https://doi.org/10.5194/egusphere-2022-427
Preprint. Discussion started: 5 July 2022

Using machine learning algorithms to analyze remote sensing and ground-truth Lake Chad's level data

Kim-Ndor Djimadoumngar[1]

[1]Ph.D. Geosciences and Computer Science, B.P. 172 Tel: (+237) 6 91 80 41 46 / (+235) 62 88 17 82, Kousseri – Cameroon

Correspondence to: Kim-Ndor Djimadoumngar (d.kimndor@gmail.com)

Abstract

Lake Chad is facing critical environmental situations since the 1960s due to the effects of climate change and anthropogenic activities on its ecosystems. The statistical analyses of remote sensing climate variables (i.e., evapotranspiration, specific humidity, soil temperature, air temperature, precipitation, soil moisture) and remote sensing and ground-truth lake level applied to the period 1993-2012 reveal that remote sensing lake level data has a skewed distribution and positive significant association with only soil moisture, whereas ground-truth lake level has a symmetrical distribution and negative significant associations with all the climate variables. The regression of remote sensing and ground-truth lake level onto climate variables using Linear Regression (LR), Support Vector Regression *(*SVR*)*, Regression Tree (RT), Random Forest Regression (RF), and Deep Learning (DL) methods show that (i) RF outperforms the other models with the highest coefficient of determination (R^2) and explained variance score (*EVS)* values and (ii) SVR has the lowest Mean Absolute Error (*MAE)*, Mean Squared Error (*MSE)*, and *k*-fold cross-validation (*k*-fold CV*)* values. The RF feature ranking function shows that soil temperature is the major driver of remote sensing lake level fluctuations, whereas precipitation is the first factor for ground-truth lake level. This study provides more in-depth knowledge of the factors influencing Lake Chad's level and perspectives for an integrated and forward-looking water management system for connecting climate change, vulnerability, human activities, and water balance research in the Lake Chad human-environment system. We cannot get the necessary ground truth data at this time because of the challenging security situations in the region. However, the development of the data analysis methodology reported here is of fundamental importance in understanding the water cycle dynamics in this important basin, even under challenging field conditions. Verification studies can be performed when more ground-truth data eventually become available.

Keywords: machine learning algorithms; remote sensing; ground-truth; Lake Chad's level.

1 Introduction

According to Magrin (2016) and the Food and Agricultural Organization (FAO, 2012), Lake Chad, located in the Lake Chad Basin (LCB), is undergoing environmental crises over the past half century due to climate change and anthropogenic activities, as it has been progressively shrinking since the 1960s. People in the surrounding areas have taken advantage of the lake region which has acted as a trading hub, offering economic opportunities and natural resources. There were cross-border economic activities in agricultural produce and fishing as well as other commodities (Nagarajan et al., 2018).

Because of the importance of water resource management for agriculture and the well-being of both humans and livestock, we believed that more accurate, reliable, qualitative, and quantitative predictions of water availability in such vulnerable regions are important. Moreover, most previous studies (Servant and Servant, 1983; Lévêque, 1987; Adamu, 2007) have focused on rainfall and evaporation as the factors affecting the lake yield in water. Our approach expanded on prior studies by examining the accuracy power of remote sensing climate variables in studying lake level fluctuations. Therefore, we decided to also address the driving forces affecting the changes in the lake level using both remote sensing and ground-truth data based on qualitative and quantitative methods.

The application of physical models was very limited in developing countries because of the high computational costs and data scarcity (Mohanty et al., 2009; Taormina et al., 2012). Alternatively, data-driven modeling (DDM) could be a solution.

It is advisable to apply various types of DDM and compare and/or combine the results because water-related applications are often characterized by noisy and poor quality data (Solomatine and Ostfeld, 2008).

Shiri et al. (2016) used the extreme learning machine (ELM) approach to predict daily water levels in the Urmia Lake. They found that ELM could accurately forecast the water level. They also found that ELM outperformed genetic programming (GP) and artificial neural networks (ANNs). Hipni et al. (2013) compared the support vector machine (SVM) model with the adaptive network-based fuzzy inference system (ANFIS) in forecasting daily dam water levels of the Klang gate. They found that SVM was a superior model to ANFIS when using the metrics such as root mean squared error (*RMSE*), mean absolute error (*MAE*), and mean absolute percentage error (*MAPE*). Coulibaly et al. (2001) modeled monthly water table depth fluctuations (1986 – 1996) in the Gondo aquifer in Burkina Faso using ANNs. They found that modeling was very important for groundwater management in areas where inadequate groundwater monitoring networks existed.

We decide to apply DDM in this study due to few hydrological data in the region. We assume that Lake Chad's level is a function of precipitation, soil moisture, air temperature, soil temperature, evapotranspiration, and specific humidity factors; precipitation is the only and most important climate variable on which all the other climate variable variations depend.

The principal goals of this study are to determine (i) how accurately remote sensing data can help study ground-truth data and (ii) what machine learning model(s) may be of best use to analyze both remote sensing and ground-truth data. Specifically, we aim to (i) examine the relationships between the respective aforementioned climate variables and remote sensing and ground-truth lake level; (ii) investigate the performances of different machine learning algorithms in estimation of remote sensing and ground-truth lake level data; and (iii) determine the major climate variable drivers of the fluctuations of remote sensing and ground-truth lake level.

This research will contribute to the general understanding of the hydrological processes in the Lake Chad basin and benefit both the scientific community and the decision-makers in (i) providing knowledge on the current state of environmental factors affecting Lake Chad's level, and (ii) developing data-driven models for future prediction and projection of lake level fluctuations. Ultimately, this will help make better predictions of water availability in the lake for local populations and other stakeholders.

1.1. Study area

According to the United Nations Environment Programme (UNEP, 2004), the Lake Chad sub-basin is shared among Chad, Niger, Nigeria, and Cameroon. It is located between 6° and 20° N, 7° and 25° E (Figure 1). It is a relic of a vast lacustrine surface area, which is equivalent to the Caspian Sea that existed in 600 AD. Its elevation ranges from 278 to 286 meters. Depending on the climate fluctuations, the lake filled all or a part of an endorheic basin of 25,000 km^2. The Chari-Logone river system is the main supply of water to Lake Chad and is located in the southern part of the basin. Another tributary of the lake is the Komadogou Yobe River in the western part of the basin (UNEP, 2004; USGS, 2018). This river system drains more than 610,000 km^2 in Southern Chad and the Central African Republic as well as areas of Cameroon and Western Sudan (Lévêque; 1987; Nagarajan et al., 2018).

As landscape, Lake Chad occupies a part of an erg oriented southeast-northwest. Its eastern shore was surrounded by dunes. The surrounding basin relief was extremely flat except for the Hadjer el Hamis rocks of volcanic origin. There were three major types of landscape: (i) many islands located along the eastern bank which matched the emerged summit of immerged erg dunes; (ii) rooted or floated vegetation islands called bench islands (mostly *Cyperus papyrus* and water reeds); and (iii) areas of open water (LCBC, 2014; Lévêque, 1987).

The close interaction between rainfalls and evaporation, the generation of lateral inflow to the lake, the groundwater leakage under the body of the lake, and human abstraction influence the overall lake water balance (UNEP, 2004). The persistent change in the rainfall patterns over the whole basin in the last 30 years has led to a shift of mean annual rainfall from 320 mm to less than 210 mm (Adamu, 2007; IAEA, 2017). The volume of the lake (72×10^9 m^3 in average) resulted from an equilibrium between water supplies essentially from rivers and losses particularly due to evaporation (Lévêque, 1987). The

United States Geological Survey (USGS, 2018) stated that approximately 90% of the rain falls from June to September with the lake quickly rising in November. The highest lake levels are in December, declining slowly for several months.

Crétaux and Birkett (2006) used satellite radar altimetry to estimate lake water variations. They concluded that altimetry is an additional useful tool for estimating river discharge and lake water prediction. Using the simple linear correlation methods, the estimated height of the permanent waters of the lake (i.e., 600 km downstream) 39 days in advance had a coefficient of determination (R^2) of about 93%. The prediction of water height on the western marshes of the lake-bed was poorer with an R^2 of 79% due to a change in response time of the local stage to the seasonal floods. According to Internationale Zusammenarbeit (GIZ, 2016), groundwater recharge depends on surface water, temporal and spatial distribution of rainfall, total annual rainfall, and the volume of runoff flowing towards topographic depressions. As a result of droughts and a significant decline in surface water over the past 40 years, groundwater levels have decreased in these areas, and some wells and boreholes have dried up (GIZ, 2016).

The World Food Programme (WFP, 2016), Magrin (2016), and FAO (2012) stated that when exemplifying the disproportionate effects of global climate change, the lake's recession increased water stress within an area battling drought and experiencing intense competition for multi-usage of the hydro-system. The consequent degradation of natural resources widely affected the subsistence livelihoods, regional food security, and quality of life of people inhabiting the LCB. The "disappearing lake" creates serious ecological issues, as land exposed by the receding shores is being used for farming or new settlements to accommodate the expanding population growth (FAO, 2012; Magrin, 2016; WFP, 2016).

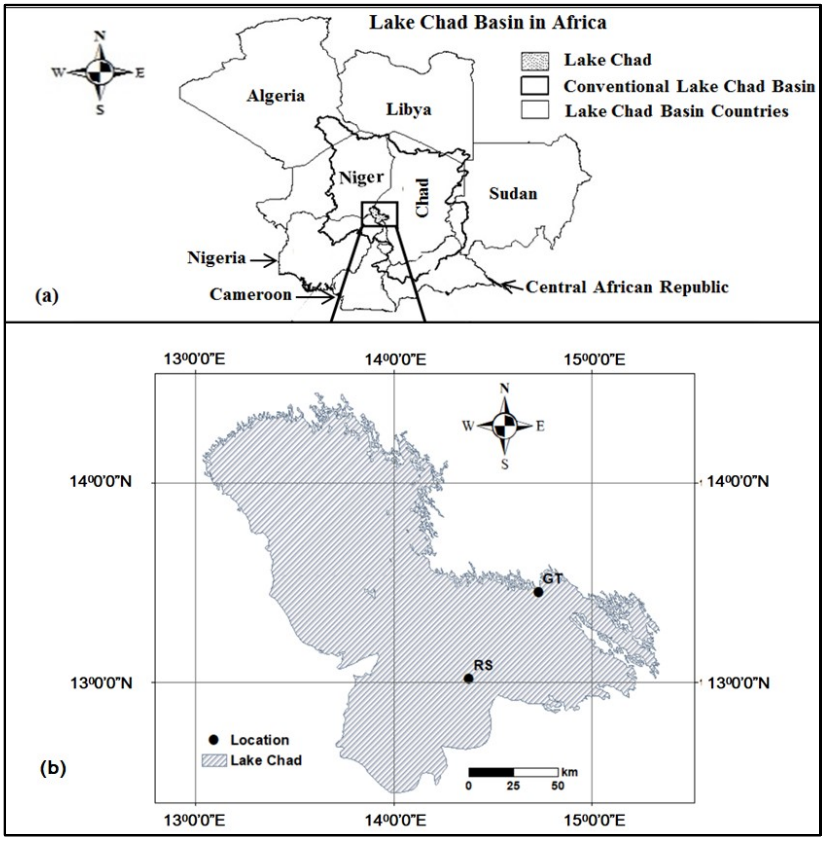

Figure 1. Map of the study area: (a) the Lake Chad in the LCB, (b) Lake Chad. GT is the location where ground-truth data are measured; RS is the location where remote sensing data were processed.

https://doi.org/10.5194/egusphere-2022-427
Preprint. Discussion started: 5 July 2022

2. Materials and methods

2.1. Data source

The dataset (Table 1) includes precipitation, air temperature, evapotranspiration, soil temperature, specific humidity, soil moisture, and remote sensing and ground-truth lake level data. We downloaded precipitation data from the Global Precipitation Climatology Center, GPCC (https://opendata.dwd.de/climate_environment/GPCC/html/fulldata-monthly_v2020_doi_download.html), monthly land-surface precipitation data from rain-gauges built based on global telecommunication system (GTS) and historical data (Schneider et al. (2020). GPCC is operated by Deutscher Wetterdienst (https://www.dwd.de/EN/ourservices/gpcc/gpcc.html) with the sponsorship of the World Meteorological Organization (WMO). The other input variables are from the Global Land Data Assimilation (GLDAS) model (Hiroko and Rodell, 2015; Rodell et al., 2004; Wharton, 2016). GLDAS data are accessible in the Giovanni data system, and are developed and maintained by the National Aeronautics and Space Administration Goddard Earth Sciences Data and Information Services Center, NASA DISC GES (https://giovanni.gsfc.nasa.gov/giovanni/). We downloaded the remote sensing lake level data from the Global Reservoirs/Lakes (https://ipad.fas.usda.gov/cropexplorer/global_reservoir/gr_regional_chart.aspx?regionid=wafrica&reservoir_name=Chad&lakeid=000068) of the United States Department of Agriculture's Foreign Agricultural Service (USDA-FAS). The data contains monthly Lake Chad height variations. We obtained the ground-truth lake level data from Société de Développement du Lac (SODELAC), a Chadian governmental agency in charge of developing Lake Chad resources. The time period was from 1993 through 2012, totaling 20 years, which provided 240 observations.

Remote sensing lake level data is processed at latitude 13.02 and longitude 14.38. To convert satellite product datum to an orthometric/mean sea level datum, we added 281.26 meters to each elevation in the lake product as instructed in the file downloaded from Global Reservoirs/Lakes. So, we have:

$$LL\_R_i = h_i + 281.26\ m \tag{1}$$

where LL_R_i is the remote sensing lake level of the ith day, h_i is the matching elevation.

For ground-truth lake level, the elevation data was collected at a hydrometric station located at latitude 13.45, longitude 14.73, and altitude of 277.87 meters. Therefore, we have:

$$LL\_G_i = h_i + 277.87\ m \tag{2}$$

where h_i is the ground-truth elevation of the ith day and LL_G_i its corresponding lake level, and $277.87\ m$ is the mean sea level of the limnimetric scale.

In both above cases, daily lake level data is average into monthly data.

Table 1. Label specifications table

Column Name	Tag	Unit	Temp. Resolution	Spatial Resolution	Type
AT	Air Temperature	°C	monthly	100 km x 100 km	Feature
ET	Evapotranspiration	kg/m^2/s	monthly	100 km x 100 km	Feature
P	Precipitation	mm	monthly	100 km x 100 km	Feature
SH	Specific Humidity	kg/kg	monthly	100 km x 100 km	Feature
SM	Soil Moisture	kg/m^2	monthly	100 km x 100 km	Feature
ST	Soil Temperature	°C	monthly	100 km x 100 km	Feature
LL_R	Remote sensing lake level	m	monthly	Not Applicable	Target
LL_G	Ground-truth lake level	m	monthly	Not Applicable	Target

2.2. Methods

The methods employed in this study are correlation coefficient analysis (r), Multiple Linear Regression (MLR), Support Vector Regression (SVR), Regression Trees (RT), Random Forests Regression (RF), and Deep Learning (DL). Table 2 has the brief descriptions of these respective methods.

Table 2. Methods' descriptions

Abbreviation	Purpose	Data Structure	Technique	Justification	Advantages
r [1, 2, 3]	Assessing the strength of the linear relationship between two variables.	Numerical variables	Finding the covariance of the variables, and divide it by the product of their standard deviations.	$\alpha = 0.05$, $p - value$, Confidence Interval, and $t - value$	Quantify the strength of the relationship between two variables.
MLR [1, 3]	Assessing the strength of the relationship between a target and many predictors individually.	A single target variable and many feature variables.	Finding the correlation and directionality of the data; then model fitting and assessment.	$p - value <$ significance level; functions minimizing the prediction error criterion.	A deeper knowledge of the association of each feature with the target variable.
SVR [4, 5]	Finding a hyperplane in an n-dimensional space that estimates a continuous-valued multivariate function.	Hyperplane to fit the data; support vectors: data nearest the hyperplane; kernel functions to transform an input data; boundary lines around the hyperplane .	Given data points SVR finds a curve to match the vector and the position of the curve.	A symmetrical loss function equally penalizing high and low estimates.	Excellent generalization powers and a high prediction accuracy
RT [6]	Forming incrementally a tree -with decision nodes and leaf nodes- from breaking down a dataset into smaller subsets.	A binary tree created by recursively splitting the data on the predictor values.	The prediction in each leaf is based on the weighted mean for node.	Splits that minimize the prediction squared error criterion.	Very interpretable models and low computational running time and storage needs.
RF [7, 8]	A prediction which is the average of the predictions made by the trees in the forest.	A combination of tree predictors such that each tree depends on the values of a random vector sampled independently.	Formed by growing trees depending on a random vector such that the tree predictor takes on numerical values.	A loss function that penalizes the predicted values that are far from the observed	Work well with default parameters; can be used for feature selection.
	Extraction of high-level, complex abstraction through a hierarchical learning process	A network of many layers of models, where each layer receives an input from the previous.	Initial layers extract low-level features; next layers combine features to form a full representation.	A loss function that minimizes the absolute or squared error criterion.	Great at revealing complex structures and a higher flexibility

DL	[9, 10]

1: Gareth et al. (2013); 2: Hayes (2013); 3: Kutner et al. (2004); 4: Awad and Khanna (2015); 5: Drucker et al. (2000); 6: Torgo (1997); 7: Breiman (2001); 8: Cutler et al. (2011); 9: Wehle, (2017); 10: Matthew et al. (2021);

2.3. Model accuracy evaluation criteria

The model error assessment measures (Table 3) used in this study are: coefficient of determination (**R^2**)**,** mean absolute error **(***MAE***),** mean squared error **(***MSE***)**, root mean squared error **(***RMSE***)**, explained variance score **(***EVS***),** and $k - fold$ cross-validation **(**$k - fold\ CV$**).**

Table 3: Performance metrics typology

Abbreviation	Name	Formula	Purpose
R^2	Coefficient of determination [1, 2]	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Measures the proportion of variability in the target variable explained by the predictors.
MAE	Mean Absolute Error [2, 3]	$MAE = \frac{\sum_{i*1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$	Calculates the average of the absolute error between the observed and fitted values of the target variable over n samples.
MSE	Mean Squared Error [2, 3]	$MSE = \frac{\sum_{i=}^{n}(y_i - \hat{y}_i)^2}{n}$	Computes the average of the squared difference between the observed and fitted values of the target variable for total n samples.
RMSE	Root Mean Squared Error [4]	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	The squared root of the *MSE*; it is the measure of the standard deviation of the residuals.
EVS	Explained Variance Score [2]	$EVS\ (y, \hat{y}) = \frac{Var\{y - \hat{y}\}}{Var\{y\}}$	Evaluates the strength of the relationship between the feature and the target variables.
$k -$ fold CV	$k -$ fold Cross-validation [5, 6, 7]	$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$	Provides the ability to estimate the model performance and its generalization to unseen data.

y_i is the true value of the target variable of the $i -$ th sample and $\hat{y}_i$ the corresponding fitted value; $\bar{y}$ represents the mean of the true values. k denotes the number of groups.

1: Hahn (1973); 2: Pedregosa et al. (2011); 3: Torgo (2014); 4: Neill and Hashemi (2018); 5: Gareth et al. (2013); 6: Hayes (2013); 7: Kutner et al. (2004).

Figure 2 shows our methodological approach from the data sources, the techniques used, to the results.

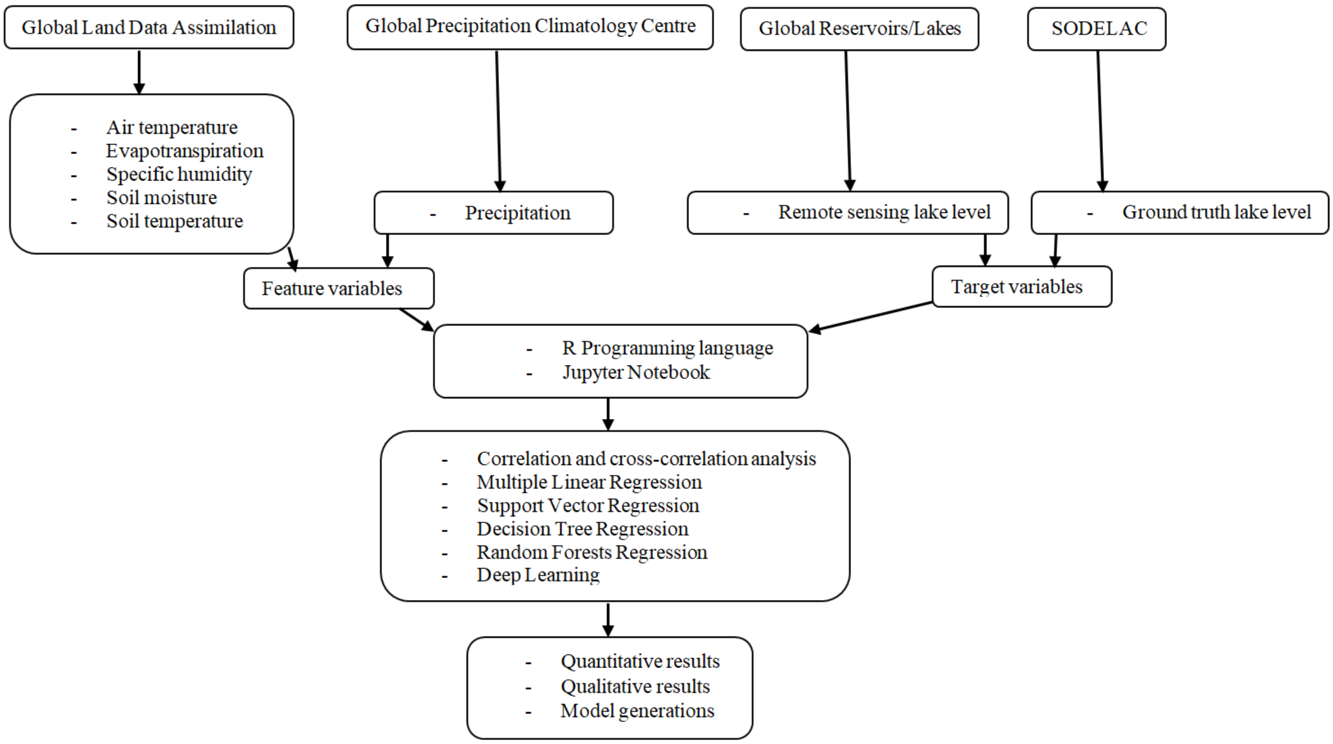

Figure 2. Flowchart diagram of the methodology

2.4. Settings and parametrization

The steps to proceed with our studies were as follows:

- Standardized the feature variables so that each of them was properly scaled and none of the them overdominated
- Performed the Variance Inflation Factor (VIF), a measure of collinearity between feature variables to check multicollinearity
- Split the dataset into training dataset (75%) and testing dataset (25%) using the *scikit-learn train_split*t function
- Trained the model using the training set. This neural network ran in regression mode, meaning it returned values
- Predicted the target variable using the testing dataset
- Computed the model performance metrics (*MAE*, *MSE*, *RMSE*, and *EVS*)
- Conducted the $k - fold$ CV to evaluate the prediction performance and compared the models with each other.

For MLR, SVR, RT, and RF, we used the functions with the default parameters from *Python scikit-learn* library. For deep learning, we used *Sequential* function from *Keras* library. The deep learning model created has two hidden layers with *Rectified Linear Unit* (*ReLU*) activation function and one output layer with *Linear* activation function. The hidden layers have 128 and 64 neurons, respectively, with all the six feature variables. We set Mean absolute error as the loss

https://doi.org/10.5194/egusphere-2022-427
Preprint. Discussion started: 5 July 2022

function. We replaced the input layers activation function by *Sigmoid* and *Tanh* and compared their metrics with the ones of *ReLU*.

3. Results and discussions

3.1. Data exploration and analysis

Figure 3 shows the seasonal pattern of all eight variables for the period of 1993 to 2012. The lake level data have almost the same temporal patterns as the feature variables. The similarity in the patterns of these variables is related to that of the rainy and dry seasons in the region since precipitation is the primary source of water in addition to Chari River and Logone River that feed Lake Chad. Soil moisture (SM), specific humidity (SH), and evapotranspiration (ET) timeseries follow the same pattern as that of precipitation but are shifted in time. Air temperature (AT) and soil temperature (ST) present bimodal distributions. Remote sensing and ground-truth lake level data follow the same patterns except for the years from 2003 to 2009 where remote sensing lake level data seems to have multimodal distribution while ground-truth lake level data have greater variations.

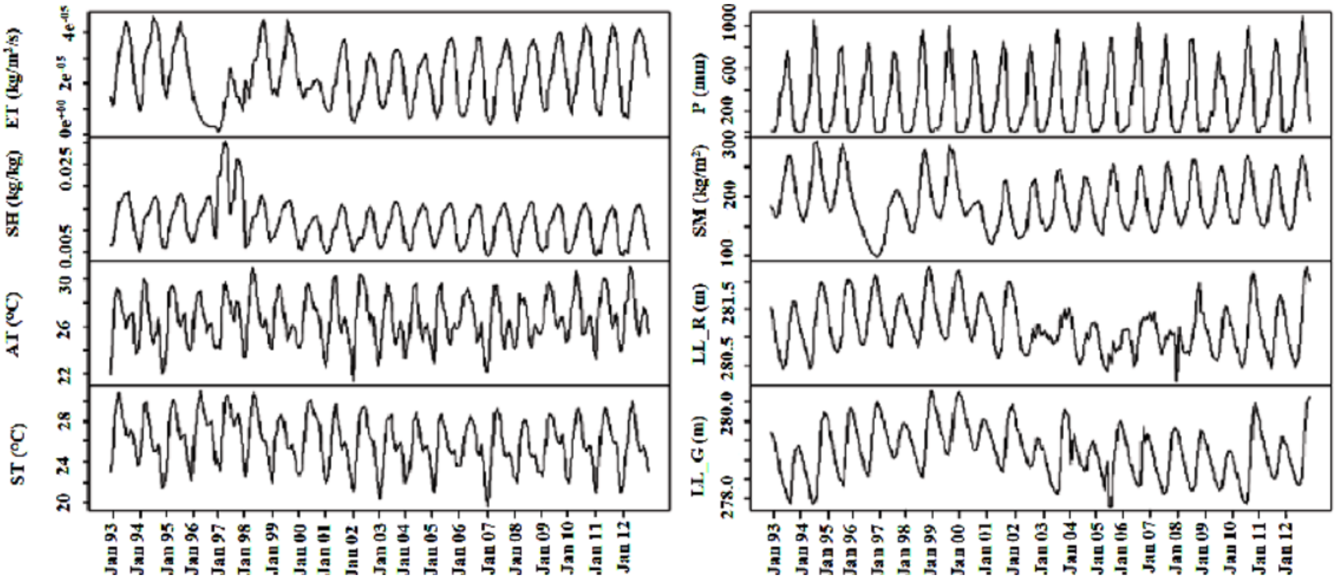

Figure 3 . Time series plot of: left panel ST, AT, SH, and ET; right panel: LL_G, LL_R, SM, and P. The abbreviation and units corresponding to the variables are listed in Table 1.

Figure 4 shows the distribution visualization of our target variables. The distributions are both unimodal (Figures 4a and b). Remote sensing lake level data have a positively highly right-skewed distribution, meaning its mean value is greater than its median. Ground-truth lake level data, on the other hand, have a symmetrical distribution; its mean and median are equal. The boxplots (Figures 4c and d) show that neither remote sensing nor ground-truth lake level data have outliers.

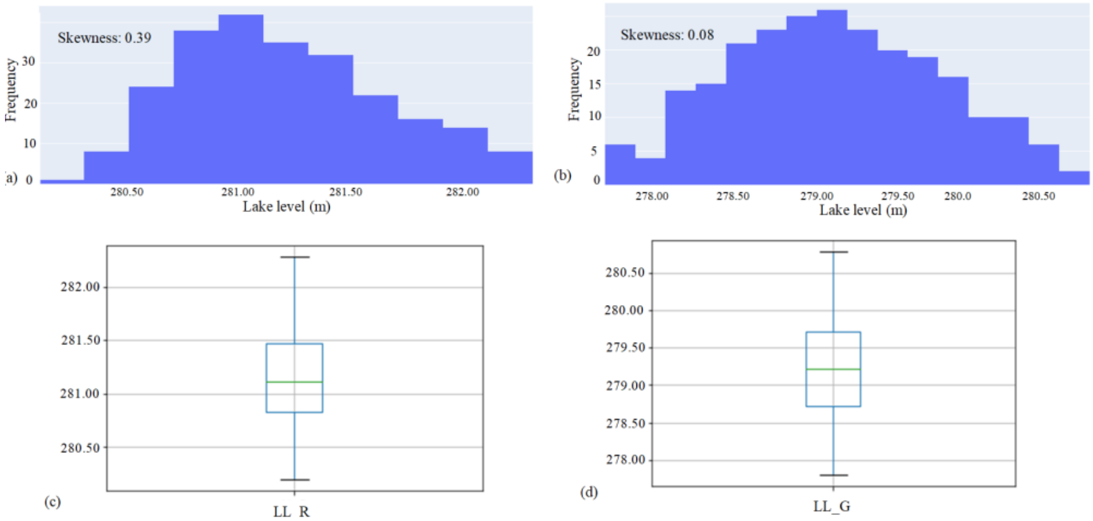

Figure 4: Distribution visualization of (a) remote sensing and (b) ground-truth lake level data

Table 4 shows the summary of the dataset. For the minima, maxima, and means of the target variables, we observed that min_{LL_R} is greater than min_{LL_G} by 2.40 meters, max_{LL_R} is greater than max_{LL_G} by 1.50 meters, and $mean_{LL\_R}$ is greater than $mean_{LL\_G}$ by 1.98 meters.

For the measure of the spread, considering the standard deviation, we have sd_{LL_R} is less than sd_{LL_G}. This signifies that remote sensing lake level data are clustered around its mean while ground-truth lake level data are more dispersed.

Based on the range, the difference between the maximum and the minimum, $range_{LL\_R} = max_{LL\_R} - min_{LL\_R} = 282.28 - 280.20 = 2.08$ is less than
$range_{LL\_G} = max_{LL\_G} - min_{LL\_G} = 280.78 - 277.80 = 2.98$. This is another evidence that ground-truth lake level data are more spread out than remote sensing lake level data.

The interquartile range (IQR) evaluation shows that:
$IQR_{LL\_R} = Q_{3(LL\_R)} - Q_{1(LL\_R)} = 281.47 - 280.83 = 0.64$ is less than
$IQR_{LL\_G} = Q_{3(LL\_G)} - Q_{1(LL\_G)} = 279.71 - 278.72 = 1.00$; this means that ground-truth lake level data vary a lot while remote sensing lake level data tend to be more or less the same.

The coefficient of variation (CV) –the ratio of the standard deviation to the mean- gives:
$CV_{LL\_R} = {}^{sd_{LL\_R}}/_{mean_{LL\_R}} = {}^{0.45}/_{281.19} = 0.0016$ is less than
$CV_{LL\_G} = {}^{sd_{LL\_G}}/_{mean_{LL\_G}} = {}^{0.69}/_{279.21} = 0.0025$. There is a greater dispersion of ground-truth lake level data around its mean since CV_{LL_G} is higher. The more precise will the estimates of remote sensing lake level data will be since CV_{LL_R} is lower.
We also see from Table that $mean_{LL\_R} > median_{LL\_R}$ while $mean_{LL\_G} = median_{LL\_G}$ confirming the skewed and symmetrical distributions of remote sensing and ground-truth lake level data, respectively.

Table 4: descriptive and quantile statistics of the dataset

	ET	SH	AT	ST	P	SM	LL_R	LL_G
Count	240.00	240.00	240.00	240.00	240.00	240.00	240.00	240.00

Min.	0.000001	0.004	21.37	19.63	0.87	98.57	280.20	277.80
1st Qu.	0.000013	0.007	25.20	24.44	16.51	158.04	280.83	278.72
Median	0.000023	0.012	26.47	25.96	193.50	182.27	281.10	279.21
Mean	0.000023	0.011	26.61	25.98	297.57	190.32	281.19	279.21
Sd.	0.000012	0.005	2.04	2.47	308.28	43.43	0.45	0.69
3rd Qu.	0.000033	0.015	28.18	28.07	504.83	222.01	281.47	279.71
Max.	0.000046	0.030	31.08	30.99	1086.74	292.34	282.28	280.78

3.2. Correlation coefficient and correlation test analyses

Table 5 shows the correlation coefficient and correlation test values. Evapotranspiration has positive and statistically significant correlation coefficients with specific humidity, soil temperature, precipitation, and soil moisture. It has a negative and statistically insignificant correlation with air temperature. Specific humidity has positive and statistically significant correlations with air temperature, soil temperature, precipitation, and soil moisture. Air temperature has a positive and statistically significant correlation with soil temperature, a positive statistically insignificant correlation with precipitation, and a negative and statistically significant correlation with soil moisture. Soil temperature has a positive and statistically significant correlation with precipitation; it has a negative and statistically insignificant correlation with soil moisture. Precipitation and soil moisture have a positive and statistically significant correlation. The couples soil moisture-evapotranspiration (0.89) and air temperature-soil temperature (0.85) are highly correlated; a regression analysis of lake level using them may cause multicollinearity. Therefore, we standardized the feature variables and performed variance inflation factor (VIF) before running the regression analysis.

The correlation test (ρ) helps verify whether the sample (i) has sufficient evidence to reject the null hypothesis. Therefore, we accept the alternative hypothesis and resolve that there is relationship between the variables in the population or (ii) has not satisfactory proof to reject the null hypothesis. Therefore, we accept the null hypothesis and state that there is no association between the variables in the population. Our two variables of interest are remote sensing (LL_R) and ground-truth lake level (LL_G). The significance level in this study is 5% ($\alpha = 0.05$). We interpreted only the $p - values$ less than the significance level (i.e., $p < 0.05$).

At 5% significance level, the correlation between remote sensing lake level and ground-truth lake level is positive, statistically significant and very large. The correlation coefficient is significantly different from 0 in both the sample and the population. The two target variables vary in the same direction. Thus, we do reject the null hypothesis and state that there is positive linear relationship between the two variables in the population.

Remote sensing lake level and evapotranspiration have a negative, statistically insignificant, and low correlation coefficient. Although the correlation coefficient is different from 0 in the sample, it is insignificantly different from 0 in the population. So, we accept the null hypothesis concluding that there is no linear association between remote sensing lake level and evapotranspiration. Remote sensing lake level has negative and statistically significant correlations with specific humidity (low), soil temperature (very large), air temperature (very large), and precipitation (medium). Remote sensing lake level decreases when the four feature variables increase and vice versa. On the other hand, it has positive and statistically significant correlation with soil moisture (low). The increase of soil moisture induces the increase of lake level. In both situations, the correlation coefficients are significantly different from 0 in the sample and in the population. Therefore we accept the alternative hypothesis that there are linear associations (negative in the first, positive in the latter) between our target and feature variables.

Ground-truth lake level has negative and statistically significant correlations with all the features variables, evapotranspiration, specific humidity, soil temperature, air temperature, precipitation, and soil moisture. The correlation coefficients are significantly different from 0 in both the sample and the population. We accept the alternative hypothesis and state that there are negative linear relationships between ground-truth lake level and the feature variables. Our target and feature variables change in opposite direction. An increase of the feature variables causes a decrease of ground-truth lake level and vice versa.

Table 5. Correlation coefficients (r) and *p-value* values between lake level and the climate variables

Parameter1	Parameter2	*r*	95% *CI*	*t*(238)	*p-value*
ET	SH	0.48	[0.38, 0.57]	8.46	< .001***
ET	AT	-0.03	[-0.15, 0.10]	-0.42	> .999
ET	ST	0.18	[0.05, 0.30]	2.77	0.043*
ET	P	0.70	[0.63, 0.76]	15.03	< .001***
ET	SM	0.89	[0.86, 0.92]	30.42	< .001***
ET	LL_R	-0.09	[-0.21, 0.04]	-1.36	0.597
ET	LL_G	-0.49	[-0.58, -0.38]	-8.56	< .001***
SH	AT	0.19	[0.06, 0.31]	2.96	0.027*
SH	ST	0.51	[0.41, 0.60]	9.19	< .001***
SH	P	0.63	[0.55, 0.70]	12.59	< .001***
SH	SM	0.42	[0.31, 0.52]	7.14	< .001***
SH	LL_R	-0.17	[-0.29, -0.04]	-2.63	0.054
SH	LL_G	-0.43	[-0.53, -0.33]	-7.44	< .001***
AT	ST	0.85	[0.81, 0.88]	24.80	< .001***
AT	P	0.00124	[-0.13, 0.13]	0.02	> .999
AT	SM	-0.28	[-0.39, -0.15]	-4.43	< .001***
AT	LL_R	-0.46	[-0.55, -0.35]	-7.97	< .001***
AT	LL_G	-0.27	[-0.38, -0.15]	-4.31	< .001***
ST	P	0.29	[0.17, 0.41]	4.76	< .001***
ST	SM	-0.09	[-0.22, 0.03]	-1.45	0.597
ST	LL_R	-0.50	[-0.59, -0.40]	-9.00	< .001***
ST	LL_G	-0.43	[-0.53, -0.32]	-7.42	< .001***
P	SM	0.63	[0.54, 0.70]	12.41	< .001***
P	LL_R	-0.24	[-0.35, -0.11]	-3.76	0.002**
P	LL_G	-0.66	[-0.72, -0.58]	-13.45	< .001***
SM	LL_R	0.13	[0.01, 0.26]	2.08	0.039
SM	LL_G	-0.32	[-0.43, -0.21]	-5.27	< .001***
LL_R	LL_G	0.66	[0.59, 0.73]	13.70	< .001***

p-value adjustment method: Holm (1979). Observations: 240

- **Variance Inflation Factor**

After standardizing the features, we ran variance inflation factor (VIF) since evapotranspiration and soil moisture as well as air temperature and soil temperature are highly correlated. The purpose is to check the multicollinearity between features using the VIF rules of thumb. For standardized data, a $VIF_i > 10$ is a sign of harmful collinearity (Kennedy, 1992). There is multicollinearity if (i) the largest VIF exceeds 10 and (ii) the mean of all of the VIF's largely exceeds 1. A VIF of 10 and 4 are considered as a proof of excessive or serious multicollinearity (Chatterjee and Price, 1991), and are often used to question the results of analyses that are pretty strong on statistical bases (O'Brien, 2007). With these VIF values, the attempt is to eliminate one or more variables from the analysis to reduce multicollinearity; VIF exceeding a threshold value can even lead to a rejection of paper by a manuscript reviewer. O'Brien (2007), when analyzing these rules of thumb, showed that they should be put into the context of the effects of other factors that affect the variance of the regression coefficients. VIF values of 10 or even higher do not, by themselves, discredit the results of regression analysis, do not call for the elimination of one or more feature variables, do not advise using ridge regression, nor do they require combining feature variables into a single index. We can much assuredly draw conclusions from regression analysis even with VIF values over the rules of 4 or 10. The conviction can depend on *t-values* and/or *confidence interval*. Our null and alternative hypotheses in this analysis are as follows:

H_0: there is no multicollinearity between the features

H_1: there is multicollinearity between the features

Table 6 has the VIF values corresponding to each of our feature variables. Using the rule of thumb of 10, our results are acceptable since all the VIF values are well below 10, and the associations are statistically very significant ($p < 0.001$). So, we conclude that there is no multicollinearity between all the features; therefore we accept the null hypothesis.

If we consider the rule of thumb of 4, evapotranspiration and soil moisture, although highly correlated, have a statistically very significant relationship ($p < 0.001$)**,** with the narrowest confidence interval, and the largest $t - value$. Thus, we are confident about the VIF values of evapotranspiration and soil moisture, we and accept the null hypothesis. Therefore, we state that there is no collinearity between these two features. Likewise, though air temperature and soil temperature are highly correlated, their association is statistically very significant with the second narrowest confidence and the second largest $t - value$. We accept the null hypothesis and conclude that there is not collinearity between air temperature and soil temperature.

Since we standardized our data and all the VIF_i values are less than 10, we state that there is no collinearity. In addition, our maximum VIF being less than 10, we conclude that there is inconsequential multicollinearity affecting the estimates of our regression analysis.

Table 6. VIF table

Feature	ET	SH	AT	ST	P	SM
VIF	7.97	2.63	5.45	7.97	2.66	8.23

3.3. Predictions

Figure 5a shows the predictions of remote sensing lake level data. Figure 5b presents the results from predicting ground-truth lake level data. For both remote sensing and ground-truth data, the magnitude and general variation of predicted lake level from LR, SVR, RT, and RF are significantly closer to the observed lake level as compared to the predictions from DL.

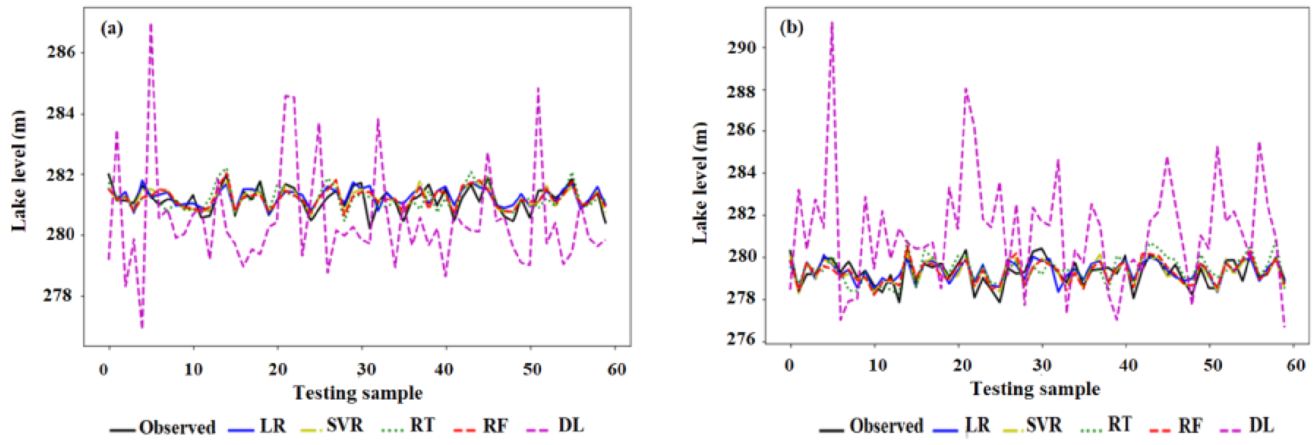

Figure 5: Comparison of observed and predicted lake level using (a) remote sensing and (b) ground-truth lake level data with linear regression (LR), support vector regression (SVR), regression tree (RT), random forest (RF), and deep learning (DL).

Table 7 shows the predictive performances of our five models on training and testing datasets for remote sensing and ground-truth lake level data, respectively. The best performances with regard to each model are in bold type.

The results from both remote sensing and ground-truth lake level data analysis show that RT has a perfect performance on training dataset, followed by RF, SVR, and LR; DL gives the worst performances with negative R^2 and EVS values. However, the CV_{MSE} results indicate that SVR outperforms the other models followed by RF, LR, RT, and DL, respectively. On the testing dataset, it is rather SVR that has better performances, followed by RF, LR, and RT in terms of R^2 values. LR and RF models are approximately equal in terms of MAE, MSE, and EVS values. The CV_{MSE} results show that SVR and LR perform equally, followed by RF, RT, and DL, respectively.

Table 7. Statistical performances of regression models on training and testing datasets for remote sensing and ground-truth lake level data.

	Remote sensing lake level					Ground-truth lake level				
Training model	LR	SVR	RT	RF	DL	LR	SVR	RT	RF	DL
R^2 (%)	0.33	0.73	**100**	0.93	-1.50	0.52	0.71	**100**	0.94	-15.44
MAE	0.29	0.18	**0.00**	0.09	0.61	0.40	0.30	**0.00**	0.14	2.30
MSE	0.14	0.06	**0.00**	0.02	0.53	0.24	0.14	**0.00**	0.03	8.08
$RMSE$	0.37	0.24	**0.00**	0.12	0.73	0.49	0.37	**0.00**	0.17	2.84
EVS	0.33	0.73	**1.00**	0.93	-1.34	0.51	0.71	**1.00**	0.94 -	-7.30
CV_{MSE}	0.15	**0.09**	0.18	0.11	1.03	0.26	**0.22**	0.40	**0.22**	6.43
Testing model										
R^2	0.38	**0.49**	0.09	0.40	-19.05	0.51	**0.62**	0.17	0.56	-25.58
MAE	0.25	**0.23**	0.32	0.25	1.56	0.37	**0.32**	0.46	0.34	2.56
MSE	0.11	**0.09**	0.16	0.11	3.65	0.20	**0.15**	0.31	0.18	11.17
$RMSE$	0.33	**0.30**	0.40	0.33	1.91	0.45	**0.39**	0.56	0.42	3.34
EVS	0.43	**0.51**	0.13	0.40	-16.62	0.51	**0.62**	0.27	0.56	-16.66
CV_{MSE}	**0.11**	**0.11**	0.24	0.15	2.13	**0.20**	**0.20**	0.33	0.24	13.86

The results from Table 7 above suggest that RF and SVR, respectively, seem the best models to further investigate remote sensing and ground-truth lake level data because of their respective highest R^2 and *EVS* and least *MAE* and *MSE* values. These results would recommend LR as a third model of choice instead of RT because LR training and testing metrics are very close, whereas RT has a perfect fit on training dataset but poorly predicts on testing dataset. DL, despite its worst metrics in this present case, should not be ignored for further studies with increased data because of its ability to improve its performance with the increase in the training sample size.

Figures 6 and 7 show the residuals plots for LR, SVR, RT, and RF models using remote sensing and ground-truth lake level data, respectively. LR model using remote sensing data (Figure 6a) has a quite random and uniform distribution of residuals against the target in two dimensions on both training and testing datasets. The histogram also shows normally and multimodal distributed errors around zero for both datasets. This seems to show that LR performs well. Using ground-truth data (Figure 7a) with both training and testing datasets, residuals are more dispersed below the horizontal axis with an outlier (for testing dataset) above. In addition, the errors distribution is right-skewed for both datasets. This appears to prove that LR is not a well fitted model for ground-truth lake level data.
SVR model using remote sensing data (Figure 6b) presents a random and uniform dispersion of residuals around the horizontal axis for both training and testing datasets; the errors distributions around zero are normal. This seems to indicate that SVR fits well on training and testing remote sensing data. In the case of ground-truth data (Figure 7b), we see that, although the dispersion is random and uniform, the training data residuals are mostly closer to the horizontal axis, whereas the testing data residuals are more dispersed with an outlier. The histogram also shows that the training errors are normally distributed and closer to zero whereas the testing errors are widely spread. This may indicate that SVR fits well on training ground-truth data but poorly predicts on testing data.
Using both remote sensing and ground-truth data, RT model (Figure 6c and Figure 7c) has perfect fits on training datasets, meaning the residuals line up with the horizontal line. On testing datasets, however, the residuals show random and uniform distributions. The histograms show residuals lining with zero for both training and testing datasets. This illustrates that RT does not fit well the testing datasets.
RF models, using remote sensing and ground-truth lake level data, respectively (Figure 6d and Figure 7d), show uniform distributions of residuals around the horizontal axis for both training and testing dataset. However, testing data residuals are more widely spread than the training data residuals. Furthermore, the histograms show that the training residuals have normal distributions while testing residuals do not. This seems to explain that RF performs well on training datasets but not on testing datasets.

https://doi.org/10.5194/egusphere-2022-427
Preprint. Discussion started: 5 July 2022

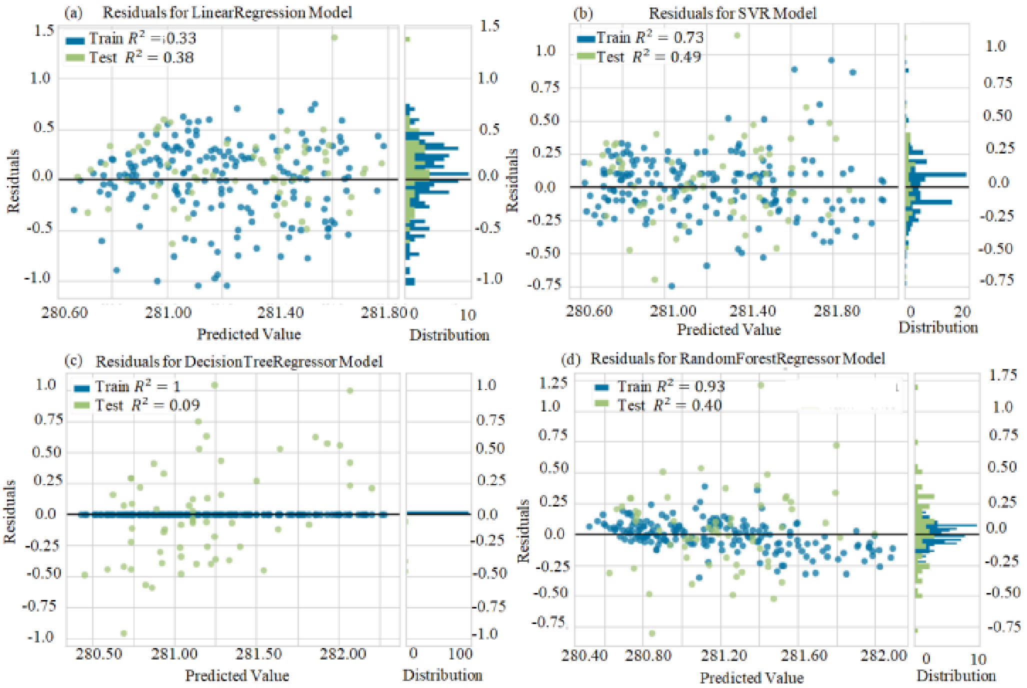

Figure 6: Plots of residuals from remote sensing lake level data models (a) LR, (b) SVR, (c) RT, and (d) RF

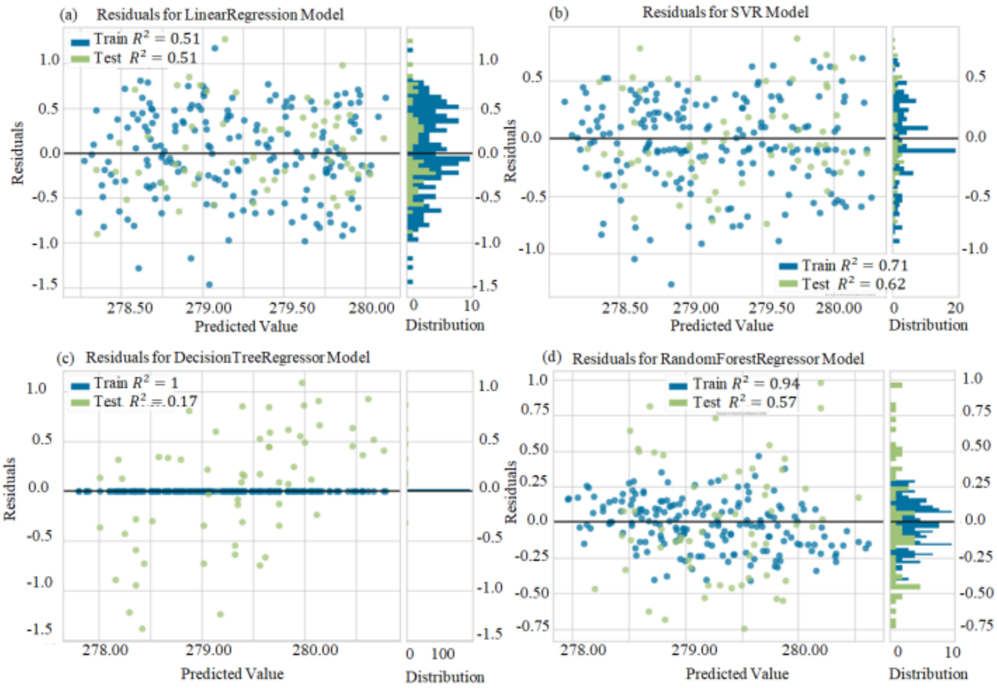

Figure 7: Plots of residuals from ground-truth lake level data models (a) LR, (b) SVR, (c) RT, and (d) RF

Figure 8 shows the plots of deep learning analysis, using *ReLU* activation function, of remote sensing (row 1) and ground-truth (row 2) lake level. Our loss function is the mean absolute error. Both training and validation losses decrease exponentially with the increase of epochs and quickly converge at around 40 epochs (Figure 8a) and 20 epochs (Figure 8d). The losses decrease to a level of stability with small gaps between them. This seems to be a good fit (meaning there is neither overfit nor underfit). The *MSE* values in both cases (Figures 8b and 8e) follow the same patterns and reach a plateau from around 60 epochs to the end. The R^2 values (Figures 8c and 8f), although exponentially increase with the increase in number of epochs, are always negative and reach a stability at around 60 epochs.

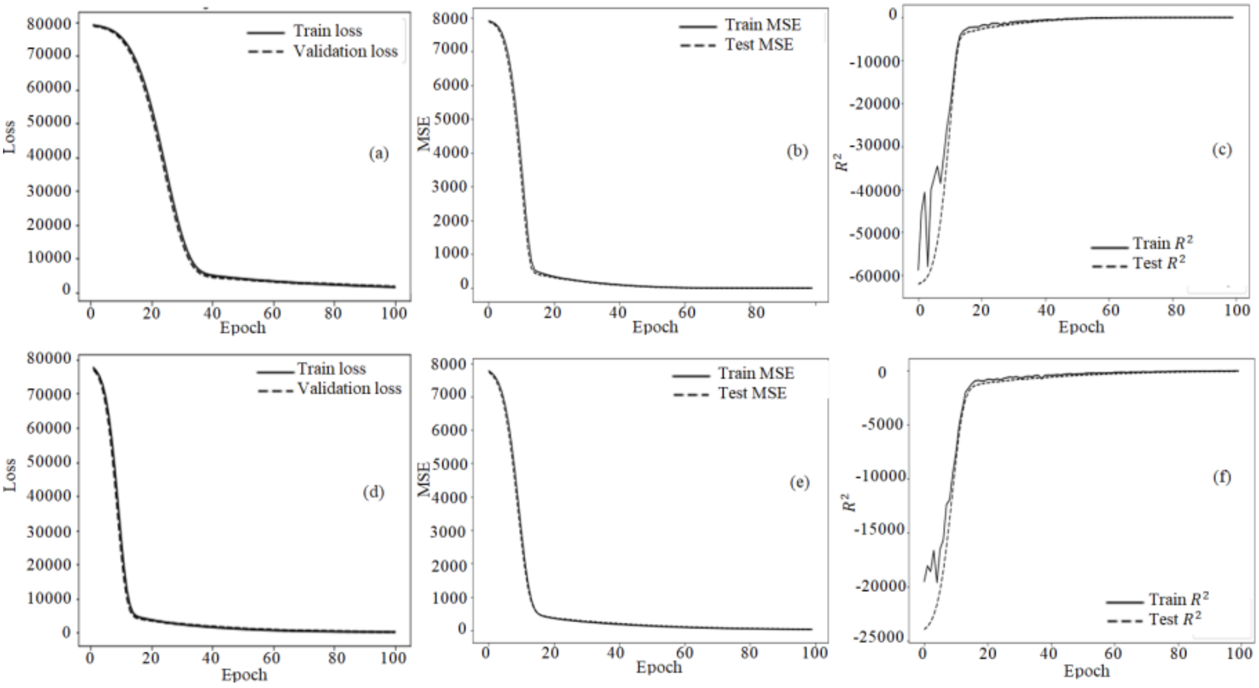

Figure 8: Plots of deep learning model metrics using *ReLU* activation function

a. Activation functions comparison

We also compared *ReLU* activation function with *Sigmoid* and *Tanh* activation functions to see which one performs better in predicting remote sensing and ground-truth lake level data. The results (Table 8) show that, for the present study, although deep learning using *ReLU* performs worse than the other models in consideration (i.e., LR, SVR, RT, and RF), changing the activation function to *Sigmoid* and *Tanh* does not improve the model. The ordering of metrics, except for *EVS* which has $Sigmoid > Tanh > ReLU$ as order, is $ReLU > Tanh > Sigmoid$.

Table 8. Statistical performances of *ReLU*, *Sigmoid*, and *Tanh* activation functions using remote sensing and ground-truth lake level data

Remote sensing data

	R^2	*MAE*	*MSE*	*EVS*	CV_{MSE}
ReLU	-16910.38	42.18	2065.40	-15386.43	728.22
Sigmoid	-352931.11	253.52	64274.93	-0.0055	44831.77
Tanh	-297346.09	232.70	54151.95	-0.44	41122.68

Ground-truth data

ReLu	-4785.58	34.40	2011.05	-4411.00	779.87
Sigmoid	-146713.14	248.27	61641.21	-0.002	44238.34
Tanh	-126664.70	230.69	53217.96	-0.003	40370.29

b. Feature importance

In Table 9, we have the feature ranking for remote sensing and ground-truth lake level data, respectively. The most important feature for remote sensing lake level is soil temperature, which is followed by soil moisture, specific humidity, air temperature, precipitation, and evapotranspiration, respectively. For ground-truth lake level, precipitation is the most important feature, closely followed by specific humidity, then air temperature. Soil moisture, soil temperature, and evapotranspiration, respectively, are the least important.

We can notice that the first two most important (0.50) features (ST and SM) for remote sensing lake level are, in reverse order, the first two least important (0.13) features (SM and ST) for ground-truth lake level. Precipitation is among the less important features for remote sensing lake level while it is most important feature for ground-truth lake level. Specific humidity and air temperature are more important for ground-truth lake level. Evapotranspiration is the least important for both targets.

Table 9. The importance of features for remote sensing and ground-truth lake level, respectively.

Remote sensing lake level		Ground-truth lake level	
Feature	Importance	Feature	Importance
ST	0.34	P	0.31
SM	0.16	SH	0.29
SH	0.14	AT	0.21
AT	0.13	SM	0.07
P	0.12	ST	0.06
ET	0.11	ET	0.05

3.4. Algorithms comparison

Figure 9 shows the plots comparing the *MAE* and *MSE* values for LR, SVR, RT, and RF algorithms, respectively, using the whole dataset of remote sensing (row 1) and ground-truth (row 2) lake level data. The results suggest that *SVR* and *RF* seem worthy to use for further study on lake level data since they have the first two lowest *MAE* and *MSE* values. The ordering of these algorithms, based on their *MAE* and *MSE* scores for both remote sensing and ground-truth data is: SVR > RF > LR > RT.

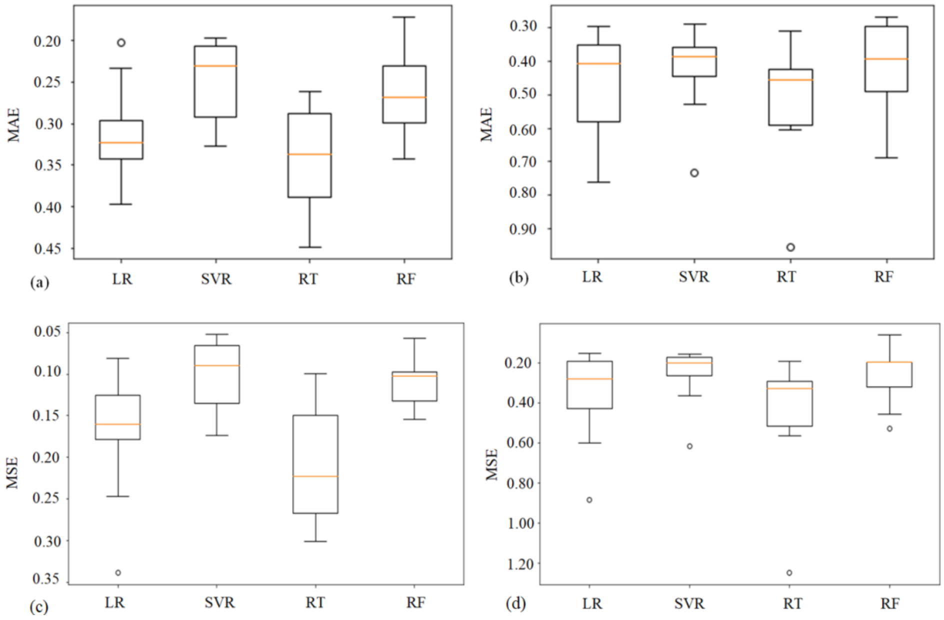

Figure 9. Comparison of LR, SVR, RT, and RF algorithms based on *MAE* and *MSE* metrics.

These algorithm comparison results confirm the statistical performances of regression models (Table 7) for which RF and SVR seem to be the best models based on their higher R^2 and lower *MAE* and *MSE* values.

Table 10 summarizes the training and testing *MAE* and *MSE* values for LR, SVR, RT, and RF algorithms using remote sensing and ground-truth lake level data, respectively. Using remote sensing data, SVR outperforms RF*,* followed by LR and RT, respectively, based on both training and testing *MAE* and *MSE* values (i.e. SVR > RF > LR > RT). When using ground-truth data, based on both training *MAE* and *MSE*, the relationships are SVR = RF > LR > RT; the testing *MAE* gives SVR > LR > RF > RT as relationships, whereas the testing is *MSE* is SVR = LR > *RF* > *RT*.

Once more, these results confirm those from Table 7 and Figure 9 based on which SVR and RF seem to be the most suitable algorithms to study lake level data, followed by LR and RT.

Table 10. Statistical performances of LR, SVR, RT, and RF algorithms based on training and testing *MAEs* and *MSEs* metrics using remote sensing and ground-truth lake level data

Remote sensing lake level

Training dataset

	LR	SVR	RT	RF
MAE	0.30	0.23	0.32	0.26
MSE	0.15	0.09	0.18	0.12

Testing dataset

	LR	SVR	RT	RF
MAE	0.25	0.24	0.37	0.28
MSE	0.11	0.11	0.25	0.14

Ground-truth lake level

Training dataset					
		LR	SVR	RT	RF
	MAE	0.42	0.38	0.52	0.38
	MSE	0.26	0.22	0.40	0.22
Testing dataset					
		LR	SVR	RT	RF
	MAE	0.37	0.36	0.52	0.38
	MSE	0.20	0.20	0.32	0.25

4. Conclusion

The study of remote sensing and ground-truth Lake Chad level data from 1993 to 2012 using remote sensing climate variables (i.e., evapotranspiration, specific humidity, soil temperature, air temperature, precipitation, soil moisture) shows that remote sensing lake level has a skewed distribution, whereas ground-truth lake level has a symmetrical distribution. Remote sensing lake level shows a positive and significant relationship with only soil moisture, and ground-truth lake level has a negative and significant relationship with all the feature variables. These associations do not undeniably mean causation.

The results also reveal that Random Forest Regression and Support Vector Regression seem to be the most suitable models to analyze remote sensing and ground-truth lake level data. Deep Learning should, however, be considered if more data is available. Linear Regression and Regression Tree do not fit well the two target variables. Soil temperature and soil moisture are the most important input features for remote sensing lake level, however they are the first two least important for ground-truth lake level. Precipitation, specific humidity, and air temperature are the factors influencing the most ground-truth lake level, while their importance is less for remote sensing lake level.

This indicates that ground-truth data, despite their scarcity, should be always and carefully considered to validate data-driven environmental models. These findings can serve as the basis for understanding how the remote sensing and ground-truth data can be used in the study of hydrologic processes in the basin. Validation studies are needed when a greater amount of remote sensing and ground-truth data is available.

Author contribution

Kim-Ndor Djimadoumngar conceptualized the research ideas, formulated the research goals and interests, collected and processed the data need for the study, applied statistical and machine learning methods to develop models, analyzed and interpreted the results, and prepared and wrote the manuscript.

Acknowledgements

I am grateful to Ms. Nancy Lorenz and Ms. Ann Marie Scahill language help in writing this manuscript. I also thank Mr. Pismon Doussoué, Mr. Roger Mbangossoum, and Mr. Ngarmadji Sangar from Société de Développement du Lac (SODELAC) for providing me with the ground-truth data.

Code and/or data availability

The input data used to produce the results used in this paper was downloaded from the Global Precipitation Climatology Centre, GPCC (https://opendata.dwd.de/climate_environment/GPCC/html/fulldata-monthly_v2020_doi_download.html) operated by Deutscher Wetterdienst (https://www.dwd.de/EN/ourservices/gpcc/gpcc.html) and the Global Land Data Assimilation (GLDAS) model accessible in the Giovanni data system, developed and maintained by the National Aeronautics and Space Administration Goddard Earth Sciences Data and Information Services Center, NASA DISC GES (https://giovanni.gsfc.nasa.gov/giovanni/). The output data (lake level) is available at the Global Reservoirs/Lakes (https://ipad.fas.usda.gov/cropexplorer/global_reservoir/gr_regional_chart.aspx?regionid=wafrica&reservoir_name=Chad&lakeid=000068) of the United States Department of Agriculture's Foreign Agricultural Service (USDA-FAS). The dataset and Jupyter Notebook and R scripts to run the models and produce the plots for all the simulations presented in this paper are found in the Dryad data repository (https://datadryad.org/stash/share/kn8fq8cdn1o1MFftLT7xmpzlqQe6IZsDudfJ2rbwGD0) with the corresponding unique identifier doi:10.5061/dryad.w6m905qms.

References

Adamu M.S.: Groundwater as part of the work of Lake Chad Basin Commission (LCBC). Lake Chad Basin Commission, 2007.Available at https://www.bgr.bund.de/EN/Themen/Wasser/Politikberatung_GW/Downloads/pbgw_africaAdamuP.html (accessed 22 October 2018).

Awad, M and Khanna, R.: Support vector regression. In: Efficient Learning Machines. 67-80, Apress, Berkeley, CA, DOI: 10.1007/978-1-4302-5990-9_4, 2015.

Breiman, L, Friedman, J., Olshen, R., and Stone, C.: Classification and Regression Trees. Wadsworth Int. Group, 1984.

Breiman, L.: Random Forest. Mach. Learn. 45, 5–32. Kluwer Academic Publishers. Manufactured in The Netherlands, 2001.

Chatterjee, S. and Price, B.: Regression analysis by example, 2nd.Ed., New York: Wiley, 1991.

Coulibaly, P., Anctil, F., Aravena, R., and Bobee, B.: Artificial neural network modeling of water table depth fluctuations. Water Resour. Res., 37 (4), 885-896, 2001.

Crétaux, J. F. and Birkett, C.: Lake studies from satellite radar altimetry. C.R. Geosci., 338(14-15), 1098-1112, 2006.

Cutler, A., Cutler, D. R., and Stevens, R. J.: Random Forests. In: Ensemble machine learning, 157-175, Springer, Boston, MA, DOI: 10.1007/978-1-4419-9326-7_5, 2011.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V.: Support vector regression machines. Neural Information Processing Systems 9, Eds. MC Mozer, JI Jordan, T Petscbe. pp. 155-161, MIT Press, 2000.

FAO, Food and Agricultural Organization: Climate change implications for fishing communities in the Lake Cha d Basin. What have we learned and what can we do better? FAO/Lake Chad Basin Commission Workshop, 18-20 November, 2012.

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R.: An introduction to statistical learning with applications in R. ISBN 978-1-4614-7138-7 (eBook) Springer, 2013.

GIZ, Internationale Zusammenarbeit, GmbH: Report on the State of the Lake Chad Basin Ecosystem, 2016. Available at www.cblt.org/sites/.../report_on_the_state_of_the_lake_chad_basin_ecosystem.pdf (accessed 17 February 2018).

Hahn, G. J.: The coefficient of determination exposed! CHEMITECH, 3(10), 609-612, 1973.

Hayes, A. F.: Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Publications, 2013.

Hipni, A., El-shafie, A., Najah, A., Karim, O. A., Hussain, A., and Mukhlisin, M.: Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). Water Resour. Manage., 27(10), 3803-3823, 2013.

Hiroko, B. and Rodell, M.: NASA/GSFC/HSL. GLDAS Noah Land Surface Model L4 monthly 1.0 x 1.0 degree V2.0, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), 2015. Available at 10.5067/QN80TO7ZHFJZ (accessed 12 December 2017).

IAEA, International Atomic Energy Agency: Integrated and sustainable management of shared aquifer systems and basins of Sahel region. Report of the IAEA-supported regional. Technical cooperation project RAF/7/011, 2017.

Kennedy, P.: A guide to Econometrics. Oxford: Blackwell, 1992.

Keras. Available at https://keras.io/about/(accessed 3 February2022).

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W.: Applied linear statistical models, 5th. Edition, McGraw-Hill, 2004.

LCBC, Lake Chad Basin Commission: The Lake Chad Basin, 2014. Available at http://www.cblt.org/en/lake-chad-basin (accessed 27 July 2018).

Lévêque, C.: *Lac Tchad. African wetland and shallow water bodies. Zones humides et lacs peu profonds d'Afrique*, Repertoire, Region 4, Chad Basin, 233-251, Editions de L'ORSTOM. Institut Français de la Recherche Scientifique pour le Développement en Coopération, Collection Travaux et Documents N° 211, Paris. Burgis, M. J., and Symoens, J. J., 1987.

Magrin, G.: The disappearance of Lake Chad: history of a myth. J. Political Ecol., 23, 204-222, 2016.

Matthew, A., Amudha, P., and Sivakumari, S.: Deep Learning Techniques: An Overview. In Advanced Machine Learning Technologies and Applications. AMLTA 2020. Advances in Intelligent Systems and Computing, 1141. 599-608, Springer, Singapore,, DOI: 10.1007/978-981-15-3383-9_54, 2021.

Mohanty, S., Jha, M. K., Kumar, A., and Sudheer, K. P.: Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. Water Resour. Manage., 24(9), 1845-1865, 2009.

Nagarajan, C., Pohl, B., Rüttinger, L., Sylvestre, F., Vivekananda, J., Wall, M., and Wolfmaier, S.: Climate-Fragility Profile: Lake Chad Basin. Berlin: Adelphi*,* 2018.

Neill, S. P. and Hashemi, M.R.: Root Mean Squared Error. Chapter 8. Ocean modelling for resource characterization in Fundamentals of Ocean Renewable Energy, 2018.

O'Brien, R. M.: A caution regarding rules of thumb for variance inflation factors. Qual. & Quant., 41, 673–690, DOI 10.1007/s11135-006-9018-6, 2007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python JMLR, 12(85), 2825−2830, 2011.

Rodell, M., Houser, P. R., Jambor, U. E. A., Gottschalck, J., Mitchell, K., Meng, C. J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., and Entin, J. K.: The global land data assimilation system. Bull. Am. Meteorol. Soc., 85(3), 381-394, 2004.

Schneider, U., Becker, A., Finger, P., Rustemeier, E., and Ziese, M.: GPCC Full Data Monthly Product Version 2020 at 1.0°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historical Data. DOI: 10.5676/DWD_GPCC/FD_M_V2020_100, 2020.

Scikit learn linear model, 2007. Available at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (accessed 3 February2022).

Scikit learn support vector regression, 2007. Available at https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html#sklearn.svm.SVR (accessed 3 February 2022).

Servant, M. and Servant, S.: Paleolimnology of an Upper Quaternary Endorheic Lake in Chad Basin. In: Cannouze, J. P., Durand, J. R.,. and Leveque, C. (Eds.): Lake Chad. Monogr. Biol., 53, 11-26, 1983.

Shiri, J., Shamshirband, S., Kisi, O., Karimi, S., Bateni, S. M., Nezhad, S. H. H., and Hashemi, A.: Prediction of water-level in the Urmia Lake using the extreme learning machine approach. Water Resour. Manage., 30(14), 5217-5229, 2016.

Solomatine, D. and Ostfeld, A.: Data-driven modeling: Some past experiences and new approaches, J. Hydroinf., 10 (1), 3-22, 2008.

Taormina, R., Chau, K., and Sethi, R.: Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. Eng. Appl. Artif. Intell., *25*(8) 1670–1676, 2012.

Torgo, L.: Functional models for regression tree leaves. LIACC - University of Porto, R. Campo Alegre, 823 - 4150 Porto – Portugal, Conference paper, Jan. 1997.

Torgo, L.: Predictive Analytics. DCC, Faculdade de Ciências / LIAAD-INESC TEC, LA, Universidade do Porto, Dec. 2014.

UNEP, United Nations Environment Programme: Lake Chad Basin, Global International Waters Assessment. GIWA Regional Assessment 43, ISBN 1651-9401, 2004.

USDA, United States Department of Agriculture: Global Reservoirs/Lakes. Lake products courtesy of the USDA/NASA G-REALM program, 2018. Available at *https://ipad.fas.usda.gov/cropexplorer/global_reservoir* (accessed 20 May 2018).

USGS, United States Geological Survey: *Earthshots:* Satellite images of environment change, Lake Chad, West Africa, 2018. Available at https://earthshots.usgs.gov/earthshots/Lake-Chad-West-Africa#ad-image-0-0 (accessed 2 March 2018).

Wehle, H. D.: Machine Learning, Deep learning, AI: What is the difference? Conference Paper**,** July 2017.

WFP, World Food Programme: Lake Chad Basin. Socio-economic analysis of the Lake Chad Basin Region, with focus on regional environmental factors, armed conflict, gender and food security issues. Desk Review, 2016.

Wharton, S.: Tendency, specific humidity, monthly mean, NASA/GSFC, Greenbelt, MD, USA, (01.20.2016), GLDAS CLM Land Surface Model L4 Monthly 1.0 x 1.0 degree, NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), 2016. Available at https://giovanni.gsfc.nasa.gov/giovanni/#service=TmAvMp&starttime=&endtime=&variableFacets=dataProductPlatformInstrument%3AGLDAS%20Model%3BdataProductSpatialResolution%3A1%20deg.%3B (accessed 28 July 2018).