Neural Computing and Applications (2023) 35:303–343
https://doi.org/10.1007/s00521-022-07699-z

ORIGINAL ARTICLE

Neurocomputing intelligence models for lakes water level forecasting: a comprehensive review

Vahdettin Demir[1] · Zaher Mundher Yaseen[2]

Received: 21 February 2022 / Accepted: 2 August 2022 / Published online: 10 September 2022

Abstract

Hydrological processes forecasting is an essential step for better water management and sustainability. Among several hydrological processes, lake water level (LWL) forecasting is one of the significant processes within a particular catchment. The complexity of the LWL fluctuation is owing to the diversity of the influential parameters including climate, hydrology and some other morphology. In this study, several versions of neurocomputing intelligence models are developed for LWL fluctuation forecasting at five great lakes Lake Superior, Lake Michigan, Lake Huron, Lake Erie, and Lake Ontario, located at the north of USA. The applied models are including M5-Tree, multivariate adaptive regression spline (MARS) and least square support vector regression (LSSVR). The models are developed using several input combinations that are configured based on the correlated lags in addition to the periodicity of time series. The sequential influence of the lakes order is considered in the modeling development. Also, cross-station modeling where lag time series of upstream lakes are used to forecast downstream LWL. Results are assessed using several statistical metrics and graphical visualization. Overall, the results indicated that the applied forecasting models efficient and trustworthy. The component of the periodicity time series enhances the forecasting performance. Cross-station modeling revealed an optimistic modeling strategy for learning transfer modeling of using information of nearby site.

Keywords Lake water level · Neurocomputing models · Lead time influence · Cross stations modeling

1 Introduction

The better understanding of lake water level (LWL) fluctuations can benefit for multiple applications of water resources management and the ecosystem [1, 2]. The changes in water level either for lakes, groundwater, or other water bodies can have a highly impact on socio-economic and environmental applications [3]. Naturally, catchments or basins have multiple sources of water inputs that might cause lakes water level rise and hence this requires much attention by hydrologists and climatologists to have more informative vision of the water mechanism and attempting to set a programming technology for LWL monitoring [4]. Worth to mention, on the other hand, decline of LWL due to like for example climate change can affect the lacustrine of the ecosystem [5, 6]. As a results, the accurate prediction of LWL can be considered as an essential element for the hydrological cycle understanding, catchment water balance, hydraulic structure design, groundwater level, contamination intrusion, flood control and several others [7]. In addition, although models containing hydrological and hydrometeorological variables such as precipitation, temperature, and evaporation can be found in the literature, it is economically more advantageous to use a model that simulates level changes based on historical level records [8, 9]. The motivation of the current research is to develop a computation data intelligence model with accurate prediction of LWL.

Based on the physical meaning, LWL fluctuation is caused by several hydrological and climatological

✉ Zaher Mundher Yaseen
z.yaseen@kfupm.edu.sa

Vahdettin Demir
vahdettin.demir@karatay.edu.tr

1 Faculty of Engineering and Natural Sciences, KTO Karatay University, Konya 42020, Turkey

2 Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, 31261 Dhahran, Saudi Arabia

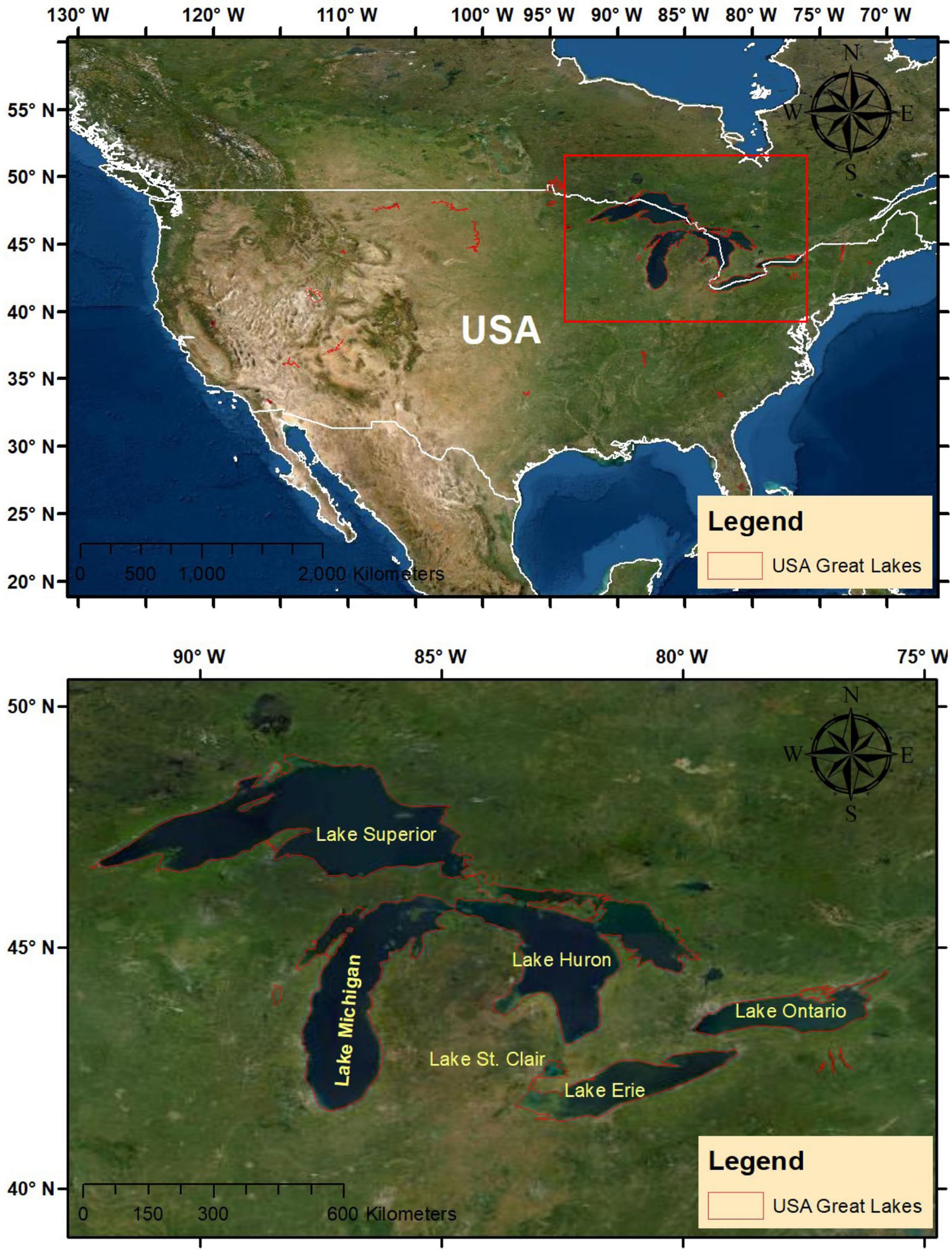


Fig. 1 Geographical location of the study area

processes experienced in particular catchment of basin [10]. Hence, the nature of this fluctuation is highly non-linear and stochastic and not easily can be comprehended. The literature evidenced the implications of several methodologies and among several, the water balance in which several parameters are implicated such as ground-water drops, catchment rainfall and runoff, inlet and outlet water discharge to the lake, water evaporation from the lake and several other causally impacted on LWL. However, this methodology is associated with several limitation such as time consume computationally, calculations errors, and required huge amount of data [11]. Hence, alternative methodologies for this purpose are highly recommended for easy catchment simulation [12]. The development of machine learning (ML) models for LWL forecasting and modeling have been adopted over the literature by several researchers [12]. For instance, support vector machine (SVM) [13], artificial neural network (ANN) [14], adaptive neuro-fuzzy inference system (ANFIS) [15], gene expression programming (GEP) [16], hybrid version of ANN using nature inspired optimization algorithm [17], conjugated ANFIS and SVM with wavelet preprocessing time series data [18], deep learning [19], minimax probability machine regression [20], random forest (RF) [21], extreme gradient boosting tree (EGBT) [22].

The discovery of new variant of ML models for forecasting LWL has been always the motive for researchers. This is inspired from the fact, there is no single ML model can be generalized as master for all types of LWL modeling. This is due to the known statement every ML model behave in a different way from one case to another. In addition, LWL is differ from one catchment to another and thus the stochasticity is totally varied. There are several ML models newly explored on their application within hydrological processes, among them, M5-Tree, multivariate adaptive regression spline (MARS) and least square support vector regression (LSSVR). They have been applied successfully in diverse hydrology processes such as river flow [23], rainfall [24], evaporation [25], drought [26], sediment transport [27], groundwater level [28] and several others [29].

The motivation of the current research was inspired from recognized gap of the adopted literature. The main research aims are (i) using of relatively new neurocomputing intelligence models (i.e., M5-Tree, MARS and LSSVR) for LWL forecasting at five great lakes located at USA, (ii) the models predictability performances were tested using several input combinations that incorporate lead time series data in addition to the periodicity of the time series data, (iii) cross-stations modeling procedure was investigated in this research for the purpose of using upstream dataset to forecast downstream LWL. A comprehensive assessment and evaluation were conducted for the initial research aim for the better understanding of the feasibility of the adopted methodology.

The reminder of the article as follows: Sect. 2 explained the case study and the utilized dataset. Sect. 3 reported the adopted ML models. Sect. 4 exhibited the modeling development procedure. Section 5 focused on the elaboration of the model results and analysis. Discussion of the obtained modeling results is revealed in Sect. 6. Finally, the research conclusion presented in the last section of the current article.

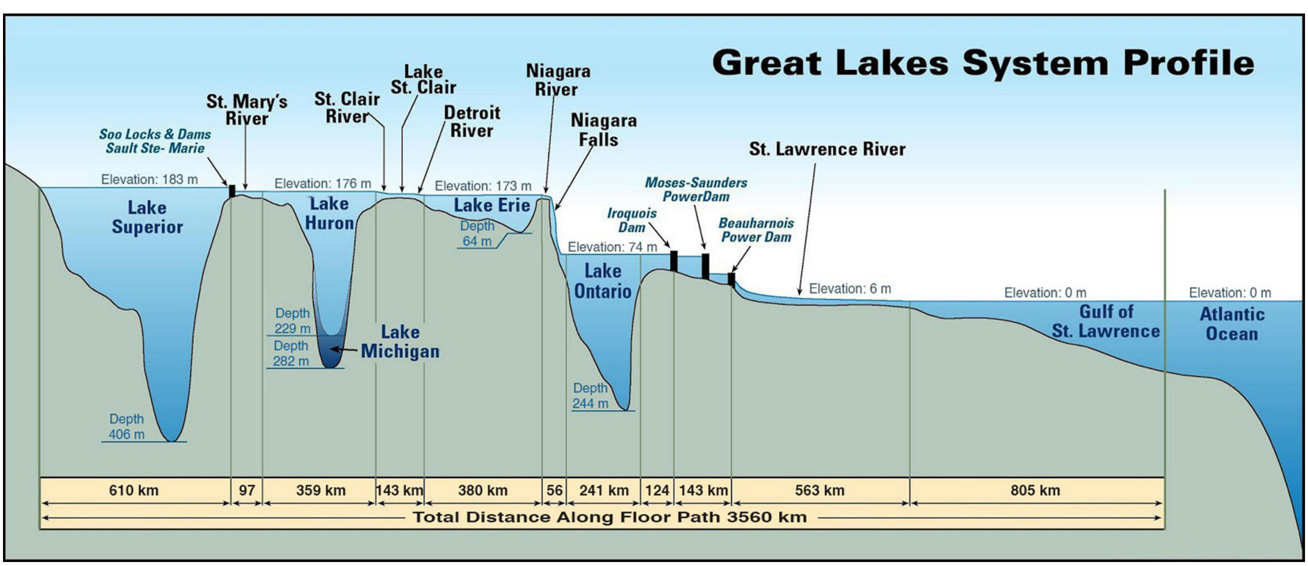


Fig. 2 Great lakes profile modified from (NOAA, 2021)

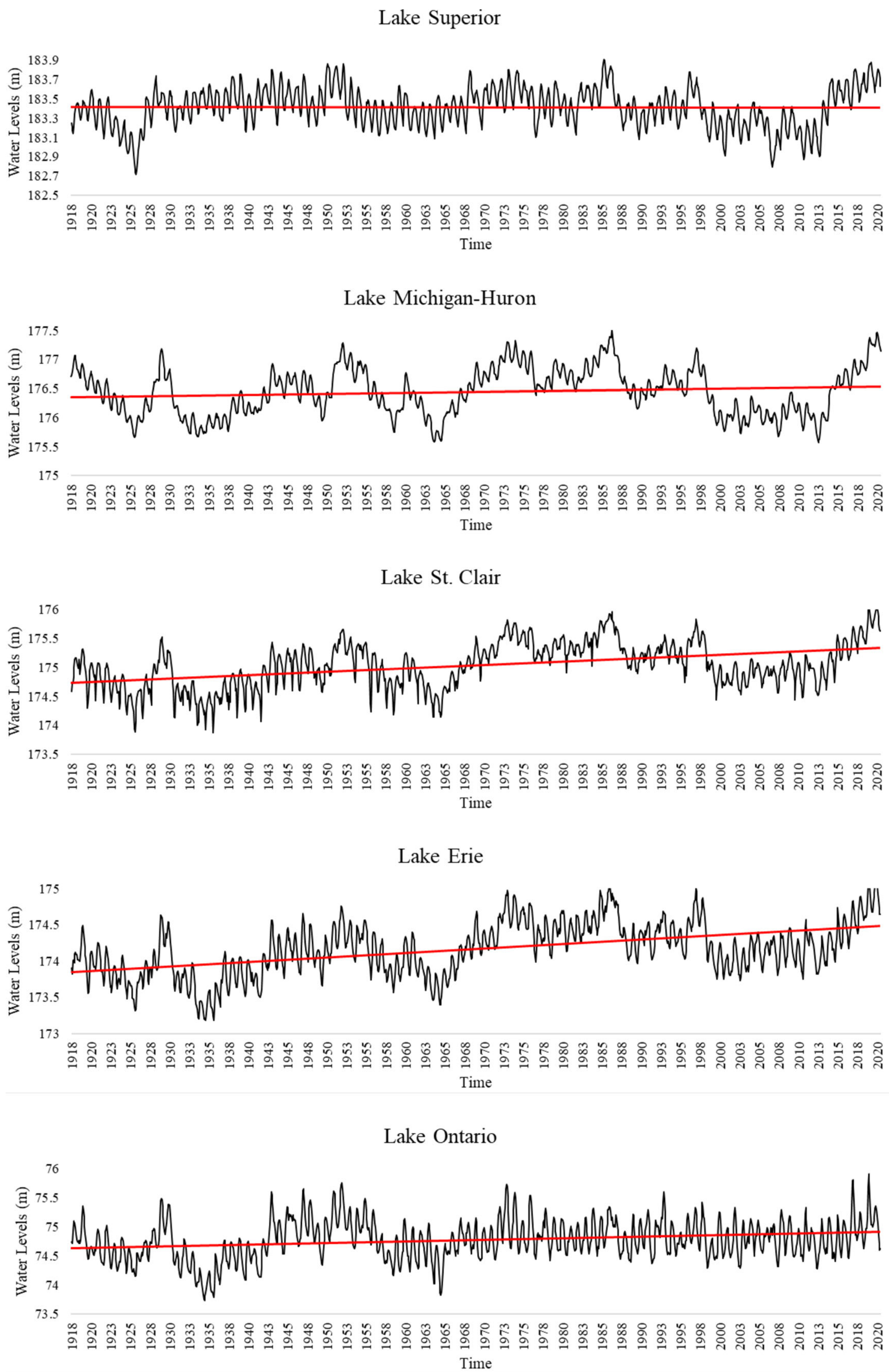


Fig. 3 The time series of the great lakes

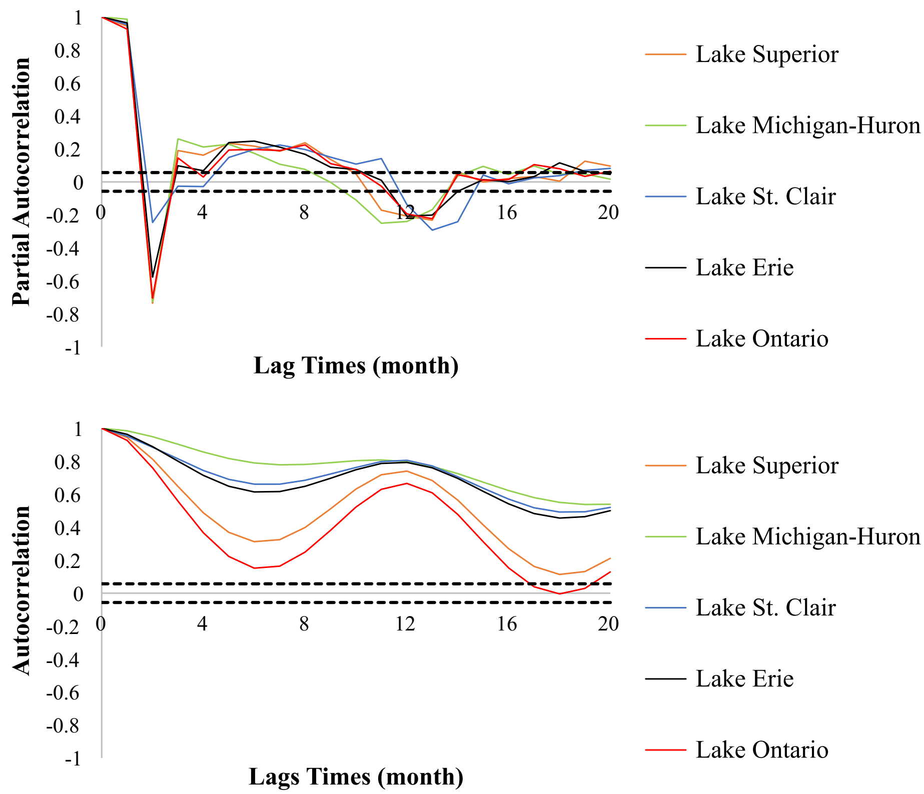


Fig. 4 Autocorrelation and partial auto-correlation function for the monthly lake levels

2 Case study and data description

The case study selceted for this investigation knows as the largest fresh lake surface water which is called as the Great lake that consisted of six lakes (Lake Ontario, Lake Erie, Lake St. Clair, Lake Michigan, Lake Huron and Lake Superior), as well as their connected canals [30]. This lakes cover 94,000 mi^2 and contain an anticipated 6 quadrillion gallons of water, accounting for nearly one-fifth of the world's fresh surface water supply and nine-tenths of the United States'. The Great Lakes provide drinking water to more than 40 million people in the United States and Canada. More than 1.5 million jobs and $60 billion in salaries are directly generated by the lakes each year. They also house over 3,500 animal species and plant, some of which are unique to the Great Lake area. The Great Lakes provide more than $52 billion in annual revenue for the region, thanks to world-class boating, hunting, and fishing options [31] The geographical position of the study area is given in Fig. 1.

The main characteristics of the lakes are as follows. Lake Superior is the world's largest freshwater lake, also the deepest and coldest of the Great Lakes, and various minerals such as copper, silver, gold, and nickel are mined around the lake. Although the Michigan-Huron lakes are separate lakes, the Mackinac Strait connects these two lakes. Manitoulin Island, the world's largest lake island, is surrounded by Lake Huron. Lake St. Clair is part of the Great Lakes system, and it connects Lake Huron (to the north) with Lake Erie via the St. Clair River and the Detroit River (to the south). Lake Erie is the southernmost located lake, shallow, can freeze in winter, and is the most polluted of the Great Lakes. Lake Ontario is the easternmost lake, has the smallest surface area, and the surrounding land is ideal for growing fruit [32]. The Great lakes profile is shown in Fig. 2.

Many systems in this region rely on forecasting changes in lake level. Flood control, reservoir management, water infrastructure management, commerce, drinking water distribution, coastal erosion, and transportation are just a few of the issues. In this study, monthly lake levels provided by the US Army Corps of Engineers for the years 1918–2020, a period of 103 years, are compiled for Lakes Superior, Huron-Michigan, St. Clair, Erie, and Ontario. Because the Lakes Huron and Michigan are connected by the Straits of Mackinac and have similar hydrologic characteristics, they are often referred to as Lake Michigan-Huron. All water levels (m) in this article are based on the International Great Lakes Datum 1985 (IGLD). The observed lake levels for Great Lakes are shown in Fig. 3. It should be noted that the lake level data used is continuous for all lakes and there is no data on missing monitoring events during the study period.

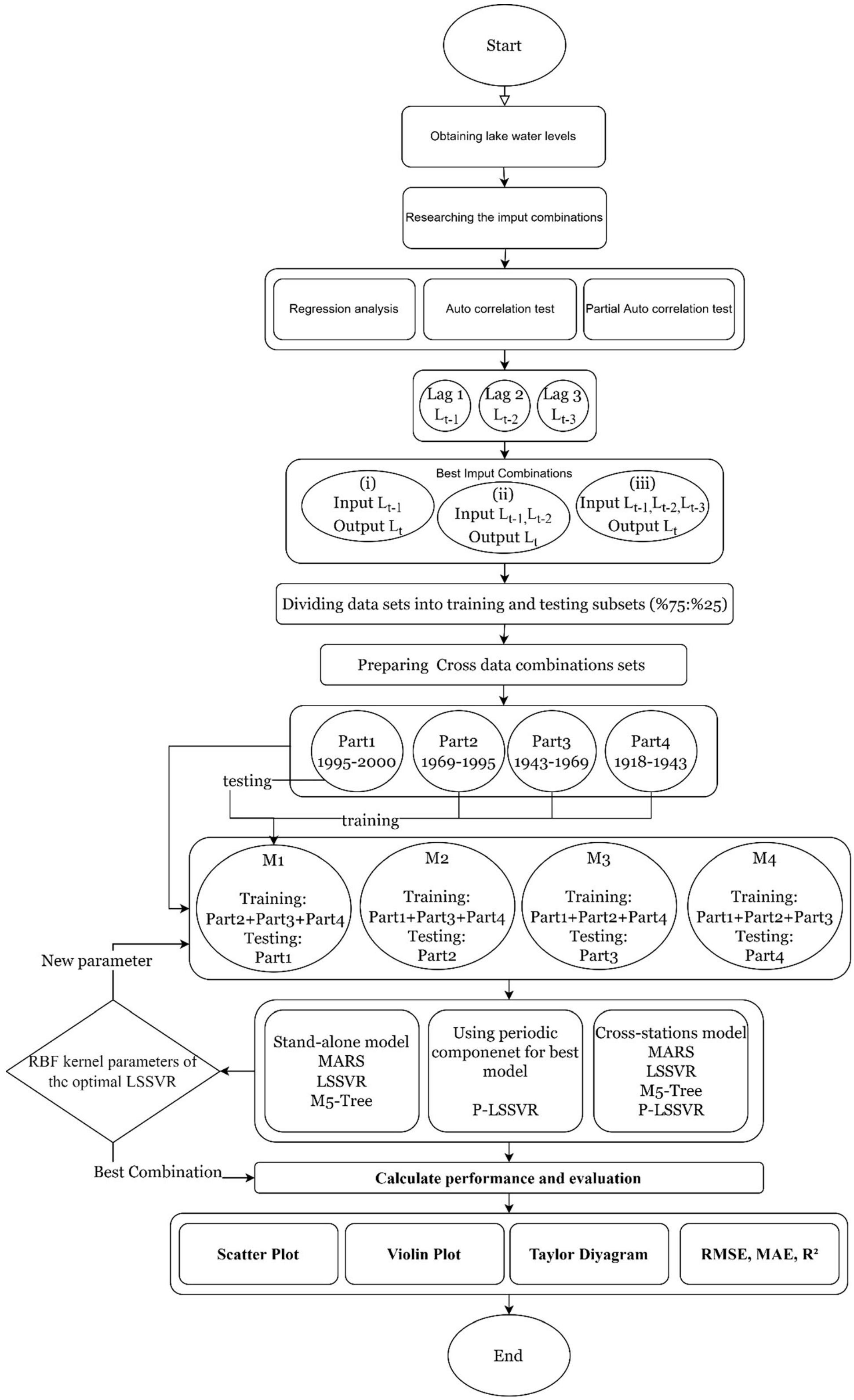


Fig. 5 The flow chart of the study

Table 1 The statistical performance of the developed MARS model

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
Lake Superior						
RMSE	M1	1995–2020	0.066	0.052	0.051	0.056
	M2	1969–1995	0.064	0.043	0.042	0.050
	M3	1943–1969	0.073	0.048	0.047	0.056
	M4	1918–1943	0.064	0.044	0.042	0.050
		Mean	0.067	0.047	0.046	0.053
MAE	M1	1995–2020	0.053	0.040	0.039	0.044
	M2	1969–1995	0.055	0.035	0.035	0.041
	M3	1943–1969	0.061	0.038	0.037	0.045
	M4	1918–1943	0.053	0.035	0.033	0.040
		Mean	0.056	0.037	0.036	0.043
R^2	M1	1995–2020	0.928	0.956	0.957	0.947
	M2	1969–1995	0.853	0.937	0.937	0.909
	M3	1943–1969	0.827	0.929	0.930	0.895
	M4	1918–1943	0.891	0.950	0.955	0.932
		Mean	0.875	0.943	0.945	0.921
Lake Michigan						
RMSE	M1	1995–2020	0.070	0.050	0.048	0.056
	M2	1969–1995	0.066	0.048	0.047	0.054
	M3	1943–1969	0.071	0.048	0.046	0.055
	M4	1918–1943	0.070	0.046	0.044	0.053
		Mean	0.069	0.048	0.046	0.054
MAE	M1	1995–2020	0.056	0.040	0.038	0.045
	M2	1969–1995	0.055	0.038	0.036	0.043
	M3	1943–1969	0.058	0.037	0.035	0.043
	M4	1918–1943	0.056	0.036	0.034	0.042
		Mean	0.056	0.038	0.036	0.043
R^2	M1	1995–2020	0.974	0.987	0.988	0.983
	M2	1969–1995	0.941	0.970	0.971	0.961
	M3	1943–1969	0.960	0.983	0.984	0.976
	M4	1918–1943	0.960	0.983	0.984	0.976
		Mean	0.959	0.981	0.982	0.974
Lake St. Clair						
RMSE	M1	1995–2020	0.101	0.096	0.092	0.096
	M2	1969–1995	0.093	0.101	0.099	0.097
	M3	1943–1969	0.130	0.124	0.125	0.126
	M4	1918–1943	0.151	0.150	0.153	0.151
		Mean	0.119	0.118	0.117	0.118
MAE	M1	1995–2020	0.080	0.073	0.069	0.074
	M2	1969–1995	0.069	0.069	0.067	0.069
	M3	1943–1969	0.094	0.088	0.088	0.090
	M4	1918–1943	0.111	0.109	0.111	0.110
		Mean	0.089	0.085	0.084	0.086
R^2	M1	1995–2020	0.923	0.929	0.937	0.929

Table 1 (continued)

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
	M2	1969–1995	0.842	0.822	0.843	0.836
	M3	1943–1969	0.835	0.851	0.849	0.845
	M4	1918–1943	0.784	0.785	0.780	0.783
		Mean	0.846	0.847	0.852	0.848
Lake Erie						
RMSE	M1	1995–2020	0.105	0.085	0.078	0.089
	M2	1969–1995	0.096	0.092	0.088	0.092
	M3	1943–1969	0.101	0.079	0.079	0.086
	M4	1918–1943	0.100	0.130	0.067	0.099
		Mean	0.101	0.096	0.078	0.092
MAE	M1	1995–2020	0.086	0.066	0.062	0.071
	M2	1969–1995	0.077	0.073	0.069	0.073
	M3	1943–1969	0.081	0.060	0.060	0.067
	M4	1918–1943	0.078	0.097	0.050	0.075
		Mean	0.080	0.074	0.060	0.072
R^2	M1	1995–2020	0.894	0.930	0.942	0.922
	M2	1969–1995	0.840	0.870	0.881	0.864
	M3	1943–1969	0.868	0.921	0.921	0.903
	M4	1918–1943	0.884	0.844	0.950	0.893
		Mean	0.872	0.891	0.923	0.895
Lake Ontario						
RMSE	M1	1995–2020	0.142	0.098	0.098	0.113
	M2	1969–1995	0.137	0.096	0.098	0.110
	M3	1943–1969	0.128	0.091	0.089	0.103
	M4	1918–1943	0.116	0.091	0.09	0.099
		Mean	0.131	0.094	0.094	0.106
MAE	M1	1995–2020	0.117	0.078	0.078	0.091
	M2	1969–1995	0.111	0.076	0.077	0.088
	M3	1943–1969	0.105	0.073	0.071	0.083
	M4	1918–1943	0.089	0.068	0.068	0.075
		Mean	0.106	0.074	0.074	0.084
R^2	M1	1995–2020	0.770	0.896	0.895	0.854
	M2	1969–1995	0.773	0.894	0.889	0.852
	M3	1943–1969	0.871	0.936	0.939	0.915
	M4	1918–1943	0.894	0.943	0.943	0.927
		Mean	0.827	0.917	0.917	0.887

Table 2 The statistical performance of the developed M5-Tree model

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
Lake Superior						
RMSE	M1	1995–2020	0.066	0.051	0.056	0.058
	M2	1969–1995	0.064	0.043	0.046	0.051
	M3	1943–1969	0.073	0.049	0.049	0.057
	M4	1918–1943	0.065	0.043	0.043	0.050
		Mean	0.067	0.046	0.048	0.054
MAE	M1	1995–2020	0.053	0.039	0.043	0.045
	M2	1969–1995	0.055	0.034	0.035	0.042
	M3	1943–1969	0.061	0.038	0.038	0.046
	M4	1918–1943	0.054	0.033	0.034	0.040
		Mean	0.056	0.036	0.038	0.043
R^2	M1	1995–2020	0.929	0.958	0.946	0.945
	M2	1969–1995	0.853	0.936	0.929	0.906
	M3	1943–1969	0.827	0.925	0.926	0.893
	M4	1918–1943	0.888	0.952	0.952	0.931
		Mean	0.874	0.943	0.938	0.919
Lake Michigan						
RMSE	M1	1995–2020	0.074	0.050	0.045	0.056
	M2	1969–1995	0.074	0.048	0.047	0.056
	M3	1943–1969	0.071	0.052	0.048	0.057
	M4	1918–1943	0.074	0.049	0.044	0.056
		Mean	0.073	0.050	0.046	0.056
MAE	M1	1995–2020	0.059	0.039	0.036	0.045
	M2	1969–1995	0.059	0.038	0.036	0.045
	M3	1943–1969	0.057	0.040	0.036	0.044
	M4	1918–1943	0.059	0.038	0.034	0.044
		Mean	0.059	0.039	0.036	0.044
R^2	M1	1995–2020	0.971	0.987	0.989	0.982
	M2	1969–1995	0.925	0.970	0.971	0.956
	M3	1943–1969	0.961	0.979	0.982	0.974
	M4	1918–1943	0.954	0.980	0.984	0.973
		Mean	0.953	0.979	0.982	0.971
Lake St. Clair						
RMSE	M1	1995–2020	0.101	0.100	0.109	0.103
	M2	1969–1995	0.093	0.103	0.106	0.100
	M3	1943–1969	0.131	0.132	0.138	0.134
	M4	1918–1943	0.146	0.156	0.157	0.153
		Mean	0.118	0.123	0.127	0.123
MAE	M1	1995–2020	0.080	0.076	0.080	0.078
	M2	1969–1995	0.068	0.071	0.076	0.072
	M3	1943–1969	0.093	0.094	0.095	0.094
	M4	1918–1943	0.108	0.116	0.114	0.113
		Mean	0.087	0.089	0.091	0.089
R^2	M1	1995–2020	0.922	0.924	0.916	0.920

Table 2 (continued)

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
	M2	1969–1995	0.843	0.826	0.823	0.830
	M3	1943–1969	0.833	0.835	0.821	0.830
	M4	1918–1943	0.790	0.765	0.755	0.770
		Mean	0.847	0.837	0.829	0.838
Lake Erie						
RMSE	M1	1995–2020	0.096	0.087	0.087	0.090
	M2	1969–1995	0.092	0.085	0.097	0.092
	M3	1943–1969	0.103	0.087	0.094	0.095
	M4	1918–1943	0.100	0.097	0.095	0.097
		Mean	0.098	0.089	0.093	0.094
MAE	M1	1995–2020	0.081	0.068	0.067	0.072
	M2	1969–1995	0.075	0.067	0.075	0.072
	M3	1943–1969	0.083	0.068	0.073	0.075
	M4	1918–1943	0.078	0.075	0.073	0.075
		Mean	0.079	0.070	0.072	0.074
R^2	M1	1995–2020	0.909	0.927	0.928	0.921
	M2	1969–1995	0.850	0.890	0.863	0.868
	M3	1943–1969	0.861	0.906	0.892	0.886
	M4	1918–1943	0.884	0.905	0.909	0.900
		Mean	0.876	0.907	0.898	0.894
Lake Ontario						
RMSE	M1	1995–2020	0.146	0.108	0.109	0.121
	M2	1969–1995	0.136	0.108	0.115	0.119
	M3	1943–1969	0.135	0.098	0.113	0.115
	M4	1918–1943	0.118	0.098	0.102	0.106
		Mean	0.134	0.103	0.110	0.115
MAE	M1	1995–2020	0.119	0.084	0.085	0.096
	M2	1969–1995	0.111	0.083	0.088	0.094
	M3	1943–1969	0.108	0.077	0.085	0.090
	M4	1918–1943	0.090	0.073	0.077	0.080
		Mean	0.107	0.079	0.084	0.090
R^2	M1	1995–2020	0.757	0.873	0.871	0.834
	M2	1969–1995	0.776	0.877	0.852	0.835
	M3	1943–1969	0.858	0.929	0.904	0.897
	M4	1918–1943	0.887	0.933	0.925	0.915
		Mean	0.820	0.903	0.888	0.870

Table 3 The statistical performance of the developed LSSVR model

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
Lake Superior						
RMSE	M1	1995–2020	0.066	0.050	0.048	0.054
	M2	1969–1995	0.064	0.042	0.042	0.049
	M3	1943–1969	0.073	0.049	0.045	0.056
	M4	1918–1943	0.064	0.042	0.040	0.049
		Mean	0.067	0.046	0.044	0.052
MAE	M1	1995–2020	0.053	0.038	0.037	0.043
	M2	1969–1995	0.055	0.034	0.034	0.041
	M3	1943–1969	0.061	0.039	0.036	0.045
	M4	1918–1943	0.053	0.034	0.032	0.040
		Mean	0.056	0.036	0.035	0.042
R^2	M1	1995–2020	0.929	0.963	0.964	0.952
	M2	1969–1995	0.853	0.939	0.939	0.910
	M3	1943–1969	0.827	0.926	0.935	0.896
	M4	1918–1943	0.891	0.954	0.958	0.934
		Mean	0.875	0.945	0.949	0.923
Lake Michigan						
RMSE	M1	1995–2020	0.070	0.048	0.046	0.054
	M2	1969–1995	0.066	0.048	0.046	0.053
	M3	1943–1969	0.069	0.047	0.045	0.054
	M4	1918–1943	0.068	0.045	0.043	0.052
		Mean	0.068	0.047	0.045	0.053
MAE	M1	1995–2020	0.056	0.038	0.036	0.044
	M2	1969–1995	0.054	0.038	0.036	0.043
	M3	1943–1969	0.055	0.037	0.035	0.042
	M4	1918–1943	0.055	0.036	0.034	0.042
		Mean	0.055	0.037	0.035	0.043
R^2	M1	1995–2020	0.974	0.988	0.989	0.984
	M2	1969–1995	0.941	0.971	0.973	0.962
	M3	1943–1969	0.963	0.983	0.984	0.976
	M4	1918–1943	0.962	0.983	0.985	0.977
		Mean	0.960	0.981	0.983	0.975
Lake St. Clair						
RMSE	M1	1995–2020	0.100	0.090	0.092	0.094
	M2	1969–1995	0.092	0.094	0.094	0.093
	M3	1943–1969	0.130	0.121	0.118	0.123
	M4	1918–1943	0.144	0.139	0.135	0.140
		Mean	0.117	0.111	0.110	0.112
MAE	M1	1995–2020	0.079	0.068	0.068	0.072
	M2	1969–1995	0.068	0.068	0.067	0.068
	M3	1943–1969	0.093	0.085	0.082	0.087
	M4	1918–1943	0.107	0.100	0.097	0.101
		Mean	0.087	0.080	0.079	0.082
R^2	M1	1995–2020	0.922	0.939	0.936	0.932

Table 3 (continued)

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
	M2	1969–1995	0.841	0.841	0.847	0.843
	M3	1943–1969	0.835	0.858	0.865	0.853
	M4	1918–1943	0.797	0.821	0.831	0.816
		Mean	0.849	0.865	0.870	0.861
Lake Erie						
RMSE	M1	1995–2020	0.097	0.076	0.076	0.083
	M2	1969–1995	0.091	0.081	0.081	0.085
	M3	1943–1969	0.100	0.077	0.076	0.084
	M4	1918–1943	0.101	0.086	0.084	0.090
		Mean	0.097	0.080	0.079	0.085
MAE	M1	1995–2020	0.081	0.060	0.060	0.067
	M2	1969–1995	0.074	0.066	0.066	0.069
	M3	1943–1969	0.080	0.059	0.058	0.066
	M4	1918–1943	0.078	0.065	0.063	0.069
		Mean	0.078	0.062	0.062	0.067
R^2	M1	1995–2020	0.909	0.944	0.944	0.932
	M2	1969–1995	0.850	0.891	0.893	0.878
	M3	1943–1969	0.868	0.925	0.927	0.907
	M4	1918–1943	0.884	0.926	0.929	0.913
		Mean	0.878	0.922	0.923	0.907
Lake Ontario						
RMSE	M1	1995–2020	0.142	0.092	0.093	0.109
	M2	1969–1995	0.136	0.089	0.090	0.105
	M3	1943–1969	0.128	0.088	0.086	0.101
	M4	1918–1943	0.114	0.088	0.085	0.096
		Mean	0.130	0.089	0.089	0.103
MAE	M1	1995–2020	0.117	0.074	0.074	0.088
	M2	1969–1995	0.111	0.071	0.070	0.084
	M3	1943–1969	0.105	0.070	0.068	0.081
	M4	1918–1943	0.087	0.066	0.064	0.072
		Mean	0.105	0.070	0.069	0.081
R^2	M1	1995–2020	0.770	0.907	0.905	0.860
	M2	1969–1995	0.776	0.908	0.906	0.863
	M3	1943–1969	0.871	0.942	0.945	0.919
	M4	1918–1943	0.895	0.948	0.951	0.931
		Mean	0.828	0.926	0.927	0.894

Table 4 The statistical performance of the developed P-LSSVR model

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
Lake Superior						
RMSE	M1	1995–2020	0.039	0.037	0.037	0.038
	M2	1969–1995	0.036	0.034	0.034	0.035
	M3	1943–1969	0.038	0.035	0.035	0.036
	M4	1918–1943	0.034	0.032	0.032	0.033
		Mean	0.037	0.034	0.034	0.035
MAE	M1	1995–2020	0.029	0.028	0.028	0.028
	M2	1969–1995	0.028	0.026	0.027	0.027
	M3	1943–1969	0.028	0.027	0.027	0.027
	M4	1918–1943	0.027	0.025	0.025	0.025
		Mean	0.028	0.026	0.027	0.027
R^2	M1	1995–2020	0.975	0.978	0.977	0.977
	M2	1969–1995	0.953	0.960	0.959	0.957
	M3	1943–1969	0.954	0.960	0.961	0.958
	M4	1918–1943	0.969	0.974	0.974	0.972
		Mean	0.963	0.968	0.968	0.966
Lake Michigan						
RMSE	M1	1995–2020	0.042	0.036	0.036	0.038
	M2	1969–1995	0.041	0.036	0.037	0.038
	M3	1943–1969	0.042	0.035	0.036	0.038
	M4	1918–1943	0.038	0.034	0.034	0.035
		Mean	0.041	0.035	0.036	0.037
MAE	M1	1995–2020	0.034	0.029	0.030	0.031
	M2	1969–1995	0.031	0.028	0.029	0.029
	M3	1943–1969	0.033	0.028	0.028	0.030
	M4	1918–1943	0.030	0.026	0.027	0.028
		Mean	0.032	0.028	0.028	0.030
R^2	M1	1995–2020	0.991	0.993	0.993	0.992
	M2	1969–1995	0.977	0.982	0.981	0.980
	M3	1943–1969	0.986	0.990	0.990	0.989
	M4	1918–1943	0.988	0.990	0.991	0.990
		Mean	0.985	0.989	0.989	0.988
Lake St. Clair						
RMSE	M1	1995–2020	0.076	0.078	0.077	0.077
	M2	1969–1995	0.081	0.079	0.079	0.080
	M3	1943–1969	0.095	0.094	0.091	0.093
	M4	1918–1943	0.101	0.104	0.101	0.102
		Mean	0.088	0.089	0.087	0.088
MAE	M1	1995–2020	0.056	0.058	0.058	0.057
	M2	1969–1995	0.055	0.055	0.055	0.055
	M3	1943–1969	0.062	0.063	0.059	0.061
	M4	1918–1943	0.069	0.071	0.070	0.070
		Mean	0.061	0.062	0.060	0.061

Table 4 (continued)

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
R^2	M1	1995–2020	0.955	0.953	0.954	0.954
	M2	1969–1995	0.879	0.890	0.885	0.885
	M3	1943–1969	0.911	0.912	0.920	0.914
	M4	1918–1943	0.898	0.893	0.896	0.895
		Mean	0.911	0.912	0.914	0.912
Lake Erie						
RMSE	M1	1995–2020	0.064	0.061	0.061	0.062
	M2	1969–1995	0.065	0.065	0.064	0.065
	M3	1943–1969	0.063	0.060	0.060	0.061
	M4	1918–1943	0.068	0.064	0.064	0.065
		Mean	0.065	0.063	0.062	0.063
MAE	M1	1995–2020	0.049	0.047	0.047	0.048
	M2	1969–1995	0.050	0.050	0.050	0.050
	M3	1943–1969	0.048	0.045	0.045	0.046
	M4	1918–1943	0.050	0.047	0.047	0.048
		Mean	0.049	0.047	0.047	0.048
R^2	M1	1995–2020	0.960	0.963	0.963	0.962
	M2	1969–1995	0.926	0.925	0.927	0.926
	M3	1943–1969	0.948	0.953	0.953	0.951
	M4	1918–1943	0.947	0.953	0.953	0.951
		Mean	0.945	0.949	0.949	0.948
Lake Ontario						
RMSE	M1	1995–2020	0.090	0.080	0.079	0.083
	M2	1969–1995	0.079	0.072	0.073	0.075
	M3	1943–1969	0.085	0.074	0.074	0.078
	M4	1918–1943	0.077	0.069	0.068	0.071
		Mean	0.083	0.074	0.073	0.077
MAE	M1	1995–2020	0.071	0.062	0.061	0.064
	M2	1969–1995	0.062	0.056	0.057	0.058
	M3	1943–1969	0.068	0.058	0.058	0.061
	M4	1918–1943	0.058	0.053	0.051	0.054
		Mean	0.065	0.057	0.057	0.060
R^2	M1	1995–2020	0.906	0.927	0.928	0.921
	M2	1969–1995	0.924	0.936	0.935	0.932
	M3	1943–1969	0.944	0.957	0.957	0.953
	M4	1918–1943	0.952	0.964	0.965	0.960
		Mean	0.931	0.946	0.946	0.941

Weather extremes and seasonal oscillations in general have had an impact on the temporal variability of lake levels, which characterize the annual hydrologic cycle from winter lows to summer highs. The temporal fluctuations shown in Fig. 3 indicate that the lake level series are not stationary [32]. The characteristics of the Great Lakes along with the statistics (period, maximum, mean, minimum, skewness coefficient (Cs) standard deviation (Sx)) of the water levels are given in Appendix 1. Pearson correlation coefficients between lake water levels are shown in Appendix 2.

The ideal model input scenario was determined using auto-correlation and partial auto-correlation functions. The auto-correlation functions and partial auto-correlation of the lake levels for Great Lakes are given in Fig. 4. Lake levels are significantly connected with previous month levels, as shown in the graph. For the Great Lakes, the partial autocorrelation function indicates a substantial association until lag3 and then stays within the confidence interval. Consequently, to simulate the L_t outflow, the input combination considered in analyzing the lake level process is L_{t-1}, L_{t-2} and L_{t-3}.

3 Applied machine learning models

3.1 M5-Tree

Quinlan developed the M5-Tree technique in 1992 as a novel regression method [33]. This model's backbone is a two-component decision tree. The technique describes the connection between variables by applying the linear function to the last leaf nodes. The M5-Tree outperforms traditional tree models for data that are similar or related in any way [34].

The M5-Tree tree consists of 2 stages. To create the decision schema, the data is separated into subsets in the first stage. To categorize clases, the class value's standard deviation attained at a node is employed. The error that occurs when the elements acting on this node are tested is used to calculate the predicted decrease [35, 36]. The following equation shows how the standard deviation reduction (SDR) is calculated.

$$SDR = sd(T) - \sum \frac{|Ti|}{|T|} sd(Ti) \quad (1)$$

"sd" stands for standard deviation in this formula. T is a set of instances that act on the node. Ti represents subset samples. These subset samples "i" belong to potential data findings. [33].

3.2 MARS

The MARS model was proposed by Friedman [37]. MARS is a model for forecasting nonlinear numeric outputs that are continuous. There are two elements to the MARS algorithm: forward and backward steps. The forward step method is used to pick a collection of relevant input variables [38]. It removes extraneous variables in the pre-selected collection using the backward step method. The following fundamental equations are used to draw a function from variable X (input) to variable Y (output). The new Y values are obtained using either the two base functions defined at the deviation point on the input range, or both variable values. [39].

$$Y = \max(0, X - c) \quad (2)$$

$$Y = \max(0, c - x) \quad (3)$$

The lower limit (threshold) value is denoted by c. In management and planning systems, time series data, and a variety of other disciplines, the MARS model is widely utilized [40–43].

3.3 LSSVR

Suykens and Vandewalle developed the LSSVR model in 1999 as an extension of the Support Vector Regression (SVR) [44]. It is used to find the best function between the input (X) and output (Y) by statistically comparing current water levels to water levels in previous time series [23]. It achieves this procedure using a multidimensional feature space and a nonlinear relationship function. The regression function can be expressed in the following way.

$$y(x) = w^1 \varphi(x) + b \quad (4)$$

Here, w is the coefficient vector, y is the output value, x is the input paramters, b is the bias term [44].

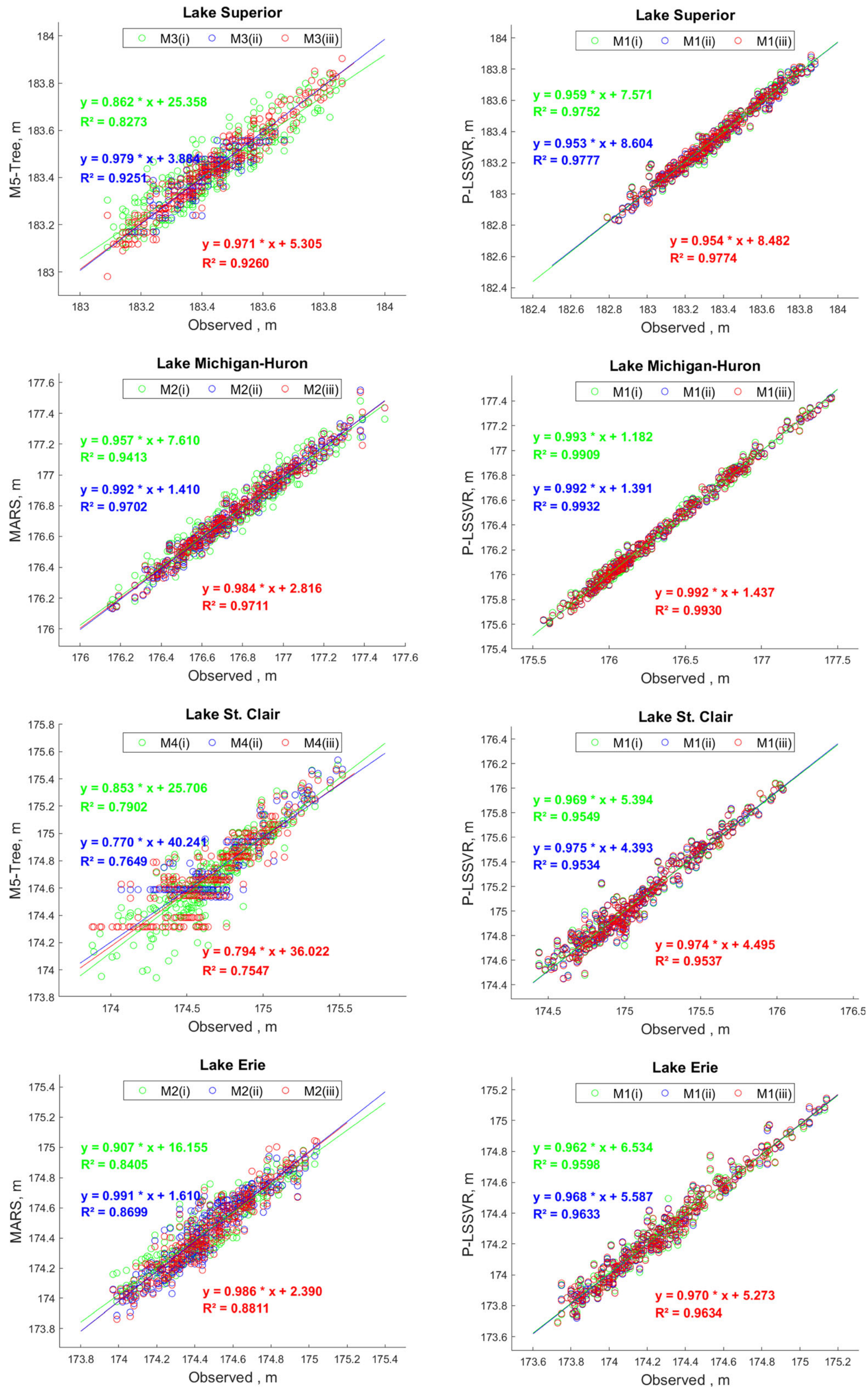


Fig. 6 Scatter plots of the observed and predicted lake level values during testing phase, produced by models for the great lakes; a worst b best

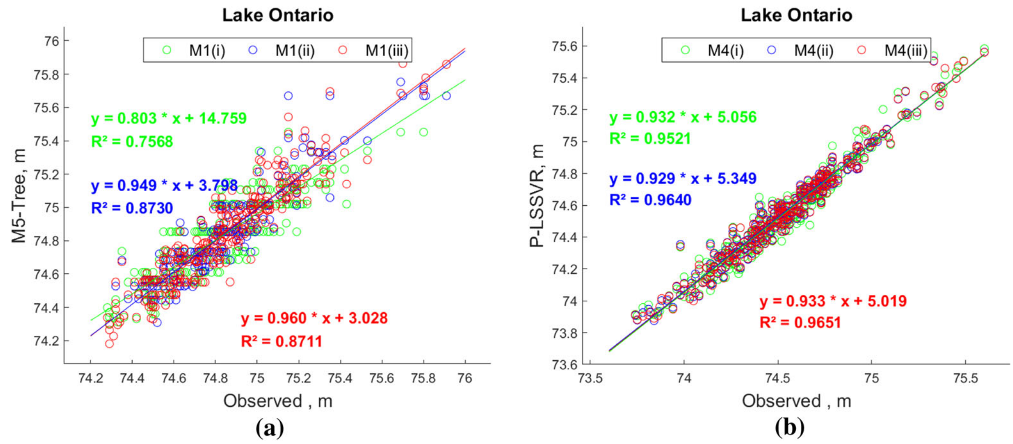


Fig. 6 continued

4 Modeling development

The effectiveness of the proposed neurocomputing intelligence approaches was investigated using data on actual LWLs obtained from authorized official organizations. The efficacy of the models in predicting the lake level for one month ahead was tested in the first part of this study. The influence of the periodical component of time series dataset on the forecasting performance was also inspected. The applicability of the data-driven prediction for the lake levels is investigated using time series data from the upstream lake station. Various input combinations based on present and preceding lake water levels were used to model the forecast. In other words, L_t gives the level of the lake at time "t", and the input variables are: L_{t-1} (i), L_{t-1}, L_{t-2} (ii), L_{t-1}, L_{t-2} and L_{t-3} (iii).

To obtain the most successful model formulation, LWL were split into four divisions (training and testing) for the Great lakes. Three data splits were utilized to initiate the modeling development on the training phase for both forecasting and cross-stations predicting. Whereas, the fourth data division was used to test the applied models. In all applications, the test dataset was varied; as a result, four different scenarios were studied.

MARS, M5-Tree and LSSVR methods were used for modeling, and Taylor and Violin diagrams were used to evaluate the results. The LSSVR model was created using open-source software [44]. For sigma and gamma values, different numbers varying from 1 to 100 in increments of 1 were tried, the parameters giving the lowest RMSE value were accepted as the best model parameter, and the RBF kernel was used in the LSSVR model. There are no control parameters in the M5-Tree and MARS models. These methods were also implemented using open-source software [45]. For all presented stations, lake-level data series were split into four training/testing divisions to achieve the best effective model. Three divisions of the data were used to train the models for both forecasting and predicting, while the fourth was used to validate (test) the model's network [22]. The testing data phase was changed in all applications; therefore, four different scenarios were investigated. For the Taylor and the Violin diagrams, an open-source MATLAB code [46] and [47] was used, respectively. The flow chart of the study is given in Fig. 5.

Evaluating hydrological applications, quantitative indicators are frequently used [48, 49]. According to Legates and McCabe [50], "goodness-of-fit" e.g., coefficient of determination (R^2) and error performance criteria (such as root mean square error (RMSE) and mean absolute error (MAE)) should be used to evaluate predictive models in hydrology (MAE). For each input combination, the suggested models were evaluated in terms of R^2, MAE and RMSE. The linear correlation between estimated and observed values is measured by R^2, which runs from − 1 to

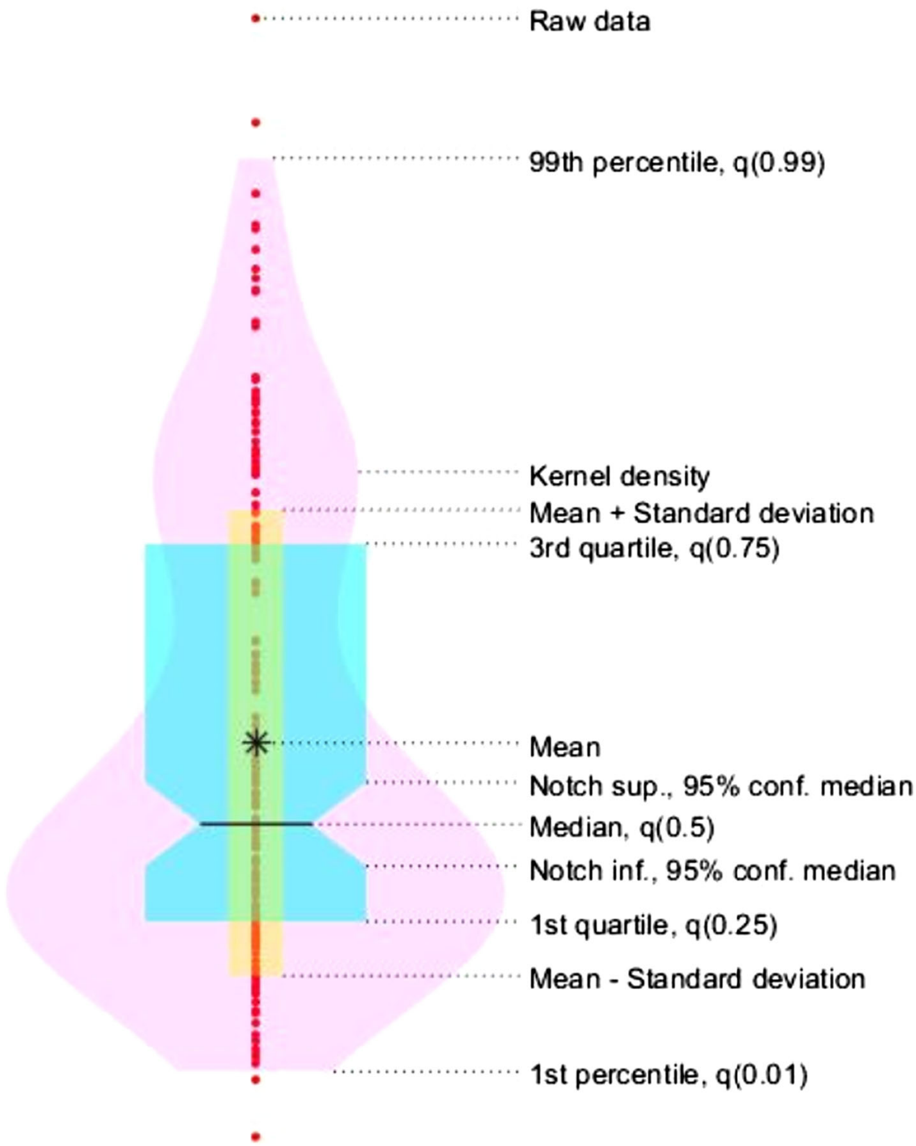


Fig. 7 Violin diagram statistical parameters

1 [51]. Values of 1 and 0 imply an ideal match and no statistical correlation, respectively. By squaring the errors, the RMSE is utilized to estimate prediction precision, resulting in a positive number. When the differences between predictions and observations grow significant, thse RMSE rises from zero for perfect predictions to huge positive values. The MAE measures the average magnitude of the errors in a set of predict, without considering their direction. When R^2, RMSE, and MAE are near to 1, 0, and 0, respectively, the best model forecasts are obtained [52]. The mathematical expression of the performance metrics can be written as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(L_e - L_o)^2} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|L_e - L_o| \tag{6}$$

$$R^2 = \left(\frac{N * (\sum L_o * L_e) - (\sum L_o) * (\sum L_e)}{\sqrt{\left[N * \sum L_o^2 - (\sum L_e)^2\right] * \left[N * \sum L_e^2 - (\sum L_e)^2\right]}}\right)^2 \tag{7}$$

Table 5 Comparison of the LSSVR, MARS, M5-Tree models in predicting monthly lake levels of the Michigan Station by using the data of Superior station

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
LSSVR	RMSE	M1	1995–2020	0.249	0.247	0.238	0.245
		M2	1969–1995	0.381	0.381	0.377	0.379
		M3	1943–1969	0.320	0.323	0.324	0.323
		M4	1918–1943	0.424	0.428	0.430	0.427
			Mean	0.343	0.345	0.343	0.344
	MAE	M1	1995–2020	0.203	0.201	0.193	0.199
		M2	1969–1995	0.344	0.344	0.342	0.343
		M3	1943–1969	0.257	0.261	0.263	0.260
		M4	1918–1943	0.342	0.346	0.348	0.345
			Mean	0.287	0.288	0.286	0.287
	R^2	M1	1995–2020	0.754	0.724	0.737	0.739
		M2	1969–1995	0.620	0.623	0.640	0.628
		M3	1943–1969	0.264	0.256	0.258	0.260
		M4	1918–1943	0.082	0.080	0.075	0.079
			Mean	0.430	0.421	0.428	0.426
MARS	RMSE	M1	1995–2020	0.254	0.254	0.250	0.253
		M2	1969–1995	0.383	0.383	0.381	0.382
		M3	1943–1969	0.328	0.324	0.326	0.326
		M4	1918–1943	0.434	0.432	0.435	0.433
			Mean	0.350	0.348	0.348	0.349
	MAE	M1	1995–2020	0.209	0.210	0.205	0.208
		M2	1969–1995	0.348	0.348	0.347	0.347
		M3	1943–1969	0.263	0.261	0.264	0.263
		M4	1918–1943	0.351	0.351	0.354	0.352
			Mean	0.293	0.292	0.292	0.293
	R^2	M1	1995–2020	0.751	0.752	0.752	0.751
		M2	1969–1995	0.627	0.631	0.636	0.631
		M3	1943–1969	0.258	0.264	0.260	0.260
		M4	1918–1943	0.074	0.085	0.069	0.076
			Mean	0.427	0.433	0.429	0.430
M5-Tree	RMSE	M1	1995–2020	0.249	0.260	0.279	0.263
		M2	1969–1995	0.381	0.396	0.395	0.391
		M3	1943–1969	0.334	0.348	0.370	0.351
		M4	1918–1943	0.441	0.445	0.462	0.450
			Mean	0.352	0.363	0.376	0.364
	MAE	M1	1995–2020	0.203	0.217	0.227	0.216
		M2	1969–1995	0.346	0.347	0.342	0.345
		M3	1943–1969	0.266	0.278	0.294	0.279
		M4	1918–1943	0.362	0.369	0.385	0.372
			Mean	0.294	0.303	0.312	0.303
	R^2	M1	1995–2020	0.738	0.685	0.599	0.674
		M2	1969–1995	0.622	0.456	0.437	0.505
		M3	1943–1969	0.244	0.216	0.150	0.203
		M4	1918–1943	0.063	0.057	0.044	0.054
			Mean	0.417	0.354	0.307	0.359

Table 5 (continued)

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
P-LSSVR	RMSE	M1	1995–2020	0.236	0.238	0.239	0.237
		M2	1969–1995	0.374	0.374	0.375	0.374
		M3	1943–1969	0.319	0.321	0.322	0.321
		M4	1918–1943	0.426	0.432	0.435	0.431
			Mean	0.339	0.341	0.342	0.341
	MAE	M1	1995–2020	0.189	0.193	0.195	0.192
		M2	1969–1995	0.339	0.340	0.341	0.340
		M3	1943–1969	0.257	0.261	0.262	0.260
		M4	1918–1943	0.346	0.351	0.354	0.350
			Mean	0.283	0.286	0.288	0.286
	R^2	M1	1995–2020	0.774	0.772	0.770	0.772
		M2	1969–1995	0.652	0.658	0.660	0.657
		M3	1943–1969	0.271	0.268	0.263	0.267
		M4	1918–1943	0.072	0.065	0.061	0.066
			Mean	0.442	0.441	0.439	0.440

Here, N represent number of lake level data, L_o denotes the actual (observed) lake level values, and L_e denotes the model output (estimation).

5 Applications results and analysis

5.1 Lake water level prediction using stand-alone models

In this subsection, the prediction analysis for the adopted three neurocomputing intelligence models (i.e., M5-Tree, MARS, LSSVR) was reported for each investigated lake. It is essential for the readers to comprehend the pattern of the investigated LWL of the current research case study and thus, Appendix 3 reports the statistical characteristics including skewness, mean, min and max records, standard deviation and the antecedent correlation values for each investigated lake.

The first scenario was used to forecast monthly lake levels, as described in the previous section. The input combinations were cross validated through data time series segmentation "four sets" in which the statistical analysis was adopted for each independent data collection. As the regularization of the learning function highly influencing the learning process of the network, several regularizations for the radial basis function kernel were tested to attain the minimum RMSE indication by recalling the major parameters of the LSSVR model. For the testing phase, Appendix 4 revealed the best LSSVR model parameters for each input combination. The prediction performance over the testing phase is listed in Tables 1, 2 and 3 for the developed predictions models (i.e., LSSVR, MARS and M5) for all the inspected lakes stations (Lake Superior, Lake Michigan, Lake Huron, Lake Erie, and Lake Ontario), respectively. Apparently, the presented results showed a significant discrepancy in the outcomes, which are the values of the RMSE and MAE and theirs mean values.

Throughout the statistical performance reported in Tables 1, 2 and 3, RMSE and MAE metrics indicated that the third input combination provided the optimal forecasting value for one month ahead LWL using LSSVR and MARS models. This can be explained due to the informative details were supplied using three months lag time for building the learning process of the applied ML models for the train/test phases. On the other hand, M5-Tree model attained the best results for Lake Michigan (iii) combination, Lake Superior, Erie and Ontario (ii) combination, and Lake St. Clair (i) combination.

When the data sets were examined, the best performance results were seen in the M4 dataset for Lake Superior, Michigan and Ontario, while the worst dataset was seen in the M1 dataset for LSSVR, MARS and M5-Tree. On the

Table 6 Comparison of the LSSVR, MARS, M5-Tree models in predicting monthly lake levels of the St. Clair Station by using the data of Michigan station

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
LSSVR	RMSE	M1	1995–2020	0.236	0.246	0.238	0.240
		M2	1969–1995	0.140	0.143	0.377	0.220
		M3	1943–1969	0.137	0.129	0.324	0.197
		M4	1918–1943	0.299	0.294	0.430	0.341
			Mean	0.203	0.203	0.343	0.250
	MAE	M1	1995–2020	0.218	0.223	0.193	0.211
		M2	1969–1995	0.116	0.118	0.342	0.192
		M3	1943–1969	0.093	0.089	0.263	0.148
		M4	1918–1943	0.253	0.254	0.348	0.285
			Mean	0.170	0.171	0.286	0.209
	R^2	M1	1995–2020	0.914	0.900	0.737	0.851
		M2	1969–1995	0.833	0.836	0.640	0.770
		M3	1943–1969	0.851	0.871	0.258	0.660
		M4	1918–1943	0.729	0.767	0.075	0.524
			Mean	0.832	0.843	0.428	0.701
MARS	RMSE	M1	1995–2020	0.262	0.262	0.255	0.260
		M2	1969–1995	0.155	0.164	0.156	0.159
		M3	1943–1969	0.141	0.129	0.130	0.134
		M4	1918–1943	0.301	0.296	0.296	0.298
			Mean	0.215	0.213	0.209	0.212
	MAE	M1	1995–2020	0.235	0.235	0.228	0.232
		M2	1969–1995	0.130	0.132	0.126	0.129
		M3	1943–1969	0.102	0.089	0.093	0.095
		M4	1918–1943	0.256	0.257	0.260	0.257
			Mean	0.181	0.178	0.177	0.178
	R^2	M1	1995–2020	0.905	0.913	0.916	0.911
		M2	1969–1995	0.819	0.791	0.839	0.816
		M3	1943–1969	0.832	0.870	0.867	0.856
		M4	1918–1943	0.732	0.771	0.788	0.764
			Mean	0.822	0.836	0.853	0.837
M5-Tree	RMSE	M1	1995–2020	0.259	0.265	0.268	0.264
		M2	1969–1995	0.161	0.176	0.196	0.178
		M3	1943–1969	0.147	0.159	0.152	0.153
		M4	1918–1943	0.300	0.297	0.297	0.298
			Mean	0.217	0.224	0.228	0.223
	MAE	M1	1995–2020	0.232	0.237	0.237	0.235
		M2	1969–1995	0.136	0.141	0.155	0.144
		M3	1943–1969	0.102	0.113	0.114	0.110
		M4	1918–1943	0.254	0.253	0.256	0.254
			Mean	0.181	0.186	0.190	0.186
	R^2	M1	1995–2020	0.902	0.905	0.887	0.898
		M2	1969–1995	0.802	0.749	0.710	0.754
		M3	1943–1969	0.827	0.795	0.816	0.813
		M4	1918–1943	0.729	0.741	0.754	0.741
			Mean	0.815	0.797	0.792	0.801

Table 6 (continued)

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
P-LSSVR	RMSE	M1	1995–2020	0.228	0.240	0.245	0.238
		M2	1969–1995	0.159	0.160	0.161	0.160
		M3	1943–1969	0.118	0.117	0.118	0.118
		M4	1918–1943	0.291	0.292	0.292	0.292
			Mean	0.199	0.202	0.204	0.202
	MAE	M1	1995–2020	0.210	0.216	0.219	0.215
		M2	1969–1995	0.128	0.128	0.128	0.128
		M3	1943–1969	0.083	0.082	0.083	0.083
		M4	1918–1943	0.259	0.259	0.260	0.259
			Mean	0.170	0.171	0.172	0.171
	R^2	M1	1995–2020	0.924	0.914	0.905	0.914
		M2	1969–1995	0.826	0.823	0.821	0.823
		M3	1943–1969	0.897	0.897	0.895	0.896
		M4	1918–1943	0.813	0.811	0.811	0.812
			Mean	0.865	0.861	0.858	0.861

other hand, Lake St. Clair and Lake Erie show the worst performance in the M4 dataset. The best performances were seen in the M1 dataset according to the MARS method for St. Clair Lake and the M1 dataset according to the three methods for St. Clair Lake. In addition, the best performances in the M2 dataset for St. Clair Lake were seen in the LSSVR and M5-Tree methods, respectively. This is clearly exhibited the fact that the applied predictive models could not discover the actual lake water levels pattern using the M1 dataset over the train/test phases of the network for Lake Superior, Michigan, and Ontario. On the other hand, in Lake Erie and Lake St Clair, the M4 data set can be interpreted as the methods could not discover the lake levels. Among the three predictive models, LSSVR model revealed the superior prediction results over MARS and M5-Tree models using M4 dataset based on the third constructed input combination. It can be observed, the LSSVR average prediction value for the third input combination and M3 dataset boosted the value of the RMSE accuracy by 4.54 and 9.09 in comparison with the average MARS and M5-Tree models for the Lake Superior and boosted by 2.22 and 2.22% for the Lake Michigan-Huron and by 6.36 and 15.45 for the Lake St. Clair and by 5.06 and 17.72% for the Lake Erie and by 5.62 and 23.59% for the Lake Ontario, respectively.

5.2 Lake water level prediction using periodic component

The forecasting modeling procedure also involved the examination and evaluation of the periodicity data component. The main aim of incorporating the periodical dataset as sub-data is to support the learning process of the applied ML models with external "informative" LWL pattern that could offer a better understanding and improve the results accuracy. The findings of the P-LSSVR model's optimal kernel parameters are shown in Appendix 5, while the results of the P-LSSVR model's testing phase are shown in Table 4. Periodicity component clearly improved the average performance accuracy of the LSSVR model in terms of RMSE and MAE for Lake Superior (32.69–35.71%), Lake Michigan (by 30.19–30.23%), Lake Huron (by 21.42–25.61%), Lake Erie (by 25.88–28.35%), and Lake Ontario (by 25.24–25.92). When comparing Tables 3 and 4, the periodic LSSVR shows that the modeling accuracy is consistent, with LSSVR for lakes with M3 as the best performing model and M1 as the poorest model.

It is also good to evaluate the observed linear relationship between the predicted and observed time-series for the testing period as a way of further assessing the performance of the used data-driven models. Figure 6 presents the best and worst results in the form of scatter plots for the studied Lakes. These figures showed the MARS, M5-Tree, LSSVR, and P-LSSVR models for all the input combinations. The P-LSSVR model established a good match and reasonable agreement between the predicted and observed lake levels.

During the testing period, the Taylor diagram was utilized to show the spatial variance of the expected lake level by the assessed models over the observed value [53]. The standard deviation (SD) between the observed and expected values is established by Taylor diagrams in radial intervals with roots, with the R values being the angles of direction. It is assumed that the observed values on the Taylor diagram have their own display, and that models with greater performances tend to present prediction performance indicators that are closer to the observed values [52]. The Taylor diagrams of the predicted and observed lake level values for the Great Lakes as produced by the MARS, M5-Tree, LSSVR, and P-LSSVR models during the testing phase (Appendix 6a–e).

In the area of engineering, the violin plot is one of the most recently investigated graphical evaluations [54]. Conceptually the violin plot is made up of two plots: a box plot and a density plot with a rotating kernel density on each side. The Violin diagram was used to examine the distribution of observed and simulated lake levels [55]. In this study, instead of the classical violin diagram using the mean and median, the new warrant violin diagram drawn in the light of many statistical parameters (mean, median, kernel density, Standard deviation with mean, quartiles, etc.) proposed by Legouhy was used [56]. The structure of the diagram is presented in Fig. 7.

The modeling results of the adopted case studied were visualized using Violin diagrams of the observed and predicted lake level values throughout the testing phase as produced by the analyzed models for the Great Lakes (Appendix 7a-e). The figures showed no clear differences between the observed and model predictions; however, the simulated distribution of the lake levels was substantially closer to the observed lake levels distribution, especially in (iii) combinations. The statistics of the Violin plots for similar instances also revealed that MARS and M5-Tree models have non-uniform values and an imbalanced interquartile, whereas LSSVR and P-LSSVR models have a lower error rate.

For (ii and iii) combinations, these figures showed that the best models were P-LSSVR, LSSVR, MARS, and M5-Tree. In comparison to the other models, P-LSSVR was shown to be the best model for displaying close to the fit line (Fig. 6). In general, the P-LSSVR and LSSVR models outperformed the M5-Tree and MARS models. Periodic component was also added for MARS and M5-Tree methods in the study. But the performance ranking remained below P-LSSVR. This could be due to the linear structure of the adopted models MARS and M5-Tree in which lead to limited learning process of the prediction matrix and mimic the nonlinear relationship between the predictors and predictand.

5.3 Cross station modeling for lake water level prediction

In this section, lake level's prediction has been conducted using the P-LSSVR, LSSVR, MARS and M5-Tree based on upstream lake level data for downstream stations. This type of modeling is important in circumstances where lake levels are lacking, or discharge monitoring is of quality. Lake level prediction utilizing upstream stations can be quite helpful in predicting missing data [23]. In this study, the cross-station prediction was undertaken for the Lake Michigan, Lake Huron, Lake Erie, and Lake Ontario. Since the Superior Lake is located at the top upstream, the data were used in the training phase and predictions were made for the Michigan-Huron lake during the testing phase. Similarly, prediction was made for Lake Huron with Michigan-Huron lake training data. Lake Huron training data was tested in Lake St. Clair. Lake Erie by training Lake St. Clair levels and finally Lake Ontario was tested by training Lake Erie. Since the hydrological characteristics of the Great Lakes are similar (please see, Figs. 3 and 4) and related (please see, Appendix 2 and Fig. 2), the estimates will be based on homogeneous physical properties. The data base was likewise cross-stationed and separated into four parts here. Appendixes 8 and 9 expressed the ideal parameters of the LSSVR and P-LSSVR model in a manner similar to that of the previous subsection application technique. Comparison of the LSSVR, MARS, M5-Tree models in predicting monthly lake levels of the Michigan Station by using the data of Superior station are given Table 5. Similar results are given in Table 6 for Lake St. Clair, Table 7 for Lake Erie, and Table 8 for Lake Ontario. From the average RMSE and MAE parameters, the best score provided by P-LSSVR and LSSVR models for M1 and M4 input combination (iii) and the worst score provided by M5-Tree models for M4 input combination (iii). In addition, the three models, generally gave the M4 data set and input combination (iii) the lowest accuracy scores. Cross-station modeling gave the best performance in Lake Erie, followed by Lake Ontario, then Lake Clair, and finally Lake Michigan, according to the P-LSSVR method. Unlike other lakes, it is seen that the errors (RMSE, MAE) are more in Lake Michigan. This may be due to the fact that the lake levels of Lake Superior, located at the upstream of Lake Michigan, are controlled by Soo locks & Dams (readers can refer to Fig. 2). In other words, there may be an anthropogenic effect outside of its natural hydrology.

The effect of inserting the periodicity feature on prediction phase was investigated. This was done to find the most accurate model in the preceding applications, which was the LSSVR model. Based on the absolute error metrics (i.e., RMSE and MAE), the prediction enhancement between the applied LSSVR and P-LSSVR models are 8.85 and 8.14%, respectively. This is clearly can be justified owing the additive. To further visualize the effect of including the periodic component. Finally, LSSVR method showed the best forecast and prediction, followed by MARS and finally M5-Tree models. The performance of

Table 7 Comparison of the LSSVR, MARS, M5-Tree models in predicting monthly lake levels of the Erie Station by using the data of St. Clair station

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
LSSVR	RMSE	M1	1995–2020	0.119	0.119	0.116	0.118
		M2	1969–1995	0.100	0.099	0.096	0.098
		M3	1943–1969	0.113	0.108	0.103	0.108
		M4	1918–1943	0.153	0.148	0.141	0.147
			Mean	0.121	0.119	0.114	0.118
	MAE	M1	1995–2020	0.098	0.097	0.093	0.096
		M2	1969–1995	0.078	0.076	0.074	0.076
		M3	1943–1969	0.084	0.082	0.079	0.082
		M4	1918–1943	0.119	0.114	0.109	0.114
			Mean	0.095	0.092	0.089	0.092
	R^2	M1	1995–2020	0.906	0.916	0.918	0.913
		M2	1969–1995	0.814	0.817	0.828	0.819
		M3	1943–1969	0.841	0.861	0.876	0.860
		M4	1918–1943	0.794	0.809	0.828	0.810
			Mean	0.839	0.851	0.862	0.851
MARS	RMSE	M1	1995–2020	0.123	0.120	0.116	0.120
		M2	1969–1995	0.100	0.105	0.099	0.101
		M3	1943–1969	0.120	0.115	0.113	0.116
		M4	1918–1943	0.160	0.152	0.148	0.153
			Mean	0.126	0.123	0.119	0.122
	MAE	M1	1995–2020	0.100	0.096	0.093	0.097
		M2	1969–1995	0.078	0.077	0.074	0.076
		M3	1943–1969	0.088	0.086	0.086	0.086
		M4	1918–1943	0.125	0.117	0.115	0.119
			Mean	0.098	0.094	0.092	0.095
	R^2	M1	1995–2020	0.909	0.914	0.918	0.914
		M2	1969–1995	0.813	0.800	0.821	0.811
		M3	1943–1969	0.830	0.843	0.851	0.841
		M4	1918–1943	0.775	0.796	0.807	0.793
			Mean	0.832	0.838	0.849	0.840
M5-Tree	RMSE	M1	1995–2020	0.129	0.126	0.124	0.126
		M2	1969–1995	0.100	0.108	0.113	0.107
		M3	1943–1969	0.121	0.123	0.124	0.123
		M4	1918–1943	0.158	0.161	0.159	0.160
			Mean	0.127	0.130	0.130	0.129
	MAE	M1	1995–2020	0.103	0.102	0.099	0.101
		M2	1969–1995	0.079	0.082	0.085	0.082
		M3	1943–1969	0.090	0.092	0.093	0.092
		M4	1918–1943	0.124	0.126	0.124	0.125
			Mean	0.099	0.101	0.100	0.100
	R^2	M1	1995–2020	0.896	0.902	0.907	0.902
		M2	1969–1995	0.813	0.791	0.781	0.795
		M3	1943–1969	0.828	0.831	0.836	0.831
		M4	1918–1943	0.785	0.765	0.776	0.775
			Mean	0.830	0.822	0.825	0.826

Table 7 (continued)

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
P-LSSVR	RMSE	M1	1995–2020	0.113	0.114	0.112	0.113
		M2	1969–1995	0.087	0.088	0.088	0.088
		M3	1943–1969	0.098	0.097	0.094	0.096
		M4	1918–1943	0.137	0.135	0.130	0.134
			Mean	0.109	0.108	0.106	0.108
	MAE	M1	1995–2020	0.090	0.088	0.087	0.088
		M2	1969–1995	0.064	0.065	0.067	0.066
		M3	1943–1969	0.073	0.074	0.072	0.073
		M4	1918–1943	0.110	0.108	0.105	0.108
			Mean	0.084	0.084	0.083	0.084
	R^2	M1	1995–2020	0.921	0.921	0.923	0.921
		M2	1969–1995	0.861	0.862	0.858	0.860
		M3	1943–1969	0.890	0.894	0.901	0.895
		M4	1918–1943	0.841	0.840	0.844	0.842
			Mean	0.878	0.879	0.881	0.880

the LSSVR method increased with the addition of the periodic component. The lake levels with the least errors were obtained in the modeling performed in Lake Superior and the highest error in Lake St. Clair. In cross-station modelling, the best performance was obtained in Lake Erie, which showed the highest correlation with Lake Clair. The worst modeling was observed in the modeling where the Lake Superior levels were used in the input dataset and the Michigan lake levels were estimated. This is assumed to be due to the fact that artificial rather than natural effects dominate the hydrological relationship between the two lakes.

6 Discussion

One of the major factors affecting the model's performances is the input combination selection. Therefore, proper input selection is required prior to applying the models. For this research, an input scenario based on the autocorrelation function (ACF) and the partial autocorrelation function (PACF) was generated and analyzed for the three models used to determine the number of effective lags of antecedent lake level. Several studies have presented this strategy for determining the best inputs for data-driven methods [57–61]. In all of the lakes, the PACF shows that the first lag of lake levels has a significant effect, while the second and third delays are extremely close to the confidence limit. Therefore, lag 1, lag 2 and lag 3 were chosen as the input set for all lakes. Long trend lag times for the current study indicates the appropriate dataset for structuring the learning process of the predictive model and thus for such a case study of the great lake, this must be considered in the future research. The concept of the cross-station modeling based on the computer learning transfer research an optimistic result for the current investigation. Indeed, this is not surprising as this technology has been approved by several other research on other hydrological applications [62–66].

7 Conclusion

The main aim of the present study is to provide a valid and reliable predictive model for LWL based on the implication of neurocomputing technology. For this purpose, three ML models including LSSVR, MARS and M5-Tree were developed to forecast LWL at five lakes located within north America. At the first stage, the autocorrelation and partial autocorrelation functions were used to select the input data sets for analysis. The results of the three models' performance were compared with mean absolute error (MAE) and root mean square error (RMSE), determination coefficient (R) and different aspects of the models' accuracy were assessed using scatter plots, Taylor diagrams and violin diagrams. As a result, this study finding indicates that the P-LSSVR model is more powerful for all lake levels modeling and a better alternative to the other three neurocomputing intelligence models. Cross-station modeling strategy showed a reliable technique for LWL forecasting using nearby hydrological information.

Table 8 Comparison of the LSSVR, MARS, M5-Tree models in predicting monthly lake levels of the Ontario Station by using the data of St. Erie station

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
LSSVR	RMSE	M1	1995–2020	0.200	0.178	0.167	0.181
		M2	1969–1995	0.268	0.270	0.245	0.261
		M3	1943–1969	0.267	0.250	0.242	0.253
		M4	1918–1943	0.164	0.160	0.163	0.162
			Mean	0.225	0.214	0.204	0.214
	MAE	M1	1995–2020	0.158	0.140	0.134	0.144
		M2	1969–1995	0.225	0.229	0.207	0.221
		M3	1943–1969	0.217	0.204	0.195	0.205
		M4	1918–1943	0.125	0.124	0.128	0.126
			Mean	0.181	0.174	0.166	0.174
	R^2	M1	1995–2020	0.541	0.643	0.686	0.624
		M2	1969–1995	0.531	0.525	0.589	0.548
		M3	1943–1969	0.689	0.720	0.704	0.704
		M4	1918–1943	0.809	0.825	0.813	0.816
			Mean	0.643	0.678	0.698	0.673
MARS	RMSE	M1	1995–2020	0.203	0.181	0.175	0.186
		M2	1969–1995	0.290	0.286	0.263	0.280
		M3	1943–1969	0.271	0.255	0.251	0.259
		M4	1918–1943	0.166	0.177	0.170	0.171
			Mean	0.233	0.225	0.215	0.224
	MAE	M1	1995–2020	0.159	0.140	0.141	0.147
		M2	1969–1995	0.244	0.242	0.222	0.236
		M3	1943–1969	0.219	0.210	0.205	0.211
		M4	1918–1943	0.126	0.140	0.134	0.133
			Mean	0.187	0.183	0.176	0.182
	R^2	M1	1995–2020	0.535	0.634	0.654	0.608
		M2	1969–1995	0.466	0.426	0.536	0.476
		M3	1943–1969	0.668	0.727	0.694	0.696
		M4	1918–1943	0.792	0.734	0.762	0.763
			Mean	0.615	0.630	0.661	0.636
M5-Tree	RMSE	M1	1995–2020	0.211	0.198	0.191	0.200
		M2	1969–1995	0.297	0.287	0.275	0.286
		M3	1943–1969	0.274	0.260	0.253	0.262
		M4	1918–1943	0.189	0.191	0.212	0.197
			Mean	0.243	0.234	0.233	0.236
	MAE	M1	1995–2020	0.165	0.153	0.149	0.156
		M2	1969–1995	0.249	0.242	0.232	0.241
		M3	1943–1969	0.222	0.211	0.205	0.212
		M4	1918–1943	0.145	0.151	0.164	0.153
			Mean	0.195	0.189	0.187	0.191
	R^2	M1	1995–2020	0.494	0.559	0.595	0.549
		M2	1969–1995	0.419	0.394	0.337	0.383
		M3	1943–1969	0.641	0.651	0.634	0.642
		M4	1918–1943	0.737	0.700	0.631	0.689
			Mean	0.573	0.576	0.549	0.566

Table 8 (continued)

Model	Statistics	Cross Station	Test data set	Input combination			
				(i)	(ii)	(iii)	Mean
P-LSSVR	RMSE	M1	1995–2020	0.151	0.149	0.149	0.150
		M2	1969–1995	0.221	0.221	0.220	0.221
		M3	1943–1969	0.239	0.240	0.238	0.239
		M4	1918–1943	0.166	0.162	0.161	0.163
			Mean	0.194	0.193	0.192	0.193
	MAE	M1	1995–2020	0.118	0.117	0.116	0.117
		M2	1969–1995	0.191	0.193	0.193	0.192
		M3	1943–1969	0.193	0.192	0.191	0.192
		M4	1918–1943	0.138	0.135	0.135	0.136
			Mean	0.160	0.159	0.159	0.159
	R^2	M1	1995–2020	0.737	0.752	0.753	0.747
		M2	1969–1995	0.689	0.711	0.701	0.700
		M3	1943–1969	0.706	0.682	0.686	0.692
		M4	1918–1943	0.764	0.778	0.781	0.774
			Mean	0.724	0.731	0.730	0.728

Appendix 1

See Table 9.

Table 9 The statistical parameters of the selected lakes for the current research

Lake	Period	X_{mean} (m)	X_{min} (m)	X_{max} (m)	Sx (m)	**Cs**
Superior	1918–2020	183.41	182.72	183.91	0.204	− 0.258
Michigan-Huron	1918–2020	176.44	175.57	177.5	0.410	0.120
St. Clair	1918–2020	175.03	173.88	176.04	0.396	− 0.068
Erie	1918–2020	174.17	173.18	175.14	0.368	− 0.019
Ontario	1918–2020	74.77	73.74	75.91	0.346	0.103

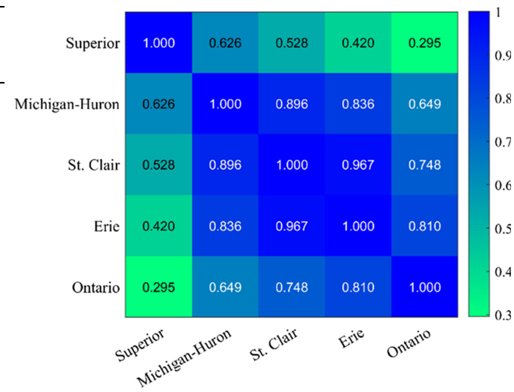
Appendix 2

See Table 10.

Table 10 Pearson correlation coefficients

Appendix 2: Pearson Correlation coefficients.

Lake	Superior	Michigan-Huron	St. Clair	Erie	Ontario
Superior	1				
Michigan-Huron	**0.626**	1			
St. Clair	0.528	**0.896**	1		
Erie	0.420	0.836	**0.967**	1	
Ontario	0.295	0.649	0.748	**0.810**	1



Appendix 3

See Table 11.

Table 11 The monthly statistical parameters of lake stations

Lake	Period	x_{mean}(m)	x_{min} (m)	x_{max} (m)	Csx(m)	Sx(m)	r1	r2	r3
Superior	1918–1943	183.386	182.720	183.800	− 0.697	0.198	0.944	0.811	0.645
	1943–1969	183.434	183.090	183.860	0.303	0.175	0.910	0.694	0.424
	1969–1995	183.478	183.080	183.910	− 0.047	0.167	0.923	0.741	0.511
	1995–2020	183.351	182.790	183.880	0.118	0.245	0.964	0.881	0.777
Michigan	1918–1943	176.252	175.660	177.180	0.493	0.352	0.981	0.935	0.872
	1943–1969	176.404	175.580	177.280	− 0.042	0.359	0.981	0.935	0.874
	1969–1995	176.767	176.150	177.500	0.262	0.269	0.970	0.900	0.811
	1995–2020	176.349	175.570	177.460	0.616	0.442	0.987	0.956	0.916
St. Clair	1918–1943	174.694	173.880	175.520	0.121	0.316	0.889	0.737	0.595
	1943–1969	174.953	174.140	175.660	− 0.232	0.320	0.913	0.788	0.661
	1969–1995	175.374	174.770	175.960	0.265	0.225	0.918	0.824	0.710
	1995–2020	175.110	174.440	176.040	0.576	0.361	0.960	0.890	0.809
Erie	1918–1943	173.832	173.180	174.640	0.270	0.293	0.940	0.817	0.678
	1943–1969	174.083	173.400	174.760	− 0.013	0.278	0.932	0.784	0.607
	1969–1995	174.468	173.970	175.040	0.171	0.231	0.922	0.772	0.598
	1995–2020	174.282	173.730	175.140	0.554	0.320	0.953	0.853	0.730
Ontario	1918–1943	74.549	73.740	75.600	0.399	0.351	0.946	0.824	0.680
	1943–1969	74.828	73.830	75.760	0.001	0.355	0.934	0.777	0.583
	1969–1995	74.862	74.360	75.730	0.511	0.285	0.881	0.613	0.286
	1995–2020	74.833	74.280	75.910	0.632	0.294	0.877	0.599	0.270

Appendix 4

See Table 12.

Table 12 Regularization constant and width of RBF kernel parameters of the optimal LSSVR models for Superior, Michigan, St. Clair, Erie and Ontario Lake stations

Cross validation	Training data set	Test data set	Input combination		
			(i)	(ii)	(iii)
Lake Superior					
M1	1918–1995	1995–2020	(100, 69)	(71, 100)	(100, 55)
M2	1918–1969 and 1995–2020	1969–1995	(3, 100)	(76, 1)	(56, 4)
M3	1918–1943 and 1969–2020	1943–1969	(3, 100)	(60, 100)	(49, 3)
M4	1943–2020	1918–1943	(100, 11)	(100, 10)	(100, 14)
Lake Michigan					
M1	1918–1995	1995–2020	(34, 100)	(100, 37)	(100, 18)
M2	1918–1969 and 1995–2020	1969–1995	(15, 2)	(36, 3)	(79, 7)
M3	1918–1943 and 1969–2020	1943–1969	(32, 100)	(100, 7)	(100, 8)
M4	1943–2020	1918–1943	(100, 3)	(100, 4)	(100, 2)
Lake St. Clair					
M1	1918–1995	1995–2020	(15, 100)	(100, 3)	(100, 49)
M2	1918–1969 and 1995–2020	1969–1995	(10, 15)	(3, 31)	(4, 42)
M3	1918–1943 and 1969–2020	1943–1969	(1, 29)	(61, 1)	(49, 1)
M4	1943–2020	1918–1943	(47, 100)	(100, 6)	(100, 9)
Lake Erie					
M1	1918–1995	1995–2020	(99, 6)	(100, 12)	(100, 35)
M2	1918–1969 and 1995–2020	1969–1995	(17, 28)	(13, 49)	(45, 100)
M3	1918–1943 and 1969–2020	1943–1969	(3, 100)	(100, 1)	(50, 1)
M4	1943–2020	1918–1943	(100, 12)	(100, 24)	(100, 31)
Lake Ontario					
M1	1918–1995	1995–2020	(2, 90)	(13, 1)	(16, 3)
M2	1918–1969 and 1995–2020	1969–1995	(8, 4)	(100, 6)	(100, 7)
M3	1918–1943 and 1969–2020	1943–1969	(33, 1)	(11, 1)	(100, 5)
M4	1943–2020	1918–1943	(100, 12)	(100, 18)	(100, 28)

Appendix 5

See Table 13. .

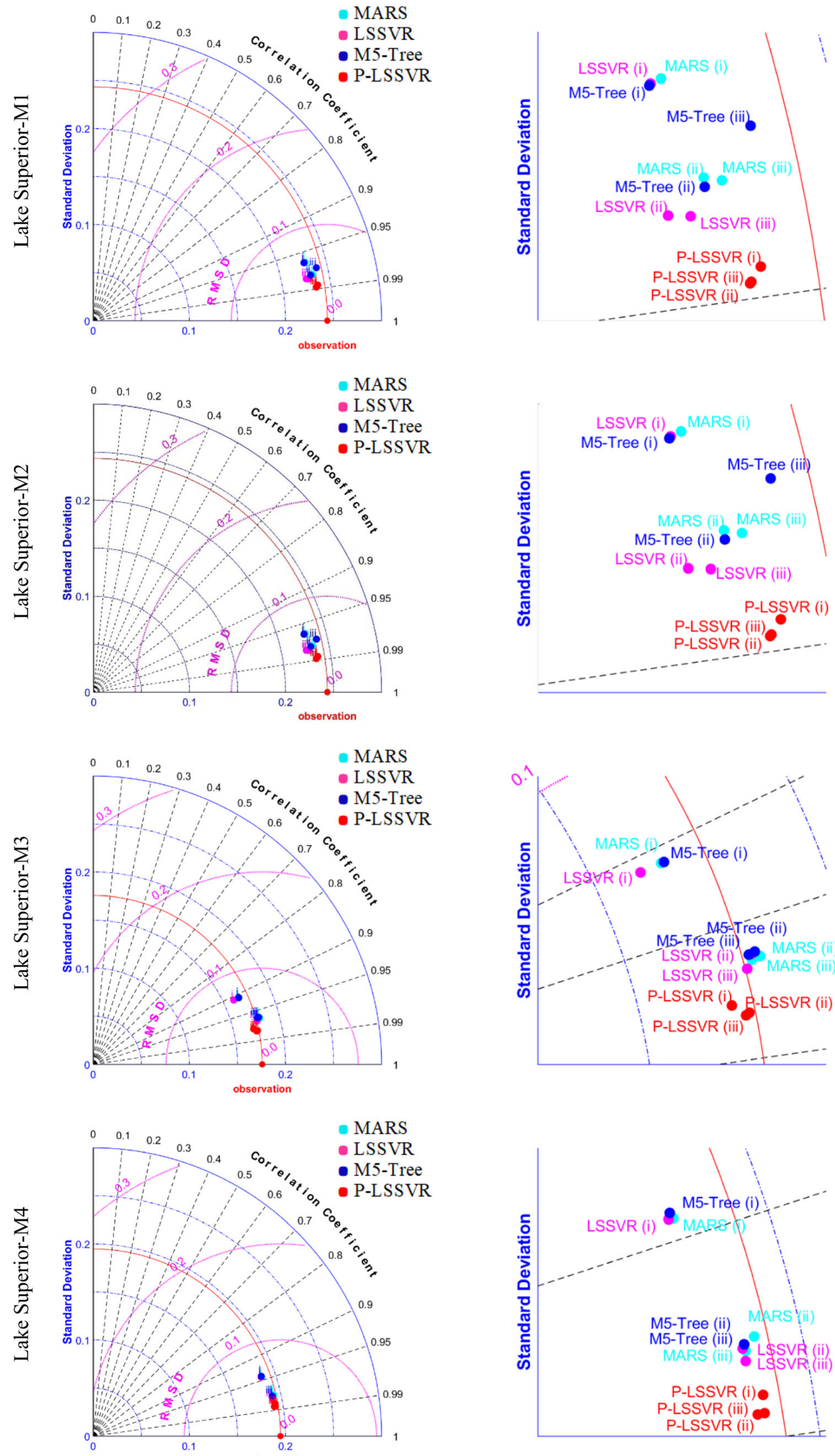
Table 13 Regularization constant and width of RBF kernel parameters of the optimal P-LSSVR models for Superior, Michigan, St. Clair, Erie and Ontario Lake stations

Cross validation	Training data set	Test data set	Input combination		
			(i)	(ii)	(iii)
Lake Superior					
M1	1918–1995	1995–2020	(30, 11)	(59, 20)	(100, 45)
M2	1918–1969 and 1995–2020	1969–1995	(100, 16)	(41, 2)	(54, 4)
M3	1918–1943 and 1969–2020	1943–1969	(8, 2)	(100, 5)	(100, 5)
M4	1943–2020	1918–1943	(100, 9)	(100, 4)	(100, 12)
Lake Michigan					
M1	1918–1995	1995–2020	(100, 13)	(100, 19)	(100, 19)
M2	1918–1969 and 1995–2020	1969–1995	(100, 5)	(100, 8)	(100, 25)
M3	1918–1943 and 1969–2020	1943–1969	(100, 7)	(100, 10)	(100, 19)
M4	1943–2020	1918–1943	(67, 7)	(100, 11)	(100, 7)
Lake St. Clair					
M1	1918–1995	1995–2020	(14, 46)	(100, 85)	(100, 91)
M2	1918–1969 and 1995–2020	1969–1995	(29, 20)	(50, 2)	(33, 63)
M3	1918–1943 and 1969–2020	1943–1969	(100, 7)	(51, 10)	(26, 4)
M4	1943–2020	1918–1943	(76, 9)	(23, 9)	(17, 9)
Lake Erie					
M1	1918–1995	1995–2020	(60, 57)	(100, 79)	(100, 77)
M2	1918–1969 and 1995–2020	1969–1995	(45, 2)	(42, 35)	(48, 45)
M3	1918–1943 and 1969–2020	1943–1969	(12, 3)	(23, 5)	(100, 9)
M4	1943–2020	1918–1943	(100, 8)	(100, 24)	(100, 39)
Lake Ontario					
M1	1918–1995	1995–2020	(90, 5)	(48, 4)	(28, 4)
M2	1918–1969 and 1995–2020	1969–1995	(8, 2)	(100, 5)	(86, 6)
M3	1918–1943 and 1969–2020	1943–1969	(69, 40)	(100, 15)	(100, 11)
M4	1943–2020	1918–1943	(100, 85)	(100, 25)	(100, 29)

Appendix 6a

See Fig. 8.

Fig. 8 Scatter plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Superior



Appendix 6b

See Fig. 9.

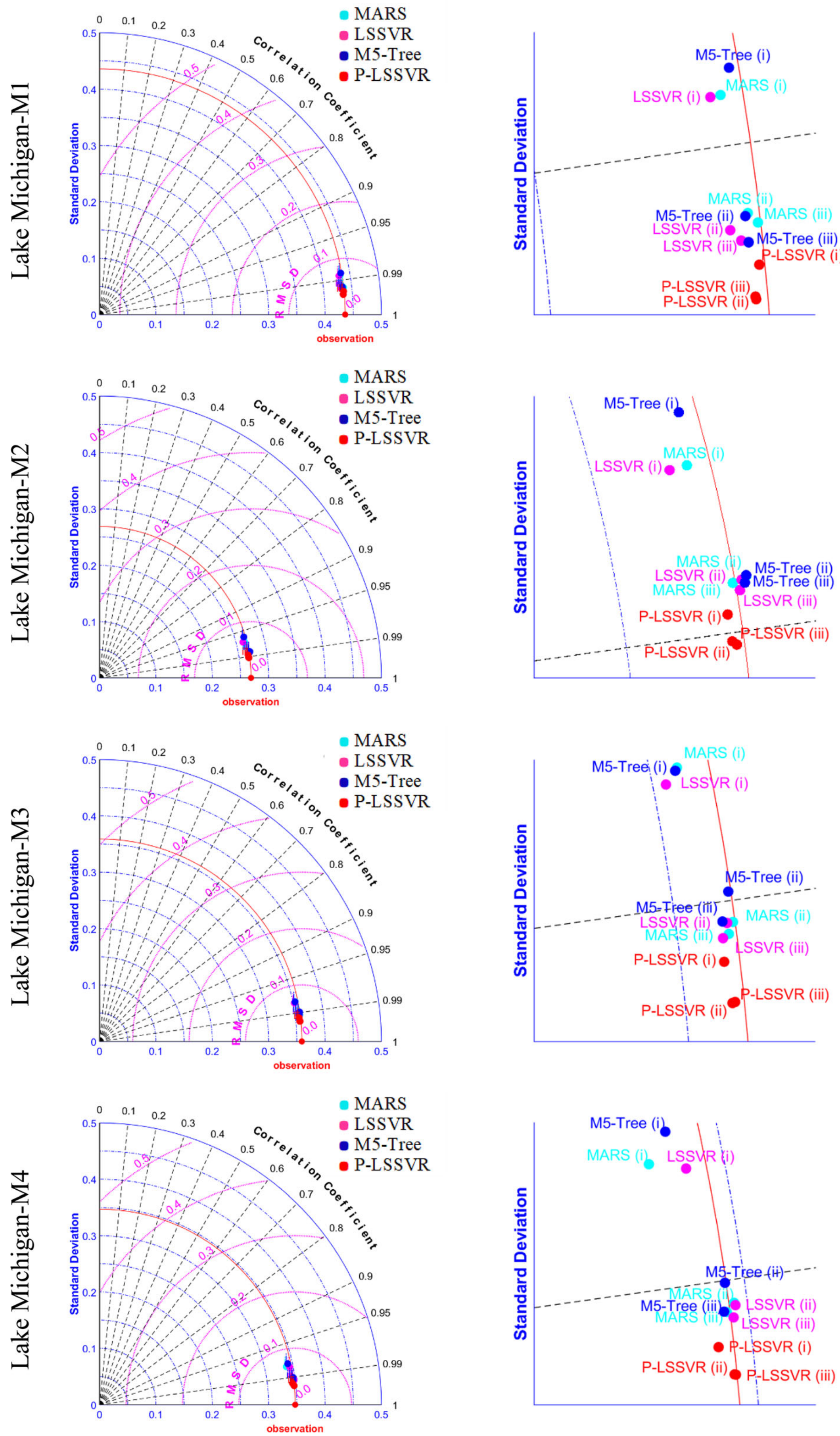


Fig. 9 Scatter plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Michigan-Huron

Appendix 6c

See Fig. 10.

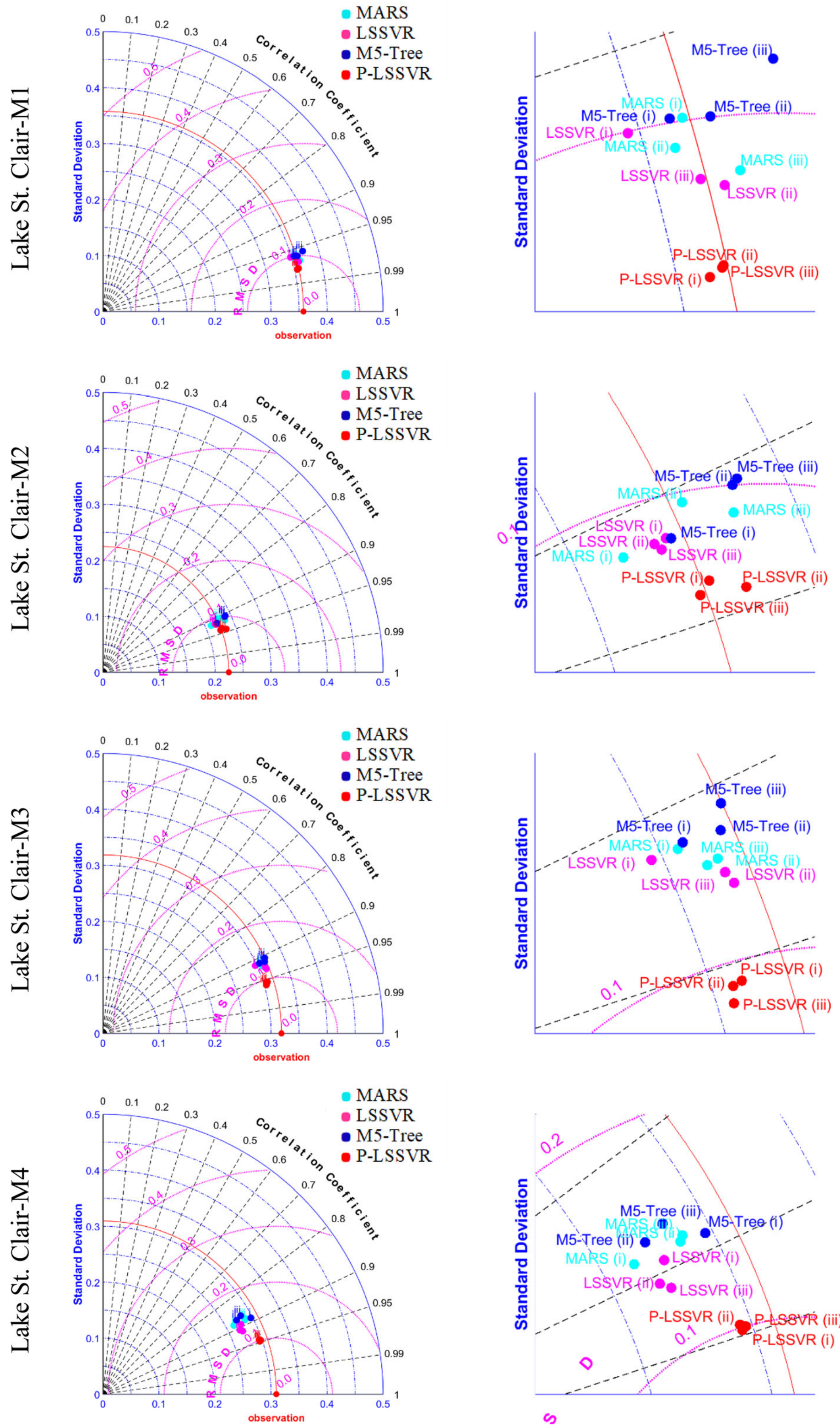


Fig. 10 Scatter plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake St. Clair

Appendix 6d

See Fig. 11.

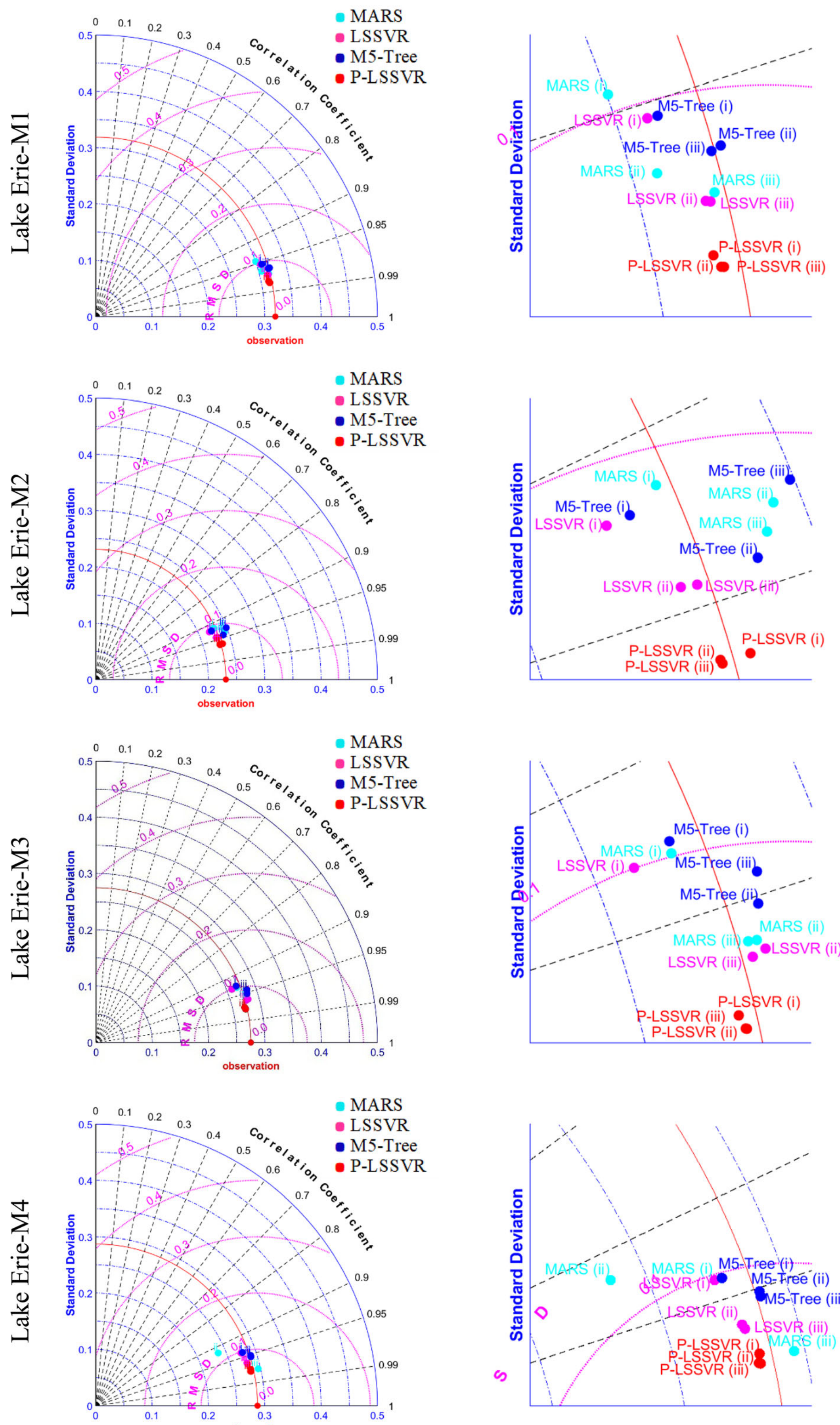


Fig. 11 Scatter plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Erie

Appendix 6e

See Fig. 12.

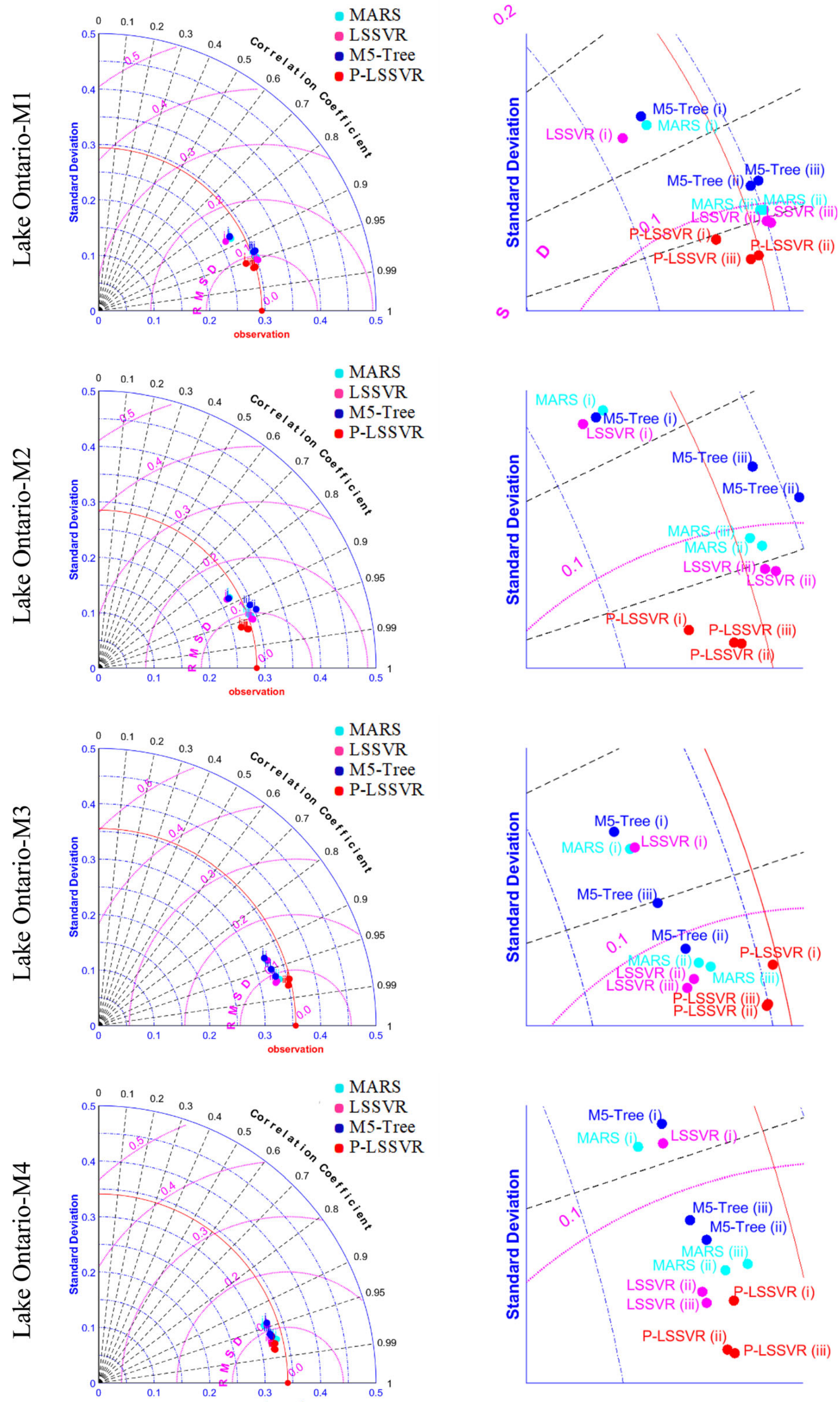


Fig. 12 Scatter plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Ontario

Appendix 7a

See Fig. 13.

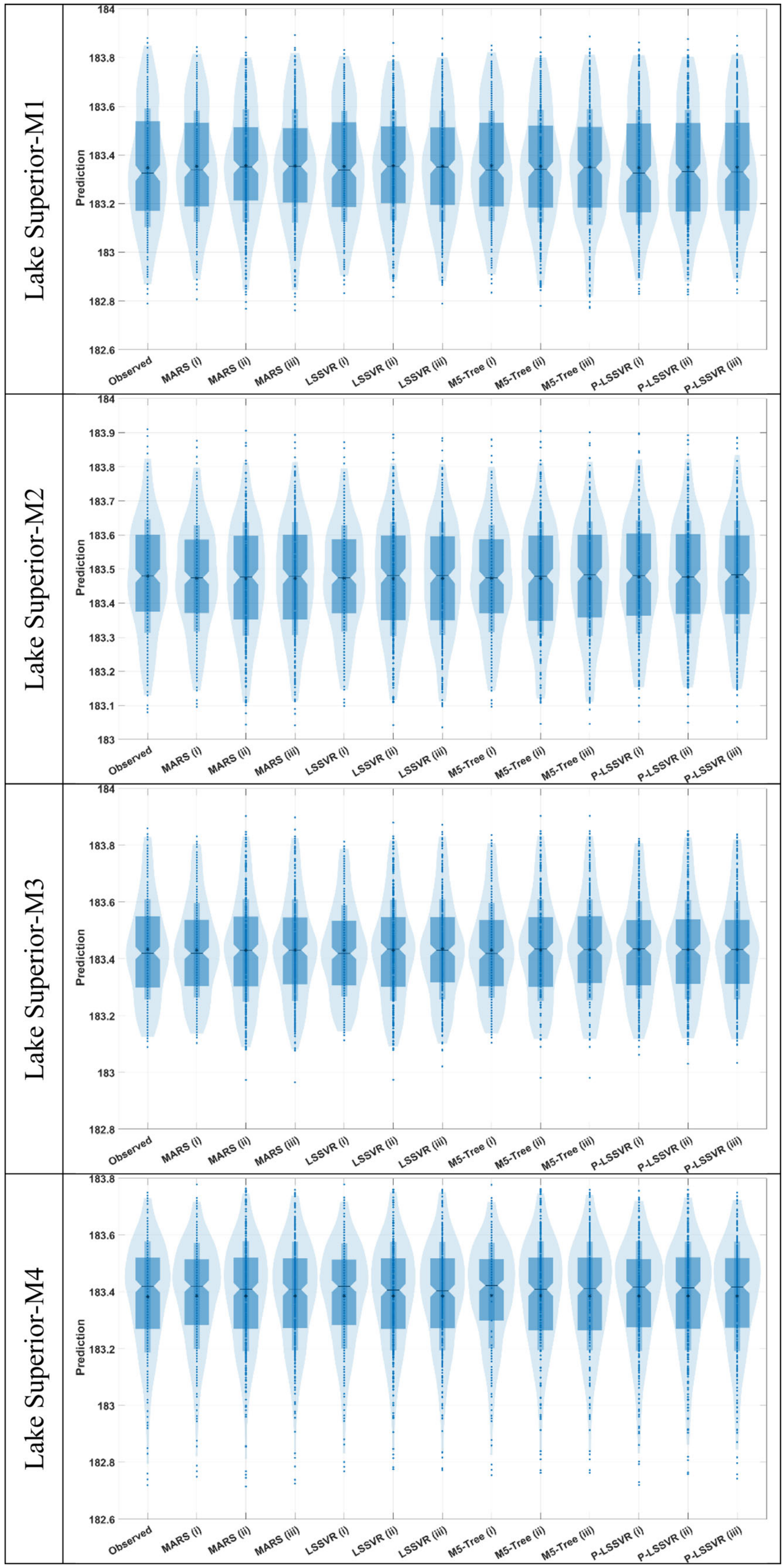


Fig. 13 Violin plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Superior

Appendix 7b

See Fig. 14.

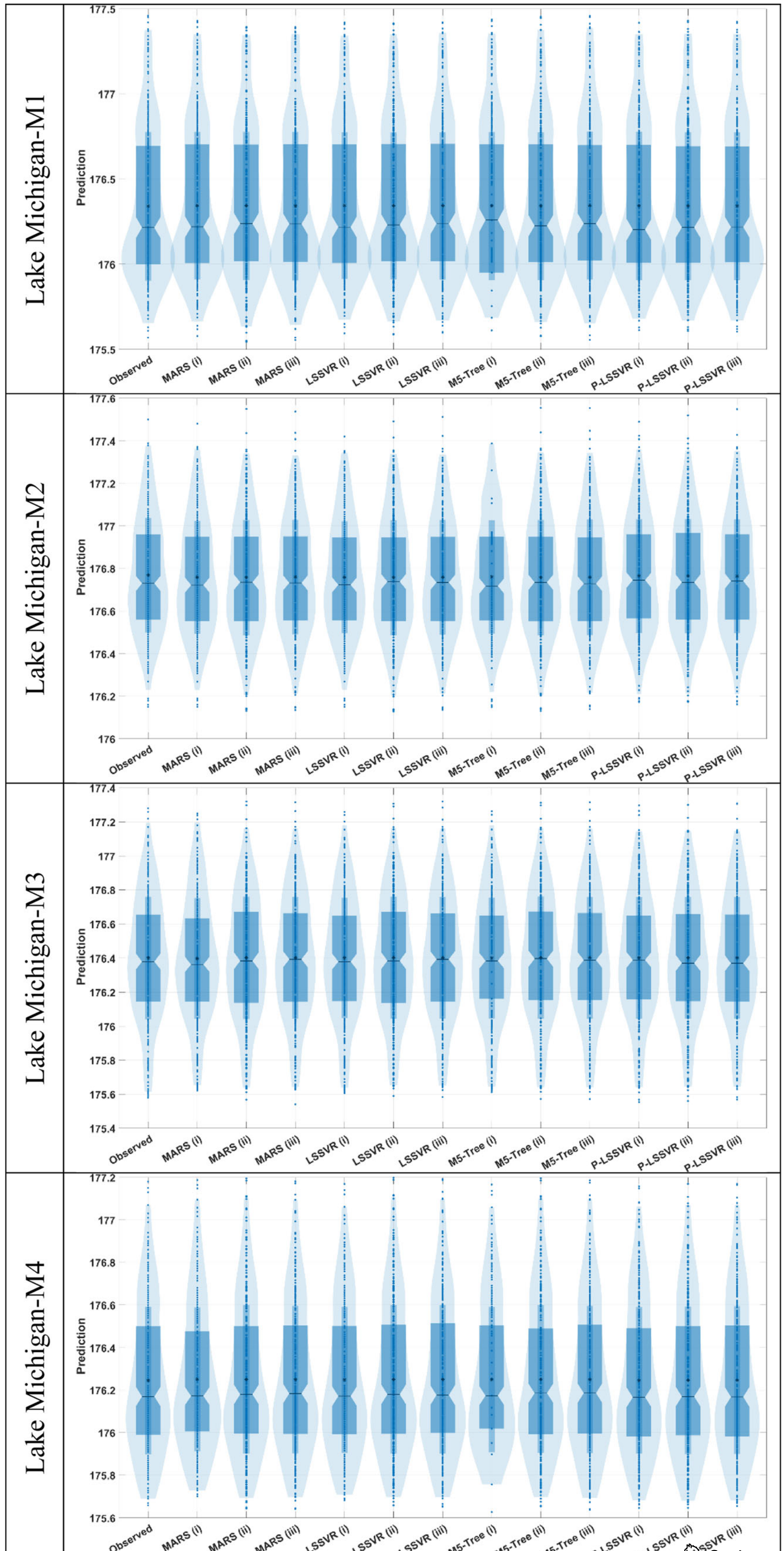


Fig. 14 Violin plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Michigan-Huron

Appendix 7c

See Fig. 15.

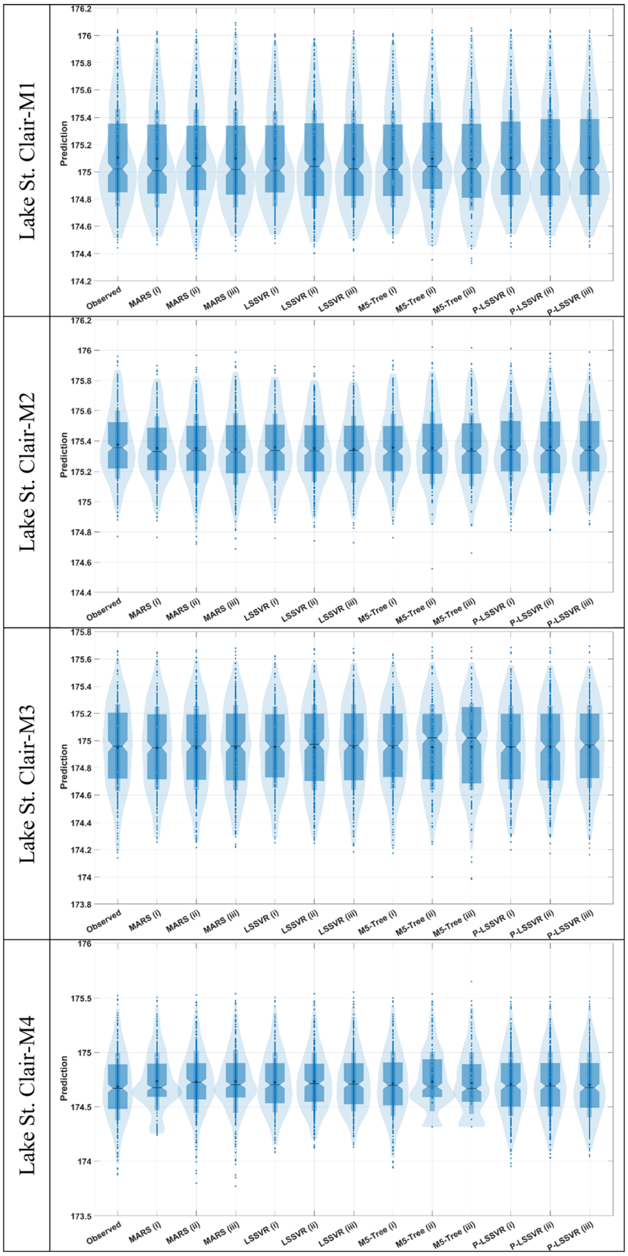


Fig. 15 Violin plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake St. Clair

Appendix 7d

See Fig. 16.

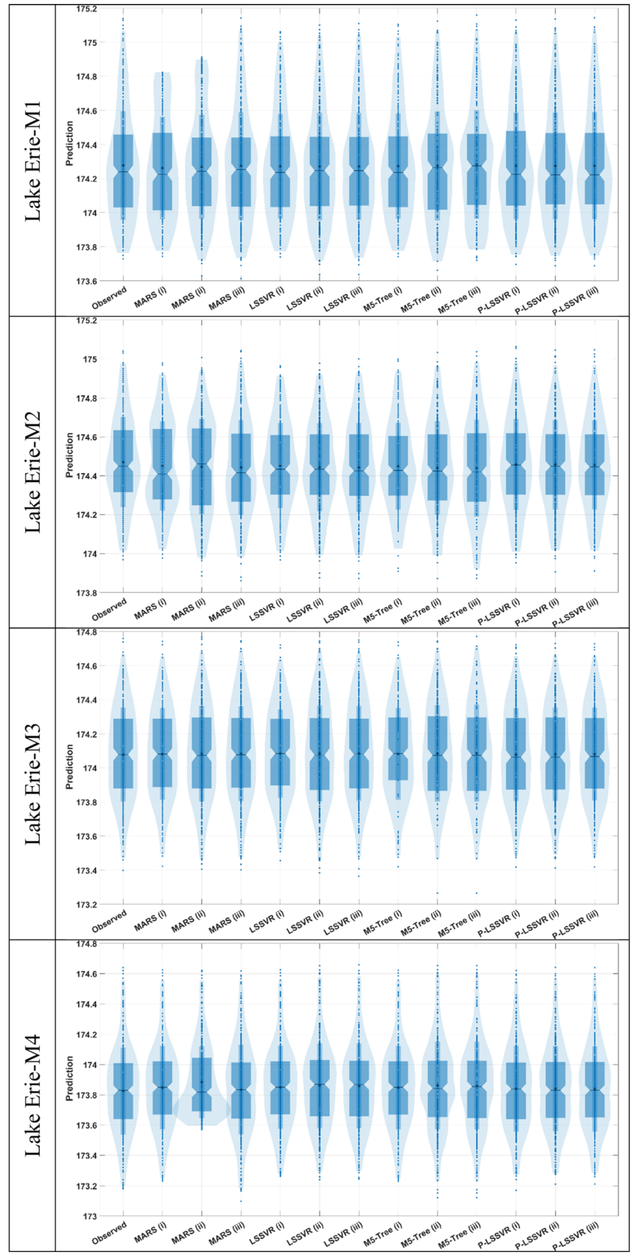


Fig. 16 Violin plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Erie

Appendix 7e

See Fig. 17.

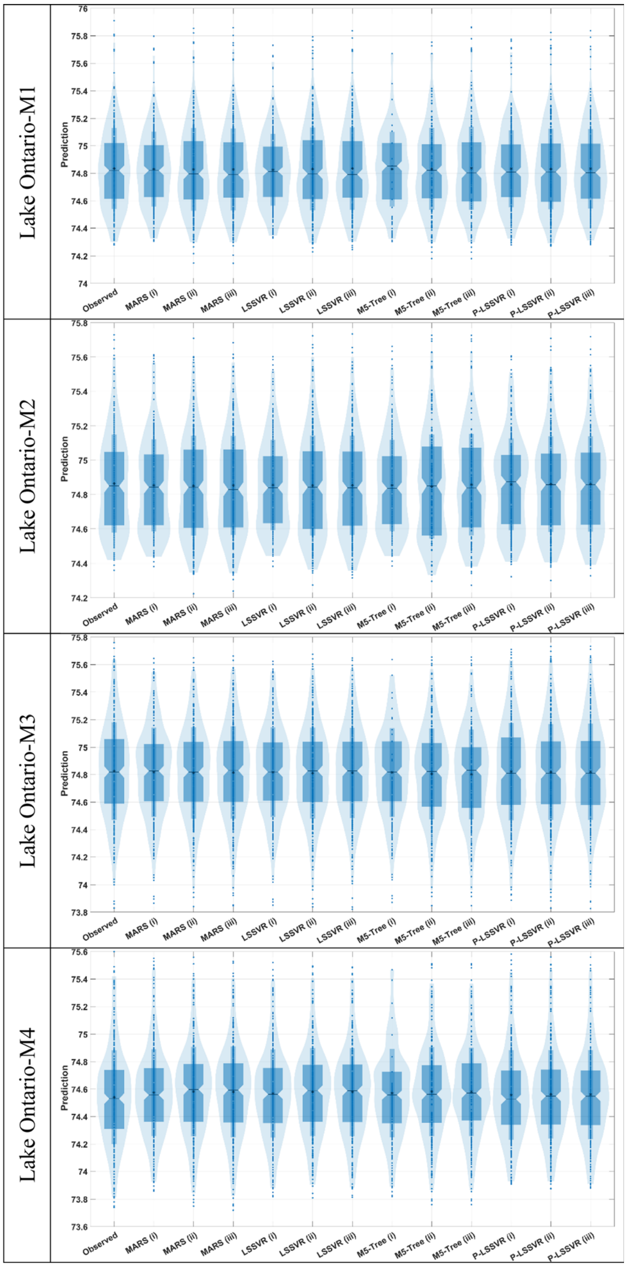


Fig. 17 Violin plots of the observed and predicted lake level values during testing phase, produced by MARS, M5-Tree, LSSVR and P-LSSVR models for the Lake Ontario

Appendix 8

See Table 14.

Table 14 The optimal parameters of the LSSVR models in cross application

Cross Station	Test data set	Input combination		
		(i)	**(ii)**	**(iii)**
Lake Michigan				
M1	1995–2020	(100, 100)	(83, 2)	(100, 4)
M2	1969–1995	(53, 100)	(29, 100)	(74, 100)
M3	1943–1969	(1, 100)	(1, 100)	(1, 24)
M4	1918–1943	(1, 100)	(1, 100)	(1, 100)
Lake St. Clair				
M1	1995–2020	(1, 100)	(1, 100)	(1, 100)
M2	1969–1995	(83, 100)	(10, 69)	(4, 69)
M3	1943–1969	(100, 2)	(100, 4)	(86, 7)
M4	1918–1943	(100, 100)	(100, 60)	(12, 1)
Lake Erie				
M1	1995–2020	(4, 100)	(100, 3)	(15, 13)
M2	1969–1995	(98, 88)	(78, 100)	(24, 50)
M3	1943–1969	(1, 79)	(100, 2)	(1, 1)
M4	1918–1943	(10, 100)	(100, 7)	(100, 12)
Lake Ontario				
M1	1995–2020	(1, 100)	(100, 8)	(100, 21)
M2	1969–1995	(1, 100)	(100, 3)	(100, 6)
M3	1943–1969	(100, 76)	(100, 1)	(63, 3)
M4	1918–1943	(84, 2)	(100, 40)	(100, 59)

Appendix 9

See Table 15.

Table 15 The optimal parameters of the P-LSSVR models in cross application

Cross Station	Test data set	Input combination		
		(i)	**(ii)**	**(iii)**
Lake Michigan				
M1	1995–2020	(100, 77)	(69, 100)	(47, 100)
M2	1969–1995	(100, 90)	(84, 100)	(68, 100)
M3	1943–1969	(1, 100)	(1, 12)	(1, 19)
M4	1918–1943	(1, 100)	(1, 100)	(1, 100)
Lake St. Clair				
M1	1995–2020	(1, 100)	(1, 100)	(1, 100)
M2	1969–1995	(62, 37)	(100, 80)	(83, 100)
M3	1943–1969	(100, 9)	(100, 12)	(37, 11)
M4	1918–1943	(57, 12)	(100, 43)	(100, 46)
Lake Erie				
M1	1995–2020	(4, 79)	(100, 16)	(100, 30)
M2	1969–1995	(69, 15)	(9, 14)	(14, 100)
M3	1943–1969	(1, 2)	(7, 5)	(100, 12)
M4	1918–1943	(100, 7)	(70, 10)	(74, 14)
Lake Ontario				
M1	1995–2020	(1, 15)	(100, 27)	(100, 35)
M2	1969–1995	(1, 100)	(100, 25)	(1, 8)
M3	1943–1969	(1, 4)	(100, 4)	(100, 6)
M4	1918–1943	(100, 22)	(100, 47)	(100, 61)

Acknowledgements The authors would like to thank the support received from KTO Karatay University.

Authors contributions Vahdettin Demir: Conceptualization; validation; investigation; data curation; methodology; project leader; editing; formal analysis; visualization; writing. Zaher Mundher Yaseen: Supervision; validation; investigation; editing, conceptualization; writing.

Funding The research received no funds.

Data and materials availability Data will be provided with final accepted version.

Declarations

Conflict of interest The authors declare no conflict of interest to any party.

Ethical approval The manuscript is conducted within the ethical manner advised by the targeted journal.

Consent to participate Not applicable.

Consent to publish The research is scientifically consent to be published.

References

1. Li XY, Xu HY, Sun YL et al (2007) Lake-level change and water balance analysis at lake Qinghai, West China during recent decades. Water Resour Manag 21:1505–1516. https://doi.org/10.1007/s11269-006-9096-1
2. Tong SL, Cui CF, Bai YL et al (2016) Application of multivariate adaptive regression spline models in long term prediction of river water pollution. Taiwan Water Conserv. https://doi.org/10.1016/j.jhydrol.2015.12.014
3. Armanuos A, Ahmed K, Shiru MS, Jamei M (2021) Impact of increasing pumping discharge on groundwater level in the nile delta aquifer. Egypt Knowledge-Based Eng Sci 2:13–23
4. Caplan B, Covitt B, Love G et al (2021) Using computational thinking and modeling to build water and watershed literacy. Connect Sci Learn, 3
5. Kelts K, Talbot M (1990) Lacustrine carbonates as geochemical archives of environmental change and biotic/abiotic interactions. In: Tilzer MM, Serruya C (eds) Large lakes. Springer, Berlin, Heidelberg, pp 288–315
6. Lücke A, Schleser GH, Zolitschka B, Negendank JFW (2003) A Lateglacial and Holocene organic carbon isotope record of lacustrine palaeoproductivity and climatic change derived from varved lake sediments of Lake Holzmaar, Germany. Quat Sci Rev 22:569–580
7. Ehteram M, Ferdowsi A, Faramarzpour M et al (2021) Hybridization of artificial intelligence models with nature inspired optimization algorithms for lake water level prediction and uncertainty analysis. Alexandria Eng J 60:2193–2208
8. Şen Z, Kadioğlu M, Batur E (2000) Stochastic Modeling of the Van Lake Monthly Level Fluctuations in Turkey. Theor Appl Climatol 65:99–110. https://doi.org/10.1007/s007040050007
9. Demir V (2022) Enhancing monthly lake levels forecasting using heuristic regression techniques with periodicity data component: application of Lake Michigan. Theor Appl Climatol. https://doi.org/10.1007/s00704-022-03982-0
10. Schulz S, Darehshouri S, Hassanzadeh E et al (2020) Climate change or irrigated agriculture–what drives the water level decline of Lake Urmia. Sci Rep 10:1–10s
11. Bengtsson L, Malm J (1997) Using rainfall-runoff modeling to interpret lake level data. J Paleolimnol 18:235–248
12. Zhu S, Lu H, Ptak M et al (2020) Lake water-level fluctuation forecasting using machine learning models: a systematic review. Environ Sci Pollut Res 27:44807–44819
13. Khan MS, Coulibaly P (2006) Application of Support Vector Machine in Lake Water Level Prediction. J Hydrol Eng. https://doi.org/10.1061/(asce)1084-0699(2006)11:3(199)
14. Altunkaynak A (2007) Forecasting surface water level fluctuations of lake van by artificial neural networks. Water Resour Manag 21:399–408. https://doi.org/10.1007/s11269-006-9022-6
15. Altunkaynak A, Sen Z (2007) Fuzzy logic model of lake water level fluctuations in Lake Van, Turkey. Theor Appl Climatol 90:227–233
16. Karimi S, Shiri J, Kisi O, Makarynskyy O (2012) Forecasting water level fluctuations of urmieh lake using gene expression programming and adaptive neuro-fuzzy inference system. Int J Ocean Clim Syst. https://doi.org/10.1260/1759-3131.3.2.109
17. Buyukyildiz M, Tezel G, Yilmaz V (2014) Estimation of the change in lake water level by artificial intelligence methods. Water Resour Manag 28:4747–4763. https://doi.org/10.1007/s11269-014-0773-1
18. Shafaei M, Kisi O (2016) Lake level forecasting using wavelet-SVR, wavelet-ANFIS and wavelet-ARMA conjunction models. Water Resour Manag 30:79–97. https://doi.org/10.1007/s11269-015-1147-z
19. Hrnjica B, Bonacci O (2019) Lake level prediction using feed forward and recurrent neural networks. Water Resour Manag. https://doi.org/10.1007/s11269-019-02255-2
20. Bonakdari H, Ebtehaj I, Samui P, Gharabaghi B (2019) Lake water-level fluctuations forecasting using minimax probability machine regression, relevance vector machine, gaussian process regression, and extreme learning machine. Water Resour Manag. https://doi.org/10.1007/s11269-019-02346-0
21. Wang Q, Wang S (2020) Machine learning-based water level prediction in lake erie. Water (Switzerland). https://doi.org/10.3390/w12102654
22. Fan C, Song C, Liu K et al (2021) Century-scale reconstruction of water storage changes of the largest lake in the inner mongolia plateau using a machine learning approach. Water Resour Res. https://doi.org/10.1029/2020WR028831
23. Yaseen ZM, Kisi O, Demir V (2016) Enhancing long-term streamflow forecasting and predicting using periodicity data component: application of artificial intelligence. Water Resour Manag 30:4125–4151. https://doi.org/10.1007/s11269-016-1408-5
24. Sattari MT, Sureh FS, Kahya E (2020) Monthly precipitation assessments in association with atmospheric circulation indices by using tree-based models. Nat Hazards 102:1077–1094
25. Deo RC, Samui P, Kim D (2016) Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. Stoch Environ Res risk Assess 30:1769–1784
26. Yaseen ZM, Ali M, Sharafati A et al (2021) Forecasting standardized precipitation index using data intelligence models: regional investigation of Bangladesh. Sci Rep. https://doi.org/10.1038/s41598-021-82977-9
27. Zahiri J, Mollaee Z, Ansari MR (2020) Estimation of suspended sediment concentration by M5 model tree based on hydrological and moderate resolution imaging spectroradiometer (MODIS) data. Water Resour Manag 34:3725–3737
28. Bahmani R, Solgi A, Ouarda TBMJ (2020) Groundwater level simulation using gene expression programming and M5 model

tree combined with wavelet transform. Hydrol Sci J 65:1430–1442. https://doi.org/10.1080/02626667.2020.1749762
29. Naganna SR, Beyaztas BH, Bokde N, Armanuos AM (2020) On the evaluation of the gradient tree boosting model for groundwater level forecasting. Knowledge-Based Eng Sci 1:48–57
30. Annin P (2018) The Great Lakes Water Wars. Island Press/Center for Resource Economics, Washington, DC
31. Vaccaro L, Read J (2011) Vital to Our Nation's economy: Great lakes jobs 2011 report. 7
32. Coulibaly P (2010) Reservoir computing approach to great lakes water level forecasting. J Hydrol 381:76–88. https://doi.org/10.1016/j.jhydrol.2009.11.027
33. Quinlan JR (1992) Learning with continuous classes. Mach Learn 92:343–348
34. Mitchell TM (1997) Machine Learning. McGraw-hill, New York
35. Solomatine DP, Xue Y (2004) M5 model trees and neural networks: application to flood forecasting in the upper reach of the huai river in China. J Hydrol Eng 9:491–501. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491)
36. Pal M, Deswal S (2009) M5 model tree based modelling of reference evapotranspiration. Hydrol Process. https://doi.org/10.1002/hyp.7266
37. Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–67. https://doi.org/10.1214/aos/1176347963
38. De Andrés J, Lorca P, de Cos Juez FJ, Sánchez-Lasheras F (2011) Bankruptcy forecasting: a hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). Expert Syst Appl 38:1866–1875
39. Sharda VN, Patel RM, Prasher SO et al (2006) Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques. Agric Water Manag 83:233–242. https://doi.org/10.1016/j.agwat.2006.01.003
40. Demir V, Çubukçu EA (2021) Digital elevation modeling with heuristic regression techniques abstract. Eur J Sci Technol. https://doi.org/10.31590/ejosat.916012
41. Bera P, Prasher SO, Patel RM et al (2006) Application of MARS in simulating pesticide concentrations in soil. Trans ASABE 49:297–307. https://doi.org/10.13031/2013.20228
42. Sephton P (2001) Forecasting recessions: can we do better on MARS? Review, vol. 83, pp.39–49
43. Al-Sudani ZA, Salih SQ, Sharafati A, Yaseen ZM (2019) Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. J Hydrol 573:1–12. https://doi.org/10.1016/j.jhydrol.2019.03.004
44. . Suykens JAK, Vandewalle J (1999) No Title. Neural Process Lett 9:293–300. Doi: https://doi.org/10.1023/A:1018628609742
45. URL1 (2022) LSSVR. http://www.esat.kuleuven.be/sista/lssvmlab/
46. URL2 MARS and M5Tree
47. URL3 Taylor Diagram
48. URL4 (2022) Boxblot & Violin plot
49. Tiyasha TTM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. J Hydrol 585:124670. https://doi.org/10.1016/j.jhydrol.2020.124670
50. Yaseen ZM (2021) An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. Chemosphere 277:130126. https://doi.org/10.1016/j.chemosphere.2021.130126
51. Legates DR, McCabe GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 35:233–241. https://doi.org/10.1029/1998WR900018
52. Tur R, Yontem S (2021) A Comparison of Soft Computing Methods for the Prediction of Wave Height Parameters. Knowledge-Based Eng Sci 2:31–46
53. Aoulmi Y, Marouf N, Amireche M et al (2021) Highly Accurate Prediction Model for Daily Runoff in Semi-Arid Basin Exploiting Metaheuristic Learning Algorithms. IEEE Access 9:92500–92515. https://doi.org/10.1109/ACCESS.2021.3092074
54. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106:7183–7192. https://doi.org/10.1029/2000JD900719
55. Sharafati A, Khosravi K, Khosravinia P et al (2019) The potential of novel data mining models for global solar radiation prediction. Int J Environ Sci Technol. https://doi.org/10.1007/s13762-019-02344-0
56. Hintze JL, Nelson RD (1998) Violin Plots: A Box Plot-Density Trace Synergism Statistical Computing and Graphics Violin Plots: A Box Plot-Density Trace Synergism. Source Am Stat 52:181–184
57. Legouhy A (2021) al_goodplot - boxblot & violin plot. In: MATLAB Cent. mathworks
58. Ebtehaj I, Bonakdari H, Gharabaghi B (2019) A reliable linear method for modeling lake level fluctuations. J Hydrol. https://doi.org/10.1016/j.jhydrol.2019.01.010
59. Yaseen ZM, Mohtar WHMW, Ameen AMS et al (2019) Implementation of univariate paradigm for streamflow simulation using hybrid data-driven model: Case study in tropical region. IEEE Access 7:74471–74481
60. Adnan RM, Liang Z, Heddam S et al (2020) Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. J Hydrol 586:124371. https://doi.org/10.1016/j.jhydrol.2019.124371
61. Hadi SJ, Abba SI, Sammen SSH et al (2019) Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. 1–16
62. Sudheer KP, Gosain AK, Ramasastri KS (2002) A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrol Process 16:1325–1330. https://doi.org/10.1002/hyp.554
63. Bhagat SK, Tung TM, Yaseen ZM (2020) Heavy metal contamination prediction using ensemble model: Case study of Bay sedimentation. Australia J Hazard Mater 403:123492. https://doi.org/10.1016/j.jhazmat.2020.123492
64. Sanikhani H, Kisi O, Maroufpoor E, Yaseen ZM (2018) Temperature-based modeling of reference evapotranspiration using several artificial intelligence models: application of different modeling scenarios. Theor Appl Climatol, Doi: https://doi.org/10.1007/s00704-018-2390-z
65. Beyaztas U, Salih SQ, Chau K-W et al (2019) Construction of functional data analysis modeling strategy for global solar radiation prediction: application of cross-station paradigm. Eng Appl Comput Fluid Mech 13:1165–1181
66. Oleiwi S, Jalal S, Hamed S et al (2018) Precipitation pattern modeling using cross-station perception: regional investigation. Environ Earth Sci. https://doi.org/10.1007/s12665-018-7898-0