Critical factors for the use of machine learning to predict lake surface water temperature

Azadeh Yousefi, Marco Toffolon

Department of Civil, Environmental and Mechanical Engineering, University of Trento, Italy

Abstract

Models based on Machine Learning (ML) are pervading all fields of science and practical applications, including the problem of forecasting water temperature in lakes, a crucial variable for ecosystems and a proxy of climate change alterations. Here, we review the most used ML algorithms in this field and highlight some physical constraints that should be carefully considered when adopting a black-box approach. To illustrate them, we refer to an artificial case study representing a temperate lake simulated by means of a physically based model, for which we have full control of input and output variables, and restrict the analysis to lake surface water temperature (LSWT). Three main factors are relevant for a successful prediction of LSWT by means of ML models: the choice of the predictors (mostly, meteorological variables), their pre-processing (we tested three approaches), and the specific ML algorithm (nine different approaches). We show that selecting the suitable physical inputs plays the most important role. In our case study, which is the product of a numerical model and not a real lake, the minimum amount of information that is needed to obtain acceptable results is to consider air temperature (AT) and day of the year. The use of additional predictors does not substantially improve the performances (the relative improvement of RMSE was 7.75% for the test data set). We also demonstrate that better results than the normal case are obtained by either pre-processing air temperature data averaging them over a time window or including values from previous days as inputs in the model. Considering the recent history of the forcing (AT) allows one to comply with the physical fact that the large water mass makes lakes acting as “filters” in their thermal response (thus, influenced also by AT from previous days), which changes depending on the lake’s depth. Eventually, we did not find a definite answer about a single optimal ML algorithm when using the same inputs (although artificial neural network had slightly better results), suggesting that the insight into the physical dynamics is still the most important factor for a successful exploitation of ML.

Keywords: Machine Learning, Lake Surface Water Temperature (LSWT), Lake thermal response, Meteorological forcing, Heat capacity, Lake depth

Highlights

1. Lake surface water temperature is simulated by means of machine learning models
2. A physically based model is used to have full control of variables
3. Nine machine learning algorithms are compared, and none of them clearly prevails
4. Best predictors are air temperature and day of the year; others have minor effects
5. Using forcing from previous days allows for considering the history of the system

1. Introduction

Lake water temperature is influenced by the local climate conditions, and the current warming trend (Woolway and Merchant, 2018; Woolway et al., 2020) represents a severe threat to the biological and chemical processes in the aquatic ecosystem (O'Reilly et al., 2003). Understanding the factors controlling the lake thermal response to changing climate conditions, and possibly quantifying the expected alterations, is of paramount importance for aquatic life. To this end, several models have been proposed ranging from very simple correlations to complex three-dimensional numerical models.

Water temperature in lakes changes with depth due to the process of thermal stratification that reaches its maximum during summer. Here we focus on lake surface water temperature (LSWT; all the acronyms used in this work are also listed in Table S1 in the Supplementary Material) as a quantity that can be easily observed and measured in situ or by remote sensing. LSWT prediction is challenging because it is influenced by a complex interaction among different meteorological factors (such as wind speed, air temperature, shortwave and longwave radiation, latent and sensible heat), lake bathymetry, inflows and outflows (Livingstone, 2003; Piccolroaz et al., 2015). Many successful models were implemented to forecast LSWT, which can be subdivided into two big categories: physically-based models, which resolve the fundamental equations and could ideally provide results in a prognostic way without any support of LSWT data (although data are needed for the calibration of the empirical relations that are always present in this kind of models); and data-driven models, whose application inevitably relies on the availability of LSWT measurements for the calibration of generic relations.

Physical models for LSWT are formulated in terms of a set of deterministic equations describing the heat budget for a bulk volume (Ragotzkie, 1978) or for one-, two- or three-dimensional discretization of the water body (Hipsey et al., 2019; Irving et al., 2006). These models demand a large amount and different types of input data, including lake bathymetry, inflow and outflow characteristics and meteorological variables (Piotrowski and Napiorkowski, 2018).

The category of data-driven models is more varied, containing the whole range of stochastic methods, from simple linear and non-linear regression to autoregressive, periodic autoregressive (Benyahya et al., 2007) and evolutionary polynomial regression (Doglioni et al., 2008). Data-driven models relate LSWT to other influential factors, mostly air temperature (AT), in order to forecast the thermal dynamics of water. Data-driven models also include machine learning (ML) algorithms, which are becoming increasingly popular as general-purpose computational models (Mohri et al., 2018). An increasing number of attempts to predict LSWT with supervised learning forecasts based on the available information has been pursued in the last decade, with a remarkable acceleration since 2019. The main references are reported in Table 1, which summarizes the features of previous studies about LSWT prediction by ML in lakes, highlighting the different ML algorithms and the features of the investigated lakes. As shown in Table 1, Artificial Neural Network (ANN) is the most used and successful ML algorithm 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., 2019; Heddam et al., 2020; Jia et al., 2019 and 2020; Quan et al., 2020; Zhu et al., 2020).

As the problem of predicting water temperature with ML in rivers has some similarities to the case of lakes, the corresponding literature has been examined and included in the Supplementary Material (Table S2). It should be mentioned that lakes are lentic water bodies that are less affected by the advection of water from upstream regions, which is often crucial in rivers, but their thermal behaviour is made more complex by the thermal stratification developing throughout the year.

At the interface of the categories of physically based and data-driven models, we can identify the so-called hybrid models, a generic definition that includes different approaches. Following Karpatne et al. (2017), there are different ways to implement hybrid models. One approach is that the input of data-driven models comes from physically based models, mostly used for downscaling of climate variables. Moreover, the physically based outputs could also be used to control the training phase of data-driven models by preparing the initial estimation of target output for the training phase (see Read et al., 2019, for an example of this kind). Another approach is to use data-driven methods in the framework of physically-based models to improve inaccurate estimates of parameters or reproducing particularly complex processes.

Read et al. (2019) used traditional physically-based models to estimate weights and biases in the pre-training of ML, as well as to calibrate the model using physical laws in the pre-training and training phase. They compared their hybrid model, consisting of a combination of ML and physically-based models, with the two methods individually. Based on their achievements, the ML outperforms the physically-based model if the number of training data exceeds a certain value (which depends on the data set), while the performance of the process-based model does not improve after a certain number of training data, confirming that ML approaches demand (and benefit from) a larger amount of data. By analysing different sets of data clustered at seasonal or climatic scales, Read et al. (2019) also emphasised that ML provides more accurate results when the data have the same pattern, and confirmed that, for their case studies, the hybrid model was the most successful approach in comparison to the ML and process-based methods separately.

A different type of hybrid model that has been developed to predict LSWT is *air2water* (Piccolroaz et al., 2013, 2018, 2020, 2021; Toffolon et al., 2014; Piccolroaz, 2016), which derives simplified equations from physical laws condensing all the details into a reduced number of parameters to be calibrated against measurements. *air2water* needs only AT as input, making it as simple to use as other data-driven models while outperforming many of them (Heddam et al., 2020). Due to these features, it can also be used in comparative tests of ML algorithms.

In this paper, we test the ML approaches to predict water temperature, with a special focus on LSWT, keeping a fundamental question in mind. How can we optimize the physical information to improve the performance of ML models, considering the choice of predictors, the pre-processing methods and the features of the ML algorithm? For instance, pre-processing is typically used to denoise the data or rescale them (Saber et al., 2019; Zhu et al., 2020): can it also help obtaining better results? To our knowledge, no study has reviewed these aspects so far, nor it has explicitly considered how physical processes (for instance, the effect of depth on the heat capacity of a lake) can influence the ML implementation. Our study aims at filling these gaps by means of a systematic analysis of the selection of predictors, their pre-processing, and the performance of different ML algorithms.

Table 1

2. Characterization of the physical problem

Before discussing the performances of the ML methods to simulate LSWT, it is interesting to understand the physics behind the investigated process. Indeed, water temperature in lakes is affected by a large number of different meteorological and limnological variables, and having a deeper insight into the heat budget yields useful information about the influential factors.

It is well-known that lakes stratify, forming layers with different temperatures (hence, density) that tend to behave as independent regions as stratification becomes stronger. This is particularly evident in temperate lakes during summer when the surface layer (epilimnion) is often well mixed and separated from deep waters (hypolimnion) by a layer where the largest temperature difference is localized (thermocline). In the well-mixed surface layer, the volume-averaged temperature is strongly correlated with LSWT, which eventually depends on the net heat flux exchanged through the lake surface. Among the factors controlling the net heat flux, we can recognize most of the variables that potentially affect the prediction of LSWT (more details in Appendix A): shortwave solar radiation; longwave radiation depending on AT and cloud cover; latent heat depending on AT, relative humidity and wind speed; sensible heat depending on AT and wind speed; and heat contribution of rainfall.

Lake's depth is another crucial parameter to be considered in modelling lake temperature. For shallow lakes, the water temperature is often uniform over the whole water column, different from deep lakes, where the hypolimnetic temperature remains significantly colder than the surface temperature during summer. Consequently, lakes behave differently depending on their depth (e.g., Calamita et al., 2021), and the behaviour of deep lakes is typically affected by the evolution of stratification (e.g., Piccolroaz et al., 2015). The role of the depth can be highlighted also by looking at minimal physical models (e.g., see Appendix A and Toffolon et al., 2014). Here, it is sufficient to recall that lakes act as ‘filters’: the input signal (AT, solar radiation, etc.) is ‘physically’ elaborated so that the output signal (e.g., LSWT) is smoothed.

3. Materials and methods

3.1. Case study

Our main purpose is to illustrate the crucial aspects for a proper application of a ML model to predict LSWT. To this end, instead of considering data from a real lake, we built a synthetic example for which we could compare and analyse the performances of the different approaches having the full control of all the involved variables. This also allows us to consider the effect of lake morphology (depth in particular) on the response of LSWT while keeping the same meteorological forcing, and hence on the capability of ML to simulate the thermal dynamics in different conditions.

The synthetic case study was obtained by means of a 1D physically based model of widespread use, the General Lake Model (GLM) (Hipsey et al., 2019). The details of the model are uninfluential for our analysis, and any other model could have been used. We acknowledge that all results and conclusions that are drawn from our work are not based on a real case, but on numerical simulations, which however are known to reproduce the main lake dynamics correctly. The advantage is that using a fully controlled input allows us to focus on the performance of the ML model without introducing additional uncertainties related to the available measurements (such as missing data and erroneous values). Moreover, it allows us to compare lakes with different depths keeping all the other factors unchanged, thus isolating a single element in the analysis of its relevance. This approach can potentially be extended to any other physical variables. The disadvantage is that the artificial case does not possess the complexity of the real world, and only the factors that are included in the lake model are evaluated.

For the general features of the synthetic case study, we selected a medium-size peri-alpine lake in the temperate region (for more details, see Table 2) as a reference. In order to consider realistic values of the lake morphology and of the climatic conditions, we chose the shape of the hypsometric curve of Lake Caldonazzo (northern Italy) and the meteorological data in the corresponding latitude and longitude extracted from ERA5 (Hersbach et al., 2020). We notice that a real lake is only used as a reference to design the artificial case study, and we do not pretend to simulate any actual dynamics. The choice of an artificial lake allowed us to consider an extended and coherent data set and to see the effect of the same forcing factors in different depths, and in particular the difference between a shallow (maximum depth 5 m) and a deep (maximum depth 60 m) case, plus two intermediate depths (20 m and 40 m). In order to obtain these cases, the hypsometric curve of Lake Caldonazzo (having a maximum depth of 49 m) was altered by stretching the vertical coordinate. The surface area was maintained to keep the same interface with the atmosphere. The inflow to the lake was artificially adjusted to compensate the evaporation and keep the water level approximately constant from year to year; this choice minimizes the thermal effect of the inflows. An example of the seasonal evolution of the thermal stratification in the shallow and deep lake is reported in Fig. S1 in the Supplementary Material. Here it suffices to say that the shallow lake is almost always well-mixed, while the deeper cases clearly stratify in summer.

The data used in the ML application include the LSWT obtained by GLM as the target variable, and the meteorological data from ERA5 as predictors. All the data sets were converted to a daily time step. The main features of the data set, including the duration, are provided in Table 2.

Table 2

3.2. Setup of ML models

The setup of a ML model requires several steps: defining the inputs (predictors), pre-processing them, and, then, apply the ML algorithm. All the aspects of the methodology shown in Fig. 1 are analysed individually in the following paragraphs.

Figure 1

The first step, the choice of the predictors, i.e. the variables used as input to the ML algorithm, is crucial. Referring to our goal of predicting LSWT, a large number of predictors have been considered as an input of ML, also based on the considerations on the heat flux components discussed above:

Air Temperature (AT), Shortwave Radiation (SWR), Longwave Radiation (LWR), Relative Humidity (RH), Wind speed (WS), Rainfall (R) and Air Pressure (AP). To these physical variables, we added the Day Of the Year (DOY), and tested them both individually and in combination. The effect of considering different predictors is discussed in section 4.1.

Care should be taken with the use of DOY, because its discontinuous behaviour at the change of year (from day 365 to day 1) is not consistent with the continuous variation of LSWT. Hence, we also investigated the sine and cosine of DOY, $\sin(DOY/n_{DOY})$ and $\cos(DOY/n_{DOY})$, where n_{DOY} the number of days in a year: these are considered as two distinct input variables, which we indicate with the single term SCDOY for simplicity, but contain an equivalent amount of information as DOY.

The second step is pre-processing. The input data cannot be considered as they are, as they need to be pre-processed to make their range of variation uniform in order to be comparable. We tested different strategies analysing the most common pre-processors (Min-Max; Robust and Standard Scaler). We anticipate that preliminary tests for our problem showed that Min-Max provided the best results, so we adopted this method for all the following analyses.

However, LSWT responds to external forcing by filtering the input due to the heat capacity of the water mass, as discussed in section 2. Thus, we included in the pre-processing step also the possibility to temporally average some of the predictors, and in particular AT. As a different strategy to obtain the same effect, we also explicitly considered the ATs from previous days, thus increasing the number of input variables of the ML model. The different pre-processing strategies are discussed in section 4.2.

Following a standard procedure, the available data were split into two subsets, typically named training and test data sets. The parameters of the ML algorithms were optimized using the training data set, which in our case was 80% of the whole data set (other proportions such as 50%, 60% and 70% were also tested, but the results did not change significantly, thanks to the availability of a large data set that reduces the risk of overparameterization). Then, the ML performances were examined using the test data set to be confident that the calibrated parameters were not the result of overfitted or underfitted training and could be used in other cases as well. The separation between training and test data sets was done chronologically (first 80%, last 20%) in order to maintain the temporal pattern of the variables within a data set. We also tested a random splitting, but the results of the two data sets were very similar in all cases reducing the possibility to identify the best parameters.

The third step is the choice of the ML algorithm, which is the kernel of the approach and contains some structural parameters, defined as hyperparameters, which are determined before training the data (e.g., Probst et al., 2019). The distinction between hyperparameters and parameters is conventional and is based on the fact that the former define the features of the algorithm (e.g., the number of layers and neurons in Artificial Neural Networks, ANN), while the latter specify the coefficients used in the equations used in the algorithm (e.g., the values of weights and biases for ANN). Both the hyperparameters and the parameters are selected and obtained for the individual case studies and combinations of input variables: the best values for the hyperparameters are identified using optimization algorithms, while the way parameters are calibrated depends on the specific ML approach.

The selection of the hyperparameters can be performed by means of different optimization methods, such as Grid Search, Random Search, Bayesian methods, Particle Swarm Optimization (PSO). In this paper, we refer to the Genetic algorithm (GA) for all ML approaches analysed: it is an evolutionary optimization method, which gives good results while keeping the computational burdening reasonable (Di Francescomarino et al., 2018).

Several ML algorithms have been used to predict water temperature in lakes (see Table 1, and Table S2 in the Supplementary Material for the case of rivers): linear and non-linear (e.g., logistic) regressions, ANN, Decision Tree (DT), Support Vector Regression (SVR), K-Nearest Neighbour algorithm (KNN), Random Forest (RF), and Gradient Boosting Machines (GBM). Although linear and logistic regression algorithms can get proper results in spite of their simplicity, more advanced ML algorithms proved to be more suitable for water temperature prediction in more general scenarios (Sharma et al., 2008; Sener et al., 2012).

In this paper, trying to consider all the options that have been proposed so far, we tested nine different ML approaches. Four of them can be traced back to the structure of ANN, although some have very peculiar features: Multilayer Perceptron Neural Network (MLPNN) and Back-Propagation Neural Network (BPNN) are two very similar approaches, that differ only for the strategy of parameter calibration; Long Short-Term Memory (LSTM) is a form of deep learning that retains the temporal memory of the system; Adaptive Neuro-Fuzzy Inference System (ANFIS) couples a network structure with fuzzy logic. Three approaches are part of DT: the standard DT and two of its branches, RF and Extremely Randomized Tree (ERT). Finally, the last two approaches are SVR and KNN. The main features of these ML algorithms are summarized and explained in Appendix B, and the details about the hyperparameters are presented in the Supplementary Material. The comparative evaluation of the nine ML is provided in section 4.3.

For the analysis of the different pre-processing strategies and predictors in sections 4.1 and 4.2, we had to refer to a single ML method: to this end, we selected BPNN, as a commonly used branch of ANN that allows for finding nonlinear correlations between dependent and independent variables and that is prone to multiple methods of training (Tu, 1996).

As a fourth and final step of the workflow, we had to evaluate the performances of the ML approaches. We adopted the root mean square error (RMSE) of the LSWT as objective function, with lower values representing the best fit. Other metrics can also be used (e.g., mean absolute error, correlation coefficient, Nash-Sutcliff efficiency index), but without significant differences in the ranking of best ML. In all cases, since we considered a synthetic case study, we assumed the GLM output as the observed truth for the present analysis.

In order to improve the robustness of the ML results and to limit the possibility that a single run can provide inconsistent results due to an unsatisfactory choice of the initial guess of the parameters, or any other numerical artefact, we decided to run each of the algorithms for 20 times (similar to what was done by Piotrowski et al., 2021, who considered 50 runs). If not specified, the metrics used to illustrate the performance are averaged among the runs; more details are provided in the Supplementary Material. Moreover, as a measure of the robustness of the method (i.e., how different are the results for the different runs), we computed the standard deviation of the 20 estimates of LSWT in each day of the simulated period, and then averaged these values along the whole record to obtain the averaged value, σ_R.

4. Results

In this section, the most influential inputs, the most effective pre-processing methods and the performance of ML approaches are discussed. The effect of the lake's depth is also analysed, and the physical characteristics of the lake are considered in the statistical analysis.

4.1. Identification of predictors

Understanding the physics behind the thermal dynamics is crucial to obtain accurate predictions, and this is especially true for approaches that are not based on physical laws. In fact, ML algorithms are generic tools, and the success of their application heavily depends on the selection of the set of predictors. According to the analysis in section 3, the potential predictors for LSWT are manifold. Previous analyses (Table 3) included, depending also on the simulated variable (daily averaged temperature, monthly or annual average or maximum): AT, DOY, Gregorian calendar, lake surface area, maximum and average water depth, altitude, WS, frozen and snowing indicators, R, percentage of cloud cover, inflow, geopotential height, LWR, SWR, pH, total dissolved solids, SWR, RH, outflow, Secchi depth, and – for seasonal evaluations – 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). We already listed most of these predictors when we discussed the factors affecting the heat flux at the lake surface. We additionally notice that, for each of them, it is necessary to choose the temporal and spatial scale of integration, which depends on the investigated problem and can affect the results significantly (e.g., Piccolroaz et al., 2016; Toffolon et al., 2020).

Table 3

The identification of the most influential predictors has often been based on statistical analysis. For instance, Zhu et al. (2020) investigated the linear and non-linear relationships between AT and LSWT. Mohseni et al. (1998) found that non-linear S-shaped functions work well for AT and LSWT correlation. Yang et al. (2020) utilized trend analysis, contribution analysis and (Pearson) correlation analysis, as methods to investigate the most influential predictors on LSWT, identifying AT and SWR.

Figure 2

It is important to recognize that there are correlations not only between the predictors and LSWT but also among the predictors themselves. We computed the Pearson and Spearman's rank-order correlations to find the relationships among the variables for our synthetic shallow and deep test cases. We recall that Spearman's rank-order measures the power and trend of monotonic relationship between two variables (Spearman, 1910), while Pearson considers the strength of linear correlation. Based on the results shown in Fig. 2, AT is the most relevant predictor for LSWT, also with a strong linear correlation, but SCDOY is also a very important predictor. The linear version of time (DOY) is significantly less correlated than SCDOY, and especially in the shallow case. Together with SCDOY and AT, also LWR and SWR follow similar trends in comparison to other predictors. Other relatively high correlations can be highlighted, e.g., R with RH and LWR, but with low scores with LSWT. We note that the influence of R might be overly small because rainfall did not affect the inflows in the simplified description used for the physically based model of the lake, a condition that is not fully representative of real cases.

Table 4

In order to analyse the effect of the choice of the predictors on the performances of ML in reproducing LSWT, we explored several combinations of the input variables based on the literature review summarized in Table 3. In this exercise, we referred only to BPNN, one of the most used ML algorithms. Table 4 reports the results of this analysis and shows the influence of each predictor, and of their combination, on the model's performances. Looking first at the single predictors considered independently, we can see that SCDOY provides the lowest RMSE (hereinafter, we use the average value among the 20 runs): 1.43°C for training set and 1.58°C for test set in the deep lake, with values around 1.78°C and 2.07°C for the shallow case. This implies that a mean year is likely a good representation of the intra-annual variability of LSWT. AT is the second choice, but the error grows significantly (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), and the choice of the other predictors does not provide acceptable results. Nonetheless, the combination of different variables allows for a significant improvement. In particular, considering AT and SCDOY already provides good performances (1.31°C – 1.46°C for the shallow case, 1.16°C – 1.29°C for the deep case), and we will use this combination as a reference for the following analyses. Indeed, adding further variables in the set of the predictors yields only minor improvements at the price of a larger amount of information to be used in the model (see the difference of RMSE in the last column of Table 4). However, we acknowledge that other variables can also be important in climatic conditions different from the temperate case investigated herein, and in the case of a real lake.

We illustrate the effect of considering AT and DOY (or SCDOY) on the model in more detail in Fig. 3, which shows the simulated LSWT (using again BPNN) for the shallow and deep cases in single runs. For the sake of clarity, we limit the plots to two years. LSWT fluctuates more in the shallow than in the deep lake since the former is more sensitive to meteorological forcing due to the smaller thermal inertia (heat capacity per unit of surface area) (Toffolon et al., 2014, 2020). It clearly appears that the range of LSWT variability predicted when AT is considered alone is too large, while the use of the single information from DOY provides a smooth description (similar to the mean year) which however is not able to reproduce the interannual variability. The use of SCDOY instead of DOY is more suitable because of its smooth transition at the end of each year. The last row of Fig. 3 shows the advantage of the combined use of SCDOY and AT to predict LSWT.

Figure 3

4.2. The importance of pre-processing

Since the input values have different units and ranges of values, pre-processing methods such as standardization and normalization are needed to remove the anomalous values and transfer the data to the same scale (in a comparable range). Sener et al. (2012) used the Min-Max transform to resize the input data (RH, AT, air pressure, and water depth) from 0 to 1 for the data to be comparable. Moving Average with Min-Max (MAMM) and Wavelet Transform (WT) are more complex pre-processing methods that not only denoise the data but also smooth them. The influence of pre-processing methods and their effect on the ML performances are analysed in this section.

We have seen that lakes act as filters, damping high-frequency inputs, and that the effect becomes stronger for deep lakes. This is a process that depends on the thermal inertia of the lake and needs to be included in the pre-processing, if the ML algorithm is not able to deal with the memory of the system. Assuming AT as a predictor, it is not possible to establish a linear relation between AT and LSWT in general, but a hysteresis exists between the two variables (Toffolon et al., 2014). Linear relationships between daily-averaged values can be valid for very shallow lakes but become critical for deep lakes, where LSWT in one day heavily depends on the previous day's value. If the ML algorithm does not have a suitable way to retain the value of the previous condition of the system, pre-processing should be exploited to have a more physically sound description of the lake system.

As a simple pre-processing approach to simulate the filtering behaviour of a lake, we considered smoothing and denoising the data using Moving Average (right-aligned window, which considers the AT of previous days) combined with Min-Max (MAMM). In this approach, different window sizes were used to minimize the RMSE. We found that the RMSE presents a minimum value for an optimal window size and that the shape of the curve is different for the shallow and deep lake (Fig. 4e,f). In the shallow case, the RMSE decreases rapidly and then increases at a similar rate for window sizes longer than the optimal value (Fig. 4e). In the deep case, the error for the test data set is less variable, different from the training data set, although it still presents a minimum for values of the windows size in the same range as that of the training data set (Fig. 4f). For this reason, the best window size was selected considering the minimum of the average of the training and test errors, and resulted in 8 days for the shallow case and 13 days for the deep case, confirming the role of the depth in enhancing the filtering effect on LSWT.

Figure 4

Further analyses would be required to assess the optimal window size as a function of the depth. However, some preliminary insight can be obtained by considering two intermediate depths (20 m and 40 m). By comparing the RMSE in the four cases, it seems that a threshold on the depth exists and determines the qualitative behaviour (shown in Fig. 5): while a clear minimum (optimal window size) appears for the shallow lake (5 m), the behaviour of the other (deeper) cases is more similar. Indeed, the 5 m lake does not stratify (Fig. S1a in Supplementary Material) and the heat capacity remains the same in all seasons. Conversely, when the depth grows, stratification develops and the thermally reactive layer varies with the season. In our case study, the thermocline is approximately on the order of 10 m in the analysed case (Fig. S1b in Supplementary Material), so any deeper lake (e.g., depth of 20 m) likely behaves in a qualitatively similar manner. We speculate that the variability of the heat capacity during the year may cause a less evident minimum in the RMSE curves because the optimal window size should change with the season.

Figure 5

A more sophisticated way to denoise and smooth the input signal is to use Wavelet Transform (WT), which fits a mother wavelet in limited periods to overcome the problems of Fourier transform of not being able to determine where and when a specific frequency occurs, as well as limitations of frequency scales caused by window length (Chun-Lin, 2010). WT are categorized into Continuous WT (CWT) and Discrete WT (DWT), which decomposes the data into specified level. In this study, the optimal mother wavelet is determined among 106 types for DWT and 21 types for CWT provided by the wavelet transform library available in Python. The DWT method decomposes the raw AT data to get the coefficients measuring the intensity of signal variations (Ranchin and Wald, 1993), and eventually leads to smoothed data. The level of decomposition of the input data was identified as $int[log_{10}(N)]$, where N is the length of the time series and the int function transfers the values to the nearest integer (Zhu et al., 2020).

The different pre-processing methods were applied to AT data, using again BPNN method as the reference ML approach. Table 5 reports the comparison, showing that, in the shallow lake, MAMM outperformed the other methods in the both the shallow and the deep lake, while WT provided the worst results. The most effective mother wavelets can change depending on the way the optimization is performed. We note that, within the same Min-Max approach, smoothing the AT gives a clear improvement with respect to the standard, 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 5

4.3. Retaining the history of the forcing

In standard ML approaches (except for LSTM), such as the BPNN method that we assume as a reference, the history of the system is not retained. Thus, a pre-processing of the input data can be useful, as demonstrated in the previous section. However, retaining the effect of previous AT by a moving average over an optimal period is a rough way to do it because it equally weighs the information from any day in the averaging window, while the last days are likely to have a stronger impact. There are various ways to overcome this deficiency, for instance by arbitrarily assigning different weights depending on the temporal distance from the simulated day. In this section, however, we explore an alternative strategy based on the explicit inclusion of previous days' inputs in the ML model.

We keep using BPNN as a reference ML algorithm and SCDOY and AT as the most relevant input data. However, while the use of SCDOY is unchanged, the AT record is repeatedly duplicated with a shift of a variable number N of days: AT in phase with LSWT, AT of the previous day, AT of two days before, etc. Thus, the number of input nodes of the BPNN is increased by $N - 1$. In this way, the previous forcing conditions (assuming that AT is a proxy) are explicitly included in the ML algorithm, which will weigh them depending on the effective relevance for LSWT. The performances of the BPNN algorithm increase dramatically (Fig. 6) by adding a number of previous days that varies with the case study, until the improvement is saturated: adding further information becomes irrelevant. Optimal performances are reached by considering approximately 15 days for the shallow lake (Fig. 6a) and about 20 days for the deep case (Fig. 6b), thus confirming that the memory of the system is relatively long and slightly increases with the depth. The last row in Table 5 shows that using AT from previous days improves the results considerably (0.96-0.94°C for shallow lake and 0.88-1.14°C for deep lakes).

There is some similarity in using the moving average and the single values in previous days as additional inputs, and this is the reason we included both in Table 5. However, it is evident that the latter approach is not a pre-processing method as the ones discussed in section 4.2.

Figure 6

Figure 7 confirms that adding AT from previous days as additional inputs can improve the model's performance more than applying a moving average to the AT. It is interesting to note that both methods provide better results than considering all the input variables together in the same day as LWST. The combination of SCDOY and of the history of AT is the most effective input to simulate LSWT in this type of lakes.

Figure 7

4.4. The choice of the ML algorithm

Based on our literature review (summarized in Table 1), the most utilized ML approach is ANN, but several other methods have been proposed. Here, we first examine the details of some ML algorithms, and then compare their performances for our synthetic case studies comprising a shallow and a deep lake.

We decided to test nine commonly used ML algorithms, also considering those used for the prediction of stream temperature: MLPNN, BPNN, LSTM, ANFIS, DT, RF, ERT, KNN and SVR (Table 6). More details about these methods are provided in the Appendix B. Trying to obtain a fair comparison among them, we set up a consistent procedure for the analysis: the same predictors were used (AT and SCDOY), the hyperparameters were optimized with the same algorithm (GA optimization), and the Min-Max pre-processing was applied to the input data without any smoothing. In the analysis of the performances, we examined the RMSE for the training and test data sets separately. We note that a low error for training and high for test implies that the model is overtrained, and does not predict other data sets properly.

Table 6

As shown in Table 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). The performances for the deep lake varied similarly: RMSE in the range 1.09-1.43°C for the training data set in most ML algorithms, again with very poor performances of ANFIS (3.74°C); and in the range 1.26-1.67°C for the test data set, always with unacceptable results from ANFIS (3.46°C). In both cases, the performances of RF deteriorated significantly in the test case, passing from the best results among the various ML algorithms for the training set to mediocre results for the test set (RMSE deteriorating of 0.35°C and 0.25°C for the shallow and deep lake, respectively). The fact that LSTM, which retains information from the previous state, works well for the deep case, and for both the training and test data set (the difference is only 0.02°C), is consistent with the fact that lakes keep the memory of previous days' conditions. This implies that the results are good in this case also without the need of a heavy pre-processing of the data that includes the smoothing of AT, or the consideration of AT values in previous days.

It is instructive to interpret the performances of the ML methods based on their specific features. We start considering the performances of two similar methods: ERT and RF. ERT is a modified version of the DT that randomly selects the parameters of the model and repeatedly train the model using the data. RF splits nodes based on optimum decision, while ERT splits the nodes randomly. Additionally, in the RF method, the inputs are divided into subsamples, and drawn with replacement each time (ending in repeated data), while in ERT the inputs are not drawn with replacement, avoiding the repetition of data (Geurts et al., 2006). Results in Table 6 show that ERT is a more robust approach because the deterioration of the performance from training to test is lower than for RF.

The results suggest that LSTM, MLPNN and BPNN are all successful methods for prediction of LSWT. In fact, these methods present the minimum difference of RMSE between the training and test data set among the various types of ML algorithms (Table 6). They share the property of being branches of Deep Learning (DL), which suggests that DL is able to extrapolate good performances to the test data set more effectively than other ML methods. DL methods are built with multiple hidden layers (layers between the input and output layer) and this introduces more degrees of freedom in the combination of weights and biases (the parameters that are calibrated in the training phase), leading to more accurate results. On the other hand, ANFIS exploits the structure of ANN, but consists of a single hidden layer of neural network, and other layers are governed by fuzzy logic (more details in Appendix B). This combination of layers makes ANFIS less accurate in comparison to the other ML methods tested in the paper.

Figure 8

The different behaviour of the shallow and deep lake can be detected by the intensity of LSWT fluctuations. The shallow lake admits a large variability on a short time scale, while fluctuations tend to be smaller in the deep lake (an example of the different behaviour is graphically illustrated in Fig. 8). However, the damped variability in the deep lake is limited to the periods when the lake is not stratified, and in summer LSWT fluctuates more because of the smaller surface volume reacting to the heat flux. Reproducing such a behaviour is not trivial for a ML method. We appreciate that DL algorithms are able to follow the target (LSWT simulated by GLM) more than other ML methods. The worst method, ANFIS, is characterized by less accurate results exactly because of the presence of large and unrealistic fluctuations.

4.5. The benchmark of a physically based data-driven model

In order to have a benchmark against which to evaluate the performances of ML, the results were compared the results obtained by the physically based, data-driven model *air2water*. The input data for the 6-parameter version of this model (Toffolon et al., 2014) are AT and DOY. The model is based on the integration of a differential equation where the time derivative of LSWT is involved, which allows for retaining the effect of the thermal inertia of the lake. By calibrating the parameters of *air2water*, we obtained a RMSE of 1.098°C and 1.078°C (training and test, respectively) for the shallow case, and 0.946°C and 1.021°C for the deep case. Therefore, *air2water* obtains a performance that is comparable with the best ML models (Table 6), and is only slightly worse than the BPNN with AT from the previous days (Table 5).

It is interesting to note that the memory of previous conditions is retained in *air2water* by means of the time derivatives of LSWT, in a way that is similar to stochastic autoregressive algorithms. Moreover, *air2water* contains a parameter that change the thermal inertia of the systems on the basis of a simplified description of the seasonal stratification (Toffolon et al., 2014). LSTM follows the same strategy by adopting a complex structure of the ANN with internal loops. Conversely, the inclusion of previous days' AT values, as in our modified BPNN application (Fig. 6g,h), includes the history of the forcing explicitly, and not that of the state of the physical system (previous day's LSWT).

5. Discussion

ML is a successful approach for predicting LSWT, and it does not need the broad variety of data (for instance, about lake morphology, meteorological forcing, heat fluxes) that are required in physically-based models. Nevertheless, the choice of the correct predictors is a crucial step for the success of this approach. We show that, in our artificial case studies, SCDOY and AT are the most influential inputs in order to forecast the LSWT; other combinations may slightly improve the accuracy, but demanding more data. The role of AT is not surprising, as it affects several components of the heat flux and is a proxy for the meteorological conditions. More interesting is analysing the influence of DOY (or SCDOY, equivalently). First, it allows to reproduce the climatological mean year, i.e. the annual cycle of LSWT that is approximately repeated every year (for a formal definition see, e.g., Toffolon et al., 2020). The range of intra-annual variability between the minimum and maximum LSWT is very large in temperate lakes, and the inter-annual variability is often a minor addition to the global RMSE (in this respect, see the use of the modified Nash-Sutcliff efficiency in Piccolroaz et al., 2016). Second, and less obvious contribution, the DOY allows the model to include the seasonal variability of stratification, which has a fundamental effect in lakes that are deep enough not to be always or periodically well-mixed.

Important factors affecting the thermal response of a lake are the morphological and hydrological features. We did not perform a comparative analysis considering a large number of lakes, as other authors did; for this reason, these lake features were not assumed as predictors. We restricted our analysis to the effect of the lake's depth focussing on two cases (shallow, 5 m, and deep, 60 m). Our results suggest that a smoother behaviour in deep lakes (due to the larger heat capacity) may lead to lower RMSE values than in shallow lakes, although deeper lakes' dynamics are in general more complex due to the nonlinear effect of stratification on LSWT.

It is well-known that optimization and pre-processing methods can effectively improve the performance of ML methods (e.g., Isik et al., 2012). We found that a moving average of the input AT was sufficient to smooth the high-frequency LSWT response, similar to the physical filtering effect of the large water mass in a lake. The other possibility to smooth the LSWT oscillations is to reduce the influence of the large day-to-day AT variability by including also values in previous days.

In our modelling exercise, we did not find clear indications about a single best ML approach. A group of methods performed reasonably well (depending also on the shallow/deep case) and only ANFIS was characterized by unacceptable errors. The lack of indications about a prevailing approach is confirmed by our literature review. We found that ANN, with its variations, represents the most used model and has become a sort of standard for LSWT predictions (Table 3). However, traditional methods sometimes performed better than advanced algorithms: for instance, Sharma et al. (2008) showed that the multiple regression algorithm, considering linear relationships, can be more effective than ANN. Moreover, SVR using GA outperformed the ANN in an application to reservoirs (Quan et al., 2020). Heddam et al. (2020) found ERT more robust than ANN and RF for the prediction of LSWT in shallow lakes, as well as demanding less computation time compared to ANN. Nevertheless, according to their study, a physically based data-driven model (*air2water*) outperformed all the ML methods, an outcome that our analysis seems to confirm. Liu and Chen (2012) studied the accuracy of ANN in comparison to a 3D hybrid model in the shallow Yuan-Yang lake (maximum depth 4.5 m), finding that ANN forecasts the water temperature at 1, 2 and 3 m below the surface not as accurately as LSWT. Besides, Samadifard et al. (2016) investigated ANN, gene expression programming (GEP) and ANFIS in the Yuan-Yang lake. Similar to our results, they showed that, generally, ANN is more robust than ANFIS. It is worth mentioning that ANN is capable of forecasting more precisely, but obtaining high performances requires a careful choice of the optimization method, and a higher number of hidden layers and neurons. Moreover, different branches of ANN are available such as MLPNN, RNN and BPNN.

Physical constraints help ML obtain more accurate predictions. In our analysis, we worked on the pre-processing of AT, including the information from previous days, to mimic the effect that the history of the forcing has in combination with the memory of the system. Other scholars included constraints directly in their model: for instance, Jia et al. (2019) added two physical constraints to their LSTM model simulating the temperature profile along the water column in Lake Mendota, the requirement that density increases with depth and the conservation of energy. Moreover, Jia et al. (2020) pretrained their LSTM model by means of a physical-based model in the case of data deficiency to get more accurate results, a procedure that transfers ML towards a hybrid model.

ML methods generally perform well with respect to pure statistical computations such as linear and nonlinear regressions. However, physically based data-driven models like *air2water* are able perform better than almost all the ML method we investigated, because they retain some physics while being flexible in adapting to the different case studies thanks to the specific calibration of the model's parameter. Thus, why using a ML approach? We feel that the flexibility of ML may represent a striking advantage when the physics of the system is not clear enough, or simplified models are not available for a specific class of problems. This is the case of global-scale analyses where different types of lakes are considered with the same approach: for instance, Piccolroaz et al. (2020) demonstrated that the performances of *air2water* are good for temperate lakes but worsen for tropical and equatorial lakes, where other factors may become as relevant as AT. A global analysis of LSWT in lakes by means of ML methods deserves some attention.

6. Conclusions

Water temperature of freshwater bodies can be predicted using different methods, among which machine learning (ML) algorithms are becoming one of the most popular approaches. In this study, we reviewed the most common ML approaches to predict lake surface water temperature (LSWT) and tried to disentangle the effect of the main factors affecting their performances. In particular, we analysed (i) the effect of considering different predictors, i.e. the variables used as input data; (ii) the improvement that can be obtained by including the effect of the history of the forcing and the memory the system on the thermal response, for instance by smoothing the air temperature input or considering previous days' values, thus mimicking the filtering effect that lakes can exert thanks to their capability to store heat for a relatively long time; and, eventually, (iii) the performances of different ML approaches. Such a comparative exercise was based on a literature review, but more importantly on the application of nine ML methods on two specifically designed case studies: a shallow lake (maximum depth 5 m) and a deep lake (maximum depth 60 m), for which the LSWT was artificially obtained by means of a physically based one-dimensional model (GLM). The use of a synthetic case allowed us to control all the variables involved in the problem and to make a comparison between the shallow and deep cases without any other factor altering the thermal response. In this respect, it should be considered that all results are based on a synthetic case and are not representative of real lakes.

The results of our analysis, which was performed using BPNN as the reference ML approach, show that a combination of air temperature (AT) and day of the year (DOY), in particular adopting a continuous description of the latter variable through the use of trigonometric functions (SCDOY), provides the minimum amount of information required to obtain reasonable results, and that adding further variables does not allow for significant improvements. Moreover, we demonstrated that smoothing the input signal for AT, for instance using a moving average, yields a noticeable improvement of the performances. As a more complicated but even more fruitful approach, previous days' AT values can be included as inputs of the ML to retain the history of the forcing conditions. This evidence confirms the need to include a memory effect in those ML approaches that cannot intrinsically consider it (the only option we tested was Long Short-Term Memory, LSTM). Finally, the comparison of nine different ML algorithms did not show striking differences in the performances, although the LSTM method seems to be slightly more robust than other approaches and ANFIS was constantly providing unacceptable results. Hence, our analysis suggests that the success of the application of ML to predict LSWT depends mostly on the identification of the correct predictors, while there is not a definitive choice of the best ML algorithm.

Acknowledgments

The authors thank Valentina Lot for her contribution to the ideas of this paper developed in a preliminary analysis on the same topic and for the suggestion to use SCDOY instead of DOY, and Hossein Amini, Marina Amadori and Sebastiano Piccolroaz for their thoughtful comments on a draft of the manuscript. The constructive criticism of four anonymous reviewers was helpful to improve the clarity of this paper.

Appendix A – Minimal model of the thermal response of a lake

In this appendix, we revisit the heat budget in a lake by initially referring to a simplified one-dimensional (1D) description,

$$\frac{\partial}{\partial t}\left(\rho\, c_p\, S\, T\right) + \frac{\partial}{\partial z}\left(S\, \phi + I\right) = 0 \tag{A1}$$

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. Eq. (A1) 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 Eq. (A1) to understand the behaviour of LSWT, assuming that it is correlated with the average temperature in the surface layer. By integrating Eq. (A1) over the thermally reactive volume V_r [m^3] at the surface, the volume-averaged temperature $\bar{T}(t)$ follows the equation

$$\frac{d}{dt}\left(\rho\, c_p\, V_r\, \bar{T}\right) = S_0\, \phi_{net} + S_D\, \phi_D + \sum_{j=1}^{N_I} \rho\, c_p\, Q_{I,j}\, T_{I,j} - \sum_{j=1}^{N_O} \rho\, c_p\, Q_{O,j}\, T_{O,j} \tag{A2}$$

where S_0 is the surface area and ϕ_{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 ϕ_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 ϕ_{net} is the combination of multiple factors:

$$\phi_{net} = H_{SW} + H_{LW\_in} - H_{LW_{out}} \pm H_L \pm H_C + H_P \tag{A3}$$

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 ϕ_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 Eq. (A2):

$$\alpha \frac{dT}{dt} = \beta \sin(\omega t) \tag{A4}$$

where α 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, β is the amplitude of the sinusoidal variability associated with a given frequency ω: 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 Eq. (A4) is straightforward:

$$T = -\frac{\beta}{\alpha\,\omega}\cos(\omega t) + c \tag{A5}$$

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 α (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.

Appendix B – Review of Machine Learning algorithms

In this section, we review the main features of the ML algorithms that we analysed in this paper. The specification of the values of the hyperparameters optimized for our application is provided in the Supplementary Material.

- Artificial Neural Network (ANN) and derived algorithms

ANN algorithms consisting of straightforward processors and their interconnections are complex systems with organizational principles that parallelly recognize the pattern, predict and optimize the weights and biases (Jain et al., 1996). The network algorithm is composed of interconnected nodes (called neurons) structured into parallel layers (Fig. B1). The first (input) layer receives the input data and, through a variable number of internal (hidden) layers, the information percolates until reaching the last (output) layer, which provides the results of the simulation. The directional links among nodes of adjacent layers transfer signals from the upstream to the downstream neurons. In each of the hidden and input layers, an additional parameter is provided to the neurons by the so-called biases. The signal exiting from one neuron is suitably weighed before entering into the subsequent one, where it is summed to the signals from the other neurons and to the bias of the same layer; then, such a sum feeds an activation function, whose output represents the new signal that is transferred downstream.

Figure B1

The number of hidden layers and the number of neurons for each layer are hyperparameters of the ANN, together with the type of activation function and the parameters of the optimization procedure for the calibration of weights and biases. The procedure for the optimization characterizes the branch of ANN. In general terms, it begins with random values of weights and biases. In the first run, the ANN provides an output that is compared with the target. Then, weights and biases are updated using gradient descent method. The iteration continues until the difference between the output and the target is minimized.

In our analyses, we considered three different types of ANN: Multi-Layer Perceptron Neural Network (MLPNN), Back-Propagation Neural Network (BPNN) and Long Short-Term Memory (LSTM) as a type of Recurrent Neural Network (RNN); and Adaptative Neuro-Fuzzy Inference System (ANFIS) which consists of one layer of neural network. These methods are further analysed hereafter.

(1) The MLPNN is an ANN consisting of more than one hidden layer. In this paper, we use this term to refer to a feed forward neural network. Its peculiarity is updating the weights based on error from input to output in just one direction.

(2) The BPNN approach is characterized by the algorithm used to update weights and biases in order to minimize the error (Li et al., 2012). First, the value of each neuron is calculated based on the initial weights. Then, the weighs are adjusted from the output layer backwards toward the input layer, knowing how much each node contributes to the error of prediction from observed data (Hecht-Nielsen, 1992). IN BPNN, the weights are updated from the output layer backwards to the input layer while in MLPNN, the weights are updated based on the error and move just forward (Dudek, 2019).

(3) The LSTM method is a type of RNN, which is a branch of ANN that has a sequential connection between time steps (Sainath et al., 2015). LSTM keeps the information for a longer period

in comparison to the standard RNN. In the hidden layers, memory blocks are special units consisting of memory cells, storing and controlling the passing of information by three gates (Fig. B2). The gates are: forget gate (for keeping or removing the information), input gate (adding information to the cell) and output gate (selecting the necessary information). The hidden state, which contains the memory of neural network, keeps the previous information (Gonzalez-Dominguez et al., 2014).

Figure B2

The ANN requires low statistical training, and identifies complex non-linear connections of dependent and independent variables and also between predictor variables, as well as being a multiple training algorithm. Liu and Chen (2012) suggested that ANN is a cost-effective approach with respect to a three-dimensional semi-implicit Eulerian-Lagrangian finite-element method (Zhang and Baptisa, 2008), and could be used for water quality prediction and ecological management. Despite the mentioned advantages, the ANN is a black box that requires computational effort and is susceptible to overfitting (Tu, 1996). The overfitting could be controlled by dropout (Piotrowski et al., 2020; Zhu and Piotrowski, 2020), whereby the nodes with their connections are randomly eliminated in the training phase (Srivastava et al., 2014).

- ANFIS

The ANFIS is a subcategory of ANN based on Takagi-Sugeno-type fuzzy inference system (Anikin and Zinoviev, 2015). The method based on the fuzzy logic principles train the model set to calculate the parameters of the membership function, which represents the degree of truth in fuzzy logic (Opeyemi, and Justice, 2012). Three methods of ANFIS are widely used: (i) ANFIS with grid partition method called ANFIS_G, (ii) ANFIS with subtractive clustering (SC) called ANFIS_S, and (iii) ANFIS with fuzzy c-means clustering (FCM) called ANFIS_F. The difference between the three algorithms is mainly related to the optimization of the fuzzy rules. The ANFIS_G needs a large number of fuzzy rules, while the two other ANFIS_S and ANFIS_F were adopted for providing a model with fewer fuzzy rules. For example, ANFIS_G cannot be applied using more than six inputs variables. We applied ANFIS_G considering its rules and limitations. Based on a first-order Sugeno model, there are two if-then rules for fuzzy logic. The network is composed by 5 layers (Fig. B3). Layer 1, named fuzzification layer, consisting of adaptive nodes, determining the degree of fuzzy membership inputs. In our study, the triangular membership function gets more accurate results than other functions. In layer 2, the signals connected to it are multiplied (w_i), and the output of this node is called the firing strength of the rule. In layer 3, node N normalizes firing strength ($\overline{w_i}$). In layer 4, the output of the adaptive nodes is the product of the results of the previous layer and a first order polynomial (f_i) due to first order Sugeno model. In layer 5, the summation of all connected signals is calculated (Opeyemi and Justice, 2012).

Figure B3

- Decision Tree (DT) and derived algorithms

DT is used to differentiate and predict in classification or regression cases through dividing, which leads to discovered features and pattern extraction. Regarding these features, DTs are widely used for prediction models and investigative data analysis (Myles al., 2004). As shown in Fig. B4, internal nodes and root nodes are filtered down along the tree to gain proper outputs, named leaf nodes, as patterns' categories (Navada et al., 2011). We tested (1) the standard DT approach and two derived algorithms, (2) Extremely Randomized Tree (ERT) and (3) Random forest (RF).

Figure B4

(1) The standard DT is drawn from root, upside, to the leaf nodes, where the nodes do not split anymore (Fig. B4). The root and internal nodes are split into branches ended to nodes, called child node. In regression cases, the attribute or feature to split is reducing the variance of each split considering the weighted average variance of child nodes (Sharma, 2020).

(2) The ERT randomly selects the features and cut point (threshold value of features) to split (Geurts et al. 2006), and trains repeatedly. The model is constructed by multiple trees using all data (Alswaina and Elleithy, 2018; Heddam et al., 2020). The ERT superiority over the DT is because of preventing overfitting, efficient computation, random selection and cut point values and more accurate prediction (Galelli and Castelletti 2013). Heddam et al. (2020) found this method more robust than ANN and RF in order to predict the water temperature of shallow lakes.

(3) The RF is in both groups of classification and regression trees implemented by selecting features randomly and training the features with bootstrap sampling, random samples with replacement (Fig. B5). The majority decision of trees for the final decision of RF is considered. RF's advantage is the accurate prediction and the combination of favourable features, evaluation of built-in performance, grading the significance of features and the relative weights due to composed similarity (Svetnik et al., 2003). In order to steadily approximates the significance of features, many trees should be utilized. Although the measurement of features significance is different in each run, their rank of importance is approximately steady (Liaw and Wiener, 2002).

Figure B5

- Support Vector Machines (SVM)

SVM is a ML approach for classification and regression analysis that for regression is called Support Vector Regression (SVR). The advantages of SVM are the good performance in the case of a small number of data, low computational effort and robustness in the case of the existence of outliers in the model (Steinwart and Christmann, 2008). Fig. B6 shows a classifier for hyperplane division, which should be optimized not only for the training case but also for any other test data (Gunn, 1998). A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset that divides the space into two disconnected parts (Veling et al, 2019). SVR minimizes the offsets between features and hyperplane, while SVM maximizes the separation distances between features (Quan et al., 2020). In Fig. B6, the circles and triangles are the features that are divided by hyperplane with best margin. Since SVR only depends on a subset of the training data, and is independent of the dimensions of input space, it is beneficial to utilize it for high dimensionality space inputs (Drucker et al., 1997). The hyperparameters of the kernel function (mapping the nonlinearity in regression algorithms in order to optimize the location of hyperplane) and the penalty could be optimized using optimization algorithms such as a GA to have more accurate results (Quan et al., 2020).

Figure B6

- K-Nearest Neighbor (KNN)

The KNN algorithm is utilized in the case of lack of prior knowledge of the data distribution (Peterson, 2009). The classification is defined as giving similarity or distance to samples (Keller et al., 1985). After detecting the nearest neighbours between the training data (see Fig. B7), the candidates of the category are scored due to their similarities to K neighbours' category (each category is determined and separated by black circles in Fig. B7). The highest score represents the sample's category (Jiang et al., 2012).

Figure B7

References

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

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

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. doi:10.1109/SIBCON.2015.7146977

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. doi:10.1139/s06-067

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.

Chun-Lin, L., 2010. A tutorial of the wavelet transform. NTUEE, Taiwan.

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. doi:10.1016/j.is.2018.01.003

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. doi:10.1002/hyp.6607

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. doi: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. doi:10.1016/j.ins.2018.12.063

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

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. doi:10.5194/hess-17-2669-2013

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

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, doi:10.21437/Interspeech.2014-483

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. doi:10.1201/b10911-3

Hecht-Nielsen, R., 1992. Theory of the backpropagation neural network. In Neural networks for perception (65-93). *Academic Press*. doi: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. Available at: doi:10.1016/j.jhydrol.2020.125130.

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. doi:10.5194/gmd-12-473-2019

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. doi: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.

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

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. doi:10.1137/1.9781611975673.63

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. doi:10.1145/3447814

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. doi: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. doi: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. doi:10.1109/tsmc.1985.6313426.

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. doi:10.1007/978-3-642-30223-7_87

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

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. doi: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. doi:10.1023/a:1022119503144

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

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. doi: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. doi:10.1002/cem.873

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. doi:10.1109/ICSGRC.2011.5991826

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. doi: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. doi:10.1038/nature01833

Peterson, L.E., 2009. K-nearest neighbor. *Scholarpedia*, *4*(2), 1883. doi: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. doi: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, doi: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. doi:10.1002/lno.10626

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. doi:10.5194/hess-17-3323-2013

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. doi:10.1002/2014WR016555.

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. doi: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. doi:10.1016/j.ejrh.2021.100780

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. doi:10.1016/j.jhydrol.2021.126629

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. doi: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. doi:10.1016/j.earscirev.2019.103076

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.

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. doi:10.1007/s00521-020-04836-4

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

Ranchin, T. and Wald, L., 1993. The wavelet transform for the analysis of remotely sensed images. *International Journal of Remote Sensing*, 14(3), 615-619. doi:10.1080/01431169308904362

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. doi:10.1029/2019WR024922

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. doi:10.1002/lno.11390

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. doi:10.15446/esrj.v20n2.43199

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. doi: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., 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. doi:10.1111/j.1365-2427.2008.01943.x

Spearman, C., 1910. Correlation calculated from faulty data. *British journal of psychology, 3*(3), 271. doi: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*. doi:10.1007/978-0-387-77242-4

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. doi: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. doi:10.1021/ci034160g

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, doi: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. doi:10.4319/lo.2014.59.6.2185

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. doi:10.1016/S0895-4356(96)00002-9

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. doi:10.22214/ijraset.2019.4624

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. doi: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. doi:10.1002/2017JD027661

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

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. doi:10.1016/j.ocemod.2007.11.005

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. doi:10.1007/s11600-020-00480-7.

Zhu, S., Ptak, M., Choiński, A. and Wu, S., 2020. 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), 7. doi:10.1007/s10661-019-7994-y

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

Figures

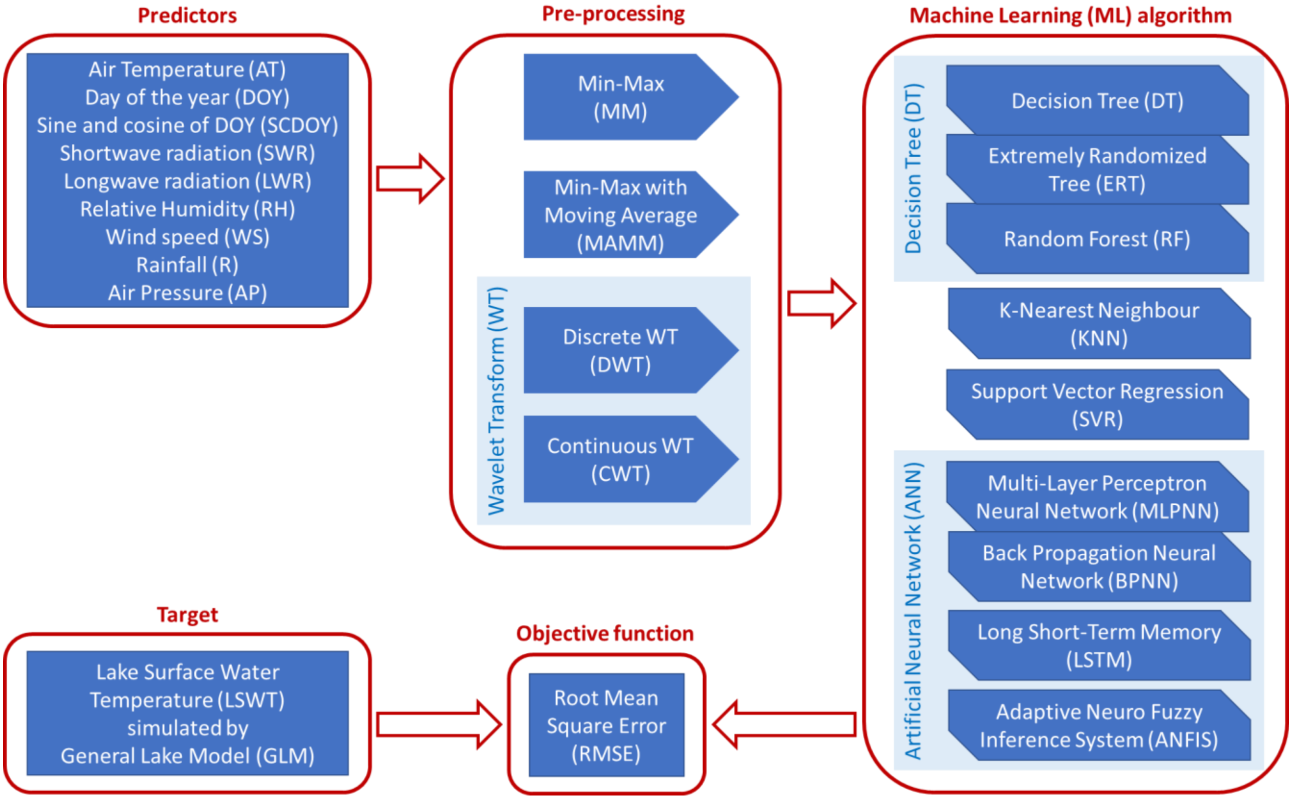

Figure 1. Workflow summarizing the steps of the comparative analysis of the performance of the different ML methods.

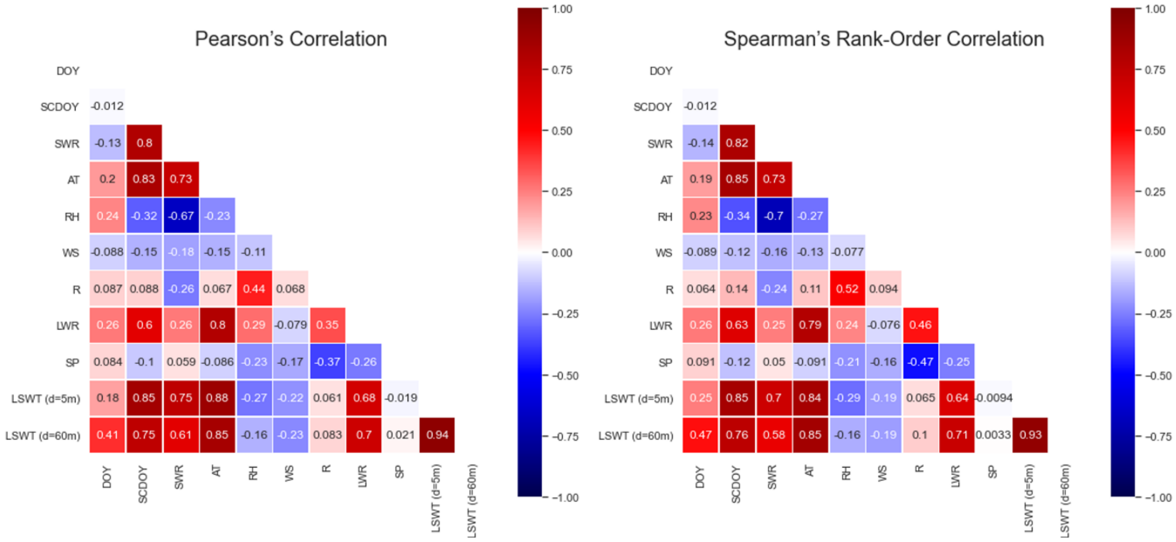

Figure 2. Correlation between variables: (a) Pearson's (linear) correlation; (b) Spearman's Rank-Order (non-linear) correlation.

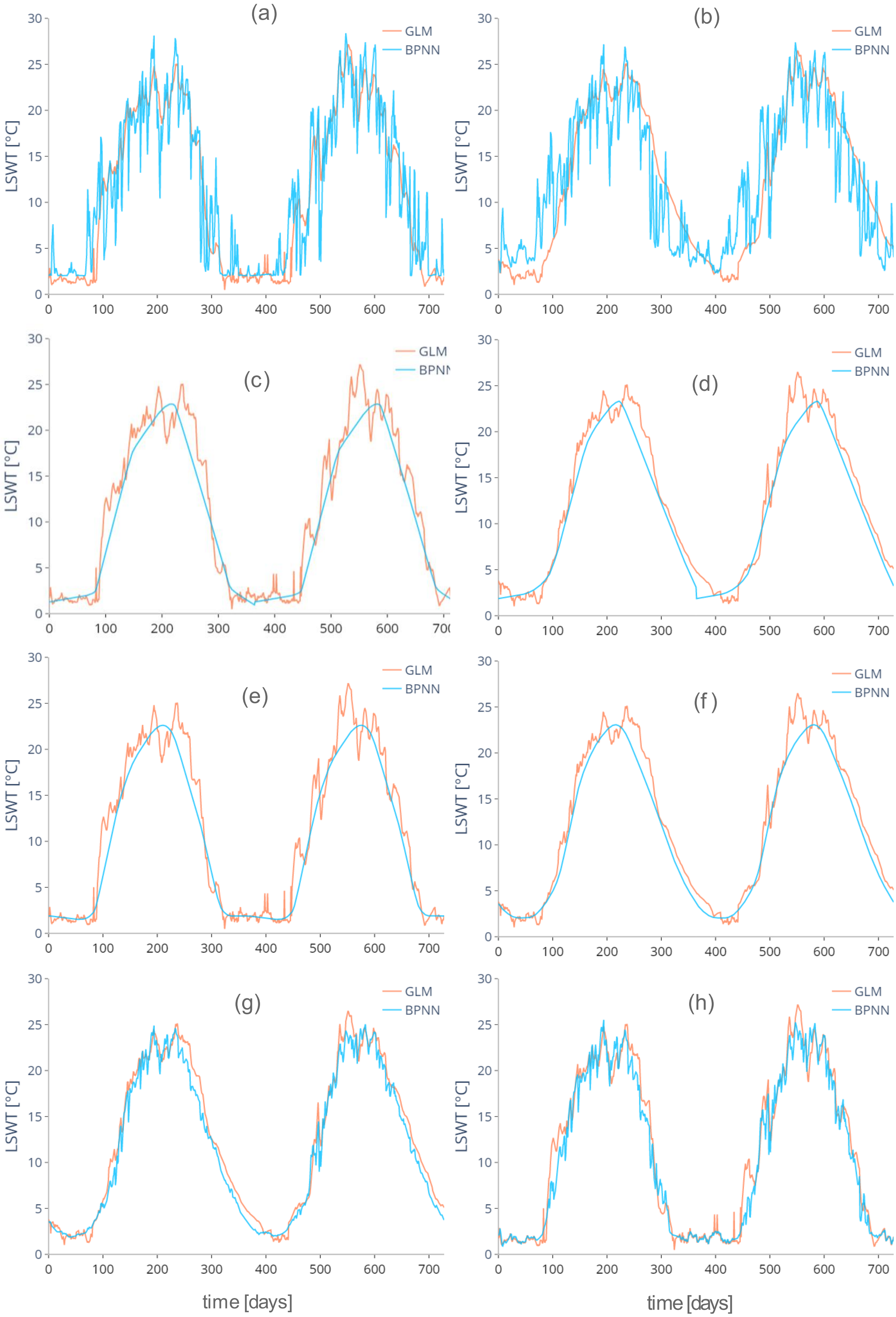

Figure 3. Effect of using AT (a,b), DOY (c,d), SCDOY (e,f), and the combinations of AT and SCDOY (g,h) on the LSWT modelled by the BPNN for a shallow (5 m, left column) and a deep (60 m, right column) lake.

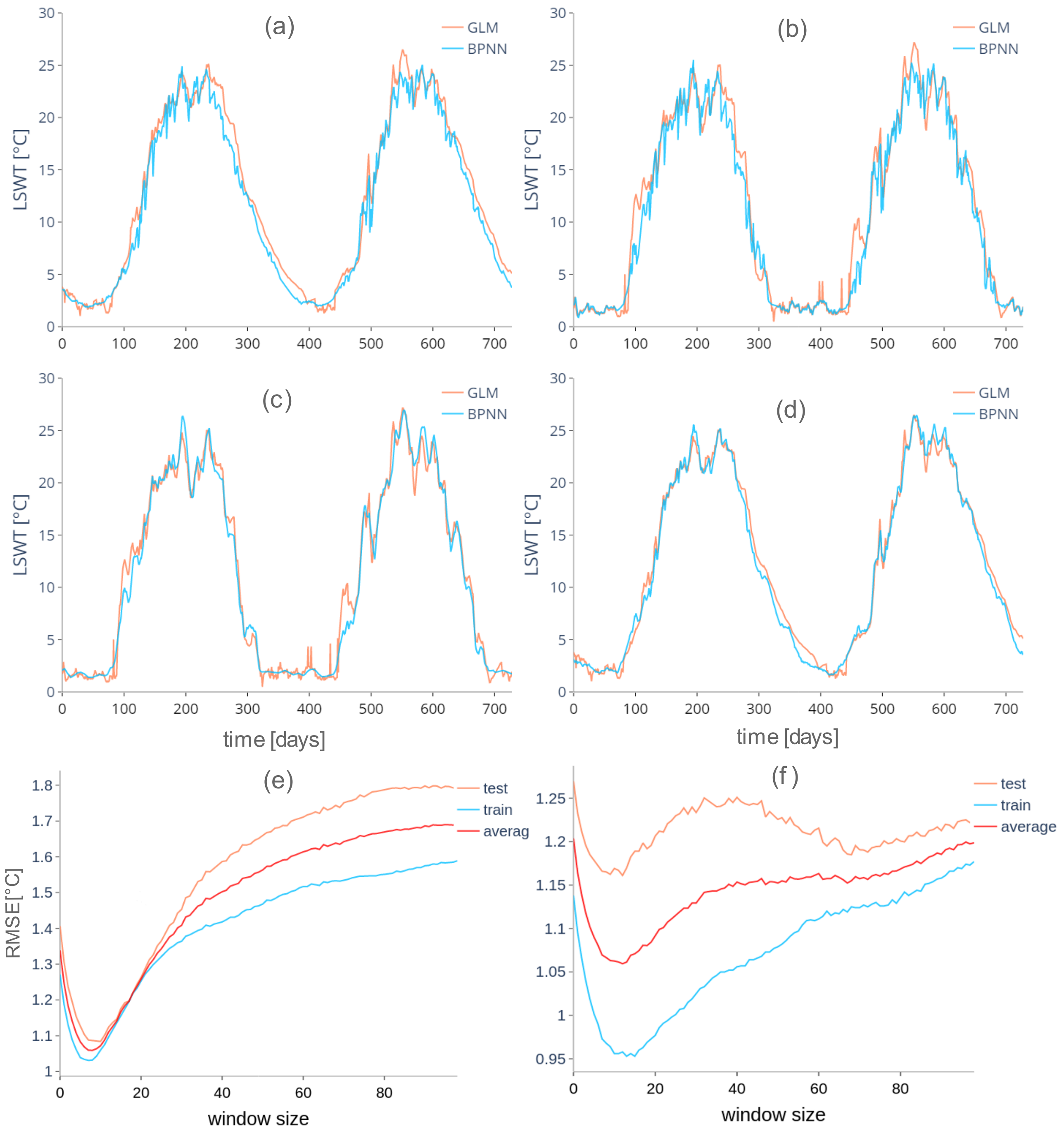

Figure 4. Effect of using a moving average for pre-processing the AT data (in addition to SCDOY of the simulated day) to predict LSWT by BPNN in a shallow (left column) and a deep (right column) lake. (a,b) Results in the standard BPNN application, with AT input only from the same day as the predicted 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.

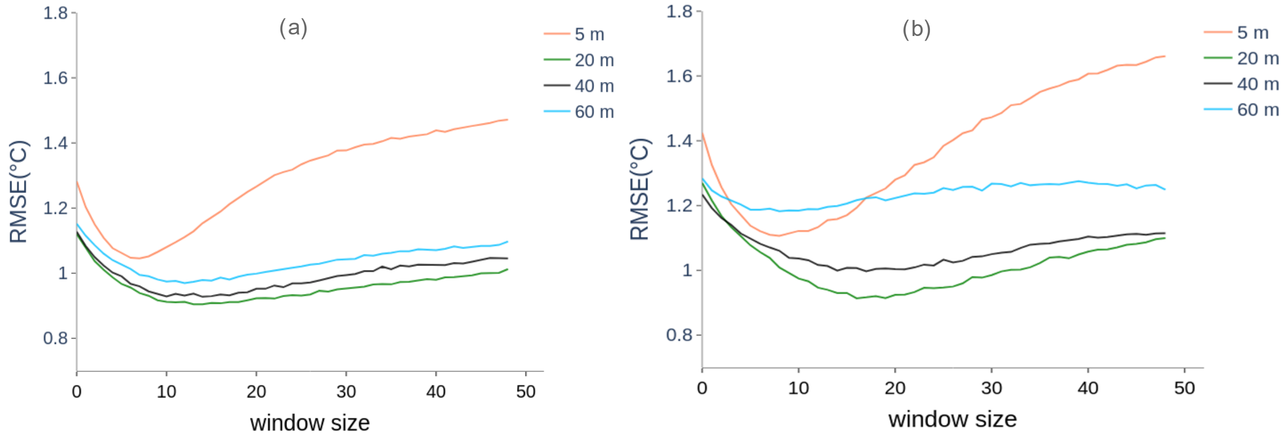

Figure 5. The effect of the depth on the window size of the moving average, using AT+SCDOY (ML method: BPNN) for the training (a) and test (b) dataset in different depths (5, 20, 40 and 60 m). All RMSE values are averaged over 20 independent runs.

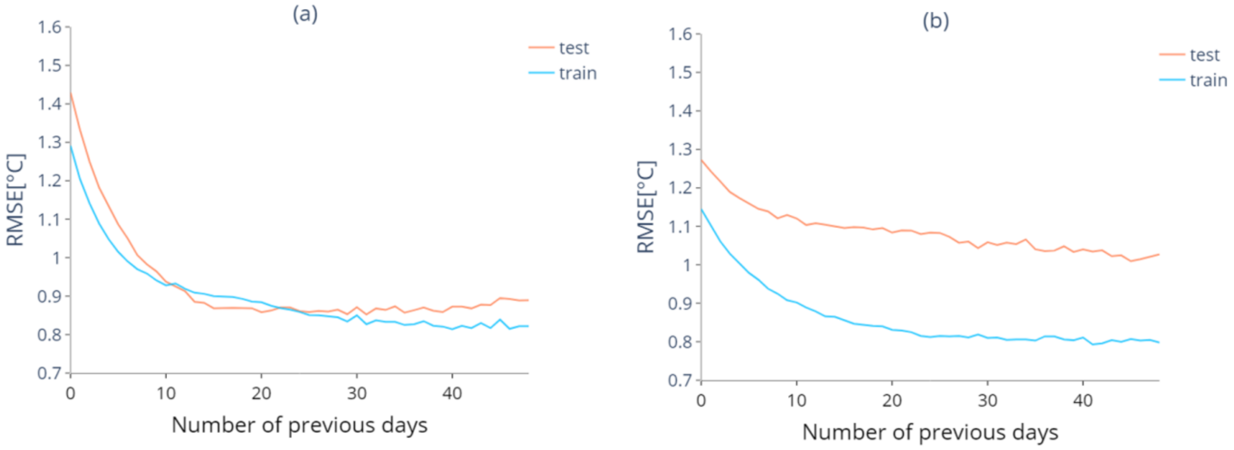

Figure 6. Variability of the RMSE using AT from a variable number of previous days in the shallow (a, 5 m) and deep (b, 60 m) lake. All RMSE values are averaged over 20 independent runs.

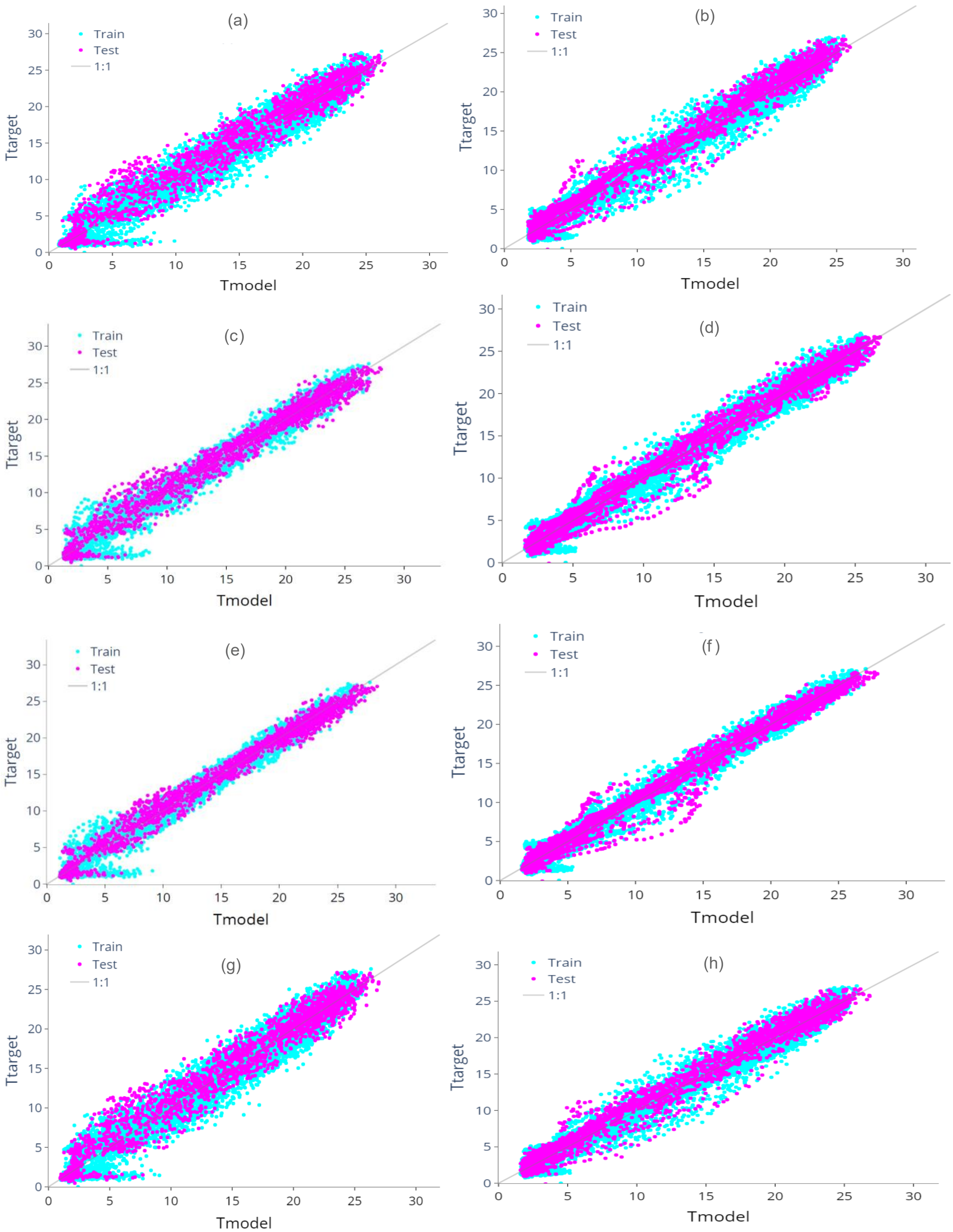

Figure 7. Scatter plots of observed and predicted LSWT using AT + SCDOY as predictors for the shallow lake (left column, subplots a, c, e, g) and the deep lake (right column, subplots b, d, f, h) for different ways to include AT in the inputs: (a, b) reference case (using AT and LSWT of the same day); (c, d) moving average of AT with windows of 10 (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.

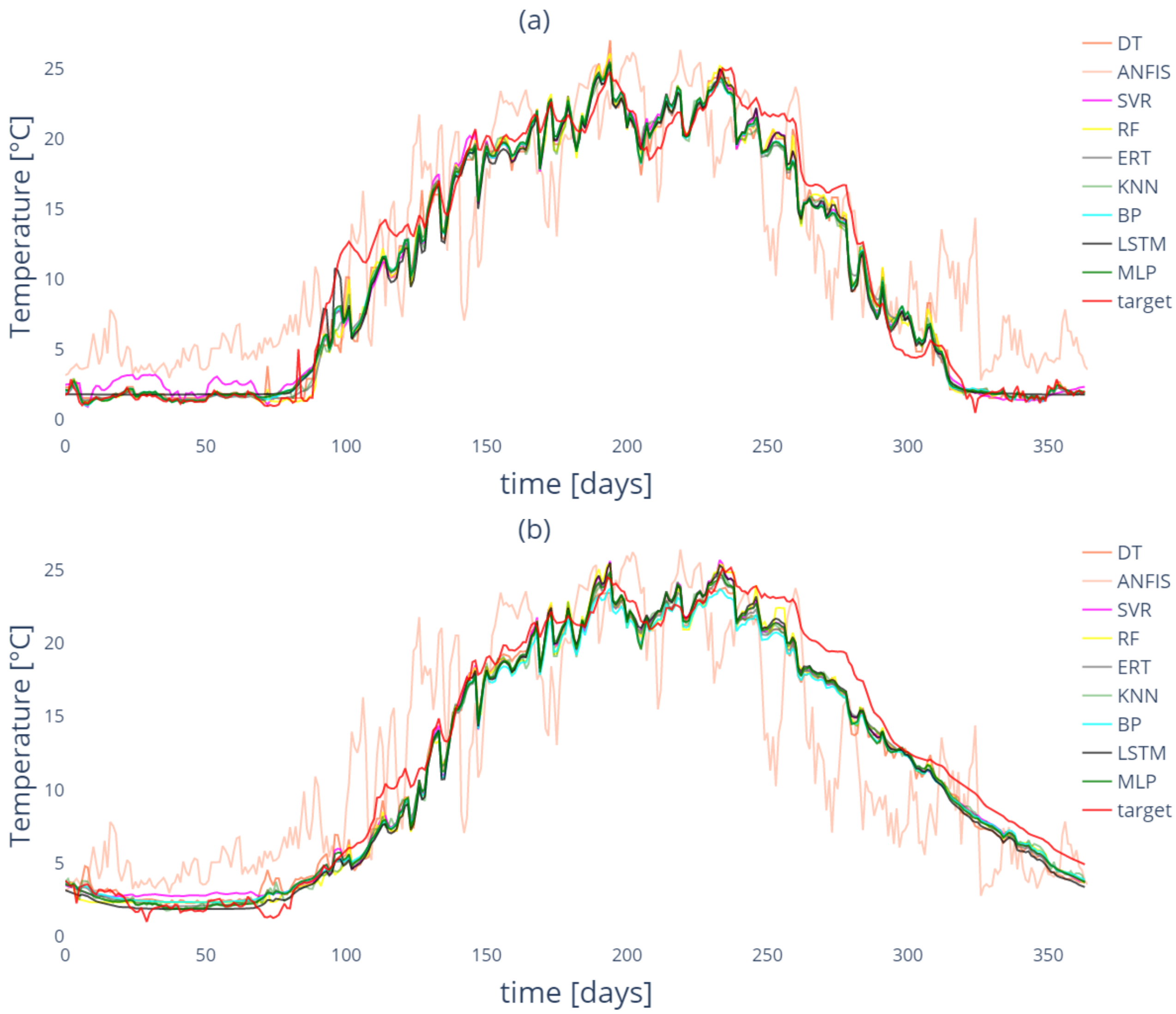

Figure 8. The evolution of LSWT in a single year (2015) as predicted by different ML approaches using AT+ SCDOY for the shallow (a) and deep (b) lake; the target is LSWT simulated by the physically based model GLM.

Figures in Appendix B

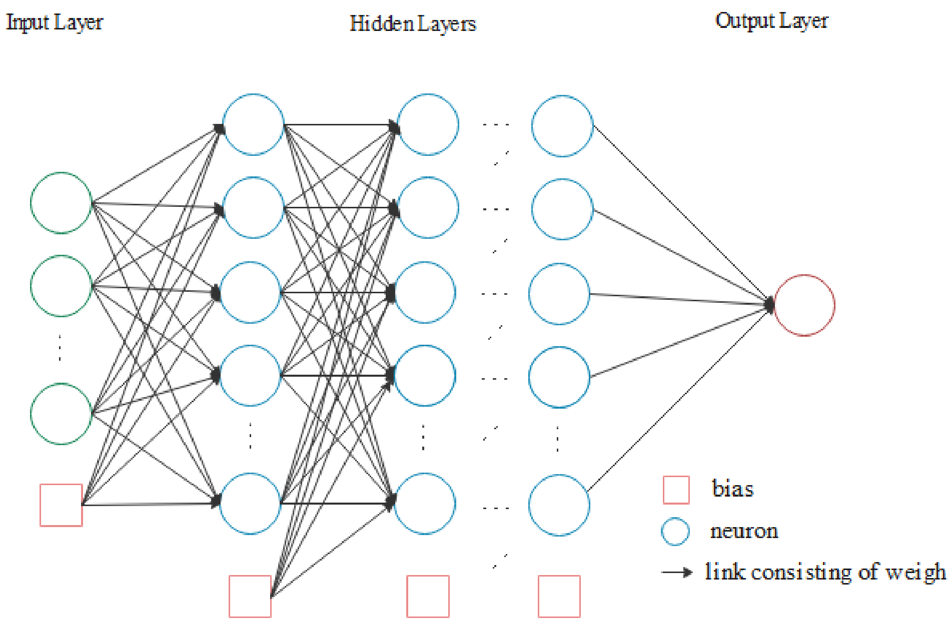

Figure B1. The general scheme of an Artificial Neural Network (ANN).

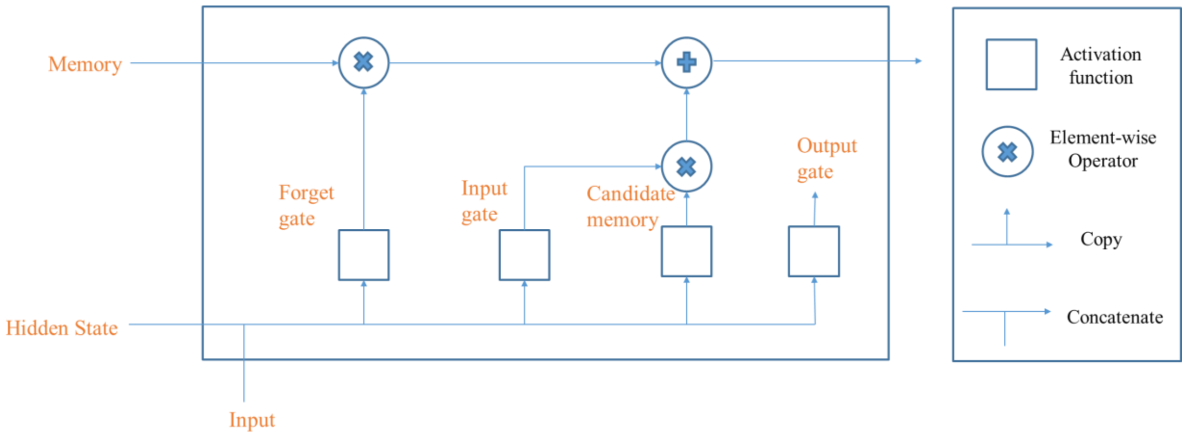

Figure B2. The scheme of Long Short-Term Memory (LSTM), a branch of ANN.

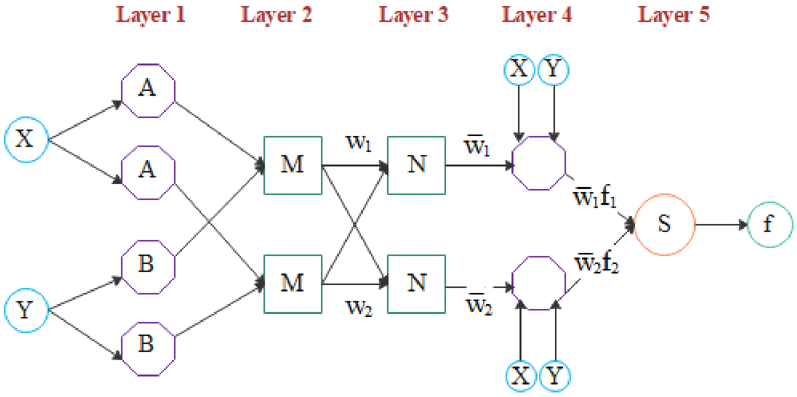

Figure B3. The scheme of Adaptative Neuro-Fuzzy Inference System (ANFIS), a branch of ANN that exploits fuzzy logic (adapted from Opeyemi and Justice, 2012).

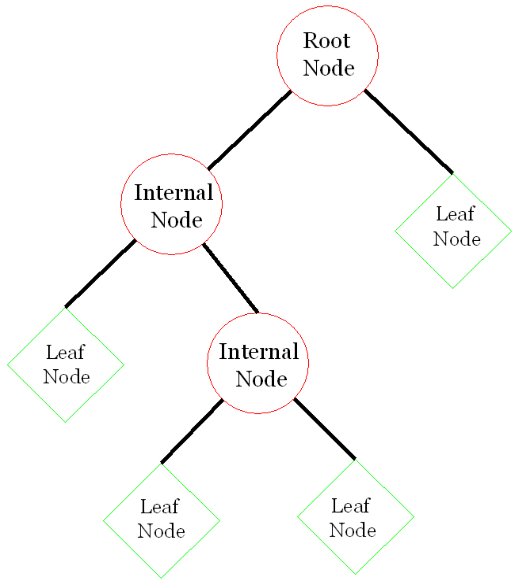

Figure B4. The scheme of Decision Tree (DT).

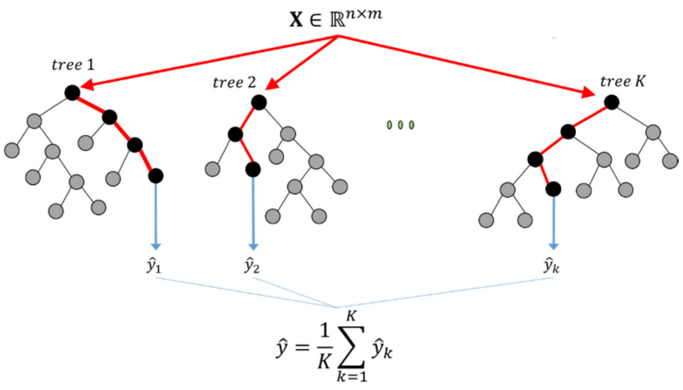

Figure B5. The scheme of Random Forest (RF) (Aldrich, 2020).

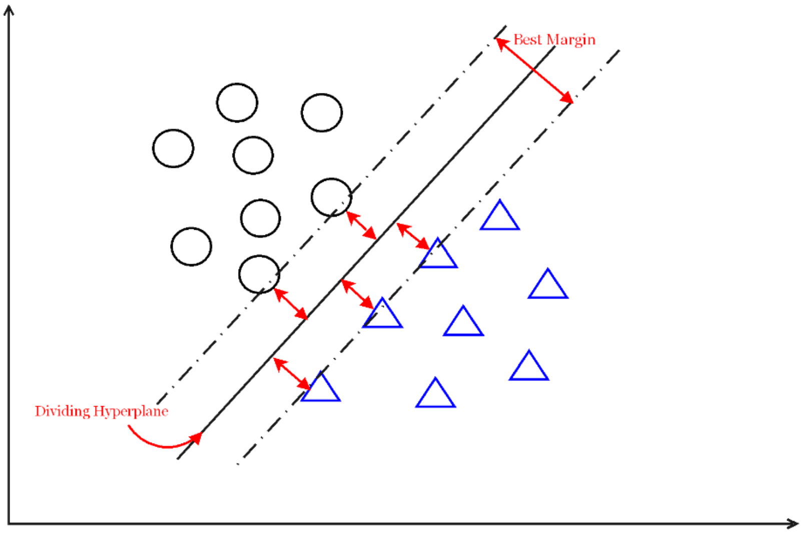

Figure B6. The scheme of Support Vector Machines (SVM). The circles and triangles are the features that need to be divided by a hyperplane with best margin.

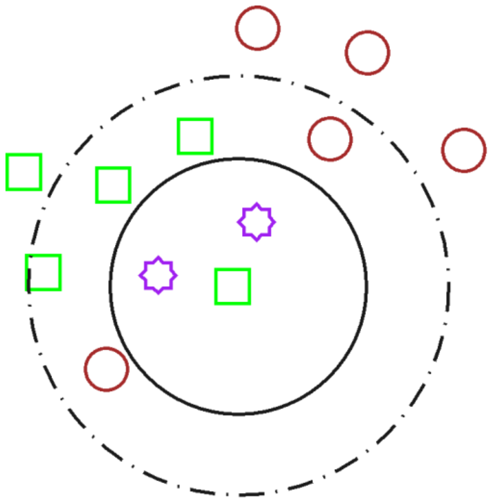

Figure B7. The scheme of KNN scheme. Small circles, squares and stars represent different categories.