Lake Surface Area Forecasting Using Integrated Satellite-SARIMA-Long-Short-Term Memory Model

Keyvan Soltani  
Razi University

Arash Azari  
Razi University

Mohammad Zeynoddin  
Laval University: Universite Laval

Afshin Amiri  
University of Tehran

Isa Ebtehaj  
Laval University: Universite Laval

Taha B.M.J. Ouarda  
Institut national de la recherche scientifique

Bahram Gharabaghi  
University of Guelph

Hossein Bonakdari (✉ hossein.bonakdari@fsaa.ulaval.ca)  
Universite Laval  https://orcid.org/0000-0001-6169-3654

Research Article

Keywords: Water resources, stochastic model, SARIMA, Tashk-Bakhtegan Lakes, hybrid model, forecasting.

DOI: https://doi.org/10.21203/rs.3.rs-631247/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Lake Surface Area Forecasting Using Integrated Satellite-SARIMA-Long-Short-Term Memory Model

Keyvan Soltani¹, Arash Azari², Mohammad Zeynoddin¹, Afshin Amiri⁴, Isa Ebtehaj³, Taha B.M.J. Ouarda⁵, Bahram Gharabaghi⁶, Hossein Bonakdari¹,*

¹Department of Civil Engineering, Razi University, Kermanshah, Iran
²Department of Water Engineering, Razi University, Kermanshah, Iran
³Department of Soils and Agri-Food Engineering, Université Laval, Québec, Canada, G1V 0A6
⁴Department of Remote Sensing and GIS, University of Tehran, Tehran, Iran
⁵Institut National de la Recherche Scientifique, Centre Eau Terre Environnement, INRS-ETE, Québec, G1K 9A9 Canada
⁶School of Engineering, University of Guelph, Guelph, Ontario NIG 2W1, Canada

*Corresponding author, Phone: +1 418 656-2131, Fax: +1 418 656-3723, E-mail: hossein.bonakdari@fsaa.ulaval.ca
Abstract

Lake Water Surface Area (WSA) plays a vital role in environmental preservation and future water resource planning and management. Accurately mapping, monitoring and forecasting Lake WSA changes are of great importance to regulatory agencies. This study used the MODIS satellite images to extract a monthly time series of WSA of two lakes located in Iran from 2001 to 2019. Following a consequence of image and time series preprocessing to obtain the preprocessed lake surface area time series, the outcomes were modeled by the Long-Short-Term Memory (LSTM) deep learning (DL) method, the stochastic Seasonal Auto-Regressive Integrated Moving Average (SARIMA) method and hybridization of these two techniques with the objective of developing WSA forecasts. After separate standardization and normalization of $A_L$TS and reevaluation of the preprocessed data, the SARIMA $(1, 0, 0) (0, 1, 1)_{12}$ model outperformed sole LSTM models with correlation index of (R) 0.819, mean absolute error (MAE) of 49.425 and mean absolute percentage error (MAPE) of 0.106. On the other hand, the hybridization (stochastic-DL) enhanced the reproduction of the primal statistical properties of WSA data and caused better mediation. However, the other accuracy indices did not change markedly (R 0.819, MAE 49.310, MAPE 0.105). The multi-step preprocessing and reevaluation also caused all LSTM models to produce their best results by less than 12 inputs.

Keywords: Water resources, stochastic model, SARIMA, Tashk-Bakhtegan Lakes, hybrid model, forecasting.
1. Introduction

Accurate mapping of lake Water Surface Areas (WSA) is essential to assess the amount of surface water available [1–5]. WSA is also helpful in determining the relationship between climate and water resources [6–9] and for assessing the impacts of changing water surfaces, which is crucial in water resources management [10-11]. The various methods for the extraction of water surface from remote sensing data fall into two general categories: single-band and multi-band techniques. The single-band technique uses a multispectral image band and identifies other ground-surface phenomena based on a threshold limit for water sources. The multi-band method helps distinguish the water masses from the differences in the reflectance properties of different bands [12]. Monitoring the water dynamics with images taken at different times can show changes in lakes, reservoirs and flood surfaces [13, 14].

Google Earth Engine (GEE) comprises a considerable amount of satellite and global data types worldwide, making it possible to analyze this data for various purposes such as change detection [15], mapping [16, 17] and ground level studies [18]. GEE has been widely used in a number of disciplines including reviewing global forest changes [19], estimating crop production [20], ground subsidence monitoring [21], coral reef mapping [22], modeling global surface water change [23, 24], flood risk assessment [25], global urban mapping [26, 27], renewable energy mapping [28], drought monitoring [29], and the reconstruction of the MODIS global vegetation index [30].

Satellite data have been commonly used in hydrological studies [31–35, 36]. Nath and Deb [37] used satellite images to detect and extract the water body of Puyang China. Abou El-Magd and Ali [38] studied surface evaporation from Lake Nasser using high-resolution radiometer satellite images. They demonstrated that robust assessments of lake evaporation can be obtained. Song et
al. [39] studied water level and lake area in the Tibetan Plateau by extracting time series from Landsat images. Moreira et al. [34] investigated and modelled water balance using satellite images and the evapotranspiration dataset in South America. Veh [40] developed an algorithm to detect the glacial lake outburst floods (GLOFs) in the Himalayas. The algorithm uses satellite images to analyze GLOFs and provide interpretable statistics for risk assessment and hazard prevention planning.

The pace of artificial intelligence (AI) models' development and their accuracy is rapidly increasing nowadays. These models are increasingly utilized in various fields of science, including water engineering and hydrology [41–43], since these models produced acceptable results in modelling sophisticated time series. Also, developments in AI and the computer industry played an important role [44] in accelerating this pace. In this field, deep learning methods produced noticeable results in modelling and forecasting hierarchical data [45-47]. The most recent deep learning model, LSTM, can utilize the unlimited historical raw data as inputs to detect the structure of the data and forecast future steps. The LSTM method is widely used in many fields like natural language understanding and speech recognition [48], image and text survey [49], hydrological data modelling such as precipitation and runoff forecasting [42,50], and modeling climatic and meteorological data [51]. Mohan and Gaitonde [52] used LSTM to model turbulent flow control and its temporal dynamics. Murad and Pyun [53] employed LSTM alongside support vector machine (SVM) and k-nearest neighbours (KNN) for human activity recognition, and they reported a higher performance of the LSTM model compared to other types of AI models. Sahoo et al. [54] used LSTM recurrent neural networks (LSTM-RNN) to model low flow hydrological time series. With a 94 percent correlation and low errors, they reported an acceptable potential of LSTM for modelling hydrological time series.
Stochastic methods are among the most renowned statistical models. These methods are popular amongst researchers because of their comprehensible principles and easy application. Seasonal Auto-Regressive Integrated Moving Average (SARIMA) uses non-seasonal and seasonal parameters to forecast time series based on historical data linearly [55–58]. Papalaskaris et al. [59] employed the SARIMA model for short-term basin rainfall forecasting in Kavala City, Greece. Mombeni et al. [60] used SARIMA for estimating one-year-ahead water demand in Iran. However, most hydrological time series have complex structures that cannot be efficiently modeled by linear methods like stochastic models or by AI models. Hence, some researchers resorted to the integration of AI and linear models to utilize both their capabilities. Hybridization of AI and linear models is one method that helps catch the complexity in time series and which has produced more accurate results [35,61–64]. Mishra et al. [65] employed a combination of stochastic SARIMA model and ANN to predict droughts in the Kansabati River basin in India. The results indicated that a hybrid model leads to higher accuracy. Shafaei et al. [66] applied wavelet pre-processing to SARIMA, ANN and hybridization of both and modelled monthly precipitation in Iran. They indicated that wavelet-SARIMA-ANN produces better results than wavelet-SARIMA and wavelet-ANN.

A novel methodology based on the integration of remote sensing and deep learning-stochastic modelling for lake surface area forecasting is proposed in the present work. To the best knowledge of the authors, no previous studies have attempted to use such hybrid model for WSA. The satellite images are downloaded, pre-processed and digitized for each time point to obtain changes in the water area. Then the achieved time series is modelled and forecasted by three methods. The modelling methods are deep learning LSTM model, stochastic SARIMA and hybridization SARIMA-LSTM. Prior to modelling, the time series structure is analysed by
stationarity and normality tests and other statistical and visual tests. If any pre-processing is needed, a standardization and/or normalization of the series is carried out to obtain the optimized modelling results. In the end, statistical and visual tools survey the methods presented in the methodology.

2 Material and Methods

2.1 Case study

The Tashk-Bakhtegan lakes (TB lakes) with a surface area of 540 km² are Iran's second-largest inland lakes. These lakes are the most important ecological habitats of Iran at an altitude of 1525 m above sea level and have a catchment area of 25,000 km². The maximum depth of Tashk-Bakhtaran lake is 2 m, and the maximum depth of Tashk lake is 3.1 m [66, 67]. These lakes are located between 29° 13'N–29° 48'N and 54° 10'E–53° 23'E. Water inflows to these lakes through the Kor and Syvand rivers. With the construction of three dams in these rivers' upper basin, the inflow of water into these lakes has decreased dramatically, causing a large area to dry out [68].

Fig.1 shows the location of the twin TB lakes in Iran.
Arid and semi-arid regions cover about one-third of the world's land area. Population growth in such areas caused an increase in the harvesting of groundwater [69]. In arid regions, lakes and wetlands play an indispensable role in the region's ecosystem, including climate change modification and food resources provision in the area. Due to growing water consumption in arid regions, water resources such as lakes ground water and other aquatic ecosystems are increasingly under stress [68].

TB lakes are under threat of complete drought due to over-harvesting of groundwater and mismanagement. In the basin of these lakes, two large rivers, Kor and Sivand, flow. Due to the vast area of TB lakes and moisture and water availability, unique plant and animal habitats exist in the surroundings [70]. In the past, TB lakes had a more fertile environment than today due to proper nutrition. At least 220 species of plants have been identified in the region's environment.
(the third largest from the species number point of view in Iran). More than 100,000 waterfowl migrate to the region in the winter [71]. There were about 5,000 Marbled Duck in 1990 [71,72].

Due to the diversity of flora and fauna in the wildlife, a refuge and a national park have been identified as protected areas. Their location is shown in Fig. 2. Three important dams that have been built in the upstream area of TB lakes: Sivand dam, Mollasadra dam and Doroordzan (Dariush) dam. The location of these dams is specified in Fig. 2, and their specifications are shown in Table 1.

![Figure 2: TB lakes watershed and location of ecological areas and distribution of dams in the area.](image)

**Table 1. Characteristics of dams located upstream of TB lakes.**

<table>
<thead>
<tr>
<th>Dam</th>
<th>River</th>
<th>H. (m)</th>
<th>Vol. (M.m$^3$)</th>
<th>Year</th>
<th>Dam Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doroordzan</td>
<td>Kor</td>
<td>85</td>
<td>960</td>
<td>1972</td>
<td>A pebble with an impermeable core</td>
</tr>
<tr>
<td>Mollasadra</td>
<td>Kor</td>
<td>75</td>
<td>440</td>
<td>2007</td>
<td>Reservoir (soil with clay core)</td>
</tr>
<tr>
<td>Sivand</td>
<td>Sivand</td>
<td>57</td>
<td>255</td>
<td>2007</td>
<td>Soil with clay core</td>
</tr>
</tbody>
</table>

1. Height; 2. Total tank volume (million cubic meters); 3. Year of operation
In Fig. 3 using MODIS satellite, the land cover changes in 2001 and 2018 are compared. This figure was provided using the MODIS Land Cover Type Product (MCD12Q1) satellite. The MCD12Q1 includes a global dataset of land cover types from 2001 to 2018. Its spatial resolution is 500 meters, and six different classification schemes have been used to produce it. The Global Earth Coverage Map provides ecological and physical characteristics of the Earth's surface.

In this study, LC_Type 1 band was employed to prepare a land cover map of the areas around TB lakes. This ground cover is based on the International Geosphere-Biosphere Program (IGBP), which is dedicated to studying global changes. The annual land cover maps around TB lakes were extracted from MCD12Q1 data in 2001 and in 2018 and are presented in Fig. 3. The reduction of agricultural coverage, pastures, and water level of the lake in the catchment area of TB lakes and the increase of barrier surface are clearly visible.
Fig. 3. Map of land cover changes between 2001 and 2018 in TB lakes watershed.

Fig. 4 shows the changes in five variables: Open shrublands, Grasslands, Barren, Croplands, and Water Bodies between 2001 and 2018. It can be observed that the area covered by Open Shrublands has been relatively stable until 2007, but since 2007, it has been increasing, while grasslands and croplands have declined with a similar trend.
Fig. 4. a) Land cover changes in 2001-2018, b) changes in TB lakes area.

Charts seem to indicate the existence of sudden changes around 2006 and 2007, particularly in the Waterbody area, which has declined since 2007 and reached its lowest surface in 2009. This reduction has had significant effects on other uses in the region. It should be noted that this decrease in water bodies in the catchment area of TB lakes has started since the construction of two dams, Mollasadra dam and Sivand dam, i.e., in 2007, and in 2009. These two dams were constructed on the two main rivers of the region, which feed the TB lakes, and resulted in the reduction of these lakes surfaces. Due to the diversity of flora and fauna in the region and protected areas around the TB lakes, these dams have caused severe damage to these genetic resources and the uses of the region. TB lakes increase the humidity of the air, and due to the
high altitude of the surrounding mountains, the resulting moisture remains in the atmosphere of the same area. This is referred to as artificial irrigation and causes better fruiting of the plants in this area.

The drought that has been observed in recent years and the significant reduction of TB lakes’ water have affected the region’s uses and caused a water crisis in the region. Croplands and grasslands have shown a significant decline, with their area shrinking to less than half its original value. Simultaneously, Shrublands and Barren soils increased, resulting in falling water levels in the region and the release of agricultural land and land-use change due to the lack of water in the area.

Considering all this background information, the question is raised on how long will the drought process of TB lakes continue, and what will be the changes in their surface in the coming years? To answer this question, we adopt the SARIMA-Long-Short-Term Memory Model to model the lake's surface changes and provide a practical model for future changes in the lake's surface.

Hence, using this model, an applied plan for water resources management in a variety of uses in the region can be developed, reducing the water crisis in the region and the abandonment of agricultural land, which has severe environmental and economic consequences in the region.

2.2. Remote sensing (RS) datasets and pre-processing

The MODIS (Moderate Resolution Imaging Spectroradiometer) tools were launched by Terra and Aqua satellites in 1999 and 2002. The MODIS sensor captures images 2230 kilometres wide and generates complete coverage of the earth in 1-2 days. By using Surface Reflectance products and their various bands (MOD09A1), the spectral reflectance of Earth's surface is estimated.
Pre-processing is a vital part of the remote sensing process. One of the problems with remote sensing images is the presence of clouds. Therefore, tools and indices like Google Earth Engine Environment (GEE) for image classification and the NDWI index are required to obtain desirable results. The NDWI index is one of the most commonly used indicators in remote sensing and is calculated from the relationships between bands (equations 1 and 2). Bands are used to obtain the water in which wavelengths have the highest and lowest spectral reflections. The NDWI relationship is computed as follows [73]:

\[
NDWI = \frac{G - \text{NIR}}{G + \text{NIR}}
\]

(1)

where the G is the green band, and the NIR is the near-infrared band. The modified NDWI relationship is as follows [12]:

\[
MNDWI = \frac{G - \text{MIR}}{G + \text{MIR}}
\]

(2)

where MIR is the mid-infrared band (wavelengths 1.2 to 2.2 µm).

The resulting image of the MNDWI index has values between -1 and +1. The pixels that indicate the presence of water have positive values. However, due to the presence of mixed pixels that cause errors in the detection of water sources, a threshold limit (MNDWI ≥ 0.3) is used to detect pure pixels with more precision [74,75]. Then, to calculate the area of water bodies in the images, the number of pure pixels identified in each image is multiplied by the area of land cover and the exact area of the water surface can be calculated.

2.3. Time series and pre-processing
A series of measurements in equal time intervals is termed time series. Each time series has a stochastic and a deterministic part. Periodical patterns, trends and jumps are the deterministic part and can exist in time series simultaneously or solely. The absence of this part in time series is called stationarity state. For any modeling, the deterministic terms can be removed, and only the stochastic part is required. Therefore, analysis methods are needed to assess the predictable pattern in time series and stationarity [76]. Applying tests to time series to extract interpretable statistics is the analysis of time series. Tests like KPSS, Mann-Whitney, Mann-Kendal, and Jarque-Berra can be employed to investigate stationarity, jump, trends and normality of time series, respectively.

In the KPSS [77] test, a regression equation is fitted to the data. If the variance of the independent variables of the relationship is null the $A_{t}$, then the series is stationary. The KPSS relationship for trend or level stationarity is as follows:

$$A_{t} = r_{t} + \beta_{t} + \varepsilon_{t}$$  \hspace{1cm} (3)

$$S^{2}(t_{j}) = \frac{1}{n} \sum_{i=1}^{n} e_{i}^{2} + \frac{2}{n} \sum_{j=1}^{1} w(j, t_{j}) \frac{1}{n} \sum_{i=j+1}^{n} e_{i} e_{i-s}$$  \hspace{1cm} (4)

$$w(s, t_{j}) = 1 - j/(t_{j} + 1)$$  \hspace{1cm} (5)

$$\text{KPSS} = \frac{1}{n^{2}} \left( \sum_{i=1}^{n} \frac{S_{i}^{2}}{S^{2}(t_{i})} \right)$$  \hspace{1cm} (6)

where $S_{i} = \Sigma e_{i}$, $t_{j}$ is the truncation lag, $e_{i}$ are the residuals, $r_{i} = r_{i-1} + u_{i}$ and $r_{i}$ is a random walk, $u_{i}$ are independent variables with equal distribution with mean zero and variance $\sigma_{i}^{2}$, $\beta_{i}$ is the deterministic term of the trend, and $\varepsilon_{i}$ the stationarity error.
In the case of non-stationarity, causing factors are investigated. Trend as a non-stationarity factor is analyzed by the Mann-Kendal test as follows [78]:

\[
\text{std}(M_T) = \begin{cases} 
(M_T - 1) \frac{\text{var}(M_T)}{0.5} & \text{MK} > 0 \\
0 & \text{MK} = 0 \\
(M_T + 1) \frac{\text{var}(M_T)}{0.5} & \text{MK} < 0 
\end{cases}
\]  

(7)

where std \((M_T)\) is the standard of Mann-Kendall statistic, MK is the Man-Kendall statistic, and var \((M_T)\) is the variance of \(M_T\). The \(M_T\) and var \((M_T)\) are defined as:

\[
M_T = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{sgn}(A_{L,j} - A_{L,i})
\]  

(8)

\[
\text{var}(M_T) = \left(2N^3 - 7N^2 - 5N\right) - \sum_j^g A_{L,j} (A_{L,j} - 1)(2L_{L,j} + 5)/18
\]  

(9)

where \(A_{L,j}\) and \(A_{L,i}\) are the lake area time series at the jth and ith group, \(g\) is the number of identical groups, sgn is the sign function, N is the number of samples and \(L_{L,j}\) is the number of the observations at the jth group. The following equation is used for seasonal changes over time, or seasonal trend:

\[
S_k = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{sgn}(A_{L,i,j} - A_{L,k,i})
\]  

(10)

\[
M_{S_k} = \sum_{k=1}^{g_0} \left(S_k - \text{sgn}(S_k)\right)
\]  

(11)

\[
\text{var}(M_{S_k}) = 2 \sum_{i=1}^{g_0-1} \sum_{j=i+1}^{g_0} \sigma_j + \sum_k^{g_0} \left(2N_k^3 - 7N_k^2 - 5N_k\right)/18
\]  

(12)
where $\omega$ represents the seasons, $k$ is the number of months, and $\sigma_{ij}$ is the covariance of stationary test in seasons $i$ and $j$. A probability corresponding to a test statistic higher than 5% means that $A_L$ is trendless.

Jumps, the second non-stationarity factor, represent sudden steps in the time series. The non-parametric Mann-Whitney (MW) test is used to evaluate this factor [79, 80]:

$$MW = \sum_{t=1}^{N} \left[ Dg(A_{L, \text{Ordered}}) - \frac{N_{m1}(N_{m1} + N_{m2} + 1)}{2} \right] / \left( \left( N_{m1}N_{m2}(N_{m1} + N_{m2} + 1) \right)^{0.5} / 12 \right)$$

where $A_{L, \text{Ordered}}$: series sorted by main series $A_L$, $Dg(A_{L, \text{Ordered}})$ the degree of $A_{L, \text{Ordered}}$ function, $N_{m1}$ and $N_{m2}$ is the number of members of the main sub-series that $N_{m1} + N_{m2} = N_{\text{total}}$. A probability related to a test statistic greater than 1% means that $A_L$ is jump-less.

Periodicity as the third deterministic factor can be surveyed by a time series graph or the auto correlation function (ACF) and the partial auto correlation function (PACF) plots. This term appears as iterative sinusoidal variations in both above graphs.

Seasonal standardization is one of the conventional stationarizing methods in hydrology. This method also reduces jumps in time series [81]. By removing the seasonal mean and standard deviation, the $A_L$ is transferred to a time series with a zero mean and a standard deviation equal to one as follows:

$$\text{std} \omega = \left( A_L(t, \omega) - \overline{A}_L(\omega) \right) / \hat{S}_d(\omega)$$
where, stdω represents the outcome of seasonal standardization, \( A_L(t, \omega) \) is the sample at \( t^{th} \)
year and the \( \omega^{th} \) season, \( \bar{A}_L(\omega) \) is the mean of the \( \omega^{th} \) season and \( S_d(\omega) \) is the standard
deviation of \( \omega^{th} \) season.

2.4. Long-Short-Term Memory (LSTM) deep learning model

Deep learning models are subclasses of artificial intelligence (AI) models enhanced for non-
linear sequence solving problems. A renowned deep learning model is the Long Short-Term
Memory (LSTM) network. The LSTM architecture is well suited for modelling sequence data
like time series and can learn long-term dependencies in series to forecast future steps. A simple
LSTM memory block is presented in Fig. 5. The LSTM model is constituted of several gates that
control the flow of information and affect the produced results. These gates are the input, the
forget, and the output gates which control the data entering to memory blocks \( c_t \), which should
be forgotten, and which are permitted to continue to further processes.

LSTM conducts a mapping [43] from an input sequence \( x \) to an output sequence \( y \) using the next
equations iteratively from \( t = 1 \) to \( t = \tau \) with initial values \( C_0 = 0 \) and \( h_0 = 0 \):

\[
f_t = \sigma (W_f A_{t,t} + U_f h_{t-1} + b_f)
\]

\[
\tilde{c}_t = \tanh (W_{\tilde{c}_t} A_{t,t} + U_{\tilde{c}_t} h_{t-1} + b_{\tilde{c}_t})
\]

where \( A_{t,t} \) is the input of the vector at time \( t \), and \( h_{t-1} \) is the hidden cell state at time \( t-1 \). The
weight matrices are \( U, W \) for input-to-hidden, and hidden-to-hidden connections, respectively. \( f_t \) is
a resulting vector with values in the range \((0, 1)\), \( \sigma(\cdot) \) represents the logistic sigmoid function and
\( W_f, U_f \) and \( b_f \) define the set of learnable parameters for the forget gate. \( \tilde{c}_t \) is an update vector
with (-1, 1) range for the cell state which calculated form $A_{L,t}$, $\tanh(*)$ is the hyperbolic tangent and $W_{\xi t}, U_{\xi t}$ and $b_{\xi t}$ are other sets of learnable parameters.

$$i_t = \sigma(W_i x_i + U_i h_{t-1} + b_i)$$  \hspace{1cm} (18)$$

$i_t$ is the forget gate with range (0,1). $W_i, U_i$ and $b_i$ are a set of learnable parameters, defined for the input gate. The results of Eqs. 16 to 18 lead to update the cell state:

$$c_t = f_t O c_{t-1} + i_t O c_{\xi t}$$  \hspace{1cm} (19)$$

where $O$ denotes element-wise multiplication. The output gate, as the last gate, controls the cell state $c_t$.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$  \hspace{1cm} (20)$$

where $o_t$ is in the range (0, 1) and $W_o, U_o$ and $b_o$ are a set of learnable parameters, defined for the output gate. $h_t$ is calculated as follows:

$$h_t = \tanh(c_t) O o_t$$  \hspace{1cm} (21)$$
2.5. Stochastic modelling concepts

Stochastic models are a subgroup of statistical models. These models are widely used in various fields of science because of their simplicity of utilization and theory. Seasonal Auto-Regressive Integrated Moving Average (SARIMA) is a stochastic model with seasonal and non-seasonal parameters that allows the model to forecast the future by using historical data [82].

In a SARIMA \((p, d, q) (P, D, Q)\) model, \(p\) and \(q\) are non-seasonal model parameters; \(P\) and \(Q\) are seasonal ones. \(d\) and \(D\) are the order of non-seasonal and seasonal differencing, respectively [83].

The simplified extension of the SARIMA equation for one step ahead forecast is as follows:

\[
(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) (1 - \Phi_1 L^{1o} - \Phi_2 L^{2o} - \ldots - \Phi_P L^P) (1 - L)^d (1 - L^\infty)^D A_L^{km^2}(t)...
\]

\[
= (1 - \theta_1 L^1 - \theta_2 L^2 - \theta_q L^q) (1 - \Theta_1 L^{1o} - \Theta_2 L^{2o} - \ldots - \Theta_Q L^{2Q}) e(t)
\]

(22)

\[
\phi(B) \Phi(B) (1 - L)^d (1 - L^\infty)^D A_L(t) = \theta(B)\Theta(B) e(t)
\]
where $\omega$ is seasonality, $\varphi$ and $\Phi$ are auto-regressive (AR) and seasonal AR (SAR) parameters, $\theta$
and $\Theta$ are the moving average (MA), $L$ is the differencing operator $A_L(t) = A_L(t-1)$. $(1 - L)^d$
equals the $d$-th non-seasonal, and $(1 - L^\omega)^D$ equals the $D$-th seasonal with the lag $\omega$. The $L$
operator helps in modelling the non-stationary series as it removes correlations in time series and
changes in mean and variance of the series. To improve the model's accuracy, each forecast is
updated with real data, and a 1-step-ahead forecast is carried out. As this model is linear,
deterministic terms must be extracted from the series, and data distribution normalized to
improve accuracy. To evaluate the distribution's normality, the Jarque-Bera test can be applied to
$A_L$ time series [84]:

$$JB = n\left(\frac{S_k^2}{6} + \frac{(K_u - 3)^2}{24}\right)$$

(23)

where $K_u$ is kurtosis $S_k$ is skewness; JB is a chi-square distribution with two degrees of
freedom that can be used to assume that data is normal. As most of the hydrological time series
are non-normal, normalizing transformation should be employed. John-Draper transform is a
normalization approach that can transform $A_L$ data. The equation is as follows:

$$A_{Ln}(\lambda) = \begin{cases} 
\text{sgn}(A_L)\frac{(|A_L|+1)^\lambda - 1}{\lambda} & \lambda \neq 0 \\
\text{sgn}(A_L)\log(|A_L|+1) & \lambda = 0 
\end{cases}$$

(24)

$$\text{sgn}(A_L) = \begin{cases} 
1 & A_L \geq 0 \\
-1 & A_L < 0 
\end{cases}$$

(25)

$\lambda$ is JD transforming parameters and $A_{Ln}$ is the normalized $A_L$ series.
2.6. Comparison measures

Correlation coefficient (R), Root mean squared error (RMSE), root mean squared relative error (RMSRE), Mean absolute percentage error (MAPE) and Mean absolute error (MAE) are used to evaluate the accuracy of models in time series obtained from pre-processing of $A_L$ data. To compare the stochastic models, corrected Akaike's Information Criterion (AICc) is used. Theil's U coefficients are also used [85–87]. The Theil's U indices compare models based on the simplicity of the model against goodness-of-fit. The lower the index, the better the model results are.

\[
R = \left( \frac{\sum_{i=1}^{N} (A_{L,O,i} - \bar{A}_{L,O}) (A_{L,P,i} - \bar{A}_{L,P})}{\sqrt{\sum_{i=1}^{N} (A_{L,O,i} - \bar{A}_{L,O})^2 \sum_{i=1}^{N} (A_{L,P,i} - \bar{A}_{L,P})^2}} \right)
\]

(26)

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (A_{L,O,i} - A_{L,P,i})^2}{N^2}}
\]

(27)

\[
\text{MAPE} = \frac{100 \sum_{i=1}^{N} \left| \frac{A_{L,O,i} - A_{L,P,i}}{A_{L,O,i}} \right|}{N}
\]

(28)

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| A_{L,O,i} - A_{L,P,i} \right|
\]

(29)

\[
\text{RMSRE} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} \left( \frac{A_{L,O,i} - A_{L,P,i}}{A_{L,O,i}} \right)^2}
\]

(30)

\[
\text{AICc} = \frac{2kn + \left( n \ln(\sigma^2) (n-k-1) \right)}{n-k-1}
\]

(31)
\[ U^I = \frac{\left[ \sum_{i=1}^{N} (A_{L,O,i} - A_{L,P,i})^2 \right]^{0.5}}{\left[ \sum_{i=1}^{N} (A_{L,O,i})^2 \right]^{0.5} + \left[ \sum_{i=1}^{N} (A_{L,P,i})^2 \right]^{0.5}} \]  

(32)

\[ U^I = \frac{\left[ \sum_{i=1}^{N} (A_{L,O,i} - A_{L,P,i})^2 \right]^{0.5}}{\left[ \sum_{i=1}^{N} (A_{L,O,i})^2 \right]^{0.5}} \]  

(33)

\[ A_{L,O,i} \text{ and } A_{L,P,i} \text{ are the } i^{th} \text{ value of observed data and predicted } A_L \text{ respectively. } N \text{ is the number of months, } \sigma_\varepsilon \text{ is the residual's standard deviation, and } k \text{ is the number of tuned parameters through the modelling process. } U^I \text{ is the accuracy of forecasting, and } U^I \text{ is the forecasting quality. Checking the stochastic models' residuals for correlations and white noise state is one of the stochastic modelling steps. For this purpose, the Ljung-Box test can be applied to model residuals as follows [88]:} \]

\[ lbq = \left( N^2 + 2N \right) \sum_{h=1}^{m} \frac{r_h}{N-1} \]  

(34)

\[ N \text{ is the number of samples, } r_h \text{ is the residual coefficient of the autoregression (} \varepsilon_t \text{) in delay } h; \text{ the value of } m \text{ is also equal to } \ln(N). \text{ If the probability related to the Ljung-Box test is greater than the } \alpha\text{-level (in this case } P_{lbq} > \alpha = 0.05), \text{ the residues series is white noise.} \]

In this research, first in the Google Earth Engine environment, the data were selected, and the necessary pre-processing was performed. MODIS MOD09A1 was used to measure the changes in the area of TB lakes. Images with a cloud coverage of less than 10% were selected to continue the process, and then the pixel value was corrected. Due to the area's characteristics, a threshold
for water identification was considered, and with the MNDWI index, water bodies were separated from other zones. Higher threshold (MNDWI ≥0.3) was identified as water bodies. The time series of changes in the extent of the lakes was calculated from 2001 to 2019. Land cover changes were extracted from MODIS MCD12Q1, and the land cover map was prepared. To determine land use, the land cover map was used to identify the changes in the area and their impact on the changes in the lake surface. Then the time series of the WSA data was extracted from the satellite data. Following, the modelling procedure was undertaken.

Initially, the WSA time series' structural characteristics were investigated by pre-processed by stationarity and normality tests. If any pre-processing is needed, a standardization and/or normalization to series is carried out to obtain the optimized modelling results. Then deep learning LSTM model, stochastic SARIMA and hybridization SARIMA-LSTM are performed.

The described procedure is depicted in the flowchart of Fig. 6.
Fig. 6. Flowchart of the analytical procedures of the study.

3. Results and discussion

3.1. RS results
In this study, MODIS data, MOD09A1 version 6 Surface Reflectance (with a resolution of 500m and 8-day from 2000 to 2019) were employed to obtain time-series variations of TB lakes water surface. The MOD09 series is one of the MODIS surface reflection products. This product has seven bands and estimates the spectral reflectance values for each band in the absence of atmospheric absorption or diffusion.

Table 2 Specifications of MOD09A1 version 6

<table>
<thead>
<tr>
<th>Band name</th>
<th>Band desc.</th>
<th>wavelength(nm)</th>
<th>Spatial resolution (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sur_refl_b01</td>
<td>S.R. Band 1</td>
<td>620-670</td>
<td>500</td>
</tr>
<tr>
<td>sur_refl_b02</td>
<td>S.R. Band 2</td>
<td>841-876</td>
<td>500</td>
</tr>
<tr>
<td>sur_refl_b03</td>
<td>S.R. Band 3</td>
<td>459-479</td>
<td>500</td>
</tr>
<tr>
<td>sur_refl_b04</td>
<td>S.R. Band 4</td>
<td>545-565</td>
<td>500</td>
</tr>
<tr>
<td>sur_refl_b05</td>
<td>S.R. Band 5</td>
<td>1230-1250</td>
<td>500</td>
</tr>
<tr>
<td>sur_refl_b06</td>
<td>S.R. Band 6</td>
<td>1628-1652</td>
<td>500</td>
</tr>
<tr>
<td>sur_refl_b07</td>
<td>S.R. Band 7</td>
<td>2105-2155</td>
<td>500</td>
</tr>
</tbody>
</table>

Band desc.: Band description; S.R.: Surface Reflectance

The necessary pre-processing, including atmospheric corrections, have been made to this product. The workflow for extracting the lake area from the MODIS images includes image preparation, image classification and statistical computation. During the preparation of the images, the location of the lakes was determined. So, at this point in the GEE Environment, images with more than 10% cloud were excluded from the lake extraction process. Images with cloud cover less than 10% were selected, and pixels suitable for classification were identified. The image classification step was also performed in the GEE environment. Fig. 7 illustrates the changes of $A_L$ from 2001 to 2019 for April Month.
Fig. 7. Changes of $A_L$ from 2001 to 2019 for April Month.

By using a function, the MNDWI index was applied to the previous step images. Water has high reflectance at the wavelength of 0.5 μm (green band) and absorbs electromagnetic waves at infrared wavelengths and has low reflectance. Therefore, in this study, band 4 (green band) and band 7 (mid-infrared) of MODIS images were used. After applying the threshold limit, the exact
area of the water surface was obtained. For better change recognition in the lake surface area, the
area has been separated from the surrounding environment, and the changes in the TB lakes
based on this model are shown in Fig. 8. Based on the calculated areas, the monthly time series
of the TB lakes area was achieved.

Fig. 8. Lake Surface changes per square kilometres from 2001 to 2019 based on MODIS satellite
imagery.
The results obtained from the annual changes in surface area of TB Lakes are shown in Fig. 9. Surface area changes have decreased dramatically from 2001 to 2019, reaching 709.487 km² in 2001. In 2002, the $A_L$ reached 975.64 km², which shows a 37% increase compared to 2001. In 2003, the lake's surface reached 821.55, and in 2004 and 2005, its value reached the highest level among the study years, occupying 1038.47 km² and 1088.07 km², respectively. After that, with a steep slope, the lake's surface shows a decrease until 2010 and this year it has reached 481.1 km². This indicates that between 2005 and 2010, the average level of lake decline was 11.16% per year. In 2011, there was an increase of 74.74 km² in the lake's water level and it fluctuated in the same range until 2013, and in 2014, it decreased by 132.192 km² compared to 2013, reaching 425,238 km². With an increase and cache, it reached 389.245 km² in 2016, which is the lowest number of observations among the study years. In 2017, the $A_L$ shows an increase of 34.26%, and in 2018 and 2019, it has reached 379,158 and 480,937 km², respectively.

Fig. 9. Annual changes in the surface area of TB lakes (2000-2019)
Differences in the $A_L$ between the study years confirm the information provided in the case study and can be considered as the main factor in reducing the water level of TB Lakes and changes in the region's ecosystem. Therefore, it is necessary to provide practical and correct solutions in the region to control the ecosystem and prevent further destruction of water resources in the region. Using applied models, the water level of TB Lakes can be modeled for better management in the future.

3.2. Obtained $A_L$ time series attributes and pre-processing

The obtained $A_L$ time-series statistical characteristics were investigated and the results are presented in Fig. 10. To survey the characteristics of the series and model it, the $A_L$ series is divided into train and test parts with 70-30% ratio. From the 224 obtained data points, 157 (from Dec 2000 to Jul Dec 2013) and 67 (from Jan 2014 to Jul 2019) were considered as train and test parts, respectively (Fig. 10a). Regarding the information provided in Table 3 the statistical features of the intervals differ considerably, which can lead to poor modelling results.
According to the information provided in Table 3, the highest $A_L$ lakes is 1292.32 km$^2$ which is related to Jan 2005 and the lowest value is related to 246.4 which is related to Jul 2018. The minimum values for train and test data are 342.52 km$^2$ and 246.4 km$^2$, respectively, and the maximum values for these two are 1292.32 km$^2$ and 733.39 km$^2$. The average value obtained for 224 data is 662.81 km$^2$ and in the train and test stage it is 757.44 and 441.08 km$^2$, respectively, and all data have positive skewness.

**Table 3. Statistical attributes of Lakes Area ($A_L$) data**

<table>
<thead>
<tr>
<th></th>
<th>Nbr.</th>
<th>Min (km$^2$)</th>
<th>Max (km$^2$)</th>
<th>1$^{\text{st}}$ Q (km$^2$)</th>
<th>Median (km$^2$)</th>
<th>3$^{\text{rd}}$ Q (km$^2$)</th>
<th>Mean (km$^2$)</th>
<th>$\sigma$ (n)</th>
<th>$\gamma_1$</th>
<th>$\gamma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>224</td>
<td>246.40</td>
<td>1292.32</td>
<td>455.76</td>
<td>605.36</td>
<td>882.24</td>
<td>662.81</td>
<td>256.94</td>
<td>0.42</td>
<td>-0.91</td>
</tr>
<tr>
<td>Train</td>
<td>157</td>
<td>342.52</td>
<td>1292.32</td>
<td>552.88</td>
<td>735.43</td>
<td>959.65</td>
<td>757.44</td>
<td>241.71</td>
<td>0.08</td>
<td>-1.08</td>
</tr>
<tr>
<td>Test</td>
<td>67</td>
<td>246.40</td>
<td>733.39</td>
<td>340.48</td>
<td>451.91</td>
<td>503.65</td>
<td>441.08</td>
<td>116.90</td>
<td>0.30</td>
<td>-0.58</td>
</tr>
</tbody>
</table>

Nbr., Number of data; Min. and Max., Minimum and Maximum of data; 1st Q. and 3rd Q., first and third Quarters; $\sigma$(n), Standard Deviation; $\gamma_1$, Skewness; $\gamma_2$, Kurtosis.
The results of the application of statistical tests to the A_L time series are provided in Table 4 and Fig. 10. According to MW, MK, SMK, KPSS tests results, the series has jumps and trends and is highly non-stationary. Furthermore, the JB test confirms the non-normality of the data. Therefore, pre-processing of A_L time series, prior to AI and stochastic modeling is mandatory.

The ACF and PACF values were calculated and the corresponding results are presented in Fig. 11 and Fig. 12. The plots plainly demonstrate the non-seasonal and seasonal trends and periodicity with lag 12. The periodicity is also observable in the time series plot (Fig. 10a) as iterative peaks and lows. This lake area data component was foreseeable as the surface water is highly impacted by solar energy’s seasonal flux and earth’s revolutions. Though this periodicity damped after two significant lags, the AL series would be more independent and better results can be obtained by removing it.

![ACF and PACF plots](image_url)

**Fig. 11.** A_L time series ACF and PACF plots.
For removing non-stationarity factors, the stdω method (\(\text{stdω}(A_L)\)) was applied to the series (Fig. 10b). After modeling, it was observed that this method only reduced the seasonality to one lag in the series and did not affect other terms. Since the stdω method contained the seasonal parameters, it was expected that it would affect mostly seasonal components. The JB transform was subsequently applied (\(\text{stdωJD}(A_L)\)). The normalization method was able to decrease the JB statistic markedly and normalize data. Also, normalization resulted in a reduction of the non-seasonal correlations from 22 to 18 lags. The corresponding results are presented in table 4 and Fig. 12 for each step.

Table 4 Lakes Area \((A_L)\) time-series stationarity and normality tests outcomes

<table>
<thead>
<tr>
<th>Tests</th>
<th>Jump</th>
<th>Trend</th>
<th>Stationarity</th>
<th>Norm.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(P_{MW})</td>
<td>(P_{MK})</td>
<td>(P_{SMK})</td>
<td>(P_{KPSS})</td>
</tr>
<tr>
<td>(A_L)</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(\text{stdω}(A_L))</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(\text{stdωJD}(A_L))</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Cons. Diff.**</td>
<td>81.21</td>
<td>53.36</td>
<td>37.30</td>
<td>98.02</td>
</tr>
</tbody>
</table>

*JB critical: 5.99; p-value > 5% = acceptable; ** Consecutive 1st order non-seasonal and seasonal differencing

3.3. LSTM Deep learning modelling

Almost all the hydrological time series, regarding their nature, have a complex structure. Therefore, studying and involving historical events in the modelling process is of high importance. The LSTM model is an enhanced model produced to cover recurrent neural networks' deficiencies (RNN). The RNNs were limited in using historical data. However, the LSTM model unlimitedly can use long-term dependencies in modelling process.

Given the seasonal correlations in time series with lag 12, the LSTM model was used for modeling pre-processed data with the hidden cell states of \(h = 12, 60, 144\) and 156 [45,89].
piecewise learning rate schedule with Initial learn rate of 0.005 was defined for the model structure. After determining the maximum epochs of 500 and learn rate drop period and drop factor of 125 and 0.2, respectively, the single LSTM layer model was defined. Computational requirements represent an important consideration. In this work, the MATLAB software and a computer with a configuration of CPU core i7, 2500 MHz and 8G RAM were used. The average time spent for modeling each input was around 100 seconds. The results of the models are provided in Table 5. The LSTM model with the seasonal standardized ($\text{std}\omega$) data and 12 inputs produced better results than inputs with higher hidden cell states with the same preprocessing.

### Table 5 LSTM results for Lake Area ($A_L$) time series

<table>
<thead>
<tr>
<th>Method</th>
<th>Inputs $h$</th>
<th>R</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE</th>
<th>RMSRE</th>
<th>UI</th>
<th>UH</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{std}\omega$</td>
<td>h12</td>
<td>0.786</td>
<td>113.227</td>
<td>92.001</td>
<td>0.230</td>
<td>0.289</td>
<td>0.114</td>
<td>0.248</td>
</tr>
<tr>
<td>$\text{std}\omega$</td>
<td>h60</td>
<td>0.790</td>
<td>144.837</td>
<td>124.816</td>
<td>0.317</td>
<td>0.380</td>
<td>0.140</td>
<td>0.317</td>
</tr>
<tr>
<td>$\text{std}\omega$</td>
<td>h144</td>
<td>0.769</td>
<td>181.314</td>
<td>164.596</td>
<td>0.418</td>
<td>0.483</td>
<td>0.169</td>
<td>0.397</td>
</tr>
<tr>
<td>$\text{std}\omega$</td>
<td>h156</td>
<td>0.746</td>
<td>200.116</td>
<td>183.361</td>
<td>0.465</td>
<td>0.532</td>
<td>0.184</td>
<td>0.439</td>
</tr>
<tr>
<td>$\text{std}\omega\text{JD}$</td>
<td>h12</td>
<td>0.806</td>
<td>109.140</td>
<td>91.571</td>
<td>0.229</td>
<td>0.281</td>
<td>0.110</td>
<td>0.239</td>
</tr>
<tr>
<td>$\text{std}\omega\text{JD}$</td>
<td>h60</td>
<td>0.893</td>
<td>116.363</td>
<td>104.381</td>
<td>0.263</td>
<td>0.304</td>
<td>0.115</td>
<td>0.255</td>
</tr>
<tr>
<td>$\text{std}\omega\text{JD}$</td>
<td>h144</td>
<td>0.770</td>
<td>157.532</td>
<td>138.723</td>
<td>0.352</td>
<td>0.416</td>
<td>0.151</td>
<td>0.345</td>
</tr>
<tr>
<td>$\text{std}\omega\text{JD}$</td>
<td>h156</td>
<td>0.852</td>
<td>146.578</td>
<td>132.055</td>
<td>0.331</td>
<td>0.380</td>
<td>0.141</td>
<td>0.321</td>
</tr>
</tbody>
</table>

$h =$ hidden states no.

In the $\text{std}\omega$ method, except for $h60$, where the value of R is improved by 2% and h12 has a better performance in other statistical parameters, and as the number of inputs increases, the accuracy of the model is affected. $h156$ has the highest error values so that the correlation coefficient has decreased by 5% compared to $h12$ and the RMSE has increased by 76.7%. RMSRE and MAPE, have increased by more than 100%. These values for LSTM models demonstrated that the models' power and quality were higher while 12 inputs were chosen for modeling, compared to the other models with more inputs. Also, it indicates that the impact of most recent historical data
is more than the oldest ones. This refers to the capability of the LSTM in modeling dependent data.

For further investigation, the pre-processed series with stationarization and normalization (stdωJD) were also modeled. Likewise, the LSTM model with 12 inputs produced the best results. The LSTM\textsubscript{stdωJD} (12) indices are as $R = 0.806$, $\text{RMSE} = 109.140$, $\text{MAE} = 91.571$, $\text{MAPE} = 0.229$, $\text{RMSRE} = 0.281$, $U^I = 0.110$, $U^{II} = 0.239$. The Theil’s coefficient also shows slight improvement in the model's quality and power while using normalization and standardization, compared to the single standardization.

The results show that in stdωJD, as in stdω, the model's accuracy decreases with increasing inputs. In h156 the value of the correlation coefficient is higher than h12 and h144. However, the statistical parameters show better performance for h12 compared to stdωJD model with other hidden cell inputs. As seen in the preprocessed data's correlogram, the seasonal correlation was damped after one seasonal lag and the dependencies were important up to one seasonal lag and few more non-seasonal lags. Therefore, the LSTM models with historical data up to previous 12 lags were investigated. Moreover, the normalization of data distribution enhanced the modeling results and decreased the errors in comparison to lone standardization. The LSTM\textsubscript{stdωJD} improved the results by $R = 2.458\%$, $\text{RMSE} = 3.610\%$, $\text{MAE} = 0.468\%$, $\text{MAPE} = 0.451\%$, $\text{RMSRE} = 2.720\%$, $U^I = 3.428\%$, $U^{II} = 3.610\%$. This improvement proves the importance of the pre-processing in AI models, regardless of their capability in modeling non-linearity.

The structure of data should be investigated prior to the preprocessing to assess the impacts of the preprocessing methods. Also, it can be concluded that using more independent inputs causes more variations that impact the final results of the deep learning method. So, limiting the LSTM model inputs to the correlated data is important.
3.4. Stochastic modeling

Stochastic models are among the most conventional modelling methods in hydrology. These models are noticed for their simple theory and application. As the basis of these models are statistical concepts, some prerequisites should be considered in modelling process. The stationarity and normalization of time series are the two necessities of stochastic models.

Concerning the results provided in section 3.2, as the pre-processed data's ACF values are damped after 18 lags and series is normal, modelling can be carried out, but higher orders of parameters are needed. Hence, a consecutive non-seasonal and seasonal differencing was applied.

Fig. 12. A\textsubscript{L} pre-processed time series ACF plots.
to series, and it was observed that all non-stationarity factors were removed from series and became stationary. The corresponding results are presented in Table 4 and Fig. 12 for each step.

The correlations in ACF plots after consecutive differencing declines considerably to one lag. But for further survey of the model’s capability, the orders of the parameters in SARIMA model are considered as: \(p = q = P = Q = \{0, 1, 2, 3, 4, 5\}\) and \(d = D = \{0, 1\}\) and seasonality \(\omega = 12\).

After coding the dynamic model in MATLAB software and considering this parameter selection, a total number of 2590 models were produced with the same computer configuration used for the LSTM models. The time spent on stochastic modeling was about two hours. The minimum values of the indices for forecasted \(A_L\) data in all were \(R = 0.01, \text{RMSE} = 68.70, \text{MAE} = 49.42, \text{MAPE} = 0.11, \text{RMSRE} = 0.14, \text{AICc} = 574.80, \text{UI} = 0.08, U^{II} = 0.15\) and the maximum values were \(R = 0.85, \text{RMSE} = 780.61, \text{MAE} = 756.47, \text{MAPE} = 1.85, \text{RMSRE} = 1.98, \text{AICc} = 862.04, \text{UI} = 0.47, U^{II} = 1.71\). With these specifications and after considering the independence of the results, simplicity and goodness of the fit of models, the superior model was chosen as SARIMA \((1,0,0)(0,1,1)_12\). The evaluation results for this model are: \(R = 0.819, \text{RMSE} = 70.217, \text{MAE} = 49.425, \text{MAPE} = 0.106, \text{RMSRE} = 0.143, \text{AICc} = 574.82, \text{UI} = 0.077, U^{II} = 0.154\). The model is the most parsimonious and adequate SARIMA model compared to the other 2589 models. It is observed that the model's correlation index is almost in the same range as the LSTM, but other indices like RMSE, MAPE are almost half. This means the linear model could forecast the variation of the AL data better than sole LSTMs after triple preprocessing and removing all the dependencies in the data. However, other model evaluation criteria should be investigated, and there are still opportunities for enhancements. Another step in the evaluation of stochastic modelling is checking the independence of the residuals. This criterion is assessed simultaneously with parsimony and other statistics to obtain a model which is not only precise
but also has uncorrelated residuals. Therefore, the Ljung-Box test was applied to the stochastic model's residuals for 60 non-seasonal or five seasonal lags. The test indicated the independence of the residuals and the adequacy of the model. The results of the independence test for the superior model are provided in Fig. 12.

![Ljung-Box residuals test results.](image)

3.5. Hybrid Deep-learning-Stochastic modelling and disparities

Hybridization of models is one of the methods of utilizing non-linear and linear models’ characteristics simultaneously. These methods allow researchers to model data and make predictions by covering the drawbacks of the single models and produce results with lower errors. For this purpose, the linear model residuals that are independent are used as inputs of the AI model. This input is assumed to be the non-linear part of the time series as the stochastic model is also assumed to be able to forecast the linear part [90]. As it can be seen in Fig. 14. The residuals of the linear model are completely independent, and no correlation remains in the residuals. However, they have the circumstances to be modeled by the AI model. Since, no correlation is found in the residuals’ series, the AI model requires less inputs to forecast future steps. However, the previous steps will be followed to provide comparison circumstances.
Fig. 14. Stochastic model residuals auto correlation function plot

B integrating SARIMA and LSTM, the superior linear model's residuals were modelled by the LSTM model with the same inputs considered for modelling in previous sections. The residuals are denoted as $\text{SARIMA}_s$. The results of the models are provided in Table 6. The $\text{SARIMA}_s$-LSTM with 12 inputs outperformed other $\text{SARIMA}_s$-LSTM hybrid models. As shown in Fig. 14, the residuals do not have correlations, therefore, the best results with the 12 inputs were expected. Using hidden cells’ inputs less than 12 could also produce these results.

Table 6 Hybrid models results for Lakes Area ($A_L$) time series

<table>
<thead>
<tr>
<th>Method</th>
<th>Inputs</th>
<th>R</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE</th>
<th>RMSRE</th>
<th>$U^I$</th>
<th>$U^{II}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SARIMA}_s$-LSTM</td>
<td>h12</td>
<td>0.819</td>
<td>70.428</td>
<td>49.310</td>
<td>0.105</td>
<td>0.143</td>
<td>0.077</td>
<td>0.154</td>
</tr>
<tr>
<td></td>
<td>h60</td>
<td>0.777</td>
<td>79.138</td>
<td>60.137</td>
<td>0.131</td>
<td>0.165</td>
<td>0.087</td>
<td>0.173</td>
</tr>
<tr>
<td></td>
<td>h144</td>
<td>0.754</td>
<td>100.928</td>
<td>82.246</td>
<td>0.198</td>
<td>0.243</td>
<td>0.104</td>
<td>0.221</td>
</tr>
<tr>
<td></td>
<td>h156</td>
<td>0.752</td>
<td>104.037</td>
<td>85.689</td>
<td>0.208</td>
<td>0.252</td>
<td>0.107</td>
<td>0.228</td>
</tr>
</tbody>
</table>

$h =$ hidden states no.

By comparing the results of the hybrid model and previously presented models, it was observed that the hybridization improved a few characteristics of the results. Compared to the single LSTM models, the Hybrid model increases the correlation of the forecast. It improved the mediation of the data by 0.061 compared to the average of the LSTM models. Also, the error
indices were almost reduced to half. However, this improvement, compared to the linear model was less noticeable than lone LSTM models. The hybridization, on the other hand, lowered the MAPE and MAE indices.

**Fig.15.** Scatter plots of the modeled $A_L$ time series. a: LSTM_{Stdω} (h12); b: LSTM_{StdωJD} (h12); c: SARIMA(1,0,0)(0,1,1)12; d: Hybrid$_S$ (h12).

Since the indices are very close and for better comparison, the scatter plots of the superior LSTM, SARIMA, and hybrid models are provided in Fig. 15. From the scatter plots, the dispersion of the modelled data can be observed. The LSTM models predicted data are more dispersed than SARIMA and hybrid models, respectively (Fig. 15 a and b). The linear model (Fig. 15c) has densified the data and brought it closer to the 10% range. However, the hybrid model was more successful than the others in bringing the forecasts closer to the median line and locating data in the 10% intervals (Fig. 15d). In other words, hybridization caused more correlation in the forecasted data and better mediation has occurred by utilizing both methods’
characteristics. The Box plot of the observed data and superior models are drawn in Fig. 16, and it can be observed that the SARIMA (1, 0, 0) (0, 1, 1) model perfectly forecasted the interquartile area of the $A_L$ time series and even were able to forecast one of the extreme values of the original series. These methods also predicted the maxima and minima of the data more accurately than other models. A potent model regenerates the statistical characteristics of the studied data. Though the linear model and the hybrid indices were slightly similar, the hybrid SARIMA-LSTM reproduced the primal statistical properties of WSA data better than sole models [91]. The hybrid model performed better in forecasting the mean and other statistical characteristics of the observed data slightly better than the SARIMA model. Therefore, hybridization was not able to produce noticeable results (Tables 5 and 6) but reproduced the original series statistical attributes. Thus, it can be considered as a superior WSA modelling methodology.
Fig. 16. Box plot of the superior models; AL: observed WSA data, Ss: SARIMA (1, 0, 0) (0, 1, 1); H: SARIMA-S-LSTM; L1: LSTMstdωJD(12), L2: LSTMstdωJD(12);

4. Conclusion

Sustainable management of freshwater inland lakes in an arid region plays a vital role in environmental preservation and quality of life. Moreover, monitoring changes in the lake's surface area due to both natural and anthropogenic stressors helps to better plan and manage water resources. Therefore, the accurate mapping and monitoring of lake surface area, and the
forecasting of these vital resources future trends are of great importance for planning and management purposes. In this study, the WSA of the TB lakes is studied. To map the lake's surface area, the MODIS satellite images were used to extract a time series depicting changes of the WSA. The images were obtained from MODIS data, MOD09A1 version 6. The pre-processing of the images included image preparation, classification, and statistical computation. The preparation and classification of the images were undertaken in GEE environment. Using the MNDWI index, the water mass was separated from the background, and the lake area was obtained from the chosen images. Finally, by repeating the process for images from 2001 to 2019 a monthly time series of lakes areas ($A_L$) was obtained. The $A_L$ time series was examined by stationarity and normality tests to investigate the structure of the timeseries. Periods with 12 lag repetition, trends and jumps with a non-normal distribution were observed in the timeseries. The timeseries was pre-processed with the conventional seasonal standardization ($\text{std}_\omega$) method and normalized with the John-Draper (JD) transform, two-time series were obtained. These timeseries were modelled with the LSTM model with $h = \{12, 60, 144, 156\}$ number of hidden cell states. The single LSTM models, with the two different preprocessing tasks, required only 12 hidden cell states to obtain the highest accuracy. LSTM$_{\text{std}_\omega}(12)$ with $R = 0.786$, RMSE = 113.227, MAPE = 0.230 and STM$_{\text{std}_\omega\text{JD}}(12)$ with $R = 0.806$, RMSE = 109.140, MAPE = 0.229 outperformed others. These results indicated that using multiple preprocessing methods and reevaluating the results of the time series structure tests is necessary since most of the time, the latter part is neglected in the AI modeling procedure.

A stochastic SARIMA model and hybridization of both deep learning and stochastic models were carried out for further investigation and surveying the possibilities to enhance the forecasting results. The superior linear model was chosen as SARIMA with $(1, 0, 0) (0, 1, 1)_{12}$
parameters based on goodness of fit and model parsimony. The stochastics models' results were better than single LSTM models and the errors were reduced by almost half, $R = 0.819$, $MAE = 49.425$, $MAPE = 0.106$. To utilize both models’ capabilities, residuals of the stochastic model were modelled by LSTM.

Results indicate that the hybrid model indices were marginally better than others. The scatter and Box plots of the models revealed that the hybridization did not produce noticeable better error indices but improved the statistical characteristics and made them closer to observational data. The hybrid SARIMA-LSTM reproduced the primal statistical properties of WSA data and caused better mediation as observed in scatter plots and the Box plot of the data compared to sole models.

In conclusion, the hybridization can reproduce model forecasts that better preserve the observed timeseries's statistical attributes compared to single models. Therefore, it is suggested that the undertaken methodology of A_L time series modelling be applied to other A_L time series and other AI methods like Extreme Learning Machine (ELM), LSTM developments like Genetic Algorithm (GA)-LSTM and a combination of these models with linear models be investigated.
5. References

https://doi.org/10.1016/j.rse.2015.10.014.

https://doi.org/10.1016/j.rse.2015.10.031.

https://doi.org/10.1016/j.pce.2018.05.005.


https://doi.org/10.1016/j.limno.2015.02.001.


https://doi.org/10.5391/JKIIS.2016.26.2.093

https://doi.org/10.1007/s00704-020-03098-3


https://doi.org/10.1007/s00704-016-1879-6


https://doi.org/10.1080/01431169608948714

https://doi.org/10.1016/j.rse.2020.111869


https://doi.org/10.3390/w12051487


[79] Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. 18:50–60


