www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN POLSKA AKADEMIA NAUK

JOURNAL OF WATER AND LAND DEVELOPMENT

e-ISSN 2083-4535

ITP-PIB

Polish Academy of Sciences (PAN) Institute of Technology and Life Sciences – National Research Institute (ITP – PIB)

JOURNAL OF WATER AND LAND DEVELOPMENT
DOI: 10.24425/jwld.2023.146608
2023, No. 58 (VII–IX): 158–170

A data-driven approach to predict hydrometeorological variability and fluctuations in lake water levels

Remziye I. Tan Kesgin[1)], Ibrahim Demir[2)], Erdal Kesgin[3)], Mohamed Abdelkader[4)] ✉, Hayrullah Agaccioglu[2)]

[1)] Fatih Sultan Mehmet Vakıf University, Faculty of Engineering, Department of Civil Engineering, Beyoglu, 34445, Istanbul, Turkey

[2)] Yıldız Technical University, Faculty of Civil Engineering, Department of Civil Engineering, Esenler, 34210, Istanbul, Turkey

[3)] Istanbul Technical University, Faculty of Civil Engineering, Department of Civil Engineering, Maslak, 34469, Istanbul, Turkey

[4)] Stevens Institute of Technology, Department of Civil, Environmental, and Ocean Engineering, 1 Castle Point Terrace, Hoboken, NJ 07030, USA

RECEIVED 18.09.2022 ACCEPTED 15.05.2023 AVAILABLE ONLINE 29.09.2023

Abstract: Beyşehir Lake is the largest freshwater lake in the Mediterranean region of Turkey that is used for drinking and irrigation purposes. The aim of this paper is to examine the potential for data-driven methods to predict long-term lake levels. The surface water level variability was forecast using conventional machine learning models, including autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and seasonal autoregressive integrated moving average (SARIMA). Based on the monthly water levels of Beyşehir Lake from 1992 to 2016, future water levels were predicted up to 24 months in advance. Water level predictions were obtained using conventional time series stochastic models, including autoregressive moving average, autoregressive integrated moving average, and seasonal autoregressive integrated moving average. Using historical records from the same period, prediction models for precipitation and evaporation were also developed. In order to assess the model's accuracy, statistical performance metrics were applied. The results indicated that the seasonal autoregressive integrated moving average model outperformed all other models for lake level, precipitation, and evaporation prediction. The obtained results suggested the importance of incorporating the seasonality component for climate predictions in the region. The findings of this study demonstrated that simple stochastic models are effective in predicting the temporal evolution of hydrometeorological variables and fluctuations in lake water levels.

Keywords: evaporation, lake water level, precipitation, stochastic time series models, water transfer

INTRODUCTION

Lake water level fluctuations play a crucial role in various aspects of human life, such as water supply, agriculture, and hydropower generation (Gownaris *et al.*, 2018). Accurate predictions of lake water level fluctuations are therefore essential for effective water resource management. In recent years, data-driven methods have emerged as powerful tools for hydrometeorological predictions, including the prediction of lake water levels. In this article, we will explore the use of data-driven methods for the prediction of lake water level fluctuations, including a discussion of the advantages and limitations of these methods and a comparison to traditional methods, such as stochastic models. We will also provide a case study demonstrating the application of these methods to real-world data. This article aims to provide a comprehensive overview of the state-of-the-art data-driven methods for the prediction of lake water level fluctuations and to highlight their potential for improving our understanding of these important water resources.

The assessment of surface water variability is a crucial aspect of regional water resource management, as it has implications for both economic and environmental policies. In light of this, the

prediction of hydrometeorological variables and lake water level fluctuations is of utmost importance for informed decision-making and the mitigation of water scarcity. In this study, a data-driven approach is proposed for the prediction of the temporal evolution of hydrometeorological variables and lake water level fluctuations. The data-driven model developed in this study aims to provide a simplified understanding of regional climate system variability. In recent years, data-driven models have become a topic of significant interest in hydrometeorological research, with an increasing number of studies utilising these models for the prediction of climate variables. The use of stochastic methods in the prediction of hydrometeorological conditions provides an accurate representation of uncertainty, reduces bias, and improves the representation of long-term climate variability. Previous research on stochastic models for hydrometeorological prediction has demonstrated the robustness of such data-driven models. In this study, stochastic methods were employed to predict the temporal evolution of hydrometeorological variables and lake water level fluctuations at Beyşehir Lake in Turkey.

Beyşehir Lake is one of the most important water resources for domestic and irrigation purposes. Additionally, Beyşehir Lake, which is surrounded by two national parks, is a site of ecological importance that has been designated a national site by the Turkish Ministry of Culture since 1991 (Nas *et al*., 2009). Thus, several studies have been conducted to evaluate the water quality and quantity in Beyşehir Lake (Nas *et al*., 2009; Aktumsek and Gezgin, 2011; Özparlak, Arslan and Arslan, 2012; Bucak *et al*., 2018; Sanli *et al*., 2021; Sanli *et al*., 2022). Predicting water levels is also crucial to ecological sustainability and resilience planning. Prediction of lake water levels has always been a challenging task for hydrologists and water resource managers. There are numerous studies regarding lake water levels in the literature determining the water spread area of lakes (Deoli *et al*., 2021; Kumar and Kuriqi, 2022).

The water surface level of a lake can be significantly impacted by human activities and climate change. Changes in precipitation, flow, evaporation, drinking water supply, and irrigation water usage can result in a decrease in water levels, leading to economic losses and irreversible changes in the lake environment (Cengiz and Kahya, 2006). In light of this, the prediction of lake water levels using statistical methods and machine learning-based algorithms has received significant attention. Accurate prediction of the mean monthly lake water level is particularly important for the planning of multiple water uses, including hydropower plants, commercial navigation, recreational boating, water quality, and the aquatic ecosystem. This study aims to demonstrate the use of data-driven methods for lake water level prediction and to analyse its variability in relation to hydrometeorological variables. The stochastic models employed in this study were developed for predicting monthly water levels up to 24 months in advance for Beyşehir Lake.

There are several nonlinear processes that contribute to the temporal variation of lake water level, including precipitation, evaporation, discharge from tributaries, interbasin water transfer, groundwater fluxes, and topography. The water level variations of a lake become even more complex when it interacts with neighbouring basins. There have been numerous forecasting techniques developed over the past few decades, including hydrodynamic models and data-driven models that are based on historical data. Physically based models require detailed input data, such as terrain information, and complex boundary conditions, and are computationally expensive. A data-driven model, on the other hand, is easier to implement and relies solely on the availability of climate data. In this study, stochastic methods, a hydrometeorological time series-based method, were used for predicting evaporation, precipitation, and lake water levels.

Stochastic methods are widely used in water resources engineering. For instance, a study by Kurunç, Yürekli and Çevik (2005) evaluated autoregressive integrated moving average (ARIMA) and Thomas–Fiering (T–F) models regarding forecasting performance for selected water quality and streamflow parameters of the Yeşilirmak River in Turkey. The study revealed that the variables evaluated had seasonal patterns, but none of them showed a significant trend over the study period. Additionally, it was found that the T–F model provided a slightly better prediction than the ARIMA model. Domenico De *et al*. (2013) compared the chaos theory with the ARIMA model in estimating sea water level on daily, weekly, 10-day, and monthly time scales at Cocos Islands based on measurement data from 1992 to 2001. The results of the study showed that the ARIMA model performed better for daily and weekly averaged time series, while the nonlinear local prediction method performed better for 10-day and monthly averaged time series.

Stochastic models like ARIMA, seasonal autoregressive integrated moving average (SARIMA), and autoregressive moving average (ARMA) are widely used for predicting hydro-meteorological data due to their several advantages. These models are flexible and can handle a variety of trends, seasonality, and autocorrelation structures in the data. They also use a parametric approach, making it easy to interpret and identify underlying patterns in the data. Furthermore, they are easy to use with a simple structure and can be fit to the data using standard statistical software packages and have well-established methods for model selection and validation. Stochastic models have been shown to perform well in a variety of applications, including the prediction of hydro-meteorological data, and can handle missing data. Additionally, they can be adapted to handle different types of data such as daily, weekly, or monthly data.

On the other hand, a variety of data-driven models can be applied to predict hydrological variables. For instance, Buyukyildiz and Tezel (2017) estimated monthly changes in the level of Beyşehir Lake using the generalised regression neural network (GRNN) method, which is an iterative training procedure. The model was constructed with five different input combinations of inflow-lost flow, precipitation, evaporation, and outflow data, with monthly water level change as the output. Further, Dimri, Ahmad and Sharif (2020) investigated the seasonal variation of temperature and precipitation in the Bhagirathi River basin in India. The study evaluated 100 years of precipitation data and concluded that the results obtained from the SARIMA model provide more accurate estimates for flood prediction, urban planning, and environmental planning.

It is pertinent to note that the reliability of forecast models relies on the accuracy and precision of the model output. Thus, comparing model results to real measurements is a critical step for model validation. For example, Coban *et al*. (2021) investigated precipitation prediction (between 2020–2024) in the Marmara region of Turkey using the ARMA, ARIMA, and SARIMA models for agricultural planning, flood control, and the

management of drinking water resources. In order to assess the models, statistical metrics such as mean absolute scaled error (*MASE*), mean absolute error (*MAE*), and root mean squared error (*RMSE*) were computed. Based on the corresponding metrics for the ARMA, ARIMA, and SARIMA models, the ARIMA model performed better than the other methods. The primary objective of this study is to develop a data-driven model for estimating future changes in lake water levels. Towards this end, water level forecasting of Beyşehir Lake, located in the southwest of Turkey, was studied as a case study based on 24 years of time series (1992–2016). Furthermore, the performance of each modelling approach was compared by using a variety of descriptive statistics.

MATERIALS AND METHODS

STUDY AREA

In the Mediterranean Sea region, Beyşehir Lake is one of the largest and most important freshwater lakes. The study area is located in the southern part of the Konya Closed Basin, Turkey, (the largest basin in Anatolia) and is approximately 90 km away from the city of Konya, at the longitude of 31°17'–31°44' E and the latitude of 37°34'–37°59' N (Fig. 1).

Since 1914, Beyşehir Lake has been used for drinking water, irrigation, fishing, commercial activities, and tourism (Buyukyildiz and Tezel, 2017). In addition to serving as a reservoir for irrigation in the Cumra Plain, the lake is also used for navigation purposes (Nas *et al.*, 2009). The lake has an average water depth of 5 m (the maximum water depth of 6.4 m was recorded in August 2006) and an area of 656 km^2 with a drainage area of 4,086.4 km^2. Tectonic Beyşehir Lake is mainly nourished by groundwater, water from the Western Taurus Mountains, and 10 small streams.

The lake water loss is mainly driven by evaporation, underground leaking, and water usage (irrigation, drinking water supply, etc.). It is pertinent to note that the Western Taurus Mountains is the most important karstic region in Turkey. As a result of the karstic ponors and dolines formations, the lake is also interconnected with the Manavgat River in the south. Therefore, the lake is either supplied with groundwater or loses water during the rainy and dry seasons (Özdemir and Özkan, 2007).

The interbasin water transfer is a man-made project that transfers water from a lake to a basin where water is less abundant or could be better utilised for social and economic development. Such schemes can be designed to alleviate water shortages in the receiving basin, generate electricity, or both (Mansouri *et al.*, 2017). Gembos derivation tunnel, which has a length of 15.5 km, is a good example of an interbasin water transfer. Before the interbasin water transfer, it is believed that the Gembos closed basin supplied water to the Manavgat River in the south (Fig. 1).

Beyşehir Lake is placed in a karst-tectonic depression, and it loses water through the sinkholes in its southwest. Doğan *et al.* (2013) developed numerical water to study the groundwater fluxes in the region and showed a significant relation between groundwater outflows and the lake water level. Therefore, the sinkholes on the lake's western shore are above this elevation, it is accepted that there is no loss from these sinkholes at median

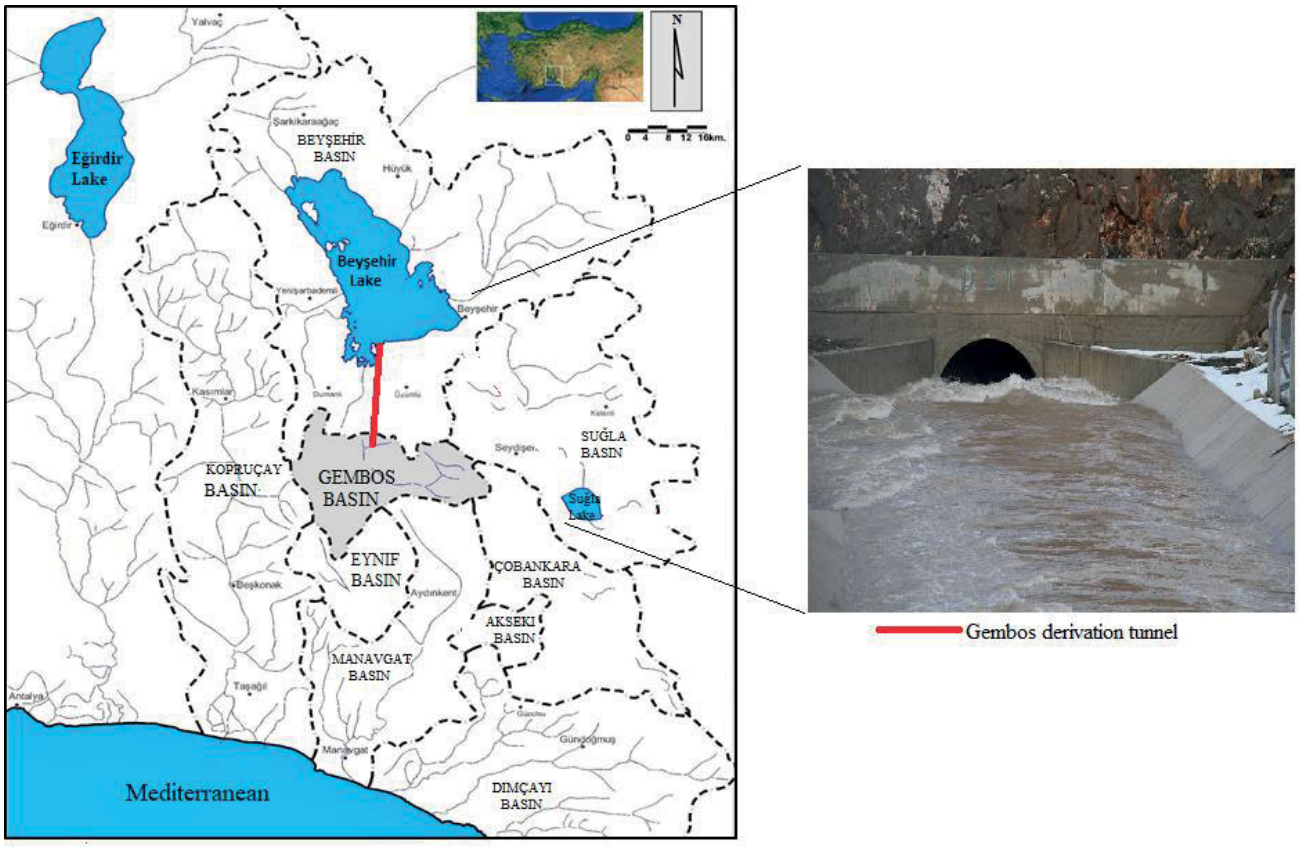

Fig. 1. The geographical location of the study area and illustration of the Gembos derivation tunnel (phot.: *H. Agaccioglu*); source: own elaboration

water levels. This situation develops depending on the water level of the leaks that will occur in the lake. The volume of water transferred to Beyşehir Lake directly affects the amount of land that can be irrigated.

Based on meteorological data, it has been observed that precipitation has decreased and average temperatures have increased in recent years. Karst formations on the surface and underground can cause unexpected water connections that are subject to change over time, and it is possible for water transfer in karst areas to affect the hydrological regimes of adjacent basins. There is a possibility that it may have positive or negative effects on water resources, which can be revealed through data-driven approaches. This comprehensive relationship should be investigated in future studies using an interdisciplinary approach, as well as extending the evaluation to other adjacent basins.

STUDY METHODS

In this study, linear stochastic models referred to as Box–Jenkins, were used to forecasting the time series of Beyşehir Lake's water level using historical records. The datasets were split into a calibration and validation period and stochastic models such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and seasonal autoregressive integrated moving average (SARIMA) were trained and validated (Srivastava, 2015). The framework data-driven modelling framework employed in this study is presented in Figure 2.

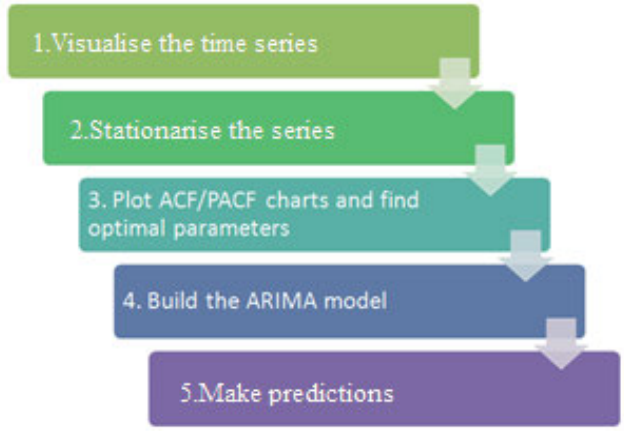

Fig. 2. Flow chart of the data-driven modelling framework; ACF = autocorrelation function, PACF = partial autocorrelation function, ARIMA = autoregressive integrated moving average; source: own elaboration

According to Mwenda, Kuznetsov and Mirau's (2015) recommendations regarding the minimum number of observations required to build an effective ARIMA model (40–50 observations), in the current study a total of 9000 observations were used to construct the model. It is important to note that the dataset consists of daily precipitation, evaporation, and water level measurements covering the period from January 1992 to October 2016. ARIMA models developed by Box and Jenkins (1976) are classified into two categories as seasonal and non-seasonal models. Seasonal ARIMA models are denoted by ARIMA $(p, d, q)\cdot(P, D, Q)_S$. Where, s is period, P, D and Q are the degrees of the seasonal autoregressive operator (SAR), seasonal differencing, and seasonal moving average operator (SMA), respectively. Non-seasonal models are denoted by the notation ARIMA (p, d, q). Where, p, d, and q are the degrees of the autoregressive operator (AR), differencing, and moving average operator (MA), respectively. Non-seasonal models are used in three different models such as autoregressive model AR (p), moving average model MA (q), and autoregressive moving average model ARMA (p, q).

Selecting appropriate values for p, d, and q can be difficult. However, the ARIMA function from the fable package will do it for you automatically. The ARIMA model is an important statistical approach to forecasting and analysing time-series data related to precipitation (Wang *et al.*, 2014).

An AR (p), MA (q), ARMA (p, q), and ARIMA (p, d, q) models can also be represented in Equations (1), (2), (3), and (4) respectively (Hannan, 1970; Mirzavand and Ghazavi, 2015; Dastorani *et al.*, 2016).

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t \quad (1)$$

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-q} \quad (2)$$

where: y_t = actual time series, ε_t = white noise, e_t = white noise (error term), t = periodic time, p = order of autoregressive term, q = order of moving average term, φ_1, ..., φ_p and θ_1, ..., θ_q = model parameters and coefficients, c = constant term.

If statistical parameters of a time series such as mean, variance, and autocorrelation remain constant versus time, the time series data is considered stationary and modelled by using ARMA (p, q) which is the combination of AR and MA models (Eq. 3) (Parvaze *et al.*, 2021).

$$y_t = \delta + \sum_{i=1}^{p} \varphi_i y_{t-i} + e_t - \sum_{j=1}^{p} \theta_j y_{t-j} \quad (3)$$

where: φ_i = i^{th} autoregressive coefficient, θ_j = j^{th} moving average coefficient, δ = stationary part of the autoregressive moving average model.

ARIMA (p, d, q) model is a combination of differencing with the ARMA model, which allows transforming a nonstationary series to a stationary series as expressed in Equation (4).

ARIMA model has three main components such as autoregressive (AR), integrated (I), and moving average (MA). AR component signifies the autocorrelation between present and past observations, while the MA component gives how the new forecasts fit the prior forecast errors and the integrated component (I) represents the degree of difference required to transform a nonstationary series into a stationary series (Asadollahfardi, Rahbar and Fatemiaghda, 2012).

$$(1 - \varphi_1 B - \cdots - \varphi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t \quad (4)$$

where: p, d, q = order of the ARIMA model, B = backward shift operator, $(1 - B)^d$ = d^{th} order difference operation.

The development of an ARIMA (p, d, q) model for forecasting typically encompasses four distinct stages, as delineated below.

- **Model identification.** The selection of optimal values for *p*, *d*, and *q* in the ARIMA model is facilitated by employing graphical techniques, such as plotting the time series and examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.
- **Parameter estimation.** Utilising the available data to estimate the parameters of the identified model.
- **Diagnostic testing.** Model residuals are examined for independence, homoscedasticity, and normality to ensure the model's validity.
- **Series comparison and forecasting.** The synthetic series generated by the best-performing model is contrasted with the original time series. The predictive capacity of hydrological models is evaluated using the Nash–Sutcliffe efficiency coefficient (Nash and Sutcliffe, 1970). Figure 3 illustrates the flowchart for identifying the best-fit model through iterative estimation of model identification and parameters until a satisfactory model is achieved (Parvaze *et al.*, 2021).

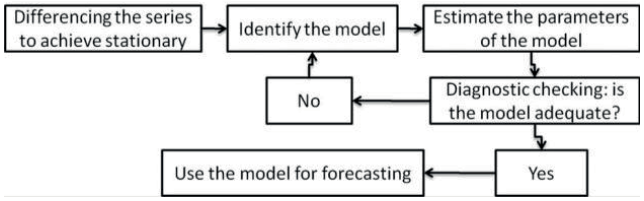

Fig. 3. Flow chart of autoregressive integrated moving average (ARIMA) model application procedure; source: own elaboration based on Box and Jenkins (1976)

The SARIMA model is designed to capture the characteristics of seasonal variations in time series data (Wang, Feng and Liu, 2013). Consequently, numerous researchers, including Han *et al.* (2013), Wang, Feng and Liu (2013), and Park, Onof and Kim (2019), have conducted investigations to estimate precipitation while taking seasonality into account.

In some cases, there is more than one model appropriate for a time series representation. Thus, some criteria should be considered for selecting an appropriate model for time series representation (Yerdelen *et al.*, 2021). Performance efficiency of prediction models may be evaluated by using such descriptive statistic parameters as root mean squared error (*RMSE*), mean absolute scaled error (*MASE*), mean absolute error (*MAE*), and Akaike's information criterion (AIC) (Mohanasundaram, Narasimhan and Kumar, 2017).

For the selection of the best-fit model, recommendations from Kurunç, Yürekli and Çevik (2005), Valipour (2015), and Sirisha, Belavagi and Attigeri (2022) were employed to identify the most suitable model. It is evident that the values of *p*, *d*, and *q* were varied within the range of 0 to 10 to ascertain the optimal ARIMA models. Subsequently, these models were categorised based on their AIC performance metrics. Upon determining the *p*, *d*, and *q* values that represent the best ARIMA models, the top SARIMA models were selected in a similar fashion, considering the seasonality of the input values. In essence, the outcomes of the best ARIMA models informed the development of SARIMA models, which entailed adjustments to the *p*, *d*, and *q* values. Ultimately, the performance of the most proficient prediction models was assessed utilising descriptive indicators, including *MAE*, *RMSE*, and *MASE*. Furthermore, the AIC was employed to gauge prediction error and the relative quality of statistical models for the given data.

RESULTS AND DISCUSSION

In this section, the potential of the stochastic methods in the long-term prediction of precipitation, evaporation, and lake water levels was assessed. Monthly time series were employed to develop the different forecast models. Further, various modelling accuracy assessment tools were computed to evaluate the prediction efficiency of autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and seasonal autoregressive integrated moving average (SARIMA) models.

First, the temporal evolution of the time series was investigated (Fig. 4). The blue lines in Fig. 4 represent the variables' temporal trend after signal decomposition. Statistical features of evaporation, precipitation, and water level time series including average, median, minimum, and maximum values and standard deviation were computed for the 24-year records (Tab. 1). The computed statistics reported that precipitation

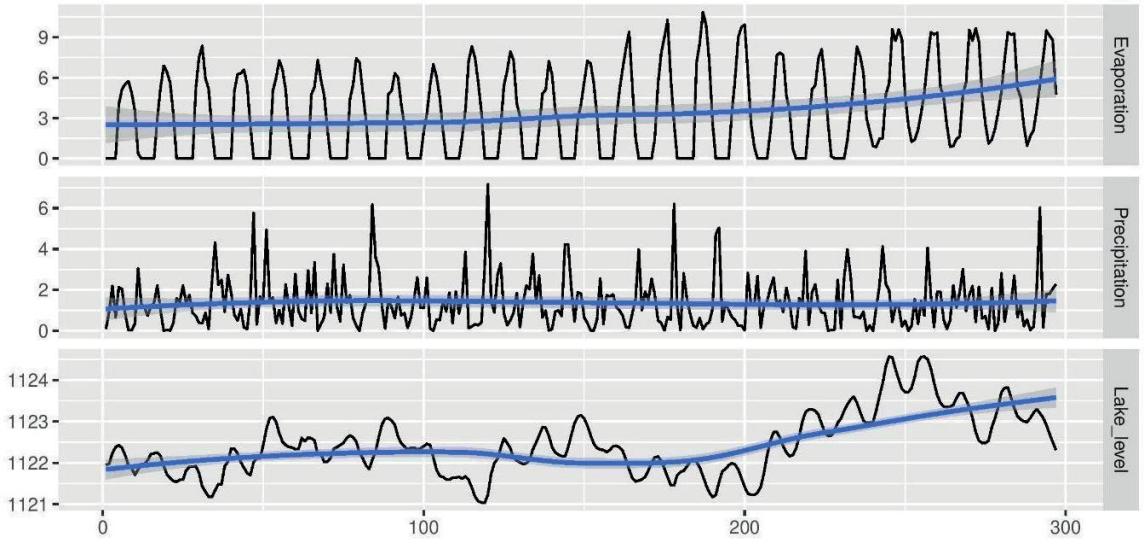

Fig. 4. Time series of evaporation (mm), precipitation (mm), and lake level (m a.m.s.l.); source: own study

Table 1. Statistical parameters of the time series data

Data	Sample size	Mean	Median	Minimum value	Maximum value	Standard deviation
Evaporation (mm)	9040	3.39	2.8	0	13.8	3.5
Precipitation (mm)	9040	1.34	0	0	90.5	4.8
Water level (m)	9040	1122.60	1122.4	1121	1124.6	0.8

Source: own study.

and evaporation values could be null. The recorded null precipitation and evaporation were mainly observed during dry and wet seasons, respectively. In other words, during summer seasons precipitation rates decrease significantly and reach zero.

Results reported in Figure 4 indicated an increasing trend of evaporation over the lake basin. It is important to note that the increasing trend has been more significant since 2012. Further, the graphical representation of the temporal evolution of the evaporation series revealed a change in the mean value where fluctuation occurs. To the knowledge of the authors, there was no modification of the evaporation measurements procedure that could result in a change point in the reported observation. Thus, the observed change can be attributed to a significant increase in evaporation rates in the region. On the other hand, precipitation data did not reveal clear long-term trends.

The data distribution was determined using the violin plot (Fig. 5). The results for the standardised data show that the most prevalent evaporation rates fall below the first quantile or around the third quantile value, suggesting a seasonal variation of evaporation with low evaporation during the wet season and high evaporation during the dry season. As with precipitation, the probability density function shows that extreme and high precipitation rates are less frequent than the first quantile. However, the most frequent water level was observed at values near the median. According to the results obtained, the lake water level appears to have remained more stable during the observation period.

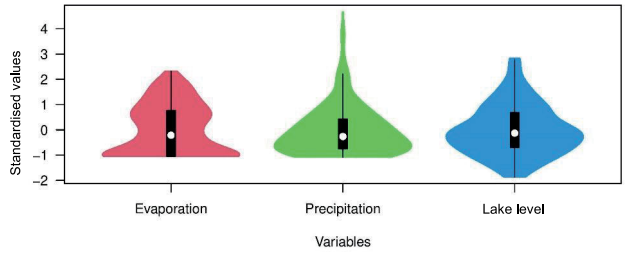

Fig. 5. Violin plot of the data used in this study; source: own study

The information presented in Figure 4 and Figure 5 is insufficient to thoroughly examine the seasonality effect of the investigated variables. To address this limitation, two statistical measures were evaluated: the autocorrelation coefficients (ACF) and partial autocorrelation coefficients (PACF) for each of the evaporation, precipitation, and water level datasets (Fig. 6). An ACF plot is instrumental in discerning the presence or absence of a trend in the data. When a significant trend is evident, ACF tends to exhibit large and positive autocorrelations for small lags. Consequently, time series with pronounced trends display positive ACF values that gradually diminish as lags increase. When data exhibit both a trend and seasonality, a combination of these effects is observed. The gradual decline in ACF is attributable to the trend, while the "scalloped" shape results from seasonality.

As depicted in Figure 5, both trend and seasonality were apparent in the ACF plots for lake water levels. To eliminate trends from the data, the first-order differential process was employed. Consequently, the seasonality effect in ACF plots was further investigated. Based on the first-order differentiated ACF plots, the data were found to be both stationary and seasonal. A seasonal plot resembles a time series plot, with the distinction that data are plotted against the individual "seasons" during which they were collected. Seasonal plots facilitate a clearer visualisation of the underlying seasonal pattern and are particularly valuable for identifying years characterised by pattern changes.

The seasonality variation of the hydrometeorological variables and lake water levels were examined in Figure 7. For each measurement year, the colours represent the average value of the monthly observation. It has been observed that the water level rises during March, April, May, and June, and reaches its lowest point in November each year. A seasonal pattern of precipitation is also evident in Figure 7. More precipitation occurs in April, October, November, and December than during the rest of the year. Evaporation rates increase dramatically in the months between May and October, with maximum evaporation occurring between June and August. The obtained results suggest that the relative increase in the lake water level rates during the months of March–June is associated with a decrease in evaporation rates and an increase in precipitation rates. On the other hand, the observed decrease in the lake water levels during September–November is associated with a significant increase in evaporation rates and a decrease in precipitation rates. However, it should be noted that a lag of one month was observed between the increase–decrease in the hydrometeorological variables and the increase–decrease in the lake water level. In other words, changes in the lake water level do not occur in the same month as the hydrometeorological variables suggesting a late response of lake dynamics to the regional climate variability.

Data from time series may exhibit a variety of patterns, and it is often helpful to separate a time series into several components (Yerdelen and Abdelkader, 2021). A time series consists of three main components namely, the trend-cycle component, the seasonal component, and a residual (remainder) component. Seasonally adjusted series include both the remainder component and the trend cycle. Therefore, they are not "smooth", and "downturns" or "upturns" may be misconstrued. In cases where seasonal variation is not of primary interest, the seasonally adjusted series can be useful (Hyndman and Athanasopoulos, 2018).

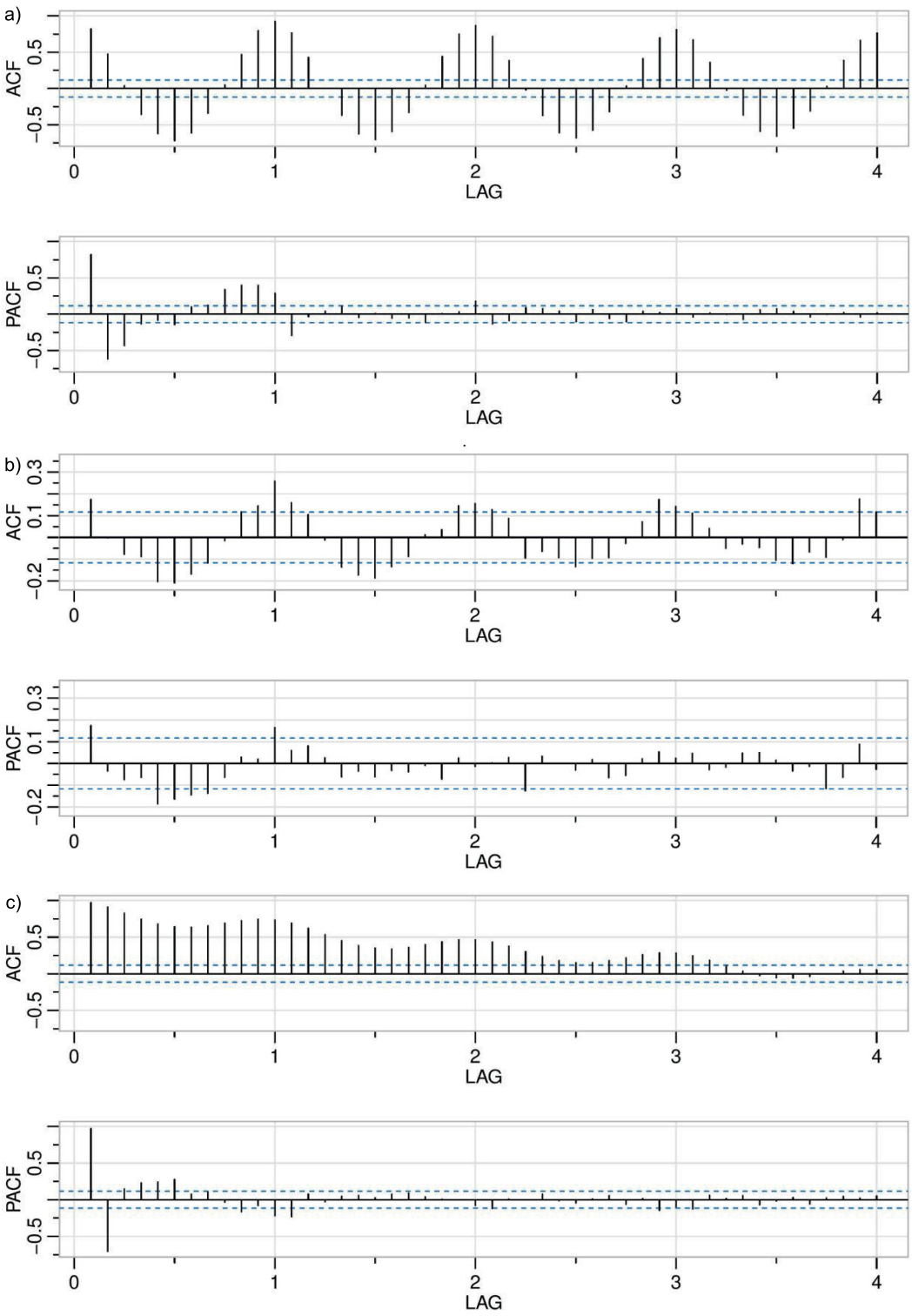

Fig. 6. Autocorrelation coefficients (ACF) and partial autocorrelation coefficients (PACF) plots of monthly: a) evaporation, b) precipitation, c) water level time series; source: own study

The precipitation, evaporation, and water level time series were decomposed to determine their seasonal, trend, and remainder components. In addition, the decomposition of the considered series was employed to understand their characteristics to meet the assumptions of the ARIMA models. Figure 8 illustrates the three components of the analysed data separately. The trend is evident in the decomposed water level and evaporation data, but not in the precipitation data. The series presented in Figure 8 suggests that the seasonal component changes over time and any two consecutive years have similar patterns for evaporation, precipitation, and water level. Further, the residuals component represents what remains after the seasonal and trend-cycle components have been subtracted.

Once the ACF plots, PACF plots, and additive components of the data have been determined, ARIMA models for evaporation, precipitation, and water level were manually defined. Using

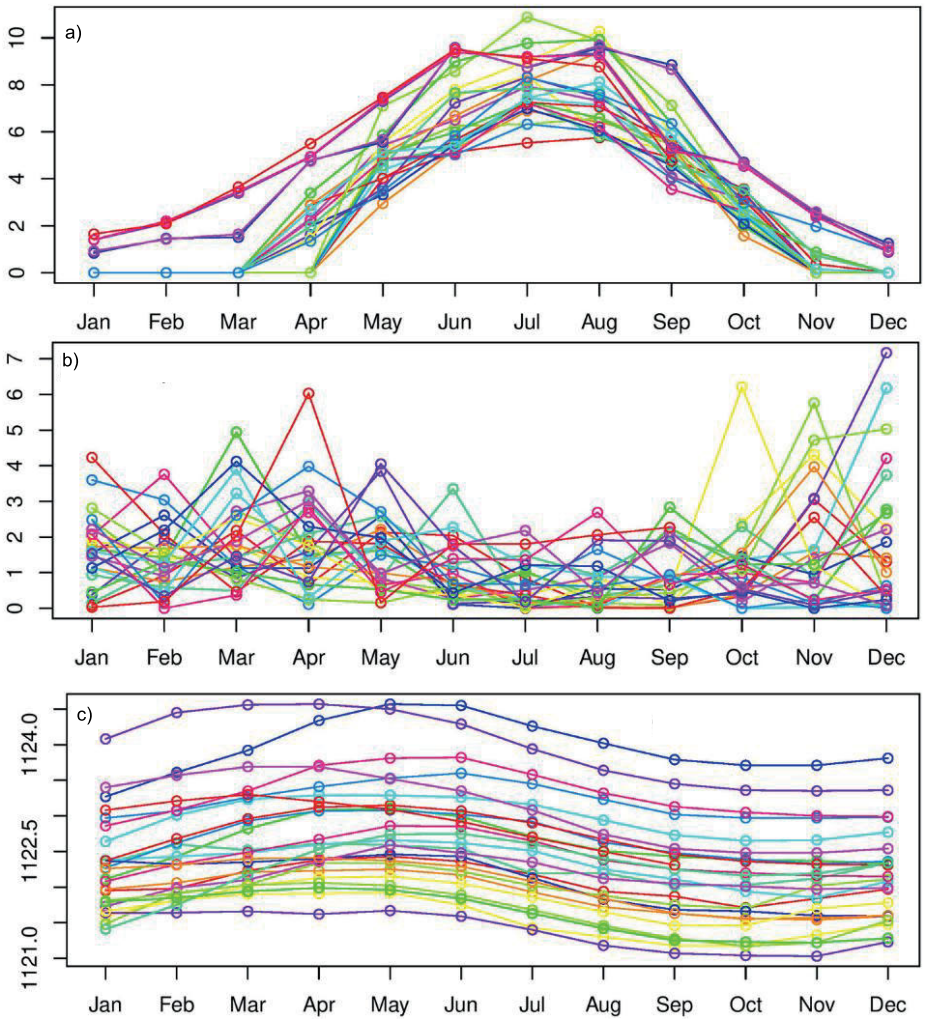

Fig. 7. Beyşehir Lake temporal variation of monthly: a) evaporation (mm), b) precipitation (mm), c) lake level (mm); source: own study

the ACF and PACF, it is possible to determine the initial values of p and q. Therefore, different ARMA models were used with p and q values ranging from 0 to 4. The results of these models were then compared using the AIC. P values were determined by considering the significant relationships of the first lag time of PACF, whereas q values were determined by using ACF. Table S1 which is demonstrated in the Supplementary materials includes the AIC values for the alternative ARMA (p, q) models. It can be stated that the ARIMA model does not respond to the assumption that the residuals are independent.

According to the findings presented in Table S1, the statistical evaluation provided the lowest Akaike's information criterion (AICc) values of –132.4292, 936.7424, and 933.7558 for the best fit ARMA (0, 3), ARMA (2, 4), and ARMA (2, 3), respectively, for lake level, precipitation, and evaporation data. In Figure 8, it is shown that the series is non-stationary. Therefore, series are transformed into stationary series by taking differentiation. In addition, the best fit ARIMA model was also found by taking the d value as 1 and testing p and q values between 0 and 4. Table S1 presents different ARIMA (p, d, q) models with their AICc values. Table S2 shows that the statistical evaluation provided the lowest AICc values as –201.8619, 966.4469, and 867.1585 for the best fit ARIMA (0, 1, 0), ARIMA (1, 1, 4) and ARIMA (2, 1, 4) model for lake level, precipitation, and evaporation data, respectively.

Since the ARIMA model limit is not taking into account time series with seasonality, the SARIMA model was utilised by using (p, d, q) (P, D, Q)$_S$ model where s is the seasonal frequency. Parameters d and D were selected as 1, 2 for the determination of appropriate SARIMA models which can be tested. The tested (p, d, q) (P, D, Q)$_S$ models were evaluated according to the lowest Akaike's information criterion (AICc) value.

An ARIMA model can be automatically parameterised and calibrated utilising the Auto ARIMA function (Bouznad *et al.*, 2020). The forecast package for R recommends selecting the parameters d and D as 1, employing the *ndiffs*() and *nsdiffs*() functions (Hyndman and Khandakar, 2008). AICc values for various tested SARIMA (p, d, q) (P, D, Q)$_S$ models are summarised in Table S3. The final row in Table S3 represents the model suggested by the ARIMA model, which is also obtained using the auto.arima() function.

Table S3 shows that the statistical evaluation provided the lowest AICc value for lake water level as –320.6726 for the best-fit SARIMA (0, 1, 0) (0, 1, 0)$_{12}$ model. The values of AICc 944.7547 and 674.5960 for SARIMA (0, 1, 1) (0, 1, 2)$_{12}$ and SARIMA (2, 1, 1) (1, 1, 1)$_{12}$ model obtained for precipitation and evaporation data, respectively.

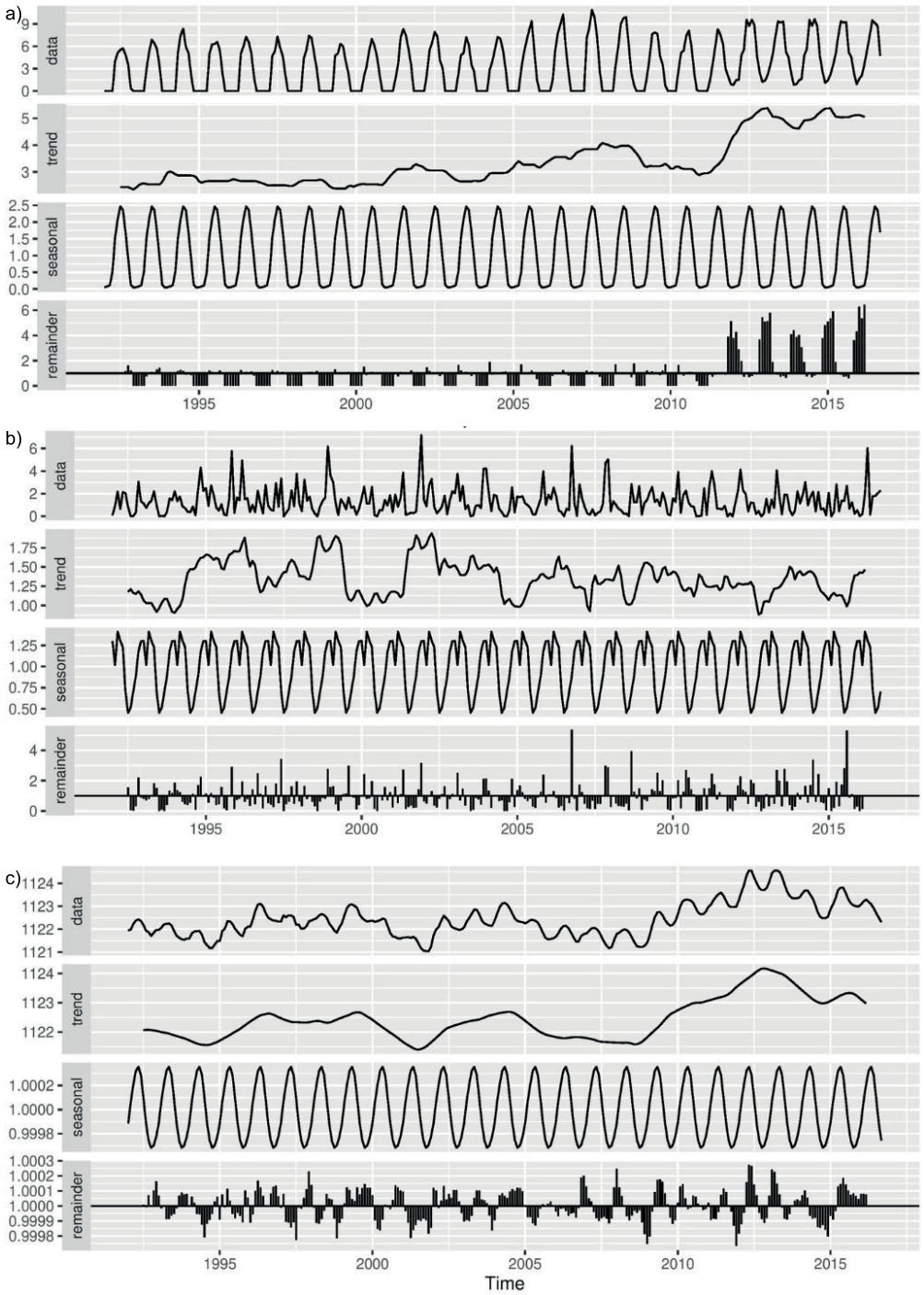

Fig. 8. Three additive components as seasonal, trend cycle, and the remainder for: a) evaporation (mm), b) precipitation (mm), c) lake level (m); source: own study

Table S1, Table S2, and Table S3 illustrate comparisons of best-fit models according to AICs values in the Supplementary materials. However, for choosing the best forecasting model statistical descriptive indicators of correlation coefficient (*R*), *MAE*, *RMSE*, and *MASE* values should be evaluated. In order to evaluate the accuracy of the models and to compare the different best-fit model approaches, four accuracy measures (*R*, *RMSE*, *MAE*, and *MASE*) were defined as illustrated in Table S4 which is given in the Supplementary materials. Moreover, these accuracy parameters were also presented according to the test and train data of these models.

In summary, the first best ARIMA models were evaluated by altering the values of p, d, and q between 0 and 10 in the presented study. These models were listed according to AIC performance criteria in Table S4. In this table, it can be shown that the best models were calculated when $p, d, q \in \langle 0\text{–}4 \rangle$. The second, according to the seasonality of the input values, best SARIMA models were also determined similarly. Due to the results of the best ARIMA models, SARIMA models were created by changing $p, d, q \in \langle 0\text{–}4 \rangle$. They were also demonstrated in Table S3. Then, the best ARIMA and SARIMA models found in Table S3 were listed according to performance criteria such as *R*, *MAE*, *RMSE*, and *MASE* in Table S4.

These indicators are expected to have lower values for the validation of the model. Thus, all models were compared and the auto SARIMA (3, 0, 1) $(0, 1, 1)_{12}$, auto SARIMA (2, 0, 2) $(2, 0, 0)_{12}$, and auto SARIMA (2, 0, 0) $(0, 1, 2)_{12}$ models were found to be the best-fit forecast models for the study area. According to the obtained values in Table S4, the SARIMA models can be evaluated as the best model in terms of *MAE*, *RMSE*, and *MASE* criteria. The obtained results suggest that the seasonality component integrated within the SARIMA model reflects the interseason variability of the studied hydroclimatological variables. For the SARIMA model, a high agreement was found between the observed and predicted data in terms of correlation, suggesting the high capability of the model to capture the temporal evolution of the predicted data. In terms of error metrics, the SIRMA model showed lower error values suggesting the model's capability to simulate low and peak values of the predicted variables.

Figure 9 shows the testing period between 2016 and 2018 for Beyşehir Lake water level, precipitation, and evaporation. The best-fit SARIMA models were used for the forecasting. In Figure 9, the time series are shown in black colour, and forecasted values are shown in blue colour while the dark grey and light grey bands represent the 80% and 95% confidence intervals of forecasts respectively. For the testing period, the SARIMA model showed high agreement with the observed data (Tab. 5). The obtained results suggest the capability of the SARIMA model to predict hydrometeorological variability and fluctuations in lake water levels.

Inland lakes are one of the most influential factors for each environment, and they have a profound impact on its surroundings. In addition to protecting humans and wildlife, they play an essential role in preserving the environment. In between, freshwater lakes are used directly for drinking water supply or crop irrigation, providing many environmental, economic, and ecological benefits. Both the environment and humans are threatened by the depletion and destruction of some of these lakes. Therefore, it is of the utmost importance to analyse and forecast the future levels of freshwater lakes. This study examined the monthly time series of Beyşehir Lake level, precipitation, and evaporation using various pre-processing methods and data-driven models, including ARMA, ARIMA, and SARIMA models.

A comparison of the results of the different stochastic models indicated that hydrometeorological variability and fluctuations in lake water can be modelled by using an appropriate method that provides accurate information on the entailing terms in time series. The SARIMA model performs better in terms of accuracy than the ARMA and ARIMA models when forecasting monthly series. This can be explained by the well representation of the seasonality component in the SRIMA models. It is suggested that linear methods should be employed in conceptual hydrological modelling, particularly lake water level modelling. In this study, it has been demonstrated that by properly understanding the components of a time series and defining the appropriate methodology, it is possible to develop an accurate and simple model based on data-driven methods. Overall, the SARIMA model definitely outperforms the other stochastic methods in short-term plans such as exploitation and consumption management, and in long-term plans such as designing and constructing hydraulic structures. The finding of this study indicated that stochastic models have the ability to portray various potential future scenarios. This helps prevent substantial shortcomings found in deterministic models, giving stochastic models an advantage.

However, the problem is that the employed method is currently designed using an assumed stationarity of the data with a constant mean and variance. This limitation should be addressed in future studies by incorporating a non-linear method under the assumption of a changing climate and a non-stationary framework. It is also important to note that data-driven models are demanding in terms of training and calibration data and their application is limited to gauged sites. Thus, the method cannot be applied to ungauged sites and unobserved regions. In addition, the findings of the investigation suggest that other stakeholders should be involved in the management of Beyşehir Lake in a cooperative and prudent manner. Consequently, anthropogenic activities will have a reduced impact on the lake's water quality, water level, aquatic ecosystems, and adjacent terrestrial ecosystems. The lake will be more resilient to the effects of climate change as a result of these measures.

Future studies should consider the fact that stochastic models, such as ARIMA, SARIMA, and ARMA, have several limitations and disadvantages to be aware of when using them for hydro-meteorological data predictions. These models make certain assumptions about the underlying structure of the data, including stationarity and linearity, which can lead to inaccurate predictions if not met. Additionally, they have limited complexity and may not be able to effectively capture more complex relationships in the data, such as non-linear relationships or multiple seasonality. The quality of the data is also a factor, as outliers, missing values, and measurement errors can all impact the accuracy of the model predictions. Choosing the correct model can be challenging, as there are many models to choose from, each with its own strengths and weaknesses. Furthermore, these models have a limited ability to account for external factors that may affect hydro-meteorological data, such as changes in land use, climate change, and human interventions. To ensure accurate predictions, it's important to carefully consider the limitations and underlying assumptions of each model when selecting the appropriate model for a given dataset.

Each of the alternative methods for hydro-meteorological data prediction, including artificial neural networks (ANNs), support vector machines (SVMs), ensemble methods, hybrid models, and physical models, has its own advantages compared to stochastic models like ARIMA, SARIMA, and ARMA. ANNs are capable of modelling complex non-linear relationships in the data and handling large amounts of data, making them well-suited for prediction problems involving large datasets. SVMs are advantageous in that they can handle both linear and non-linear relationships in the data, and can be used for both classification and regression problems. Ensemble methods combine the strengths of multiple models, leading to more robust predictions, and can account for model uncertainty. Hybrid models combine the strengths of multiple modelling approaches, resulting in more accurate predictions and can handle complex relationships in the data and large amounts of data. Physical models are based on the underlying physical processes that control hydro-meteorological variables and can lead to more accurate predictions, as well as being used for long-term predictions. The choice of method will depend on the specific characteristics of the data and the desired outcome, and in some cases, a combination of methods may be necessary to achieve the best results.

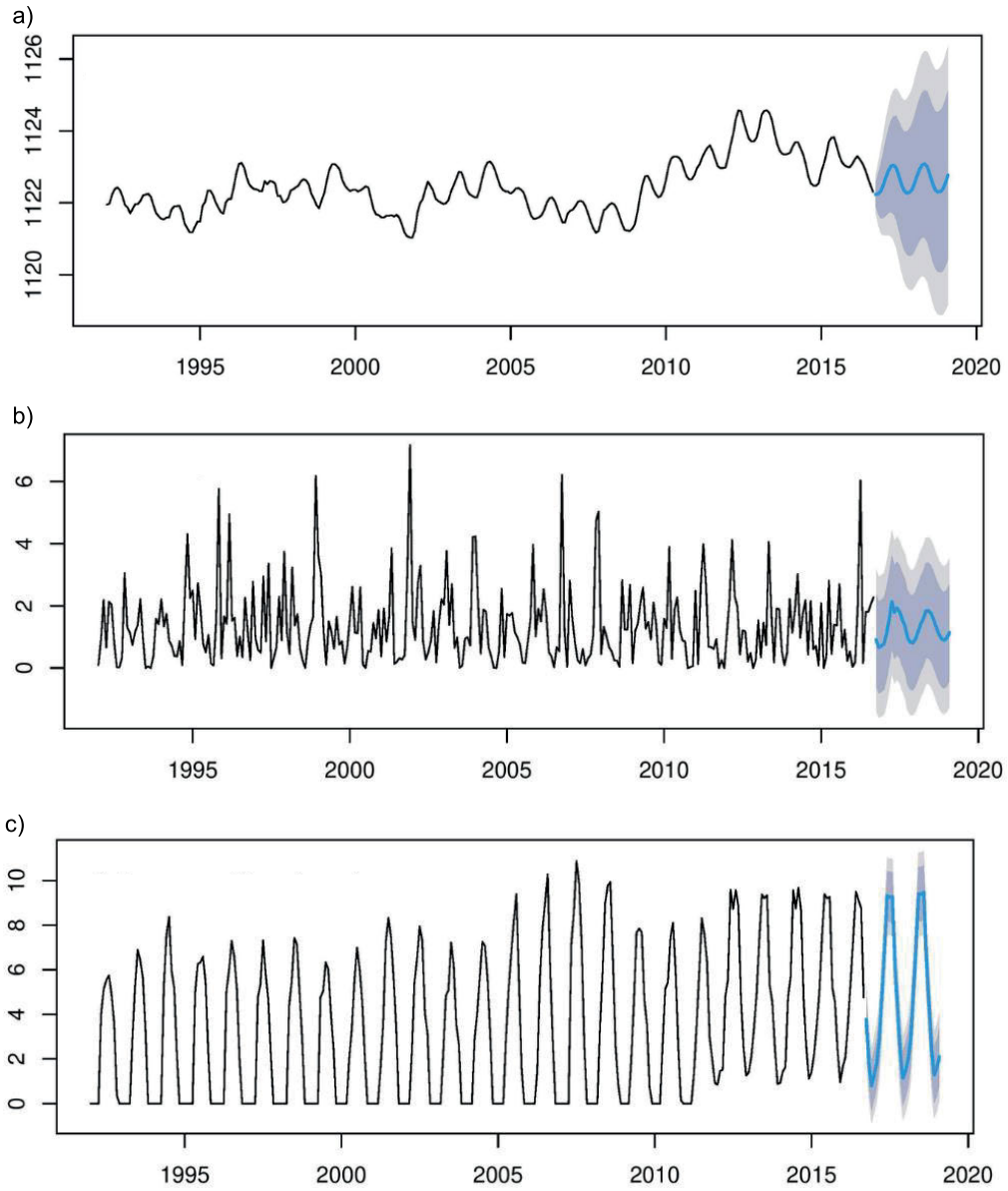

Fig. 9. Seasonal autoregressive integrated moving average (SARIMA) model forecasts of a) lake level (m); SARIMA (3, 0, 1) (0, 1, 1)$_{12}$, b) precipitation (mm); SARIMA (2, 0, 2) (2, 0, 0)$_{12}$ with non-zero mean, c) evaporation (mm); SARIMA (2, 0, 0) (0, 1, 2)$_{12}$ with drift; source: own study

CONCLUSIONS

Beyşehir Lake is the largest freshwater lake in Turkey that acts a key role in the centre of the interbasin water transfer. Many surface and underground karst forms in karst regions especially located in a lake basin result in unexpected water connections that change with time, and it is highly possible to change the hydrological regime of neighbouring basins due to the interbasin water transfer. Therefore, significant effects on water resources such as Beyşehir Lake also result in significant hydrometeorological variability and variations in lake levels. In brief, the presented study recommends forecasting and predicting the lake water level fluctuations with considering hydrometeorological parameters.

The present study also evaluates the temporal evolution of hydro-meteorological data and uses the ARMA, ARIMA, and SARIMA methods to predict precipitation, evaporation, and the level of Beyşehir Lake in the Mediterranean Sea region of Turkey. Forecasts were made for a two-year time period based on 24 years of historical data for model tuning.

In this study, *MAE*, *RMSE*, and *MASE* metrics were computed to assess the models' performance. The SARIMA model performed better than the other models in terms of forecast accuracy, where the model performance criteria and AICc provided the lowest values.

Among the three methods, ARMA, ARIMA, and SARIMA, ARIMA achieved the most favourable results. The SARIMA model is determined to be the most suitable water level forecast model based on model evaluation criteria. The ARIMA model is one of the time series models that take seasonality into account as an integral part of the modelling process. This approach models seasonality by taking into account the statistical properties of the data. In future studies, besides the precipitation parameter, a study can incorporate other variables such as solar radiation, air

pollution, snowfall, and climate indices. Various machine learning methods can be used to estimate the relationships between hydro-meteorological variables. Forecasting approaches can be combined, and various model evaluation criteria can be proposed for the forecasting model's success.

SUPPLEMENTARY MATERIAL

Supplementary material to this article can be found online at https://www.jwld.pl/files/Supplementary_material_Tan_Kesgin.pdf

ACKNOWLEDGEMENTS

We acknowledge the support from the Turkish State Meteorological Service for providing climate data used in this work.

REFERENCES

Aktumsek, A. and Gezgin, S. (2011) "Seasonal variations of metal concentrations in muscle tissue of tench (*Tinca tinca*), water and sediment in Beysehir Lake (Turkey)," *Environmental Technology*, 32(13), pp. 1479–1485. Available at: https://doi.org/10.1080/09593330.2010.540717.

Asadollahfardi, G., Rahbar, M. and Fatemiaghda, M. (2012) "Application of time series models to predict water quality of upstream and downstream of Latian dam in Iran," *Universal Journal of Environmental Research and Technology*, 2(1), pp. 26–35. Available at: http://www.environmentaljournal.org/2-1/ujert-2-1-4.pdf (Accessed: October 26, 2021).

Bouznad, I.-E. *et al.* (2020) "Trend analysis and spatiotemporal prediction of precipitation, temperature, and evapotranspiration values using the ARIMA models: Case of the Algerian Highlands," *Arabian Journal of Geosciences*, 13, 1281. Available at: https://doi.org/10.1007/s12517-020-06330-6.

Box, G.E.P. and Jenkins, G.M. (1976) *Time series analysis: Forecasting and control*. Rev. edn. San Francisco: Holden-Day.

Bucak, T. *et al.* (2018) "Modeling the effects of climatic and land use changes on phytoplankton and water quality of the largest Turkish freshwater lake: Lake Beyşehir," *Science of the Total Environment*, 621, pp. 802–816. Available at: https://doi.org/10.1016/j.scitotenv.2017.11.258.

Buyukyildiz, M. and Tezel, G. (2017) "Utilization of PSO algorithm in estimation of water level change of Lake Beysehir," *Theoretical and Applied Climatology*, 128, pp. 181–191. Available at: https://doi.org/10.1007/s00704-015-1660-2.

Cengiz, T.M. and Kahya, E. (2006) "Türkiye göl su seviyelerinin eğilim ve harmonik analizi [Trend and harmonic analysis of lake water levels in Turkey]," *itüdergisi/d*, 5(3), pp. 215–224. Available at: http://itudergi.itu.edu.tr/index.php/itudergisi_d/article/viewFile/511/441 (Accessed: May 10, 2021).

Coban, V. *et al.* (2021) "Precipitation forecasting in Marmara region of Turkey," *Arabian Journal of Geosciences*, 14, 86. Available at: https://doi.org/10.1007/s12517-020-06363-x.

Dastorani, M. *et al.* (2016) "Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition," *Natural Hazards*, 81(3), pp. 1811–1827. Available at: https://doi.org/10.1007/s11069-016-2163-x.

Deoli, V. *et al.* (2021) "Water spread mapping of multiple lakes using remote sensing and satellite data," *Arabian Journal of Geosciences*, 14, 2213. Available at: https://doi.org/10.1007/s12517-021-08597-9.

Deoli, V., Kumar, D. and Kuriqi, A. (2022) "Detection of water spread area changes in eutrophic lake using Landsat data," *Sensors*, 22(18), 6827. Available at: https://doi.org/10.3390/s22186827.

Dimri, T., Ahmad, S. and Sharif, M. (2020) "Time series analysis of climate variables using seasonal ARIMA approach," *Journal of Earth System Science*, 129, 149. Available at: https://doi.org/10.1007/s12040-020-01408-x.

Doğan, A. *et al.* (2013) "Göl-yeraltı suyu-iklim ilişkisinin yeraltı suyu akım modeli ve coğrafi bilgi sistemleri (CBS) yardımıyla belirlenerek gölün optimum dinamik işletme modelinin oluşturulması: Beyşehir gölü modeli [Investigation of the optimum dynamic operation model of the lake by determining the lake-groundwater-climate relationship with the groundwater flow model and geographic information systems (GIS): Lake Beyşehir model]," *Project No: 109Y271*. İstanbul: Tübitak, pp. 1–194.

Domenico De, M. *et al.* (2013) "Chaos and reproduction in sea level," *Applied Mathematical Modelling*, 37(6), pp. 3687–3697. Available at: https://doi.org/10.1016/j.apm.2012.08.018.

Gownaris, N.J. *et al.* (2018) "Water level fluctuations and the ecosystem functioning of lakes," *Journal of Great Lakes Research*, 44(6), pp. 1154–1163. Available at: https://doi.org/10.1016/j.jglr.2018.08.005.

Han, P. *et al.* (2013) "Application of the ARIMA models in drought forecasting using the standardized precipitation index," in D. Li and Y. Chen (eds.) *Computer and Computing Technologies in Agriculture VI. CCTA 2012. IFIP Advances in Information and Communication Technology*, 392. Berlin, Heidelberg: Springer. Available at: https://doi.org/10.1007/978-3-642-36124-1_42.

Hannan, E.J. (1970) *Multiple Time Series*. Hoboken: John Wiley & Sons.

Hyndman, R.J. and Athanasopoulos, G. (2018) *Forecasting: Principles and practice*. 2nd edn. Melbourne: OTexts.

Hyndman, R.J. and Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R," *Journal of Statistical Software*, 27(3). Available at: https://doi.org/10.18637/jss.v027.i03.

Kurunç, A., Yürekli, K. and Çevik, O. (2005) "Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilırmak River, Turkey," *Environmental Modelling and Software*, 20(9), pp. 1195–1200. Available at: https://doi.org/10.1016/j.envsoft.2004.11.001.

Mansouri, R. *et al.* (2017) "Dynamic programming model for hydraulics and water resources simulating and optimizing water transfer system (a case study in Iran)," *Aqua*, 66(8), pp. 684–700. Available at: https://doi.org/10.2166/aqua.2017.110.

Mirzavand, M. and Ghazavi, R. (2015) "A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods," *Water Resources Management*, 29, pp. 1315–1328. Available at: https://doi.org/10.1007/s11269-014-0875-9.

Mohanasundaram, S., Narasimhan, B. and Kumar, G.S. (2017) "Transfer function noise modelling of groundwater level fluctuation using threshold rainfall-based binary-weighted parameter estimation approach," *Hydrological Sciences Journal*, 62(1), pp. 36–49. Available at: https://doi.org/10.1080/02626667.2016.1171325.

Mwenda, A., Kuznetsov, D. and Mirau, S. (2015) "Analyzing the impact of historical data length in non-seasonal ARIMA models forecasting," *Mathematical Theory and Modeling*, 5(10), pp. 77–85. Available at: https://core.ac.uk/download/pdf/234680245.pdf (Accessed: June 15, 2021).

Nas, B. *et al.* (2009) "Seasonal and spatial variability of metals concentrations in Lake Beyşehir, Turkey," *Environmental Technology*, 30(4), pp. 345–353. Available at: https://doi.org/10.1080/09593330902752984.

Nash, J.E. and Sutcliffe, J.V. (1970) "River flow forecasting through conceptual models part I – A discussion of principles," *Journal of Hydrology*, 10(3), pp. 282–290. Available at: https://doi.org/10.1016/0022-1694(70)90255-6.

Özdemir, F.Y. and Özkan, Ö. (2007) "Importance of geological characteristics at determining basin conversation borders: Sample of Lake Beysehir (Konya) Basin," *International Congress on River Basin Management, Basin Resources Protection*, pp. 294–307.

Özparlak, H., Arslan, G. and Arslan, E. (2012) "Determination of some metal levels in muscle tissue of nine fish species from the Beyşehir Lake, Turkey," *Turkish Journal of Fisheries and Aquatic Sciences*, 12(4). Available at: https://doi.org/10.4194/1303-2712-v12_4_04.

Park, J., Onof, C. and Kim, D. (2019) "A hybrid stochastic rainfall model that reproduces some important rainfall characteristics at hourly to yearly timescales," *Hydrology and Earth System Sciences*, 23, pp. 989–1014. Available at: https://doi.org/10.5194/hess-23-989-2019.

Parvaze, S. *et al.* (2021) "Temporal flood forecasting for trans-boundary Jhelum River of Greater Himalayas," *Theoretical and Applied Climatology*, 144, pp. 493–506. Available at: https://doi.org/10.1007/s00704-021-03562-8.

Sanli, A. *et al.* (2022) "Effect of lake-water budget management preferences on optimum operating conditions and neighboring basins interacting: case of Lake Beyşehir (Turkey)," *Sustainable Water Resources Management*, 8(1). Available at: https://doi.org/10.1007/s40899-021-00599-5.

Sanli, A.E. *et al.* (2021) "Statistical assessment of interbasin water transfer for karst areas (Turkey)," *Arabian Journal of Geosciences*, 14, 2342. Available at: https://doi.org/10.1007/s12517-021-08693-w.

Sirisha, U.M., Belavagi, M.C. and Attigeri, G. (2022) "Profit prediction using ARIMA, SARIMA and LSTM models in time series forecasting: A comparison," *IEEE Access*, 10, pp. 124715–124727. Available at: https://doi.org/10.1109/access.2022.3224938.

Srivastava, T. (2015) *A complete tutorial on time series modeling in R*. Analytics Vidhya. Last updated: 15 May 2023. Available at: https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/ (Accessed: May 15, 2020).

Valipour, M.S. (2015) "Long-term runoff study using SARIMA and ARIMA models in the United States," *Meteorological Applications*, 22(3), pp. 592–598. Available at: https://doi.org/10.1002/met.1491.

Wang, H. *et al.* (2014) "An improved ARIMA model for precipitation simulations," *Nonlinear Processes in Geophysics*, 21(6), pp. 1159–1168. Available at: https://doi.org/10.5194/npg-21-1159-2014.

Wang, S., Feng, J. and Liu, G. (2013) "Application of seasonal time series model in the precipitation forecast," *Mathematical and Computer Modelling*, 58(3–4), pp. 677–683. Available at: https://doi.org/10.1016/j.mcm.2011.10.034.

Yerdelen, C. *et al.* (2021) "Estimation of standard duration maximum rainfall by using regression models," *Journal of Water and Land Development*, 50, pp. 281–288. Available at: https://doi.org/10.24425/jwld.2021.138184.

Yerdelen, C. and Abdelkader, M. (2021) "Hydrological data trend analysis with wavelet transform," *Comptes Rendus De L'Acade'-mie Bulgare Des Sciences*, 74(8), pp. 1194–1202. Available at: https://doi.org/10.7546/CRABS.2021.08.11.