 0.978 | 0.974 | >0.999 | >0.999 | 0.000 | -0.477 | 0.027 | 0.028 | 1.147 | 1.271 | 1.148 | 1.281 |

# D. Hypsographic curve of synthetic lakes

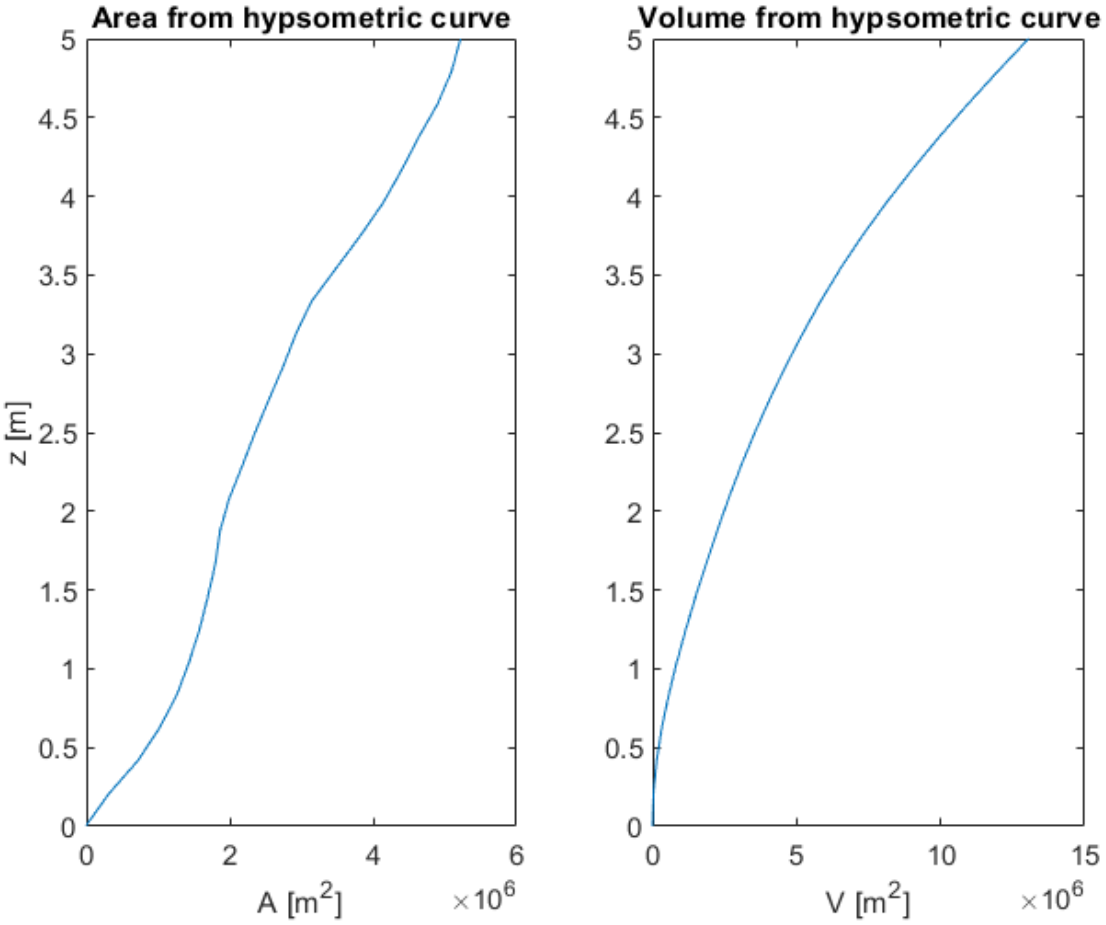



Figure D.1. Area and Volume of hypsographic curve of shallow lake with 5m depth.

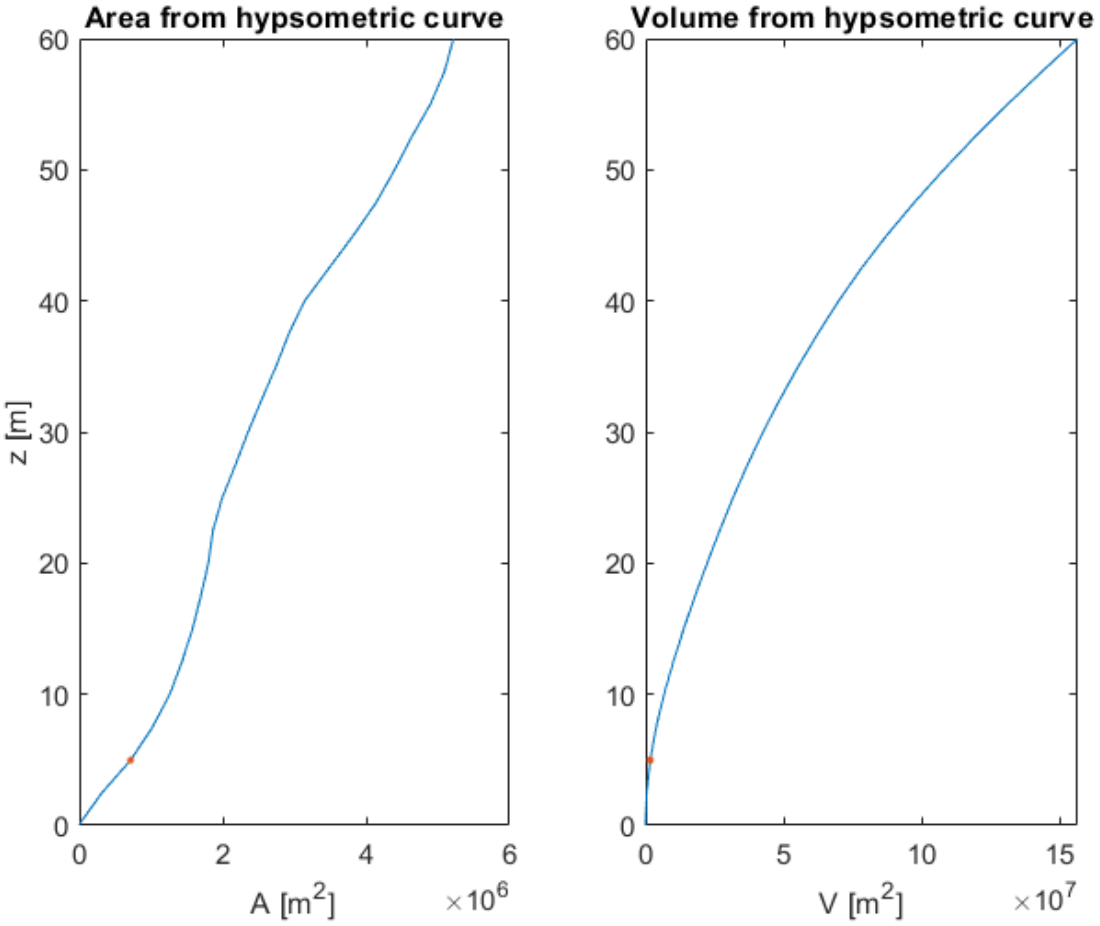



Figure D.2. Area and Volume of hypsographic curve of deep lake with 60 m depth.

# List of References

Adams, W.P. and Roulet, N.T., 1984. Sampling of snow and ice on lakes. Arctic, pp.270-275. https://doi.org/10.14430/arctic2200

Aldrich, C., 2020. Process Variable Importance Analysis by Use of Random Forests in a Shapley Regression Framework. Minerals, 10(5), 420. https://doi.org/10.3390/min10050420

Alswaina, F., Elleithy, K., 2018. Android malware permission-based multi-class classification using extremely randomized trees. IEEE Access, 6, 76217–76227. https://doi.org/10.1109/ACCESS.2018.2883975.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L. and Ridella, S., 2012, April. The'K'in K-fold Cross Validation. In ESANN (pp. 441-446).

Anikin, I.V. and Zinoviev, I.P., 2015, May. Fuzzy control based on new type of Takagi-Sugeno fuzzy inference system. In 2015 International Siberian Conference on Control and Communications (SIBCON), 1-4. IEEE. https://doi.org/10.1109/SIBCON.2015.7146977

Ashton, G.D., 2011. River and lake ice thickening, thinning, and snow ice formation. Cold Regions Science and Technology, 68(1-2), pp.3-19. https://doi.org/10.1016/j.coldregions.2011.05.004

Assel, R.A., 1976. Great Lakes ice thickness prediction. Journal of Great Lakes Research, 2(2), pp.248-255. https://doi.org/10.1016/S0380-1330(76)72289-5

Assel, R.A., 2004. Lake Erie ice cover climatology: basin averaged ice cover, winters 1898-2002.

Austin, J.A. and Colman, S.M., 2007. Lake Superior summer water temperatures are increasing more rapidly than regional air temperatures: A positive ice-albedo feedback. Geophysical Research Letters, 34(6). https://doi.org/10.1029/2006GL029021

Baker, D.G., Baker-Blocker, A., DeWitt, B.H. and Dixon, D.W., 1976. Objective prediction of ice formation, freeze-up and breakup on the Great Lakes. Journal of Applied Meteorology and Climatology, 15(10), pp.1033-1040. https://doi.org/10.1175/1520-0450(1976)015<1033:OPOIFF>2.0.CO;2

Barica, J. and Mathias, J.A., 1979. Oxygen depletion and winterkill risk in small prairie lakes under extended ice cover. Journal of the Fisheries Board of Canada, 36(8), pp.980-986. https://doi.org/10.1139/f79-136

Barrette, P.D., 2011. A laboratory study on the flexural strength of white ice and clear ice from the Rideau Canal skateway. Canadian Journal of Civil Engineering, 38(12), pp.1435-1439. https://doi.org/10.1139/l11-106

Benyahya, L., St-Hilaire, A., Quarda, T.B., Bobée, B. and Ahmadi-Nedushan, B., 2007. Modeling of water temperatures based on stochastic approaches: case study of the Deschutes River. Journal of Environmental Engineering and Science, 6(4), 437-448. https://doi.org/10.1139/s06-067

Bergmeir, C. and Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, pp.192-213. https://doi.org/10.1016/j.ins.2011.12.028

Bilello, M.A., 1964. Method for predicting river and lake ice formation. Journal of Applied Meteorology and Climatology, 3(1), pp.38-44. https://doi.org/10.1175/1520-0450(1964)003<0038:MFPRAL>2.0.CO;2

Blum, A.L. and Langley, P., 1997. Selection of relevant features and examples in ML. Artificial intelligence, 97(1-2), pp.245-271. https://doi.org/10.1016/S0004-3702(97)00063-5

Boehrer, B. and Schultze, M., 2008. Stratification of lakes. Reviews of Geophysics, 46(2). https://doi.org/10.1029/2006rg000210

Bouffard, D., Zdorovennova, G., Bogdanov, S., Efremova, T., Lavanchy, S., Palshin, N., Terzhevik, A., Vinnå, L.R., Volkov, S., Wüest, A. and Zdorovennov, R., 2019. Under-ice convection dynamics in a boreal lake. Inland Waters, 9(2), pp.142-161. https://doi.org/10.1080/20442041.2018.1533356

Brown, L.C. and Duguay, C.R., 2010. The response and role of ice cover in lake-climate interactions. Progress in physical geography, 34(5), pp.671-704.https://doi.org/10.1177/0309133310375653

Caissie, D., El-Jabi, N. and Satish, M.G., 2001. Modelling of maximum daily water temperatures in a small stream using air temperatures. Journal of hydrology, 251(1-2), pp.14-28. https://doi.org/10.1016/S0022-1694(01)00427-9

Calamita, E., Piccolroaz, S., Majone, B., Toffolon, M., 2021. On the role of local depth and latitude on surface warming heterogeneity in the Laurentian Great Lakes. Inland Waters, 11(2), 208-222. doi:10.1080/20442041.2021.1873698.

Carnicer, J., Coll, M., Ninyerola, M., Pons, X., Sanchez, G. and Penuelas, J., 2011. Widespread crown condition decline, food web disruption, and amplified tree mortality with increased climate change-type drought. Proceedings of the National Academy of Sciences, 108(4), pp.1474-1478. https://doi.org/10.1073/pnas.1010070108

Carrea, L., Crétaux, J.-F., Liu, X., Wu, Y., Bergé-Nguyen, M., Calmettes, B., Duguay, C., Jiang, D., Merchant, C.J., Mueller, D., Selmes, N., Simis, S., Spyrakos, E., Stelzer, K., Warren, M., Yesou, H., Zhang, D., 2022. ESA Lakes Climate Change Initiative (Lakes_cci): Lake products, Version 2.0.2. NERC EDS Centre for Environmental Data Analysis, 06 July 2022. https://doi.org/10.5285/a07deacaffb8453e93d57ee214676304. https://dx.doi.org/10.5285/a07deacaffb8453e93d57ee214676304

Chapra, S.C., Boehlert, B., Fant, C., Bierman Jr, V.J., Henderson, J., Mills, D., Mas, D.M., Rennels, L., Jantarasami, L., Martinich, J. and Strzepek, K.M., 2017. Climate change impacts on harmful algal blooms in US freshwaters: a screening-level assessment. Environmental Science & Technology, 51(16), pp.8933-8943. https://doi.org/10.1021/acs.est.7b01498

Chenard, J.F. and Caissie, D., 2008. Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada. Hydrological Processes: An International Journal, 22(17), 3361-3372. https://doi.org/10.1002/hyp.6928

Chollet, F., 2015, Keras, https://keras.io/

Cournapeau, D., 2007, scikit-learn, https://scikit-learn.org/

DeWeber, J.T. and Wagner, T., 2014. A regional neural network ensemble for predicting mean daily river water temperature. Journal of Hydrology, 517, 187-200. https://doi.org/10.1016/j.jhydrol.2014.05.035

Di Francescomarino, C., Dumas, M., Federici, M., Ghidini, C., Maggi, F.M., Rizzi, W. and Simonetto, L., 2018. Genetic algorithms for hyperparameter optimization in predictive business process monitoring. Information Systems, 74, 67-83. https://doi.org/10.1016/j.is.2018.01.003

Ding, F. and Mao, Z., 2021. Observation and Analysis of Water Temperature in Ice-Covered Shallow Lake: Case Study in Qinghuahu Lake. Water, 13(21), p.3139. https://doi.org/10.3390/w13213139

Doglioni, A., Giustolisi, O., Savic, D.A. and Webb, B.W., 2008. An investigation on stream temperature analysis based on evolutionary computing. Hydrological Processes: An International Journal, 22(3), 315-326. https://doi.org/10.1002/hyp.6607

Dokulil, M. T., de Eyto, E., Maberly, S. C., May, L., Weyhenmeyer, G. A., and Woolway, R. I., 2021. Increasing maximum lake surface temperature under climate change. Climatic Change, 165(3):1–17. https://doi.org/10.1007/s10584-021-03085-1

Drucker, H., Burges, C.J., Kaufman, L., Smola, A. and Vapnik, V., 1997. Support vector regression machines. Advances in neural information processing systems, 9, 155-161. https://doi.org/10.7551/mitpress/7503.003.0048

Dudek, G., 2019. Generating random weights and biases in feedforward neural networks with random hidden nodes. Information sciences, 481, 33-56. https://doi.org/10.1016/j.ins.2018.12.063

Fahle, S., Kneißler, A., Glaser, T. and Kuhlenkötter, B., 2020, September. Research on preprocessing methods for time series classification using machine learning models in the domain of radial-axial ring rolling. In Congress of the German Academic Association for Production Technology (pp. 487-496). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-62138-7_49

Ferreira, R.G., da Silva, D.D., Elesbon, A.A.A., Fernandes-Filho, E.I., Veloso, G.V., de Souza Fraga, M. and Ferreira, L.B., 2021. Machine learning models for streamflow regionalization in a tropical watershed. Journal of Environmental Management, 280, p.111713. https://doi.org/10.1016/j.jenvman.2020.111713

Galelli, S., Castelletti, A., 2013. Assessing the predictive capability of randomized tree based ensembles in streamflow modelling. Hydrology and Earth System Sciences, 17(7), 2669–2684. https://doi.org/10.5194/hess-17-2669-2013

Gebrekiros, S.T., 2016. Factors affecting stream fish community composition and habitat suitability. Journal of Aquaculture and Marine Biology, 4(2), p.00076. https://doi.org/10.15406/jamb.2016.04.00076

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning journal, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Gómez-Escalonilla, V., Martínez-Santos, P. and Martín-Loeches, M., 2021. Preprocessing approaches in machine learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali. Hydrology and Earth System Sciences Discussions, 2021, pp.1-30. https://doi.org/10.5194/hess-2021-261

Gonzalez-Dominguez, J., Lopez-Moreno, I., Sak, H., Gonzalez-Rodriguez, J., Moreno, P.J., 2014. Automatic language identification using long short-term memory recurrent neural networks. proceedings of Interspeech 2014, 2155-2159, https://doi.org/10.21437/Interspeech.2014-483

Gow, A.J., 1986. Orientation textures in ice sheets of quietly frozen lakes. Journal of crystal growth, 74(2), pp.247-258. https://doi.org/10.1016/0022-0248(86)90114-4

Goyal, M.K., Burn, D.H. and Ojha, C.S.P., 2012. Evaluation of machine learning tools as a statistical downscaling tool: temperatures projections for multi-stations for Thames River Basin, Canada. Theoretical and Applied Climatology, 108(3-4), 519-534. https://doi.org/10.1007/s00704-011-0546-1

Graf, R., Zhu, S. and Sivakumar, B., 2019. Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach. Journal of Hydrology, 578, 124115. https://doi.org/10.1016/j.jhydrol.2019.124115

Gray, E., Elliott, J. A., Mackay, E. B., Folkard, A. M., Keenan, P. O., and Jones, I. D., 2019. Modelling lake cyanobacterial blooms: Disentangling the climate-driven impacts of changing mixed depth and water temperature. Freshwater Biology, 64(12):2141–2155. https://doi.org/10.1111/fwb.13402

Grbić, R., Kurtagić, D. and Slišković, D., 2013. Stream water temperature prediction based on Gaussian process regression. Expert systems with applications, 40(18), 7407-7414. doi:10.1016/j.eswa.2013.06.077

Gunn, S.R., 1998. Support vector machines for classification and regression. ISIS technical report, 14(1), 5-16. https://doi.org/10.1201/b10911-3

Hadzima-Nyarko, M., Rabi, A. and Šperac, M., 2014. Implementation of artificial neural networks in modeling the water-air temperature relationship of the River Drava. Water resources management, 28(5), 1379-1394. https://doi.org/10.1007/s11269-014-0557-7

Hayhoe, K., VanDorn, J., Croley II, T., Schlegal, N. and Wuebbles, D., 2010. Regional climate change projections for Chicago and the US Great Lakes. Journal of Great Lakes Research, 36, pp.7-21. https://doi.org/10.1016/j.jglr.2010.03.012

Hecht-Nielsen, R., 1992. Theory of the backpropagation neural network. In Neural networks for perception (65-93). Academic Press. https://doi.org/ https://doi.org/10.1109/ijcnn.1989.118638

Heddam, S., Ptak, M. & Zhu, S., 2020. Modelling of daily lake surface water temperature from air temperature: Extremely randomized trees (ERT) versus Air2Water, MARS, M5Tree, RF and MLPNN. Journal of Hydrology, 588, p.125130. https://doi.org/10.1016/j.jhydrol.2020.125130.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D. and Simmons, A., 2020. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), pp.1999-2049. https://doi.org/10.1002/qj.3803

Hewitt, Bailey A., Lianna S. Lopez, Katrina M. Gaibisels, Alyssa Murdoch, Scott N. Higgins, John J. Magnuson, Andrew M. Paterson, James A. Rusak, Huaxia Yao, and Sapna Sharma. 2018. "Historical Trends, Drivers, and Future Projections of Ice Phenology in Small North Temperate Lakes in the Laurentian Great Lakes Region" Water 10, no. 1: 70. https://doi.org/10.3390/w10010070

Hipsey, M.R., Bruce, L.C., Boon, C., Busch, B., Carey, C.C., Hamilton, D.P., Hanson, P.C., Read, J.S., Sousa, E.D., Weber, M. and Winslow, L.A., 2019. A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON). Geoscientific Model Development, 12(1), 473-523. https://doi.org/10.5194/gmd-12-473-2019

Holland, J. H. et al., 1992. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MITpress. https://doi.org/10.7551/mitpress/1090.001.0001

Holthuijzen, M.F., 2017. A comparison of five statistical methods for predicting stream temperature across stream networks. Utah State University.

Imrit, M.A., Yousaf, Z. and Sharma, S., 2022. Quantifying the Trends and Drivers of Ice Thickness in Lakes and Rivers across North America. Water, 14(12), p.1841. https://doi.org/10.3390/w14121841

Irving, G., Guendelman, E., Losasso, F. and Fedkiw, R., 2006. Efficient simulation of large bodies of water by coupling two and three dimensional techniques. In ACM SIGGRAPH 2006 Papers, 805-811. https://doi.org/10.1145/1179352.1141959

Isik, F., Ozden, G. and Kuntalp, M., 2012. Importance of data preprocessing for neural networks modeling: The case of estimating the compaction parameters of soils. Energy Education Science and Technology Part A: Energy Science Research, 29, pp.463-74.

Izetta Riera, C.J. and Granitto, P.M., 2009. Feature selection with simple ANN ensembles. In XV Congreso Argentino de Ciencias de la Computación.

Jain, A.K., Mao, J. and Mohiuddin, K.M., 1996. Artificial neural networks: A tutorial. Computer, 29(3), 31-44. https://doi.org/10.1109/2.485891

Jenny, J.P., Anneville, O., Arnaud, F., Baulaz, Y., Bouffard, D., Domaizon, I., Bocaniov, S.A., Chèvre, N., Dittrich, M., Dorioz, J.M. and Dunlop, E.S., 2020. Scientists' warning to humanity: rapid degradation of the world's large lakes. Journal of Great Lakes Research, 46(4), pp.686-702. https://doi.org/10.1016/j.jglr.2020.05.006

Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., & Kumar, V., 2021, Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles. ACM/IMS Transactions on Data Science, 2(3), 1–26. https://doi.org/10.1145/3447814

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M. and Kumar, V., 2019, May. Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. In Proceedings of the 2019 SIAM International Conference on Data Mining, 558-566. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611975673.63

Jiang, S., Pang, G., Wu, M. and Kuang, L., 2012. An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, 39(1), 1503-1509. https://doi.org/10.1016/j.eswa.2011.08.040

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. and Kumar, V., 2017. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on knowledge and data engineering, 29(10), 2318-2331. https://doi.org/10.1109/tkde.2017.2720168

Keller, J.M., Gray, M.R. & Givens, J.A., 1985. A fuzzy K-nearest neighbor algorithm. IEEE Transactions on Systems, Man, and Cybernetics, SMC-15(4), pp.580–585. https://doi.org/10.1109/tsmc.1985.6313426.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kirillin, G., Leppäranta, M., Terzhevik, A., Granin, N., Bernhardt, J., Engelhardt, C., Efremova, T., Golosov, S., Palshin, N., Sherstyankin, P. and Zdorovennova, G., 2012. Physics of seasonally ice-covered lakes: a review. Aquatic sciences, 74, pp.659-682.https://doi.org/10.1007/s00027-012-0279-y

Knoll, L.B., Sharma, S., Denfeld, B.A., Flaim, G., Hori, Y., Magnuson, J.J., Straile, D. and Weyhenmeyer, G.A., 2019. Consequences of lake and river ice loss on cultural ecosystem services. Limnology and Oceanography Letters, 4(5), pp.119-131. https://doi.org/10.1002/lol2.10116

Koller, D. and Sahami, M., 1996, July. Toward optimal feature selection. In ICML (Vol. 96, No. 28, p. 292).

Konak, A., Coit, D. W., and Smith, A. E., 2006. Multi-objective optimization using genetic algorithms: A tutorial. Reliability engineering & system safety, 91(9):992–1007. https://doi.org/10.1016/j.ress.2005.11.018

Köppen, W., 1918. Klassifikation der klimate nach Temperatur, Niederschlag und Yahreslauf. Pet. Mitt., 64, pp.193-203.

Laanaya, F., St-Hilaire, A. and Gloaguen, E., 2017. Water temperature modelling: comparison between the generalized additive model, logistic, residuals regression and linear regression models. Hydrological sciences journal, 62(7), 1078-1093. https://doi.org/10.1080/02626667.2016.1246799

Lei, R., Leppäranta, M., Erm, A., Jaatinen, E. and Paern, O., 2011. Field investigations of apparent optical properties of ice cover in Finnish and Estonian lakes in winter 2009. Estonian Journal of Earth Sciences, 60(1), pp.50-64. . https://doi.org/10.3176/earth.2011.1.05

Leppäranta, M., 2014. Freezing of lakes and the evolution of their ice cover. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-29081-7

Li, H., Deng, X., Kim, D.Y. and Smith, E.P., 2014. Modeling maximum daily temperature using a varying coefficient regression model. Water Resources Research, 50(4), 3073-3087. https://doi.org/10.1002/2013WR014243

Li, J., Cheng, J.H., Shi, J.Y. and Huang, F., 2012. Brief introduction of back propagation (BP) neural network algorithm and its improvement. In Advances in computer science and information engineering, 553-558. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30223-7_87

Li, X., Peng, S., Xi, Y., Woolway, R.I. and Liu, G., 2022. Earlier ice loss accelerates lake warming in the Northern Hemisphere. Nature communications, 13(1), p.5156. https://doi.org/10.1038/s41467-022-32830-y

Liaw, A. and Wiener, M., 2002. Classification and regression by Random Forest. R news, 2(3), 18-22.

Liu, L., Davedu, S., Fujisaki-Manome, A., Hu, H., Jablonowski, C. and Chu, P.Y., 2022. Machine Learning Model-Based Ice Cover Forecasting for a Vital Waterway in Large Lakes. Journal of Marine Science and Engineering, 10(8), p.1022. https://doi.org/10.3390/jmse10081022

Liu, W.C. and Chen, W.B., 2012. Prediction of water temperature in a subtropical subalpine lake using an artificial neural network and three-dimensional circulation models. Computers & Geosciences, 45, 13-25. https://doi.org/10.1016/j.cageo.2012.03.010

Livingstone, D.M., 2003. Impact of secular climate change on the thermal structure of a large temperate central European lake. Climatic change, 57(1-2), 205-225. https://doi.org/10.1023/a:1022119503144

Lu, H. and Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, p.126169. https://doi.org/10.1016/j.chemosphere.2020.126169

Magee, M.R. and Wu, C.H., 2017. Effects of changing climate on ice cover in three morphometrically different lakes. Hydrological processes, 31(2), pp.308-323. https://doi.org/10.1002/hyp.10996

Magnuson, J.J., Webster, K.E., Assel, R.A., Bowser, C.J., Dillon, P.J., Eaton, J.G., Evans, H.E., Fee, E.J., Hall, R.I., Mortsch, L.R. and Schindler, D.W., 1997. Potential effects of climate changes on aquatic systems: Laurentian Great Lakes and Precambrian Shield Region. Hydrological processes, 11(8), pp.825-871. https://doi.org/10.1002/(SICI)1099-1085(19970630)11:8<825::AID-HYP509>3.0.CO;2-G

Maier, H.R., Jain, A., Dandy, G.C. and Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future

directions. Environmental modelling & software, 25(8), pp.891-909. https://doi.org/10.1016/j.envsoft.2010.02.003

Marshall, E.W., 1965. Structure of lake ice in the Keweenaw Peninsula, Michigan. Great Lakes Research Division, Univ. Michigan, Publ, (13).

Masterson, D.M., 2009. State of the art of ice bearing capacity and ice construction. Cold Regions Science and Technology, 58(3), pp.99-112. https://doi.org/10.1016/j.coldregions.2009.04.002

McCuen, R.H., Knight, Z. and Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe efficiency index. Journal of hydrologic engineering, 11(6), pp.597-602. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597)

McMahan, H.B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D. and Chikkerur, S., 2013, August. Ad click prediction: a view from the trenches. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1222-1230). https://doi.org/10.1145/2487575.2488200

Missaghi, S., Hondzo, M., and Herb, W. (2017). Prediction of lake water temperature, dissolved oxygen, and fish habitat under changing climate. Climatic Change, 141(4):747–757. https://doi.org/10.1007/s10584-017-1916-1

Mohri, M., Rostamizadeh, A. and Talwalkar, A., 2018. Foundations of machine learning. MIT press.

Mohseni, O. and Stefan, H.G., 1999. Stream temperature/air temperature relationship: a physical interpretation. Journal of hydrology, 218(3-4), 128-141. https://doi.org/10.1016/S0022-1694(99)00034-7

Mohseni, O., Stefan, H.G. and Erickson, T.R., 1998. A non-linear regression model for weekly stream temperatures. Water Resources Research, 34(10), 2685-2692. https://doi.org/10.1029/98wr01877

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. and Brown, S.D., 2004. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285. https://doi.org/10.1002/cem.873

Napiorkowski, M.J., Piotrowski, A.P. and Napiorkowski, J.J., 2014. Stream temperature forecasting by means of ensemble of neural networks: Importance of input variables and ensemble size. In River Flow, 2017-2025. https://doi.org/10.1201/b17133-269

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology, 10(3), pp.282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Navada, A., Ansari, A.N., Patil, S. and Sonkamble, B.A., 2011, June. Overview of use of decision tree algorithms in machine learning. In 2011 IEEE control and system graduate research colloquium, 37-42. IEEE. https://doi.org/10.1109/ICSGRC.2011.5991826

Nevitt, J. and Hancock, G.R., 2000. Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. The Journal of experimental education, 68(3), pp.251-268. https://doi.org/10.1080/00220970009600095.

Niimi, A. J., 1982, Economic and environmental issues of the proposed extension of the winter navigation season and improvements on the Great Lakes–St. Lawrence Seaway system, J. Great Lakes Res., 8, 532–549, doi:10.1016/S0380-1330(82)71991-4. https://doi.org/10.1016/S0380-1330(82)71991-4

Obertegger, U., Obrador, B. and Flaim, G., 2017. Dissolved oxygen dynamics under ice: Three winters of high-frequency data from Lake Tovel, Italy. Water Resources Research, 53(8), pp.7234-7246. https://doi.org/10.1002/2017WR020599

Opeyemi, O. and Justice, E.O., 2012. Development of neuro-fuzzy system for early prediction of heart attack. International Journal of Information Technology and Computer Science, 4(9), pp.22-28. https://doi.org/10.5815/ijitcs.2012.09.03

O'Reilly, C. M., Sharma, S., Gray, D. K., Hampton, S. E., Read, J. S., Rowley, R. J.,Opeyemi, O. and Justice, E.O., 2012. Development of neuro-fuzzy system for early prediction of heart attack. Information Technology and Computer Science, 9(9), 22-28. https://doi.org/10.5815/ijitcs.2012.09.03

O'Reilly, C.M., Alin, S.R., Plisnier, P.D., Cohen, A.S. and McKee, B.A., 2003. Climate change decreases aquatic ecosystem productivity of Lake Tanganyika, Africa. Nature, 424(6950), 766-768. https://doi.org/10.1038/nature01833

Perga, M.E., Minaudo, C., Doda, T., Arthaud, F., Beria, H., Chmiel, H.E., Escoffier, N., Lambert, T., Napolleoni, R., Obrador, B. and Perolo, P., 2024. Near-bed stratification controls bottom hypoxia in ice-covered alpine lakes. Limnology and Oceanography. https://doi.org/10.1002/lno.12341

Peterson, L.E., 2009. K-nearest neighbor. Scholarpedia, 4(2), 1883. https://doi.org/10.4249/scholarpedia.1883

Piccolroaz, S., 2016. Prediction of lake surface temperature using the air2water model: guidelines, challenges, and future perspectives. Advances in Oceanography and Limnology, 7(1), 36–50. https://doi.org/10.4081/aiol.2016.5791

Piccolroaz, S., Calamita, E., Majone, B., Gallice, A., Siviglia, A., and Toffolon, M. 2016. Prediction of river water temperature: a comparison between a new family of hybrid models and statistical approaches. Hydrological Processes, 30(21), 3901-3917, https://doi.org/10.1002/hyp.10913.

Piccolroaz, S., Healey, N.C., Lenters, J.D., Schladow, S.G., Hook, S.J., Sahoo, G.B., Toffolon, M., 2018. On the predictability of lake surface temperature using air temperature in a changing climate: a case study for Lake Tahoe (USA). Limnology and oceanography, 63(1), 243–261. https://doi.org/10.1002/lno.10626

Piccolroaz, S., Toffolon, M. and Majone, B., 2015. The role of stratification on lakes' thermal response: The case of Lake Superior. Water Resources Research, 51(10), 7878-7894. https://doi.org/10.1002/2014WR016555.

Piccolroaz, S., Toffolon, M., Majone, B., 2013. A simple lumped model to convert air temperature into surface water temperature in lakes. Hydrology and Earth System Sciences, 17(8), 3323–3338. https://doi.org/10.5194/hess-17-3323-2013

Piccolroaz, S., Woolway, R.I. and Merchant, C.J., 2020. Global reconstruction of twentieth century lake surface water temperature reveals different warming trends depending on the climatic zone. Climatic Change, 160(3), pp.427-442. https://doi.org/10.1007/s10584-020-02663-z

Piccolroaz, S., Zhu, S., Ptak, M., Sojka, M. and Du, X., 2021. Warming of lowland Polish lakes under future climate change scenarios and consequences for ice cover and mixing dynamics. Journal of Hydrology: Regional Studies, 34, 100780. https://doi.org/10.1016/j.ejrh.2021.100780

Piotrowski, A.P. and Napiorkowski, J.J., 2018. Performance of the air2stream model that relates air and stream water temperatures depends on the calibration method. Journal of Hydrology, 561, 395-412. https://doi.org/10.1016/j.jhydrol.2018.04.016

Piotrowski, A.P., Napiorkowski, J.J. and Piotrowska, A.E., 2020. Impact of deep learning-based dropout on shallow neural networks applied to stream temperature modelling. Earth-Science Reviews, 201, 103076. https://doi.org/10.1016/j.earscirev.2019.103076

Piotrowski, A.P., Napiorkowski, M.J., Kalinowska, M., Napiorkowski, J.J. and Osuch, M., 2016. Are evolutionary algorithms effective in calibrating different artificial neural network types for streamwater temperature prediction? Water resources management, 30(3), 1217-1237. https://doi.org/10.1007/s11269-015-1222-5

Piotrowski, A.P., Napiorkowski, M.J., Napiorkowski, J.J. and Osuch, M., 2015. Comparing various artificial neural network types for water temperature prediction in rivers. Journal of Hydrology, 529, 302-315. https://doi.org/10.1016/j.jhydrol.2015.07.044

Piotrowski, A.P., Osuch, M. and Napiorkowski, J.J., 2021. Influence of the choice of stream temperature model on the projections of water temperature in rivers. Journal of Hydrology, 601, 126629. https://doi.org/10.1016/j.jhydrol.2021.126629

Piotrowski, A.P., Osuch, M., Napiorkowski, M.J., Rowinski, P.M. and Napiorkowski, J.J., 2014. Comparing large number of metaheuristics for artificial neural networks training to predict water temperature in a natural river. Computers & Geosciences, 64, 136-151. https://doi.org/10.1016/j.cageo.2013.12.013

Probst, P., Boulesteix, A.L. and Bischl, B., 2019. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. Journal of Machine Learning Research, 20(53), 1-32. https://doi.org/10.48550/arXiv.1802.09596

Ptak, M. and Nowak, B., 2016. Variability of oxygen-thermal conditions in selected lakes in poland. Ecological Chemistry and Engineering S, 23(4):639–650. https://doi.org/10.1515/eces-2016-0045

Ptak, M., Sojka, M., and Nowak, B., 2020. Effect of climate warming on a change in thermal and ice conditions in the largest lake in poland–lake ́sniardwy. Journal of Hydrology and Hydromechanics, 68(3):260–270. https://doi.org/10.2478/johh-2020-0024

Qiu, R., Wang, Y., Wang, D., Qiu, W., Wu, J. and Tao, Y., 2020. Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River. Science of The Total Environment, p.139729. https://doi.org/10.1016/j.scitotenv.2020.139729

Quan Q, Hao Z, Xifeng H, Jingchun L., 2020. Research on water temperature prediction based on improved support vector regression. Neural Computing and Applications. Springer Science and Business Media LLC; 2020 Mar 28. https://doi.org/10.1007/s00521-020-04836-4

Rabi, A., Hadzima-Nyarko, M. and Šperac, M., 2015. Modelling river temperature from air temperature: case of the River Drava (Croatia). Hydrological sciences journal, 60(9), 1490-1507. https://doi.org/10.1080/02626667.2014.914215

Radulescu, V., 2020. Application of the Neural Networks in Prediction of the Thermal Flow on the Jiu River. In Heat Transfer Summer Conference (Vol. 83709, p. V001T04A005). American Society of Mechanical Engineers. https://doi.org/10.1115/HT2020-9043

Rafiq, M.Y., Bugmann, G. and Easterbrook, D.J., 2001. Neural network design for engineering applications. Computers & Structures, 79(17), pp.1541-1552. https://doi.org/10.1016/j.compstruc.2005.03.019

Ragotzkie, R.A., 1978. Heat budgets of lakes. In Lakes (1-19). Springer, New York, NY. https://doi.org/10.1007/978-1-4757-1152-3_1

Read, J.S., Jia, X., Willard, J., Appling, A.P., Zwart, J.A., Oliver, S.K., Karpatne, A., Hansen, G.J., Hanson, P.C., Watkins, W. and Steinbach, M., 2019. Process-guided deep learning predictions of lake water temperature. Water Resources Research, 55(11), 9173-9190. https://doi.org/10.1029/2019WR024922

Rehana, S., 2019. River water temperature modelling under climate change using support vector regression. In Hydrology in a Changing World, 171-183. Springer, Cham. https://doi.org/10.1007/978-3-030-02197-9_8

Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research, 3(Mar), pp.1371-1382.

Rivers-Moore, N.A., Bezuidenhout, C.N. and Jewitt, G.P., 2005. Modelling highly variable daily maximum water temperatures in a perennial South African river system. African Journal of Aquatic Science, 30(1), 55-63. https://doi.org/10.2989/16085910509503835

Saber, A., James, D.E. and Hayes, D.F., 2020. Long-term forecast of water temperature and dissolved oxygen profiles in deep lakes using artificial neural networks conjugated with wavelet transform. Limnology and Oceanography, 65(6), 1297-1317. https://doi.org/10.1002/lno.11390

Sahoo, G.B., Schladow, S.G. and Reuter, J.E., 2009. Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. Journal of Hydrology, 378(3-4), 325-342. https://doi.org/10.1016/j.jhydrol.2009.09.037

Sainath, T.N., Vinyals, O., Senior, A. and Sak, H., 2015, April. Convolutional, long short-term memory, fully connected deep neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4580-4584. IEEE. https://doi.org/10.1109/icassp.2015.7178838

Samadianfard, S., Honeyeh, K., Ozgur, K. and Wen-Cheng, L., 2016. Water temperature prediction in a subtropical subalpine lake using soft computing techniques. Earth Sciences Research Journal, 20(2), 1-11. https://doi.org/10.15446/esrj.v20n2.43199

Schaefli, B. and Gupta, H. V., 2007. Do Nash values have value? Hydrological Processes: An International Journal, 21(15):2075–2080. https://doi.org/10.1002/hyp.6825

Schmid, M., Hunziker, S. and Wüest, A., 2014. Lake surface temperatures in a changing climate: A global sensitivity analysis. Climatic change, 124, pp.301-315. https://doi.org/10.1007/s10584-014-1087-2

Schneider, P. and Hook, S. J., 2010. Space observations of inland water bodies show rapid surface warming since 1985. Geophysical Research Letters, 37(22). https://doi.org/10.1029/2010GL045059

Sener, E., Terzi, O., Sener, S. and Kucukkara, R., 2012. Modeling of water temperature based on GIS and ANN techniques: Case study of Lake Egirdir (Turkey). Ekoloji, 21(83), 44-52. https://doi.org/10.5053/ekoloji.2012.835

Sharma, A., 2020. 4 Simple Ways to Split a Decision Tree in Machine Learning. https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/

Sharma, S., Blagrave, K., Magnuson, J.J., O'Reilly, C.M., Oliver, S., Batt, R.D., Magee, M.R., Straile, D., Weyhenmeyer, G.A., Winslow, L. and Woolway, R.I., 2019. Widespread loss of lake ice around the Northern Hemisphere in a warming world. Nature Climate Change, 9(3), pp.227-231. https://doi.org/10.1038/s41558-018-0393-5

Sharma, S., Walker, S.C. and Jackson, D.A., 2008. Empirical modelling of lake water-temperature relationships: a comparison of approaches. Freshwater biology, 53(5), 897-911. https://doi.org/10.1111/j.1365-2427.2008.01943.x

Shuter, B.J., Finstad, A.G., Helland, I.P., Zweimüller, I. and Hölker, F., 2012. The role of winter phenology in shaping the ecology of freshwater fish and their sensitivities to climate change. Aquatic Sciences, 74(4), pp.637-657. https://doi.org/10.1007/s00027-012-0274-3

Siegert, M. J., Clarke, R. J., Mowlem, M., Ross, N., Hill, C. S., Tait, A., ... & Waugh, E., 2012. Clean access, measurement, and sampling of Ellsworth Subglacial Lake: a method for exploring deep Antarctic subglacial lake environments. Reviews of Geophysics, 50(1). https://doi.org/10.1029/2011RG000361

Spearman, C., 1910. Correlation calculated from faulty data. British journal of psychology, 3(3), 271. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Steinwart, I. and Christmann, A., 2008. Support vector machines. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-77242-4

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. and Zeileis, A., 2008. Conditional variable importance for random forests. BMC bioinformatics, 9, pp.1-11. https://doi.org/10.1186/1471-2105-9-307

St-Hilaire, A., Ouarda, T.B., Bargaoui, Z., Daigle, A. and Bilodeau, L., 2012. Daily river water temperature forecast model with a k-nearest neighbour approach. Hydrological Processes, 26(9), 1302-1310. https://doi.org/10.1002/hyp.8216

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947-1958. https://doi.org/10.1021/ci034160g

Temizyurek, M. and Dadaser-Celik, F., 2018. Modelling the effects of meteorological parameters on water temperature using artificial neural networks. Water Science and Technology, 77(6), 1724-1733. https://doi.org/10.2166/wst.2018.058

Toffolon, M., Piccolroaz, S., Calamita, E. 2020. On the use of averaged indicators to assess lakes' thermal response to changes in climatic conditions, Environmental Research Letters, 15, 034060, https://doi.org/10.1088/1748-9326/ab763e

Toffolon, M., Piccolroaz, S., Majone, B., Soja, A.M., Peeters, F., Schmid, M., Wüest, A., 2014. Prediction of surface temperature in lakes with different morphology using air temperature. Limnology and Oceanography, 59(6), 2185–2202. https://doi.org/10.4319/lo.2014.59.6.2185

Trinh, N.X., Trinh, T.Q., Phan, T.P., Thanh, T.N. and Thanh, B.N., 2019. Water Temperature Prediction Models in Northern Coastal Area, Vietnam. Asian Review of Environmental and Earth Sciences 6 (1) 2313-8173. https://doi.org/10.20448/journal.506.2019.61.1.8.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology, 49(11), 1225-1231. https://doi.org/10.1016/S0895-4356(96)00002-9

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F. and Mueller, A., 2015. Scikit-learn: Machine learning without learning the machinery. GetMobile: Mobile Computing and Communications, 19(1), pp.29-33.https://doi.org/10.1145/2786984.2786995

Veling, P.S., Kalelkar, M.R.S., Ajgaonkar, M.L.V. and Mestry, M.N.V., 2019. Mango Disease Detection By Using Image Processing. International journal for research in applied science and engineering technology, 7(4), 3717-3726. https://doi.org/10.22214/ijraset.2019.4624

Vivencio, D.P., Hruschka, E.R., do Carmo Nicoletti, M., dos Santos, E.B. and Galvao, S.D., 2007, April. Feature-weighted k-nearest neighbor classifier. In 2007 IEEE Symposium on Foundations of Computational Intelligence (pp. 481-486). IEEE. https://doi.org/10.1109/FOCI.2007.371516

Volkov, S., Bogdanov, S., Zdorovennov, R., Zdorovennova, G., Terzhevik, A., Palshin, N., Bouffard, D. and Kirillin, G., 2019. Fine scale structure of convective mixed layer in ice-covered lake. Environmental Fluid Mechanics, 19, pp.751-764. https://doi.org/10.1007/s10652-018-9652-2

Wade, J., Kelleher, C. and Hannah, D.M., 2023. Machine learning unravels controls on river water temperature regime dynamics. Journal of Hydrology, p.129821. https://doi.org/10.1016/j.jhydrol.2023.129821

Walsh, S.E., Vavrus, S.J., Foley, J.A., Fisher, V.A., Wynne, R.H. and Lenters, J.D., 1998. Global patterns of lake ice phenology and climate: Model simulations and observations. Journal of Geophysical Research: Atmospheres, 103(D22), pp.28825-28837. https://doi.org/10.1029/98JD02275

Watson, C.D., Auger, G., Tewari, M., Treinish, L.A. and Johnston, K.E., 2021. Predicting complete winter ice coverage at Lake George, New York. International Journal of Climatology, 41, pp.E1236-E1251. https://doi.org/10.1002/joc.6764

Wenxian, G., Hongxiang, W., Jianxin, X. and Wensheng, D., 2010, May. PSO-BP neural network model for predicting water temperature in the middle of the Yangtze river. In 2010 International Conference on Intelligent Computation Technology and Automation, 2, 951-954. IEEE. https://doi.org/10.1109/ICICTA.2010.501

Weyhenmeyer, G.A., Obertegger, U., Rudebeck, H., Jakobsson, E., Jansen, J., Zdorovennova, G., Bansal, S., Block, B.D., Carey, C.C., Doubek, J.P. and Dugan, H., 2022. Towards critical white ice conditions in lakes under global warming. Nature communications, 13(1), p.4974. https://doi.org/10.1038/s41467-022-32633-1

Woolway, R.I., Kraemer, B.M., Lenters, J.D., Merchant, C.J., O'Reilly, C.M. and Sharma, S., 2020. Global lake responses to climate change. Nature Reviews Earth & Environment, 1(8), 388-403. https://doi.org/10.1038/s43017-020-0067-5

Woolway, R.I., Merchant, C.J., 2018. Intralake heterogeneity of thermal responses to climate change: A study of large Northern Hemisphere lakes. Journal of Geophysical Research: Atmospheres, 123(6), 3087-3098. https://doi.org/10.1002/2017JD027661

Wu, Y., Duguay, C.R. and Xu, L., 2021. Assessment of machine learning classifiers for global lake ice cover mapping from MODIS TOA reflectance data. Remote Sensing of Environment, 253, p.112206. https://doi.org/10.1016/j.rse.2020.112206

Xie, P., Qiu, Y., Wang, X., Shi, L. and Liang, W., 2020, May. Lake Ice Phenology Extraction using Machine Learning Methodology. In IOP Conference Series: Earth and Environmental Science (Vol. 502, No. 1, p. 012034). IOP Publishing. https://doi.org/10.1088/1755-1315/502/1/012034

Yang, K., Yu, Z. and Luo, Y., 2020. Analysis on driving factors of lake surface water temperature for major lakes in Yunnan-Guizhou Plateau. Water Research, 184, 116018. doi:10.1016/j.watres.2020.116018

Yousefi, A. and Toffolon, M., 2022. Critical factors for the use of machine learning to predict lake surface water temperature. Journal of Hydrology, 606, p.127418. https://doi.org/10.1016/j.jhydrol.2021.127418

Zaier, I., Shu, C., Ouarda, T.B., Seidou, O. and Chebana, F., 2010. Estimation of ice thickness on lakes using artificial neural network ensembles. Journal of Hydrology, 383(3-4), pp.330-340. https://doi.org/10.1016/j.jhydrol.2010.01.00

Zhang, Y. and Baptista, A.M., 2008. SELFE: a semi-implicit Eulerian–Lagrangian finite-element model for cross-scale ocean circulation. Ocean modelling, 21(3-4), 71-96. https://doi.org/10.1016/j.ocemod.2007.11.005

Zhu, S. and Heddam, S., 2019. Modelling of Maximum Daily Water Temperature for Streams: Optimally Pruned Extreme Learning Machine (OPELM) versus Radial Basis Function Neural Networks (RBFNN). Environmental Processes, 6(3), 789-804. https://doi.org/10.1007/s40710-019-00385-8

Zhu, S., Bonacci, O., Oskoruš, D., Hadzima-Nyarko, M. and Wu, S., 2019. Long term variations of river temperature and the influence of air temperature and river discharge: case study of Kupa River watershed in Croatia. Journal of Hydrology and Hydromechanics, 67(4), 305-313. https://doi.org/10.2478/johh-2019-0019

Zhu, S., Hadzima-Nyarko, M., Gao, A., Wang, F., Wu, J. and Wu, S., 2019. Two hybrid data-driven models for modeling water-air temperature relationship in rivers. Environmental Science and Pollution Research, 26(12), 12622-12630. https://doi.org/10.1007/s11356-019-04716-y

Zhu, S., Heddam, S., Nyarko, E.K., Hadzima-Nyarko, M., Piccolroaz, S., and Wu, S., 2019. Modeling daily water temperature for rivers: comparison between adaptive neurofuzzy inference systems and artificial neural networks models. Environmental Science and Pollution Research 26 (1), 402–420. https://doi.org/10.1007/s11356-018-3650-2

Zhu, S., Heddam, S., Wu, S., Dai, J. and Jia, B., 2019. Extreme learning machine-based prediction of daily water temperature for rivers. Environmental Earth Sciences, 78(6), 202. https://doi.org/10.1007/s12665-019-8202-7

Zhu, S., Nyarko, E.K., Hadzima-Nyarko, M., Heddam, S. and Wu, S., 2019. Assessing the performance of a suite of machine learning models for daily river water temperature prediction. PeerJ, 7, e7065. https://doi.org/10.7717/peerj.7065

Zhu, S., Piotrowski, A.P., 2020. River/stream water temperature forecasting using artificial intelligence models: a systematic review. Acta Geophysica, 68(5), pp.1433–1442. https://doi.org/10.1007/s11600-020-00480-7.

Zhu, S., Ptak, M., Choiński, A., and Wu, S., 2020a. Exploring and quantifying the impact of climate change on surface water temperature of a high mountain lake in central europe. Environmental monitoring and assessment, 192(1):1–11. https://doi.org/10.1007/s10661-019-7994-y

Zhu, S., Ptak, M., Yaseen, Z. M., Dai, J., and Sivakumar, B., 2020b. Forecasting surface water temperature in lakes: A comparison of approaches. Journal of Hydrology,585:124809. https://doi.org/10.1016/j.jhydrol.2020.124809

# List of research outputs

## Scientific publications

Yousefi, A. and Toffolon, M., 2022. Critical factors for the use of machine learning to predict lake surface water temperature. *Journal of Hydrology*, *606*, 127418. https://doi.org/10.1016/j.jhydrol.2021.127418

Toffolon, M., Yousefi, A., & Piccolroaz, S. (2022). Estimation of the thermally reactive layer in lakes based on surface water temperature. *Water Resources Research*, *58*(6), e2021WR031755. https://doi.org/10.1029/2021WR031755

## Unpublished works

Yousefi, A., Weyhenmeyer, G., Piccolroaz, S., Toffolon, M. Selection of features to predict ice thickness in boreal lakes identified by machine learning (In preparation)

Yousefi, A., Piccolroaz, S., Amadori, M., Toffolon, M. Machine Learning unveils the factors controlling lake surface water temperature in different climates (In preparation)

## Conferences

Yousefi, A. and Toffolon, M., 2021. Machine Learning Algorithms for Lake Surface Water Temperature Prediction. Knowledge Guided Machine Learning (KGML) 2021 Conference.

Yousefi, A. and Toffolon, M., 2021.The influence of water depth and forcing factors on the performances of Machine Learning approaches for the simulation of lake surface water temperature. EGU General Assembly Conference Abstracts, EGU21-2970.

Yousefi, A., Weyhenmeyer, G., Piccolroaz, S., Toffolon, M., 2022. Lake Ice Thickness Patterns identified by Machine Learning. SIL Conference.

Yousefi, A., Amadori, M., Piccolroaz, S., Toffolon, M., 2022. Analysis of the influence of different meteorological factors on lake surface water temperature across different climates. SIL Conference.

Yousefi, A., Piccolroaz, S., Amadori, M., Toffolon, M., 2024. Ranking the meteorological factors influencing lake surface water temperature across different climates. PPNW Conference.

# Acknowledgements

I would like to express my deepest gratitude to my esteemed supervisors, Marco Toffolon and Sebastiano Piccolroaz. Their exceptional guidance, unwavering support, and immense expertise have been instrumental in the successful completion of this thesis. Their profound knowledge of the subject matter, insightful feedback, and continuous encouragement have shaped the direction of my research and enriched my understanding of the field. I am truly fortunate to have had the opportunity to work under their mentorship, and I am grateful for their dedication and commitment throughout this journey.

I would also like to extend my heartfelt thanks to Gesa Wehenmeyer for her invaluable supervision during my exchange period in Uppsala, Sweden. Her guidance, encouragement, and expertise have had a profound impact on my academic growth and have broadened my perspectives in the field. I am grateful for the opportunity to learn from her and for the support she provided during this enriching experience.

Additionally, I would like to acknowledge Marina Amadori for her co-supervision during the CCI project. Her insightful input, attention to detail, and collaborative spirit have greatly contributed to the success of our joint efforts. I am thankful for her guidance and expertise, which have been integral to the development and execution of the project.

Furthermore, I would like to express my heartfelt appreciation to my family for their unconditional love, unwavering support, and understanding throughout this endeavor. Their encouragement, belief in my abilities, and sacrifices have been a constant source of motivation and strength.

Lastly, I would like to extend my gratitude to all the individuals, colleagues, and friends who have provided assistance, shared their knowledge, and offered valuable suggestions during the course of this research. Their contributions, whether big or small, have played a significant role in shaping this thesis, and I am sincerely thankful for their support.