Advanced Hybrid Machine Learning Algorithms for Multistep Lake Water Level Forecasting

Elham Fijani (✉ efijani@ut.ac.ir)

University of Tehran <https://orcid.org/0000-0001-9390-0849>

Khabat Khosravi

Ferdowsi University of Mashhad

Rahim Barzegar

McGill University

John Quilty

University of Waterloo

Jan Adamowski

McGill University

Homa Kheyrollah-Pour

Wilfrid Laurier University

Research Article

Keywords: Lake water level forecasting, Machine learning, Iterative Classifier Optimizer, Bagging, Additive Regression, Great Lakes.

Posted Date: September 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-912383/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Advanced hybrid machine learning algorithms for multistep lake water level forecasting

Elham Fijani^{1,*}, Khabat Khosravi², Rahim Barzegar^{3,4}, John Quilty⁵, Jan Adamowski³, Homa Kheyrollah-Pour⁴

1- School of Geology, College of Science, University of Tehran, Tehran, Iran.

2- Department of Watershed Management Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

3- Department of Bioresource Engineering, McGill University, Ste-Anne-de-Bellevue, Canada.

4- Department of Geography & Environmental Studies, Wilfrid Laurier University, Waterloo, Canada.

5- Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Canada.

*Corresponding author: efijani@ut.ac.ir

Abstract

Random Tree (RT) and Iterative Classifier Optimizer (ICO) based on Alternating Model Tree (AMT) regressor machine learning (ML) algorithms coupled with Bagging (BA) or Additive Regression (AR) hybrid algorithms were applied to forecasting multistep ahead (up to three months) Lake Superior and Lake Michigan water level (WL). Partial autocorrelation (PACF) of each lake's WL time series estimated the most important lag times — up to five months in both lakes — as potential inputs. The WL time series data was partitioned into training (from 1918 to 1988) and testing (from 1989 to 2018) for model building and evaluation, respectively. Developed algorithms were validated through statistically and visually based metric using testing data. Although both hybrid ensemble algorithms improved individual ML algorithms' performance, the BA algorithm outperformed the AR algorithm. As a novel model in forecasting problems, the ICO algorithm was shown to have great potential in generating robust multistep lake WL forecasts.

Keywords: Lake water level forecasting, Machine learning, Iterative Classifier Optimizer, Bagging, Additive Regression, Great Lakes.

1. Introduction

Lakes are valuable freshwater resources and their water level (WL) fluctuations are the dominant factor controlling the functioning of their ecosystems. As a lake's WL fluctuations are important to consider in a water budget, their accurate forecasting is essential to the sustainable management of water resources, flood controls, water supply, the health of ecosystems, sustainable development and management of water resources within a basin as well as economic development (Yadav and Eliza, 2017; Yan et al., 2020).

Since lake WL fluctuations typically respond nonlinearly to many natural and human induced changes (e.g., winds, reservoir management), nonlinear models are often adopted to accurately forecast lake WLs (Zhu et al., 2020a). Accordingly, in recent years, various types of numerical, statistical, and machine-learning-algorithm-based models have been applied to the modeling of lake WLs (Kebede et al., 2006; Huang et al., 2010; Khatibi et al., 2014; Zhu et al., 2020a). However, numerical (e.g., Process-Oriented Model (Fry et al., 2020), Routing Application for Parallel computation of Discharge (RAPID) model (Lin et al., 2015)) and statistical methods have limitations for lake WL forecasting: (i) besides being time-consuming, numerical models require extensive datasets including lake bathymetry, meteorological variables, inflow, outflow, etc. (Zhu et al., 2020b) and its calibration is also difficult, while (ii) statistical models (e.g. multiple linear regression) impose overly restrictive assumptions on the data — a linear structure and errors that belong to a particular parametric family. Popular and increasingly applied worldwide to forecasting lakes' WL, machine learning (ML) models have a general model structure, are non-parametric, and can estimate nonlinear relationships between input-output data due to its flexible and nonlinear structure. For instance, artificial neural networks (ANNs) (Altunkaynak, 2007; Coulibaly, 2010), fuzzy logic (FL) (Altunkaynak and Şen, 2007), adaptive neuro-fuzzy inference systems (ANFIS) (Yarar et al., 2009; Kisi et al., 2012), support vector machine (SVM) (Çimen and Kisi, 2009; Yadav and Eliza, 2017), extreme learning machines (ELM) (Shiri et al., 2016; Bonakdari et al., 2019), and Long Short-Term Memory (LSTM) deep learning (DL) (Zhu et al., 2020b) have been used for lake WL modeling.

Although standalone ML models have been successfully applied to hydrological modeling, creating hybrid models by (i) adding preprocessing methods such as wavelet transforms to denoise or capture multiscale changes in time series data (Quilty and Adamowski, 2018; Barzegar et al., 2016)), (ii) improving ML

models' structure through the implementation of evolutionary algorithms (*e.g.*, genetic algorithm, particle swarm optimization — Cheng et al., 2006; Taormina and Chau, 2015), or *(iii)* resampling the dataset to reduce bias and variance in model output using ensemble methods (*e.g.*, Bagging (BA), boosting, stacking and Additive Regression (AR) — Bui et al., 2016; Ribeiro and Coelho, 2020) has been shown to enhance predictive performance over standalone models. In the present study, BA and AR ensemble methods were employed to develop hybrid models. As the BA method employs a resampling with replacement process to generate different random training sets to obtain a unique learner (Graczyk et al., 2010; Tiwari and Chatterjee, 2011), it is well suited to hydrological modeling, particularly when training datasets are small (Barzegar et al., 2020). In contrast, AR is a type of boosting method where each iteration fits a new base model to the error residuals of the previous iteration. Prediction is then accomplished by summing up the predictions of each base model (Baskin et al., 2017).

As mentioned above, neural-, kernel-, fuzzy-, and DL-based algorithms have been successfully applied for lake WL modeling; however, they are complex and need high computation costs. In contrast, tree-based models, Random Tree (RT), and iterative algorithms, Iterative Classifier Optimizer (ICO), have recently attracted scientists' attention for hydrological modeling, as they are comparatively novel, simple, robust and powerful (Hassan et al., 2017, Feng et al., 2017, Salam and Islam, 2020; Meshram et al., 2021). Therefore, in the present study, the lake WL forecasting accuracy of ML algorithms, ICO and RT, were compared. To the best of the authors' knowledge, although the ICO has been applied for suspended sediment load prediction (Meshram et al., 2020), it has not previously been used for time series forecasting, and in particular, for lake WL fluctuations. Moreover, the RT has also been used for reference evapotranspiration prediction (Salam and Islam, 2020). In undertaking the study, we sought to develop ICO- and RT-based multistep (*i.e.*, up to three-steps ahead) lake WL forecasts for North America's Lake Superior and Lake Michigan. Their hybrid counterparts, developed using AR and BA models (*i.e.*, BA-RT, BA-ICO, AR-RT, AR-ICO), were also explored for the first time in an effort to potentially improve the performance of the individual ICO and RT models.

2. Study area and data

The Great Lakes, a chain of deep freshwater lakes in east-central North America (lat. 41°22'N—49°00'N, long. 76°10'W—92°10'W), comprises Lakes Superior, Michigan, Huron, Erie, and Ontario (Fig. 1). The lakes serve as the boundary between the United States and Canada, except Lake Michigan, which is entirely in the United States. This research focuses on producing multistep lake WL forecasts for Lakes Superior and Michigan.

The farthest northwest, Lake Superior has the largest area (82,477 km²) of all freshwater lakes in the world. Its wide crescent form, convex to the north, extends 564 km east-west and has a breadth of 258 km. Located in the headwaters of the Great Lakes watershed, it is bordered by Ontario to the north and Minnesota, Wisconsin and Michigan to the west and south. The lake's surface is 184 m AMSL, and the point of maximum depth, at 406 m, is 223 m below sea level. The volume of the lake is 12,240 km³. It has the highest surface elevation, the largest total water volume, and the greatest depth of the five Great Lakes (Fuller and Shear, 1995). Because of its large size (relative to the watershed area) and the porous "dam" at its outlet, long-term lake-level fluctuations are much lower than in any of the other Great Lakes—generally less than 1 meter.

Lake Michigan, the third largest of the five Great Lakes, extends over an area of 58,061 km², is 494 km long in a north-south direction, and its width, including Green Bay, is 190 km. The surface is 177 m AMSL, and the point of maximum depth, 282 m, is 105 m below sea level. Approximately 100 streams flow into the lake, few of which are of appreciable size. The northern end of the lake contains several islands, the largest of which is Beaver Island, Michigan (Hough, 1968).

Fig. 1 Map of the Lakes Michigan and Superior

In this study, we selected Lakes Michigan and Superior to test the developed models on WL series with low and high variations. Note that the amplitude of WL variations is smaller for Lake Superior (the largest lake) than the other lakes while the WL variations in Lake Michigan is high compared to other Great lakes. Monthly WL data for Lakes Michigan and Superior over the period of 1918 to 2018 was obtained from the U.S. National Oceanic and Atmospheric Administration, NOAA. The WL time series was partitioned into two subsets: training (70%; from 1918 to 1988) and testing (30%; from 1989 to 2018). It is worth to note that there is no ‘rule of thumb’ for partitioning train/test sets. However, previous studies (Palani et al., 2008; Barzegar et al., 2016) generally agree that the testing subset must employ data never used in the training phase and that these data should represent approximately 10 – 40% of the size of the training set; the 70:30 dataset partitioning approach is often used for training and testing machine learning models (Huang and Wang, 2011; Li et al., 2011; Donate and Cortez, 2014; Rezaeianzadeh et al., 2017; Pérez-Chacón et al., 2020) and was thus adopted here. Summary statistics of the WL data for these subsets (Table 1) show WLs in Lake Michigan to have varied between 175.57 m and 177.50 m, with a mean and standard deviation (Std) of 176.42 m and 0.39 m, respectively. For Lake Superior the monthly WLs range between 182.72 m and

183.91 m (with a mean and standard deviation of 183.4 m and 0.2 m, respectively). The calculated skewness of the subsets indicates that data are light-tailed relative to a normal distribution. The kurtosis for all subsets, except the training data for Lake Superior, was negative, indicating that the left side is the longer tail of the distribution.

Table 1 Descriptive statistics of Lake Michigan and Superior WLs for the model training and testing datasets

Lake	Subsets	Min	Max	Mean	Standard deviation	Skewness	Kurtosis
Michigan	Train	175.58	177.50	176.47	0.41	-0.01	-0.79
	Test	175.57	177.19	176.32	0.35	0.21	-0.85
Superior	Train	182.72	183.91	183.44	0.19	-0.35	0.44
	Test	182.79	183.81	183.81	0.21	0.00	-0.43

3. Methodology

3.1. Base machine learners

In the present study, we used two ML techniques, including RT and ICO, to develop lake water level forecasting models. Each of these algorithms are described as follows:

3.1.1. Random Tree (RT)

The Random Tree (RT) model is a supervised learning algorithm that considers ‘K’ randomly selected input variables when performing a split at each node in an unpruned decision tree. RT can be used for both classification and regression problems and is the foundation of the Random Forest algorithm (Breiman, 2001). As a fast learning algorithm, RT can achieve high accuracy even in complex modeling situations (*e.g.*, when dynamic changes in the environment must be considered; Mu et al., 2017). RT’s strengths stem from its randomization when selecting input variables to split nodes, which is not considered in ‘standard’ decision trees. Section 4 includes the RT hyper-parameters and their optimal values (determined via trial-and-error).

According to Drmota and Gittenberger (1997), if T is a plane rooted tree with n nodes, which is considered as a family tree of a class, the profile of the tree may be described by the number of nodes or the number

of leaves. Supposing a class A with trees T ($T \in A$) consist of n nodes, the weight function $\omega(T)$ can be written as (Drmota and Gittenberger, 1997):

$$\omega(T) = \prod_{k \geq 0} \varphi_k^{n_k(T)} \quad (1)$$

where $(\varphi_k; k \geq 0)$ are non-negative numbers and $n_k(T)$ is the number of nodes $v \in T$ with out-degree k . Moreover, set:

$$a_n = \sum_{T: |T|=n} \omega(T) \quad (2)$$

Then the corresponding generating function (GF) $a(z) = \sum_{n \geq 0} a_n z^n$ satisfies the functional equation:

$$a(z) = z\varphi(a(z)) \quad (3)$$

where $\varphi(t) = \sum_{k \geq 0} \varphi_k t^k$, thereby equipping sets $A_n = \{T \in A : |T| = n\}$ with a probability distribution induced by the weight function $\omega(T)$ (Drmota and Gittenberger, 1997).

3.1.2. Iterative Classifier Optimizer (ICO)

The Iterative Classifier Optimizer (ICO) is used to identify the optimal number of iterations for a given base iterative classifier/regressor through cross-validation and a user-defined performance metric (Saad, 2018). At first, the base classifier/regressor is specified, along with any additional model parameters. Afterwards, ICO tests the base method for different iteration numbers and evaluates the pre-defined performance metric for each iteration number. ICO has a ‘look ahead’ mechanism that tests the base method at a higher number of iterations to improve the algorithm’s ability to locate an optimal iteration number. The number of iterations that results in the best cross-validation performance is selected as the optimal number of iterations. In this study, Alternating Model Trees (Frank et al., 2015) are used as the regressor and is briefly described below.

3.1.2.1 Alternating Model Trees (AMT)

Alternating Model Trees (AMT) are a class of regression trees that are grown using boosting, synthesizing an ensemble of trees in a single tree structure (Frank et al., 2015). In AMT, there are two types of nodes: splitter and prediction nodes. At each splitter node, two prediction nodes are generated. The splitting variable is selected by considering all input variables and splitting on the median value of the data that reaches that node. Afterwards, two linear regression models are fit on the subsets resulting from the split. The input variables used in the linear regression models are those that minimize the squared error. The two linear regression models are then placed at the prediction nodes attached to the recently created splitter node. When generating the prediction nodes at a given splitter node, it is possible to use two different input variables in the two linear regression models. To obtain the final prediction via AMT, the individual predictions made at each visited prediction node are multiplied by a shrinkage parameter (used to control overfitting) and summed together. The number of iterations (equivalent to the number of splitter nodes) and the shrinkage parameter must be set by the user. As noted in Section 3.1.2, ICO can be used to optimize the number of iterations in AMT. Additional details on AMT can be found in Frank et al. (2015).

3.2. Hybrid algorithms

To develop hybrid models, two commonly used algorithms namely AR and BA were applied. Details on these algorithms are given as follows:

3.2.1. Additive Regression (AR)

At each iteration of the Additive Regression (AR) method adopted here (Friedman, 2002), the standalone algorithm is fit to the residuals from the previous iteration, i.e., AR is a gradient boosting method (Friedman and Meulman, 2003). At the final iteration, all predictions are added to generate the final prediction (Frank et al., 2016). The AR algorithm can be represented as (Friedman, 2002):

$$F(\mathbf{X}) = \sum_{m=0}^M \beta_m h(\mathbf{X}; \mathbf{a}_m) \quad (4)$$

where $h(\mathbf{X}; \mathbf{a}_m)$ is the standalone model output considering inputs $\mathbf{X}$ and model parameters $\mathbf{a}_m$ at iteration m ($m = 0, \dots, M$), where M is the number of iterations (after the initial model is fit at $m = 0$); β_m are a set

of expansion coefficients at iteration m ; and $F(\mathbf{X})$ is the AR algorithm output. The standalone and AR algorithms' parameters (a_m and β_m) are fit in a stepwise manner (*i.e.*, each set of parameters is estimated at a particular iteration) (Friedman, 1999).

At each iteration of the AR algorithm, an 'intermediate' AR model output ($F_m(\mathbf{X})$) is generated (Friedman, 2002):

$$F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + \beta_m h(\mathbf{X}; a_m) \quad (5)$$

where $F(\mathbf{X}) = F_M(\mathbf{X})$. Further details on the AR algorithm can be found in Friedman (2002), Friedman and Meulman, (2003), and Frank et al. (2016). Since AR is coupled with RT in this study, it is considered a special case of gradient boosting known as multiple additive regression trees (Friedman and Meulman, 2003).

3.2.2. Bagging (BA)

Bagging (BA) is a robust ensemble method that is often used to improve model accuracy. Drawing random samples from the training data (input-target variable pairs) with replacement (also referred to as bootstraps), BA creates a model for each sample of the data (Bauer and Kohavi, 1999). After each model has been trained, an ensemble prediction can be made for new input data by taking the mean of the predictions across all trained models (also known as ensemble members). Since each training set is made by creating a bootstrap replicate of the original training set, this method can weaken the defects of poorly performing ensemble members and increase the overall accuracy of the ensemble (Breiman, 1996). Despite its simplicity, BA has been shown to significantly reduce the out-of-sample mean square error (MSE) when applied to classification and regression trees on a wide number of datasets (Breiman, 1996) and is equally applicable to any base algorithm. By coupling RT with BA, the result is a RF model.

3.3. Model evaluation metrics

The coefficient of determination (R^2) between observed and predicted WLs, the root mean squared error (RMSE), mean absolute error (MAE), root mean square percentage error (RMSPE), mean absolute

percentage error (MAPE), Kling-Gupta Efficiency (KGE), and Kendall's TAU (KTAU) were calculated to evaluate the performance of the developed models (Arndt et al., 1999; Gupta et al., 2009; Barzegar et al., 2019, 2020):

$$R^2 = \left[\frac{\frac{1}{N} \sum_{i=1}^N (M_i - M_m) (F_i - F_m)}{\sqrt{\frac{1}{N} \sum_{i=1}^N (M_i - M_m)^2} \sqrt{\frac{1}{N} \sum_{i=1}^N (F_i - F_m)^2}} \right]^2 \quad (5)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |M_i - F_i| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (M_i - F_i)^2} \quad (7)$$

$$RMSPE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{M_i - F_i}{M_i} \right)^2} \times 100 \quad (8)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|M_i - F_i|}{M_i} \times 100 \quad (9)$$

$$KGE = 1 - \sqrt{(r - 1) + (a - 1) + (b - 1)} \quad (10)$$

$$KTAU = \frac{2(n_c - n_d)}{N(N - 1)} \quad (11)$$

where, N is the number of data, M_i , F_i , M_m , and F_m are the observed, forecasted, mean observed, and mean forecasted data, respectively; r is the Pearson correlation coefficient between the forecasts and observations, a is the ratio between the variance of the forecasts and the variance of the observations, and b is the ratio of the means of the simulations and observations. KGE ranges from $-\infty$ to 1 with an ideal value of 1, where n_c is the number of concordant pairs and n_d is the number of discordant pairs (Knoben et al., 2019). The KTAU measures the degree of correspondence between two rankings and their significance (i.e., it measures the strength of association of the joint distribution of two or more variables). The R^2 ranges between 0 and 1, where 0 indicates that the regression line does not fit the set of data points and a value of

1 represents that the regression line perfectly fits the set of data points. The smaller values of error measuring metrics (e.g., RMSE, MAE, RMSPE and MAPE), the higher accuracy of the forecasting. Additionally, the performance of the models was visualized using different plots including time series comparative plot, bar graph, spider diagram and box-swarm plot.

4. Model development

Input variable selection, one of the most important steps in modeling, can enhance model performance and decrease the complexity of the modeling process. Although there are several input variable selection methodologies — principal component analysis; Gamma Test; forward selection technique (Noori et al., 2011); Edgeworth Approximation-based conditional mutual information; partial correlation input section input variable selection (Quilty et al., 2016) — to select the inputs having the greatest impact on the output variable, a simple Pearson correlation coefficient method was employed here. This is one of the most common and successful methods used in hydrology for selecting appropriate input variables in regression modeling (Yaseen et al., 2016; Khozani et al., 2019; Salih et al., 2019; Choubin et al., 2018). Accordingly, partial autocorrelation functions (PACF) of the monthly WL time series for Lakes Michigan and Superior were plotted and the most significant lag times were estimated (Fig. 2). For both WL time series, PACF values for a lag time of five months showed the greatest 95% confidence level partial correlation; accordingly, lag times of up to five months were considered to be the most important input variables for WL forecasting in these lakes. The correlation coefficients of these lag times were considered as input candidates [*i.e.*, $WL_{(t-5)}$, $WL_{(t-4)}$, ... $WL_{(t)}$] for forecasting multistep ahead WLs [*i.e.* $WL_{(t+1)}$, $WL_{(t+2)}$ and $WL_{(t+3)}$]. Tables 2 and 3 represent the correlation coefficients of the input and output variables for WL forecasting in Lakes Michigan and Superior, respectively. The WL at time t showed the greatest correlation with the outputs; the greater the time lag, the weaker the correlation. To construct the various input combinations, the variable with the highest correlation coefficient [*i.e.*, $WL_{(t)}$] was considered the first input variable (Combination 1). The second input combination was constructed by adding the variable with the next highest correlation coefficient [*i.e.*, $WL_{(t-1)}$] to the previous input. This procedure was continued by

successively adding the next most highly correlated of the remaining input variables, until the variable with the lowest correlation coefficient [*i.e.*, $WL_{(t-5)}$ in $WL_{(t+1)}$ forecasting for both lakes] was added to construct input combination 6 (Tables 4 and 5).

Fig. 2 PACF plot with 95% confidence bands for the WL time series in Lakes (a) Michigan and (b) Superior

Table 2 Correlation matrix heatmap of the input candidates and outputs of the models for Lake Michigan

	$WL_{(t-5)}$	$WL_{(t-4)}$	$WL_{(t-3)}$	$WL_{(t-2)}$	$WL_{(t-1)}$	$WL_{(t)}$	$WL_{(t+1)}$	$WL_{(t+2)}$	$WL_{(t+3)}$
$WL_{(t-5)}$	1.000								
$WL_{(t-4)}$	0.985	1.000							
$WL_{(t-3)}$	0.949	0.985	1.000						
$WL_{(t-2)}$	0.902	0.949	0.985	1.000					
$WL_{(t-1)}$	0.854	0.902	0.949	0.985	1.000				
$WL_{(t)}$	0.815	0.854	0.902	0.949	0.985	1.000			
$WL_{(t+1)}$	0.790	0.815	0.854	0.902	0.949	0.985	1.000		
$WL_{(t+2)}$	0.781	0.790	0.815	0.854	0.902	0.949	0.985	1.000	
$WL_{(t+3)}$	0.786	0.781	0.790	0.815	0.854	0.902	0.949	0.985	1.000

Table 3 Correlation matrix heatmap of the input candidates and outputs of the models for Lake Superior

	$WL_{(t-5)}$	$WL_{(t-4)}$	$WL_{(t-3)}$	$WL_{(t-2)}$	$WL_{(t-1)}$	$WL_{(t)}$	$WL_{(t+1)}$	$WL_{(t+2)}$	$WL_{(t+3)}$
$WL_{(t-5)}$	1.000								
$WL_{(t-4)}$	0.943	1.000							
$WL_{(t-3)}$	0.806	0.943	1.000						
$WL_{(t-2)}$	0.634	0.806	0.943	1.000					
$WL_{(t-1)}$	0.468	0.634	0.806	0.942	1.000				

$WL_{(t)}$	0.344	0.468	0.633	0.805	0.942	1.000			
$WL_{(t+1)}$	0.286	0.344	0.467	0.633	0.806	0.943	1.000		
$WL_{(t+2)}$	0.300	0.286	0.344	0.468	0.634	0.806	0.943	1.000	
$WL_{(t+3)}$	0.378	0.300	0.287	0.345	0.469	0.635	0.807	0.943	1.000

Table 4 Different input combinations applied for multistep ahead WL forecasting in Lake Michigan

No.	Different input combination	Output	Different input combination	Output
1	$WL_{(t)}$	$WL_{(t+1)}, WL_{(t+2)}$	$WL_{(t)}$	$WL_{(t+3)}$
2	$WL_{(t)}, WL_{(t-1)}$	$WL_{(t+1)}, WL_{(t+2)}$	$WL_{(t)}, WL_{(t-1)}$	$WL_{(t+3)}$
3	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}$	$WL_{(t+1)}, WL_{(t+2)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}$	$WL_{(t+3)}$
4	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}$	$WL_{(t+1)}, WL_{(t+2)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}$	$WL_{(t+3)}$
5	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-4)}$	$WL_{(t+1)}, WL_{(t+2)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-5)}$	$WL_{(t+3)}$
6	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-4)}, WL_{(t-5)}$	$WL_{(t+1)}, WL_{(t+2)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-5)}, WL_{(t-4)}$	$WL_{(t+3)}$

Table 5 Different input combinations applied for multistep ahead WL forecasting in Lake Superior

No.	Different input combination	Output	Different input combination	Output
1	$WL_{(t)}$	$WL_{(t+1)}$	$WL_{(t)}$	$WL_{(t+2)}, WL_{(t+3)}$
2	$WL_{(t)}, WL_{(t-1)}$	$WL_{(t+1)}$	$WL_{(t)}, WL_{(t-1)}$	$WL_{(t+2)}, WL_{(t+3)}$
3	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}$	$WL_{(t+1)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}$	$WL_{(t+2)}, WL_{(t+3)}$
4	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}$	$WL_{(t+1)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}$	$WL_{(t+2)}, WL_{(t+3)}$
5	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-4)}$	$WL_{(t+1)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-5)}$	$WL_{(t+2)}, WL_{(t+3)}$
6	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-4)}, WL_{(t-5)}$	$WL_{(t+1)}$	$WL_{(t)}, WL_{(t-1)}, WL_{(t-2)}, WL_{(t-3)}, WL_{(t-5)}, WL_{(t-4)}$	$WL_{(t+2)}, WL_{(t+3)}$

After constructing the input combinations, the best input variables combination for each standalone model (RT and ICO) was determined. The input variables of each combination were provided to the models and were then run in the Waikato Environment for Knowledge Analysis (WEKA 3.9) software (Witten et al., 2016) using the default model hyper-parameters. The models' outputs were then compared to measured WL values. The results for six different input variable combinations used in training and testing the models are summarized in Table 6, where the lowest training set RMSE is highlighted to indicate the optimal input variable combination. Input combination No. 2, including $WL_{(t)}$ and $WL_{(t-1)}$, resulted in the lowest RMSE for RT-based $WL_{(t+1)}$ and $WL_{(t+2)}$ forecasting in Lake Michigan, while input combination No. 1 including $WL_{(t)}$, resulted in the lowest error for $WL_{(t+3)}$ forecasting. Similar to Lake Michigan, input combination No.

2 provided the minimum RMSE for RT-based $WL_{(t+1)}$ forecasting for Lake Superior, whereas the input combination No. 4 including variables $WL_{(t)}$, $WL_{(t-1)}$, $WL_{(t-2)}$, $WL_{(t-3)}$ and combination No. 6 including variables $WL_{(t)}$, $WL_{(t-1)}$, $WL_{(t-2)}$, $WL_{(t-3)}$, $WL_{(t-5)}$, and $WL_{(t-4)}$ obtained the lowest RMSE for RT-based $WL_{(t+2)}$ and $WL_{(t+3)}$ forecasting, respectively in Lake Superior. The most effective variable(s) for RT-based forecasting models differed for the two lakes. For the ICO-based models, input combination No. 6 provided the minimum RMSE for all time horizons in both lakes.

Table 6 Identifying the best input combination based on the RMSE metric in the training period

NO.	Lake Michigan						Lake Superior					
	$WL_{(t+1)}$		$WL_{(t+2)}$		$WL_{(t+3)}$		$WL_{(t+1)}$		$WL_{(t+2)}$		$WL_{(t+3)}$	
	RT	ICO	RT	ICO	RT	ICO	RT	ICO	RT	ICO	RT	ICO
1	0.080	0.068	0.141	0.125	0.191	0.170	0.070	0.067	0.133	0.118	0.181	0.152
2	0.067	0.050	0.139	0.125	0.210	0.152	0.058	0.044	0.123	0.087	0.180	0.125
3	0.076	0.052	0.149	0.101	0.208	0.151	0.060	0.044	0.116	0.086	0.167	0.124
4	0.074	0.050	0.151	0.105	0.216	0.149	0.065	0.045	0.113	0.084	0.151	0.118
5	0.081	0.050	0.156	0.102	0.219	0.141	0.069	0.044	0.119	0.078	0.141	0.109
6	0.082	0.049	0.159	0.099	0.227	0.135	0.076	0.041	0.127	0.074	0.135	0.096

After finding the most effective input variables for each multistep step ahead WL forecast for each lake, the selected input variables served as models' inputs and the models' optimal hyper-parameters were calibrated. A WEKA software was used to determine the models' optimal values through a trial-and-error process, the most popular method for this software (de Lima et al., 2014). The model hyper-parameters' optimal values are listed in Table 7. The models were developed using those optimal hyper-parameters and then validated using the testing data. To develop the standalone models, single ICO (i.e. based on AMT as an iterative regressor) and RT algorithms were separately calibrated using training data and afterwards validated using the testing (unseen) data. While, for hybrid models, the ICO and RT models were used as base regressor of the AR and BA ensembles for construction and evaluation of the models. The same training and testing datasets were used for the AR-RT, AR-ICO, BA-RT, and BA-ICO ensemble models. In the next section, the performance of the different models is presented and discussed.

Table 7 Optimal values of the developed models' hyper-parameters for WL forecasting in Lakes Michigan and Superior

Model operators	Optimum value				
	RT	AMT	ICO	BA	AR
K-value	0*	-	-	-	-
Batch size	100	100	100	100	100
Maximum Depth of tree	4	-	-	-	-
Minimum number of instances at leaf	1	-	-	-	-
Minimum variance proportion	0.001	-	-	-	-
Number of folds	0	-	10	-	-
Evaluation Metric	-	-	RMSE	-	-
Iterative classifier	-	-	AMT	-	-
Look ahead iteration	-	-	50	-	-
Number of Iterations	-	12	-	-	-
Number of decimal places	-	2	2	2	2
Regressor	-	-	-	RT/ICO	RT/ICO
Number of execution slots	-	-	-	1	-
Shrinkage	-	1.0	-	-	2

* 'K' is set to 0 (as the default value); however, this corresponds with $\text{int}(\log_2(D) + 1)$ where D is the number of predictors used in the models (see Table 4, 5, and 6)

** AMT is coupled with ICO, the Number of Iterations value(s) listed in the table were found to be optimal by ICO.

5. Results

Two different standalone models (RT and ICO) and their corresponding hybrid versions developed using the AR and BA algorithms, drawing on their optimal input variable combinations, were tested and their performance evaluated using quantitative metrics and visual methods.

Comparative time series and scatter plots of measured vs. forecasted WL values for $WL_{(t+1)}$, $WL_{(t+2)}$, and $WL_{(t+3)}$ forecasting for Lake Michigan are shown in Fig. 3 and Figs. S1-S2 (in the supplementary material).

For test phase $WL_{(t+1)}$ forecasts (Fig. 3), the BA-ICO model was the most accurate ($R^2 = 0.979$), while the standalone RT models showed greater scatter between measured and forecasted WLs ($R^2 = 0.965$). For $WL_{(t+2)}$ forecasting (Fig. S1), the BA-ICO ($R^2 = 0.876$) and AR-ICO ($R^2 = 0.877$) models showed the best WL forecasts, compared to the poorer predictions and greater scatter of the RT models ($R^2 = 0.854$). The scatter plots of the forecasted $WL_{(t+3)}$ and the equivalent measured WLs for Lake Michigan showed the ICO and AR-ICO ($R^2 = 0.77$) and BA-ICO ($R^2 = 0.768$) to have obtained the highest accuracy and RT ($R^2 = 0.722$) model to have provided the lowest accuracy in their WLs forecasts. All developed models, for all

time horizons for Lake Michigan, can be categorized as having very good performance ($0.7 < R^2 \leq 1.0$; Moriasi et al., 2007).

For Lake Superior, the plots of the measured vs. forecasted $WL_{(t+1)}$, $WL_{(t+2)}$, and $WL_{(t+3)}$ scatter plots, presented in the supplementary material (Figs. S3-S5), show, based on the R^2 , that for all steps the forecasted WL obtained using the ICO and its hybrid counterparts with AR and BA algorithms had higher accuracy than the other models. Moreover, in the test phase, WLs forecast by the RT model showed greater scatter compared to the other models, along with the poorest fitted linear regression equation for all steps ahead. All models developed for $WL_{(t+1)}$ forecasting for Lake Superior showed a very good performance ($0.7 < R^2 \leq 1.0$), while for $WL_{(t+2)}$ RT models performed satisfactorily ($0.5 < R^2 \leq 0.6$), and other developed models well ($0.6 < R^2 \leq 0.7$). For $WL_{(t+3)}$, all developed models had an unsatisfactory performance ($0.0 < R^2 \leq 0.5$); however, the ICO-based models (single ICO and hybrid BA-ICO and AR-ICO) had much better performance compared to the RT models. The comparative plots for both lakes illustrated that coupling the standalone ML model (i.e. RT and ICO) with BA and AR results in capturing the WL extremes (both minimum and maximum) much better than the standalone models. Moreover, the scatter plots showed that the forecasted WL using the hybrid models are much closer to the 1:1 line (i.e. forecasted WL = measured WL) than the standalone models for both lakes.

Fig. 3 Comparative temporal series and scatter plots of measured and test period forecasted $WL_{(t+1)}$ for Lake Michigan

The error metrics (*i.e.*, RMSE and MAE) for the developed Lake Michigan WL models (Fig. 4a,b) show the standalone ICO model to have generated lower errors compared to the standalone RT model for all time

horizon [$WL_{(t+1)}$, $WL_{(t+2)}$, and $WL_{(t+3)}$] forecasts. The RT model generally yielded the highest RMSE and MAE for all time horizons, indicating the worst performance, while, based on having the lowest RMSE and MAE, the BA-ICO model (RMSE = 0.05 m, MAE = 0.039 m) provided the best $WL_{(t+1)}$ forecast for Lake Michigan. For $WL_{(t+2)}$ and $WL_{(t+3)}$ forecasting, the AR-ICO and BA-ICO hybrid model along with the standalone ICO provided the lowest error, indicating it to be the most efficient model for Lake Michigan WL forecasting. Although the AR and BA algorithms can no improve the ICO models in forecasting two and three-months ahead WL of Lake Michigan, they can enhance the accuracy of the RT models. The percentage error of the WL forecasting models in terms of RMSPE and MAPE are provided in Table 8. Based on these model accuracy statistics the BA- and AR-ICO showed no improvement over their corresponding standalone ICO models for $WL_{(t+2)}$, and $WL_{(t+3)}$ forecasting for Lake Michigan, except for $WL_{(t+1)}$ forecasting. The maximum improvement of the models belongs to the AR-RT for $WL_{(t+1)}$ forecasting in Lake Michigan.

Based on the KTAU indicator, the RT model (KTAU = 0.982, 0.924 and 0.850 for $WL_{(t+1)}$, $WL_{(t+2)}$, and $WL_{(t+3)}$, respectively), showed the lowest KTAU values, indicating the worst performance for WL forecasting in Lake Michigan (Fig. 4c). In contrast, the hybrid BA-ICO [KTAU = 0.990, 0.936, and 0.877 for $WL_{(t+1)}$, $WL_{(t+2)}$, and $WL_{(t+3)}$, respectively] had the best performance amongst the models developed. In general, the hybrid models improved performance compared to the standalone models with respect to the KTAU metric, particularly the standalone RT model.

The RT [KGE = 0.977 and 0.846 for $WL_{(t+1)}$ and $WL_{(t+3)}$, respectively] and BA- and AR-ICO [KGE = 0.918 for $WL_{(t+2)}$] showed the lowest KGE values for Lake Michigan (Fig. 4d), while the hybrid AR-RT for $WL_{(t+2)}$ (KGE = 0.989), and $WL_{(t+3)}$ (KGE = 0.937) and BA-ICO for $WL_{(t+1)}$ (KGE = 0.868), outperformed all other developed models (Fig. 4d).

Fig. 4 Bar graph of the (a) RMSE and (b) MAE metrics and Spider diagram of the calculated (c) KTAU and (d) KGE metrics for WL forecasting models in Lake Michigan

Models' errors for Lake Superior WL forecasting show that, similarly to Lake Michigan, the RMSE and MAE of the individual ICO model were lower than those of the RT model for all time horizons (Fig. 5a, b). The BA-ICO model provided the best performance for $WL_{(t+1)}$ forecasting in terms of RMSE and MAE, while the BA-RT models were the best model for $WL_{(t+2)}$ and $WL_{(t+3)}$ forecasting. Generally, the BA hybrid algorithm generates more accurate predictions than the AR hybrid algorithm.

The MAPE and RMSPE of the models developed for WL forecasting for Lake Superior (Table 8) show that for WL forecasting, both AR and BA algorithms decrease the individual RT model's errors in terms of RMSPE and MAPE (up to 14% and 27%, respectively), while the those algorithms led to decreases of up to 4% for the ICO-based WL forecasting models. The error reduction achieved with the BA algorithm was higher than that of the AR algorithm for the ICO-based model (*i.e.*, up to 4% and 1% decrease in error

based on the RMSPE and MAPE, respectively). These statistics showed that for Lake Superior, similar to $WL_{(t+2)}$ forecasting, for $WL_{(t+3)}$ forecasting the BA algorithm performs better than the AR algorithm not only in improving the ICO models but also in improving the individual RT model. It is worth noting that the AR and BA algorithms not only improve model performance compared to the standalone models, but they also reduce overfitting, which is a crucial issue when applying machine learning models to real-world data.

For Lake Superior, the BA-ICO model (0.973) showed the highest KTAU value for $WL_{(t+1)}$ while the highest KTAU for the $WL_{(t+2)}$ and $WL_{(t+3)}$ were 0.835 and 0.686 for AR-ICO and BA-RT models, respectively, indicating the superior performance of the BA-based models for this lake WL forecasting (Fig. 5c). Overall, based on the KTAU, the hybrid models performed better than their corresponding single models.

The KGE metric shows that AR-RT, with values of 0.950 and 0.760, outperforms the other models for $WL_{(t+1)}$ and $WL_{(t+2)}$ forecasting of Lake Superior, respectively, while the BA-RT (0.642) had the best performance among the standalone and hybrid models for $WL_{(t+3)}$ forecasting (Fig. 5d).

Fig. 5 Bar graph of the calculated (a) RMSE and (b) MAE metrics and Spider diagram of the calculated (c) KTAU and (d) KGE metrics for WL forecasting models in Lake Superior

Table 8 Performance of the developed models based on MAPE and RMSPE metrics

Lake	Model	WL _(t+1)		WL _(t+2)		WL _(t+3)	
		MAPE (%)	RMSPE (%)	MAPE (%)	RMSPE (%)	MAPE (%)	RMSPE (%)
Michigan	ICO	2.3	2.9	5.7	7.1	8.0	9.7
	RT	3.0	3.8	6.4	7.9	8.8	10.9
	BA-ICO	2.2	2.9	5.7	7.1	8.0	9.7
	BA-RT	2.6	3.5	5.5	7.3	8.3	10.6
	AR-ICO	2.3	2.9	5.7	7.1	8.0	9.7
	AR-RT	2.3	3.3	5.2	7.2	7.8	10.2
Superior	ICO	2.3	2.9	5.6	6.6	7.4	9.1
	RT	2.9	3.8	6.1	7.5	8.1	10.2
	BA-ICO	2.2	2.8	5.5	6.5	7.3	9.0
	BA-RT	2.5	3.6	5.0	6.7	5.9	8.8
	AR-ICO	2.3	2.9	5.6	6.6	7.4	9.0
	AR-RT	2.5	3.6	5.1	6.9	7.3	9.7

384

385 To further evaluate differences in the models' performance in forecasting WLs for Lakes Michigan and
386 Superior, we present combined box-swarm plots (Fig. 6). Although the statistical metrics indicated,
387 generally, the BA-ICO outperformed the other developed models in forecasting $WL_{(t+1)}$ for Lake Michigan,
388 the box plot also confirms that the BA-ICO performed better than the other developed models, and
389 considerably better than the AR-ICO model, in capturing the minimum WL. The swarm plot also confirmed
390 the superiority of the BA-ICO over the AR-ICO in a detail comparison. It is observed that the main
391 difference of the Lake Michigan's models was attributed to the capturing the minimum values, particularly
392 for the three months ahead forecasts. Swarm patterns for different models developed for Lake Michigan
393 vary slightly, indicating the low variability of the model's error. However, it is evident that the hybrid could
394 model the measured WLs better than the standalone models. For Lake Superior, the box plots illustrated
395 that developed models mainly performed differently for different step aheads WL forecasting. Moreover,
396 outlier WL forecasts could be seen for this lake. The plots showed that the ICO-based models could capture
397 these extremes along with the median WLs much better than the RT-based models. Similar to Lake
398 Michigan, it was observed that the BA-ICO followed by BA-RT outperformed among the developed models
399 for Lake Superior WL forecasting.

Fig. 6 Box-swarm plots of the measured and forecasted WLs in Lakes Michigan and Superior

6. Discussion

Although both hybrid BA- and AR-based algorithms can improve the performance of individual ML algorithms, generally, the BA outperforms the AR. However, the performance enhancement provided by the hybrid algorithms differed for the different forecast lead times, and further differed according to whether AR or BA algorithms were employed at particular lead times. In most steps ahead, the BA algorithm's

performance was much better than that of the AR; however, the AR did perform better for some steps ahead. It is worth noting that in generating multistep forecasts in hydrology it is recommended to explore (and potentially use) different hybrid algorithms at different steps to obtain the best possible performance. The poor performance of the standalone models compared to their counterpart hybrid models was the result of greater bias and variances in the single forecasts. This occurred as the hybrid algorithms decrease bias and/or variance of the standalone models by incorporating several of them together, thereby creating a better performing strong learner (or ensemble model). Apart from reducing the bias and variance of the standalone models, hybrid algorithms can also address overfitting issues in the modeling process, particularly when a small dataset is used to train the models (Erdal and Karakurt, 2013; Barzegar et al., 2019). Compared to neural networks, tree-based models have been shown to demonstrate improved generalization (Mihelich et al., 2020). However, the BA and AR-based learning algorithms decrease the predictive variance and/or bias of the base regressor result in decreasing the overfitting and improving the generalization. In a study by Mihelich et al. (2020), they concluded that BA can significantly enhance the generalization performance of the tree- and neural network-based ML models.

Previous studies have shown that using hybrid ensemble models can reduce the error of standalone models. For example, Bui et al. (2020) showed that BA-based hybrid models [*e.g.*, BA- RF, -M5P, - RT, and - reduced error pruning tree (REPT)] improved the performance of individual models in predicting water quality indices. Omran et al. (2016) applied three ML models (*e.g.*, multilayer perceptron, support vector machines, and Gaussian processes regression), four regression tree models (*e.g.*, M5P, REPT, M5-Rules, and decision stump), and two hybrid methods (*e.g.*, AR and BA) with each of the seven standalone models as base classifiers, to predict the compressive strength of concrete. They concluded that the hybrid algorithms improved prediction accuracy for the four regression tree models, but had less success with the other three ML predictive models. Accordingly, the current study concurs with the findings of Omran et al. (2016).

As a model newly applied to forecasting problems, the ICO algorithm using an iterative base regressor (*e.g.* AMT) is capable of generating accurate multistep WL forecasts. Moreover, the ICO algorithm was shown

to be more accurate than the RT model in WL forecasting for the study sites. Differences in model performances were attributed to the different algorithms' structures. Currently, no published study has used ICO for hydrological forecasting, making comparisons difficult; however, Pak and Teh (2016) did employ several ML algorithms for classification purposes, including ICO and RT. Their study indicated that the ICO algorithm performed better than the RT in their particular application, as it did in the current study's specific application. Although several previous studies have used ML [*e.g.*, ANN, ANFIS, gene expression programming (GEP) (Talebizadeh and Moridnejad, 2011; Kisi et al., 2012) and deep learning [*e.g.* LSTM (Zhu et al., 2020b)] for WL forecasting in different lakes, to the best of our knowledge, the lack of any published studies regarding hybrid ensemble forecasting models for lake WL forecasting makes a comprehensive comparison with the results of this study difficult. Moreover, there are limited works for ML-based water level forecasting in Lakes Michigan and Superior. In one of those rare studies, Altunkaynak (2014) established standalone fuzzy logic (FL) and multilayer perceptron (MLP) along with their wavelet-based hybrid model to forecast Michigan Lake WL up to 12 months ahead lead times (*i.e.* $t+3$, $t+6$, $t+9$ and $t+12$ time steps). In terms of RMSE, developed models for three months step ahead WL forecasting gave RMSE of 0.494 m for both standalone FL and MLP and RMSE of 0.139 and 0.136 m for wavelet-FL and wavelet-MLP models. Compared to the results of the current study, it is found that the models developed herein have higher accuracy than standalone models in Altunkaynak (2014) in terms of RMSE. While the wavelet-based models appear to outperform the models presented herein, it is important to note that the wavelet based models in Altunkaynak (2014) were developed using an approach that is inappropriate for real-world forecasting application (since the forecasts incorporate 'future data'). Details on (how to address) the incorrect usage of wavelet-based ML models for forecasting can be found in Quilty and Adamowski (2018). Statistical (*e.g.* Advanced Hydrologic Prediction System (AHPS) (Andrew et al., 2011)) and numerical (Process-Oriented Model (Fry et al., 2020)) models have also been used successfully to forecast WLs in both lakes. However, as indicated in the literature, these models have several disadvantages over the ML models (*e.g.* time consuming, restrictive assumptions, so on). The results of the

developed hybrid models in this study show that they can be used as accurate alternatives to these is better than those models.

Lake WL forecasting could also be carried out using mechanistic models, focused on the causality of input–output relationships (Baker et al., 2018). It has been frequently used for rainfall-runoff process, streamflow and river water level modeling (Lees, 2000; Romanowicz et al., 2008; Young et al., 2014; Smith et al., 2014). However, ML models have several advantages over the mechanistic models. The mechanistic models use assumption and oversimplify those assumptions while there is no assumption in ML modeling to limit the usage of the ML. The mechanistic hydrological modeling should be a numerical representation of the hydrological catchment/basin that shows the catchment/basin dynamics adopting a minimum degree of model complexity recognized directly from the existing data (Smith et al., 2014). Therefore, it may not be able to model complex processes and may require a huge amount of data while ML models, particularly used algorithms in this study, can successfully simulate any linear/nonlinear complex relationships between input and output.

The source of the uncertainty in machine learning models includes uncertainty in input data, model structure, parameters, and output. Although Petropoulos et al. (2018) successfully applied BA for exploring the uncertainty in time series forecasting, it is recommended to investigate the capability of the hybrid models used in the current study for uncertainty of hydrological time series forecasting. In Petropoulos et al. (2018), they suggested a bootstrap model combination strategy to deal with different sources of uncertainty. They concluded that bootstrapping can inherently address the data uncertainty. Also, model uncertainty can be considered by fitting various models to each bootstrap series. Even if a standalone model is recognized as the best over the different bootstraps, the various sets of the model’s parameters (estimated through this process) should be considered when generating model predictions, to reflect parameter uncertainty.

7. Conclusions

483 The accurate forecasting of lake water levels (WLs) is important for sustainable water resources planning
 484 and management. Given the nonlinear behaviour of WL fluctuations in lakes with response to driving
 485 factors (rainfall, evaporation, etc.), machine learning (ML) algorithms are especially useful to generate
 486 accurate forecasts of such processes. Drawing upon data collected from Lakes Superior and Michigan, the
 487 present research evaluated, for the first time, novel hybrid ensemble ML algorithms by investigating the
 488 feasibility of newly introduced methods to forecast WLs. The ability to forecast multistep ahead [$WL_{(t+1)}$,
 489 $WL_{(t+2)}$, and $WL_{(t+3)}$] monthly lake WL using standalone ML algorithms (ICO (i.e. based on AMT as an
 490 iterative regressor) and RT), as well as ICO- and RT-based hybrid ensemble algorithms integrated with AR
 491 and BA models, was assessed. Prior to the modeling process, input time series data analysis was carried out
 492 to identify important time lags. Then, in order to find the most effective input variables for each step ahead
 493 lake WL forecasting, different input combinations were constructed. The models developed drew on these
 494 input variables and optimal model hyper-parameters (determined by trial-and-error). The performance of
 495 the developed models was assessed in terms of statistical metrics: R^2 , RMSE, MAE, RMSPE, MAPE, as
 496 well as KTAU and KGE.

497 The statistics showed that the ICO-based models (single ICO and hybrid BA-ICO and AR-ICO) had better
 498 performance compared to RT-based models (both single and corresponding hybridized models) in WL
 499 forecasting for both lakes Michigan and Superior. All models developed for all time horizons for Lake
 500 Michigan had an excellent performance. In contrast, the performance of models for Lake Superior varied
 501 between unsatisfactory and excellent. All $WL_{(t+1)}$ forecasts for Lake Superior were excellent, whereas the
 502 standalone RT and the corresponding hybrid algorithms (BA-RT and AR-RT) showed unsatisfactory
 503 performances for $WL_{(t+3)}$ forecasting.

504 Overall, hybrid ML techniques provided forecasts of greater accuracy than did standalone models.
 505 However, the performance enhancement of these hybrid algorithms differed for different forecast lead times
 506 and their performance was also tied to whether AR or BA algorithms were used. Based on the RMSPE and
 507 MAPE, the BA-ICO and AR-ICO improved their corresponding standalone ML models for $WL_{(t+1)}$, $WL_{(t+2)}$,
 508 and $WL_{(t+3)}$ forecasts of Lake Michigan WL more so than the other developed hybrid models. For Lake

Superior, the BA-ICO in $WL_{(t+1)}$ and $WL_{(t+2)}$ forecasting and BA-RT in $WL_{(t+3)}$ forecasting improved their corresponding standalone ML models more so than other developed hybrid models. Therefore, in order to obtain the best possible hydrological forecasts, examining different hybrid algorithms at all relevant forecast lead times is important. In general, the hybrid ensemble models provided better performance than the standalone models since they reduced forecast bias and variance by including the strengths of multiple models. This study also demonstrated that the ICO algorithm, as a novel algorithm in forecasting problems, offers great potential to produce accurate multistep WL forecasts.

Acknowledgements

Funding for this research was provided by an NSERC Discovery Grant held by Jan Adamowski.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRedit authorship contribution statement

Elham Fijani: Writing – original draft, Visualization. **Khabat Khosravi:** Formal analysis, Software, review & editing. **Rahim Barzegar:** Conceptualization, Writing – original draft, Visualization. **John Quilty:** Writing - review & editing. **Jan Adamowski:** Supervision, Writing - review & editing. **Homa Kheyrollah Pour:** Writing - review & editing.

References

- Altunkaynak, A., 2007. Forecasting surface water level fluctuations of Lake Van by artificial neural networks. *Water Resources Management* 21, 399–408.
- Altunkaynak, A., Şen, Z., 2007. Fuzzy logic model of lake water level fluctuations in Lake Van, Turkey. *Theoretical and Applied Climatology* 90, 227–233.

535 Arndt, S., Turvey, C., & Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings:
536 Spearman's r versus Kendall's τ correlation. *Journal of Psychiatric Research*, 33(2), 97-104.

537 Baker, R. E., Peña, J. M., Jayamohan, J., & Jérusalem, A. (2018). Mechanistic models versus machine
538 learning, a fight worth fighting for the biological community?. *Biology letters*, 14(5), 20170660.

539 Barzegar, R., Aalami, M. T., & Adamowski, J. (2020). Short-term water quality variable prediction using
540 a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*,
541 1-19.

542 Barzegar, R., Aalami, M.T., Adamowski, J., 2020. Short-term water quality variable prediction using a
543 hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*
544 34, 415–433.

545 Barzegar, R., Adamowski, J., & Moghaddam, A. A. (2016). Application of wavelet-artificial intelligence
546 hybrid models for water quality prediction: a case study in Aji-Chay River, Iran. *Stochastic*
547 *environmental research and risk assessment*, 30(7), 1797-1819.

548 Barzegar, R., Fijani, E., Moghaddam, A.A., Tziritis E., 2017. Forecasting of groundwater level fluctuations
549 using ensemble hybrid multi-wavelet neural network-based models. *Science of the Total Environment*
550 599–600, 20-31.

551 Barzegar, R., Ghasri, M., Qi, Z., Quilty, J., & Adamowski, J. (2019). Using bootstrap ELM and LSSVM
552 models to estimate river ice thickness in the Mackenzie River Basin in the Northwest Territories,
553 Canada. *Journal of Hydrology*, 577, 123903.

554 Baskin, I.I., Marcou, G., Horvath, D., Varnek, A., 2017. Bagging and boosting of regression models.
555 *Tutorials in Chemoinformatics* 28, 249–255.

556 Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging,
557 boosting, and variants. *Machine Learning* 36, 105-139.

558 Bonakdari, H., Ebtehaj, I., Samui, P., Gharabaghi, B., 2019. Lake water-level fluctuations forecasting using
559 minimax probability machine regression, relevance vector machine, gaussian process regression, and
560 extreme learning machine. *Water Resources Management* 33 (11), 3965-3984.

561 Breiman, L., 1996. Bagging predictors. *Machine Learning* 24, 123–140.

562 Bui, D.T., Ho, T-C., Pradhan, B., Pham, B-T., Nhu, V-H., Revhaug, I., 2016. GIS-based modeling of
563 rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost,
564 Bagging, and MultiBoost ensemble frameworks. *Environmental Earth Sciences* 75, 1101.

565 Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N., 2020. Improving prediction of water
566 quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment* 721,
567 137612.

Chen, W., Hong, H., Li, S., Shahabi, H., Wangi, Y., Wang, X., Bin Ahmad, B., 2019. Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. *Journal of Hydrology* 575, 864-873.

Cheng, C-T., Zhao, M-Y., Chau, K.W., Wu, X-Y., 2006. Using genetic algorithm and TOPSIS for Xinanjing model calibration with a single procedure. *Journal of Hydrology* 316 (1-4), 129-140.

Cimen, M., Kisi, O., 2009. Comparison of two different data driven techniques in modeling surface water level fluctuations of Lake Van, Turkey. *Journal of Hydrology* 378(3-4), 253-262.

Coulibaly, P., 2010. Reservoir computing approach to Great Lakes water level forecasting. *Journal of Hydrology* 381 (1-2), 76-88.

de Lima, T. P., da Silva, A. J., Ludermir, T. B., & de Oliveira, W. R. (2014). An automatic methodology for construction of multi-classifier systems based on the combination of selection and fusion. *Progress in Artificial Intelligence*, 2(4), 205-215.

Donate, J. P., & Cortez, P. (2014). Evolutionary optimization of sparsely connected and time-lagged neural networks for time series forecasting. *Applied Soft Computing*, 23, 432-443.

Drmot, M., & Gittenberger, B. (1997). On the profile of random trees. *Random Structures & Algorithms*, 10(4), 421-451.

Fry, L. M., Apps, D., & Gronewold, A. D. (2020). Operational seasonal water supply and water level forecasting for the Laurentian Great Lakes. *Journal of Water Resources Planning and Management*, 146(9), 04020072.

Feng, Y., Cui, N., Gong, D., Zhang, Q., & Zhao, L. (2017). Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. *Agricultural Water Management*, 193, 163-173.

Frank, E., Mayo, M., & Kramer, S. (2015, April). Alternating model trees. In *Proceedings of the 30th annual ACM symposium on applied computing* (pp. 871-878). <https://doi.org/10.1145/2695664.2695848>.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), 367-378.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in medicine*, 22(9), 1365-1381.

Fuller, K., Shear, H., 1995. *The Great Lakes: An Environmental Atlas and Resource Book*. U. S. Environmental Protection Agency/Government of Canada; 3rd edition. 46 p.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, 377(1-2), 80-91.

Hassan, M. A., Khalil, A., Kaseb, S., & Kassem, M. A. (2017). Exploring the potential of tree-based ensemble methods in solar radiation modeling. *Applied Energy*, 203, 897-916.

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- Hough J.L., 1968. Great lakes (North America). In: Geomorphology. Encyclopedia of Earth Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-31060-6_170.
- Huang, A., Rao, Y.R., Lu, Y., Zhao, J., 2010. Hydrodynamic modeling of Lake Ontario: an intercomparison of three models. *Journal of Geophysical Research*, 115 (12), C12076.
- Huang, G., & Wang, L. (2011). Hybrid neural network models for hydrologic time series forecasting based on genetic algorithm. In 2011 Fourth International Joint Conference on Computational Sciences and Optimization (pp. 1347-1350). IEEE.
- Kebede, S., Travi, Y., Alemayehu, T., Marc, V., 2006. Water balance of Lake Tana and its sensitivity to fluctuations in rainfall, Blue Nile basin, Ethiopia. *Journal of Hydrology* 316 (1–4), 233-247.
- Khatibi, R., Ghorbani, M.A., Naghipour, L., Jothiprakash, V., Fathima, T.A., Fazelifard, M.H., 2014. Inter-comparison of time series models of lake levels predicted by several modeling strategies. *Journal of Hydrology* 511, 530-545.
- Khosravi, K., Cooper, J.R., Daggupati, P., Pham, B.T., Bui, D.T., 2020. Bedload transport rate prediction: Application of novel hybrid data mining techniques. *Journal of Hydrology* 585, 124774.
- Kisi, O., Shiri, J., Nikoofar, B., 2012. Forecasting daily lake levels using artificial intelligence approaches. *Computers & Geosciences* 41, 169-180.
- Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10), 4323-4331.
- Lees, M. J. (2000). Data-based mechanistic modelling and forecasting of hydrological systems. *Journal of Hydroinformatics*, 2(1), 15-34.
- Li, S. T., Kuo, S. C., Cheng, Y. C., & Chen, C. C. (2011). A vector forecasting model for fuzzy time series. *Applied Soft Computing*, 11(3), 3125-3134.
- Lin, P., Yang, Z. L., Cai, X., & David, C. H. (2015). Development and evaluation of a physically-based lake level model for water resource management: A case study for Lake Buchanan, Texas. *Journal of Hydrology: Regional Studies*, 4, 661-674.
- Manikandan, G and Aravind, Vasudev and Anitha, Balasubramanian. (2018). A Survey to Identify an Efficient Classification Algorithm for Heart Disease Prediction. *International Journal of Pure and Applied Mathematics*, 119(12), pp. 13337-13345.
- Meshram, S. G., Safari, M. J. S., Khosravi, K., & Meshram, C. (2021). Iterative classifier optimizer-based pace regression and random forest hybrid models for suspended sediment load prediction. *Environmental Science and Pollution Research*, 28(9), 11637-11649.

635 Mihelich, M., Dognin, C., Shu, Y., & Blot, M. (2020, June). A Characterization of Mean Squared Error for
636 Estimator with Bagging. In International Conference on Artificial Intelligence and Statistics (pp. 288-
637 297). PMLR.

638 Mu, X., Ting, K.M., Zhou, Z.-H., (2017). Classification under streaming emerging new classes: a solution
639 using completely-random trees. IEEE Transactions on Knowledge and Data Engineering 29, 1605–
640 1618.

641 Noori, R., Karbassi, A. R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M. H., Farokhnia, A., &
642 Gousheh, M. G. (2011). Assessment of input variables determination on the SVM model performance
643 using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. Journal
644 of Hydrology, 401(3-4), 177-189.

645 Omran, B. A., Chen, Q., & Jin, R. (2016). Comparison of data mining techniques for predicting compressive
646 strength of environmentally friendly concrete. Journal of Computing in Civil Engineering, 30(6),
647 04016029.

648 Pak, I., & Teh, P. L. (2016). Machine learning classifiers: Evaluation of the performance in online reviews.
649 Indian Journal of Science and Technology, 9(45), 1-9.

650 Pérez-Chacón, R., Asencio-Cortés, G., Martínez-Álvarez, F., & Troncoso, A. (2020). Big data time series
651 forecasting based on pattern sequence similarity and its application to the electricity demand.
652 Information Sciences, 540, 160-174.

653 Quilty, J., & Adamowski, J. (2018). Addressing the incorrect usage of wavelet-based hydrological and
654 water resources forecasting models for real-world applications with best practices and a new
655 forecasting framework. Journal of hydrology, 563, 336-353.

656 Quilty, J., Adamowski, J., 2018. Addressing the incorrect usage of wavelet-based hydrological and water
657 resources forecasting models for real-world applications with best practices and a new forecasting
658 framework. Journal of Hydrology 563, 336-353.

659 Quilty, J., Adamowski, J., Khalil, B., & Rathinasamy, M. (2016). Bootstrap rank-ordered conditional
660 mutual information (broCMI): A nonlinear input variable selection method for water resources
661 modeling. Water Resources Research, 52(3), 2299-2326.

662 Rezaeianzadeh, M., Kalin, L., & Anderson, C. J. (2017). Wetland water-level prediction using ANN in
663 conjunction with base-flow recession analysis. Journal of Hydrologic Engineering, 22(1), D4015003.

664 Ribeiro, M.H.D.M., and Coelho, L.D.S., (2020). Ensemble approach based on bagging, boosting and
665 stacking for short-term prediction in agribusiness time series. Applied Soft Computing, 86, 105837.

666 Romanowicz, R. J., Young, P. C., Beven, K. J., & Pappenberger, F. (2008). A data based mechanistic
667 approach to nonlinear flood routing and adaptive flood level forecasting. Advances in Water
668 Resources, 31(8), 1048-1056.

669 Saad, I.A., 2018. An Efficient Classification Algorithms for Image Retrieval Based Color and Texture
670 Features. *Journal of Al-Qadisiyah for Computer Science and Mathematics* 10(1), 42 - 53.

671 Salam, R., & Islam, A. R. M. T. (2020). Potential of RT, Bagging and RS ensemble learning algorithms for
672 reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh.
673 *Journal of Hydrology*, 590, 125241.

674 Sharafati, A., Khosravi, K., Khosravinia, P., Ahmed, K., Salman, S.A., Yaseen, Z.M., 2019. The potential
675 of novel data mining models for global solar radiation prediction. *International Journal of*
676 *Environmental Science and Technology* 16, 7147–7164.

677 Shiri, J., Shamshirband, S., Kisi, O., Karimi, S., Bateni, S.M., Hosseini Nezhad, S.H., Hashemi, A., 2016.
678 Prediction of water-level in the Urmia Lake using the extreme learning machine approach. *Water*
679 *Resources Management* 30 (14), 5217-5229.

680 Smith, P. J., Panziera, L., & Beven, K. J. (2014). Forecasting flash floods using data-based mechanistic
681 models and NORA radar rainfall forecasts. *Hydrological Sciences Journal*, 59(7), 1403-1417.

682 Taormina, R., Chau, K-W., 2015. Data-driven input variable selection for rainfall–runoff modeling using
683 binary-coded particle swarm optimization and Extreme Learning Machines. *Journal of Hydrology* 529,
684 1617–1632.

685 Tiwari, M.K., Chatterjee, C., 2011. A new wavelet–bootstrap–ANN hybrid model for daily discharge
686 forecasting. *Journal of Hydroinformatics* 13 (3), 500–519.

687 Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). The WEKA workbench. online appendix for “Data
688 Mining: Practical machine learning tools and techniques”. In Morgan Kaufmann.

689 Xu, B., Lin, B., 2017. Does the high–tech industry consistently reduce CO2 emissions? Results from
690 nonparametric additive regression model. *Environmental Impact Assessment Review* 63, 44-58.

691 Yadav, B., Eliza, K., 2017. A Hybrid Wavelet-Support Vector Machine model for Prediction of Lake water
692 level Fluctuations Using Hydro-Meteorological Data. *Measurement* 103, 294-301.

693 Yan, S., Wang, X., Zhang, Y., Liu, D., Yi, Y., Li, C., Liu, Q., Yang, Z., 2020. A hybrid PCA-GAM model
694 for investigating the spatiotemporal impacts of water level fluctuations on the diversity of benthic
695 macroinvertebrates in Baiyangdian Lake, North China. *Ecological Indicators* 116, 106459.

696 Yarar, A., Onucyildiz, M., Coptu, N.K., 2009. Modelling level change in lakes using neuro-fuzzy and
697 artificial neural networks. *Journal of Hydrology* 365 (3–4), 329-334.

698 Young, P. C., Romanowicz, R. J., & Beven, K. J. (2014). A data-based mechanistic modelling approach
699 to real-time flood forecasting. *Applied Uncertainty Analysis for Flood Risk Management*, edited by:
700 Beven, KJ and Hall, JW, Imperial College Press: London, 407-461.

701 Zhu, S., Hrnjica, B., Ptak, M., Choiński, A., Sivakumar, B., 2020b. Forecasting of water level in multiple
702 temperate lakes using machine learning models. *Journal of Hydrology* 585, 124819.

703 Zhu, S., Lu, H., Ptak, M., Dai, J., Ji, Q., 2020a. Lake water-level fluctuation forecasting using machine
704 learning models: a systematic review, *Environmental Science and Pollution Research*. 27, 44807–
705 44819.
706

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.docx](#)