Improving the Prediction of Evaporation Variable in Mosul Dam Using ARIMA Model and Time Series Analysis

Khalid MK Khafaji[1,*], Bassem Ben Hamed[1]

[1]National School of Electronics and Telecommunications, University of Sfax, Tunisia

Emails: alkhafajik@gmail.com; bassem.benhamed@enetcom.usf.tn

Abstract

Evaporation plays a significant role in managing water resources and is an important indicator in risk and crisis management, particularly in operating reservoirs and dams. Precise predictions of evaporation rates are crucial to effective water resource management, and various modelling methods, including AI and autoregression, have been employed to create accurate models. This makes it more important to use innovative technology to continuously monitor this phenomenon with accurate scientific results, allowing decision-makers to be aware of and prepare for potential drought risks and crises. In this study, therefore, we propose the establishment of a mechanism that would include analyzing and exploring the data used in this study (Evaporation) and cleaning up the impurities of actual and lost values to obtain accurate data that would serve as actual inputs to ARIMA model that will adopt in this study, This mechanism would contribute to the performance and efficiency of this model using time series data to accurately predict future trends of evaporation plants in the water of the Mosul dam. Our objective is to explain the diversity of climate policies and actions using a data-based approach to analyzing integrated parameters over the years, etc. This is complemented in depth by how different methods of extracting data behaviour are used to study model forecasts. This collaborative study aims to enhance future studies by using more comprehensive datasets with more learning models. The researchers believe in the power of sharing knowledge and are thus committed to sharing the results of other causes outside of global warming that contribute to climate change.

Keywords: Arima; Time series; Data analysis; Prediction; Evaporation; Risk

1. Introduction

The Mosul dam is the largest dam in northern Iraq and the fourth dam in the Middle East, located on the Tigris stream from Turkish territory. The area of water reserved for this dam is approximately 11.1 km2 (2.7 cubic miles) and an estimated 1.7 million people in Mosul have access to electricity [1]. In view of the recent problems between Iraq and Turkey caused by the water crisis and the low number of releases of Iraq water share by Turkey, the dry climate in Iraq and the high summer temperatures caused an increase in evaporation, leading to a reduction in the actual water stock in the dam. The evaporation losses can have a significant effect on the water budget of reservoirs or lakes; this will lead to a decrease in the water level. In managing dams, construction design is critical in reducing evaporation loss; for example, a deep and narrow dam will minimize the surface area, decreasing the evaporation loss. Additionally, the ratio of surface area to volume and the presence of a salt subterranean water table can influence the effectiveness of strategies to reduce dam evaporation [2]. Additionally, the impact of climate change is to increase the evaporation rate from dams; while the volume of water stored in dams is decreasing, this leads to an increase in the loss of water evaporated from the dam [3]. Given the importance of the evaporation element in the open water storage process (dams and lakes) and the current scarcity of water in Iraq, it has become necessary to focus attention on this critical situation by using smart learning models and serving their results in decision-making to manage future risks. Therefore, there is an ever-escalating need for appropriate dam risk management strategies to elucidate the factors associated with changes in water magnitude and frequency [4], suggesting that a deeper knowledge of nature, timing and extent of these influences are required to understand changes in dam behavior [5]. Before a predefined model is developed, the paradoxes must be found in the original data set and

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

removed to provide the intended prediction. This research provides one electrical analysis to reduce noise. In addition, they are affected by many environmental conditions. This study will reveal the impact of data analysis and exploration on the efficiency and accuracy of the learning models used in the proposed methodology as well as the low error rate through the RMSE scale [6], and a measurement of errors between the two observations that reflect the same phenomenon, as we will clean up the data, isolate the missing values and then prepare them through a series of statistical procedures to obtain high-quality data that can be used as key inputs to the (deep and machine) smart learning models. The data entered will be ready for training, verification and testing in order to produce accurate, scientific and logical outputs by which the desired results can be obtained from the methodologies used for the same purpose. In addition to the importance of this methodology in exploring and analyzing the data set used in this study, its importance and the need to cleanse it of undesirable values for obtaining high-quality, pure data that preserves the effective performance of learning models in order to produce tangible outputs that can be relied upon in decision-making processes or in accessing relevant facts with scientific value. This study aims to predict the amount of water evaporation in the Mosul dam, which suffers from problems in its stock due to drought and the low release of the water share from the Turkish side, which could cause a crisis of water scarcity for the population and land in the future and a lack of electricity generation. Also, test the effectiveness of the proposed ARIMA model. The results showed the practical efficacy of the model used for sound and accurate forecasting of the integrated framework and its potential use in disaster management and risk prevention decisions, by incorporating machine learning techniques, deep learning algorithms, and ensemble methods. The next generation of AI-based prediction models aims to provide more reliable and efficient predictions of dam evaporation rates. The continuous advancement of AI technologies offers significant potential for optimizing water resource management strategies and enhancing the resilience of water systems in the face of climate change.

2. Related Work

In [7] AI algorithms can predict future evaporation trends by utilizing historical time-series data on evaporation rates with high accuracy and adaptability. Studies have shown that AI models, such as the MIC-RF model, exhibit high robustness and accuracy in predicting dam seepage. Challenges in utilizing artificial intelligence (AI) for dam evaporation prediction are multifaceted. One of the primary challenges lies in the quality of data inputs, as inaccuracies or inconsistencies in the data can significantly impact the performance of AI models. Additionally, the complexity of integrating AI algorithms with existing prediction frameworks poses a technical challenge for researchers and practitioners. Overcoming these challenges requires a concerted effort to improve data quality, develop robust AI algorithms, and enhance the compatibility of AI systems with dam evaporation prediction models. The integration of AI techniques into evaporation prediction has the potential to revolutionize water resource management practices.

In [8] the impact of time series analysis on dam evaporation prediction is profound. By exploring the relationship between climatic conditions and reservoir evaporation rates, researchers have developed models that enhance predictive capabilities. Artificial intelligence (AI) has emerged as a powerful tool for improving dam evaporation prediction models. The future prospects of artificial intelligence (AI) in enhancing dam evaporation prediction are promising. Researchers are exploring the impact of climatic conditions on evaporation prediction and developing advanced AI models to improve forecasting accuracy. Time series analysis identifies patterns and trends in evaporation data, enabling more accurate forecasting of future evaporation rates. Understanding how climatic variables such as temperature, humidity, and wind speed affect evaporation is vital for effective water resource management.

In [9] time series modeler (TSM) for rainfall forecasting in an Indian coastal region is presented in this paper. This model is based on a five-year dataset (2009–2013) that includes key attributes such as temperature, dew point, wind speed, maximum temperature, minimum temperature, visibility, and rainfall. An innovative approach was used by training and testing this dataset using TSM of SPSS, the Statistical Package for the Social Sciences. A dependable model for rainfall prediction is achievable because the model's performance is assessed using statistical performance measures, such as squared error (MSE), root mean squared error (RMSE), mean absolute percent error (MAPE), and mean absolute deviation (MAD). With an accurate range of 80% for predictions, the model's output is quite respectable. The SPSS 20.0 autoregressive integrated moving average (ARIMA) model of TSM is the foundation of this approach.

In [10][11] Applying the autoregressive integrated moving average (ARIMA) model has significantly improved dam evaporation prediction by leveraging time series data; the ARIMA model can forecast future trends in evaporation rates with a high degree of accuracy. Moreover, a hybrid approach that combines linear and nonlinear models has been proposed to further enhance prediction capabilities. This integration of statistical machine learning techniques with traditional ARIMA modelling has shown promising results in predicting dam evaporation

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

rates. Time series analysis, mainly through applying the autoregressive integrated moving average (ARIMA) model, has provided valuable insights into evaporation patterns. Artificial intelligence (AI) has emerged as a powerful tool for improving prediction models, although data quality and integration challenges persist.

In [12] the data quality used in dam evaporation prediction models directly impacts their accuracy. Variations in data quality can lead to differences in model performance, as demonstrated in field-scale applications. Therefore, ensuring the integrity and reliability of the data inputs is crucial for developing robust and precise prediction models. The researchers continuously strive to improve data collection methods and quality to enhance the overall accuracy of dam evaporation predictions.

In [13][14] The root mean square error (RMSE) serves as a crucial metric for evaluating the accuracy of dam evaporation prediction models. It measures the average difference between predicted and actual values, providing insights into the model's predictive performance. Understanding the relationship between RMSE and prediction accuracy is essential for refining prediction models and enhancing their reliability. By minimizing RMSE values, researchers can improve the overall accuracy of dam evaporation predictions and optimize water resource management strategies.

Crisis management is one of the best practices necessary for a sound corporate governance structure. In addition, the decision-making process is necessary at every stage of the effectiveness of the rescue effort and events are impacted by crisis management. Big data analysis is required for this decision-making, which is more difficult than regular data analysis. This emphasizes the requirement for computational intelligence, real time algorithms that can decide quickly, analyse various data formats, extract the facts, and present it using visualization strategies [15].

Time is the most crucial factor in deciding the success or failure of an institution in data science and machine learning. These organizations devote years to data analysis to gain a deeper understanding of data behaviour, allowing us to make better judgments during crises and prepare for future threats. Despite your title as a data scientist, you must prepare for this study. You can stop stressing about building a time machine right now. Time series modelling is a powerful tool for finding and forecasting trends and patterns. Yet, many additional dimensions are included in a time series model. The bulk of the visuals you see online are of univariate time series. Unfortunately, it doesn't work in a way in the actual world. A data scientist's worth is based on how well they handle multiple variables simultaneously [16].

3. Methodology

3.1 Exploring Data Analysis (E.D.A)

By examining the data on evaporation over time, valuable information can be gleaned regarding fluctuations in seasonality and geographical location, which can, in turn, significantly assist in the effective management of water resources [17][18]. Examining evaporation data over time provides valuable insights into fluctuations of evaporation rates across different seasons and regions. Research conducted on the hydrological balance of urban areas has specifically investigated evaporation from paved surfaces to better understand how urbanization affects the water cycle [17]. Furthermore, examining reservoir surface evaporation plays a crucial role in estimating water loss and optimizing water distribution in arid regions [18]. We will work to clean the data, locate any missing values, isolate them, and then get them ready to be transferred to the second practical section (Model) as actual data. Inputs that are prepared for training, verification, and testing so that the prediction approach suggested in this study can provide the required outcomes. The data set is historical data for the two variables (Level and Evaporation) over a period of approximately 13 years (1993 - 2006), where the main variable in the dam is the water level variable, as well as one of the most important variables of the Mosul Dam (the evaporation variable), which directly affects the behavior of The dam by affecting the actual water storage in the dam basin. (Dataset was approved in coordination with the Dams and Water Resources Research Center in Mosul University). In this part we're going to calculate the values and extract the values that are set for each column in the data set, as shown in Table 1.

Table 1: Information of data

#	Colum	Non-Null	Count	Dtype
0	data	non-null	4784	Datetime64[ns]
1	Evaporation	non-null	4784	Object 1)

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

Here we will discover the lost and extreme values in the adopted data set and then clean up the data from the extremes and replace those lost values so that we can make the data free of impurities and properly sequenced as valid inputs to the model used in our methodology. Missing Values: We will note in Table 2 that the data collection used from two main columns (date and evaporation) shows that the first column (date) is very sound and does not contain any lost value, while the second column (evaporation) contains four missing values in its data.

Table 2: Missing statistics

Colum	Name	Missing Value	Total Rows	% Missing
0	data	0	4784	0.0000
1	Evaporation	4	4784	0.0800

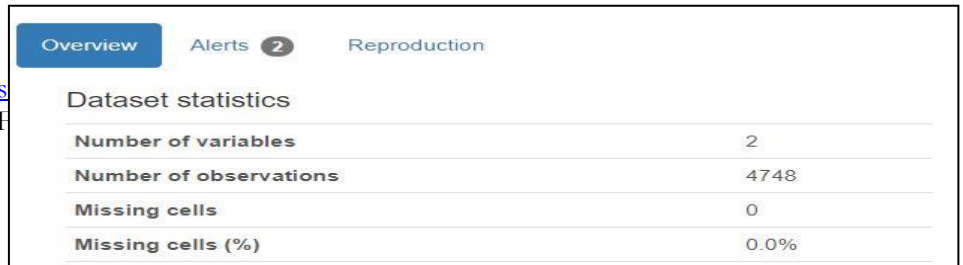
Undefined Values: We estimate the missing values in a data set based on known values. Interpolation is a statistical method for filling in the gaps between data points. When dealing with gaps in time series data, it is possible to utilize extrapolation. This is especially handy when working with nonsystematic time series data because the time intervals between consecutive data items may be arbitrary. After completing the necessary conFigure urations, we have the data in the format we need for our work. Next, we'll perform statistical analysis on the conFigure ured values to learn more about the Dada behaviour, and then we'll use that knowledge to refine our approach and finish the study. In this section, we're going to replace all the extreme values with 'nan' (not a number) value, which denote missing or extreme values that can't be calculated and don't interfere with the deep learning model's ability to process data. This will prevent errors and filter the values in the data set. In addition to correcting a few mistakes, some values. As shown in Figure. 1.

```
df = df.replace(['_____'],np.nan)
df = df.replace(['_'],np.nan)
df = df.replace(['–'],np.nan)
df = df.replace(['_____________'],np.nan)
df = df.replace(['___________'],np.nan)
df = df.replace(['_____'],np.nan)
df = df.replace(['____'],np.nan)
df = df.replace(['_________'],np.nan)
df = df.replace(['_____________'],np.nan)
df = df.replace(['___________'],np.nan)
df = df.replace(['________'],np.nan)
df = df.replace(['_____________'],np.nan)
df = df.replace(['_______'],np.nan)
df = df.replace(['_________________'],np.nan)
df = df.replace(['__________'],np.nan)
df = df.replace(['\\'],np.nan)
df = df.replace(['_____________ '],np.nan)
df = df.replace(['/'],np.nan)
df = df.replace(['Device malfunction'],np.nan)
df = df.replace(['N.A'],np.nan)
df = df.replace(['.0.29'],'0.29')
df = df.replace(['12.4.'],'12.4')
df = df.replace(['9.62.'],'9.62')
```

Figure 1. Replace undefined values with "np.nan" (screenshot)

3.2 Pandas Profiling Report

The Pandas will give us a detailed report on the analysis of the data and processing that we have done and will provide us with a set of features of the data collection used in this study. The Python Pandas Library is important in the area of statistics and data science. After processing the data and extracting the missing values, it was assumed that each of the values lost in the evaporation coefficient would be zero, which was missing.



DOI: https
Received: F

Figure 2. Overview (screenshot)

In Figure. 2. Statistical data set adopted. The Figure ure shows the statistical data required for the two variables studied (date and evaporation). The number of observations (4748) is also recorded, which represents the number of rows with actual values in the data set. This corresponds to the number of values that appeared in Table 7. Prior to the statistical procedure after the data were processed. We also note the absence of missing data and the appearance (0) after processing compared to Figure. 2.

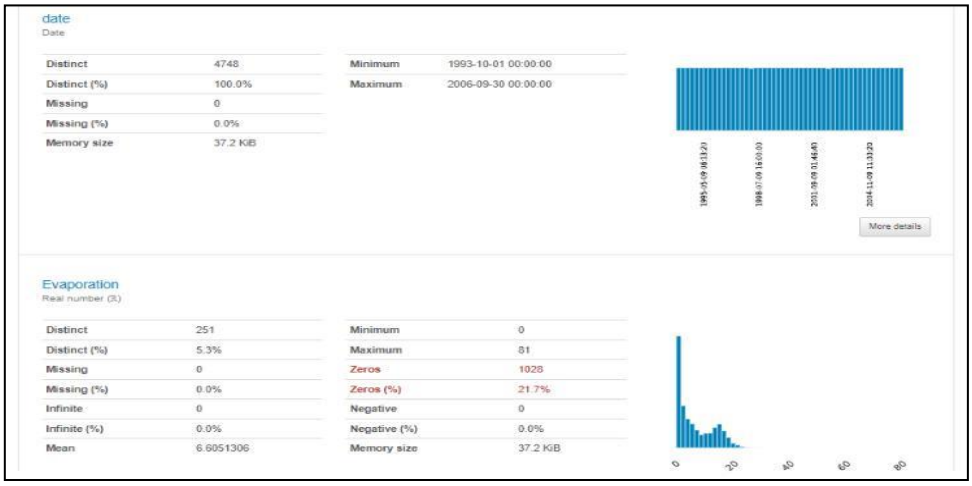


Figure 3. Variables (screenshot)

Figure. 3. Shows the information on the main variables (date, evaporation) and the number of values used for each variable. It notes that the data variable starts from 01.10.1993 to 30.09.2006 and that the evaporation variable values range from 0 to 20.

3.3 Interpolation of Missing Values

Since the missing values are naturally empty in the data set, we consider this missing value to be equal to 0, so that all attributes have value, and this will increase the performance of the model used because the lost and extreme values cause confusion in the model's performance, which has a negative impact on the accuracy of the model's output, as shown in Table 3.

Table 3: Interpolation of missing values

Colum	Name	Missing Value	Total Rows	% Missing
0	data	0	4784	0.0000
1	Evaporation	0	4784	0.0000

4. Statistics the Normal Distribution of Data

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

Upon obtaining the outcome of the requisite data format in our study and ensuring its appropriate structure, statistical analysis will be conducted on the prepared data values to ascertain the data's behavior. The background and pertinent information contained within this data will be examined to enhance the methods offered in this study. Initially, a series of statistical analyses were conducted on our computed data. Our data exhibits a unique characteristic, as we have determined that the sum of the values equals zero. This observation indicates that the data does not conform to a Gaussian distribution. The presence of a tail is evident, as depicted by the red line in Figure 4.

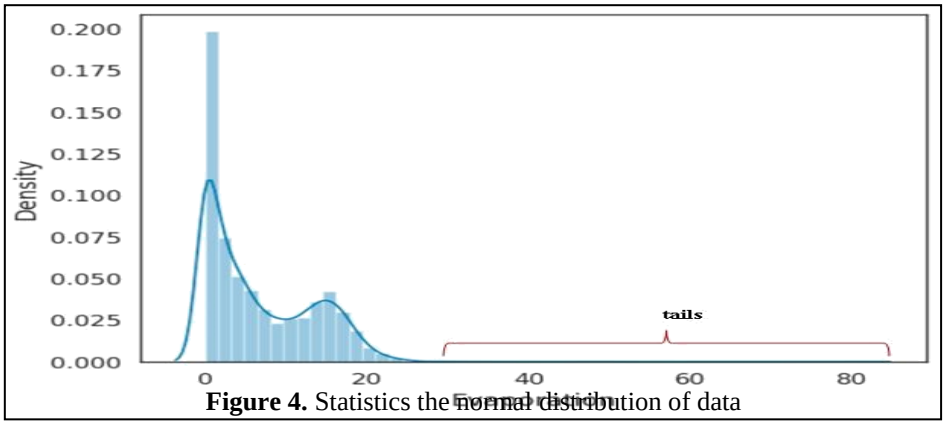


Figure 4. Statistics the normal distribution of data

The Figure. 4. illustrates the distribution of evaporation values, which predominantly fall within the range of 0 to 20 or somewhat higher. Additionally, there is evidence of a tail in the distribution, indicating the presence of noneffective terminal values. The present analysis will employ the Kurtosis of a normal distribution and the dispersion of a normal distribution to ascertain this phenomenon. The Kurtosis of a normal distribution measures the shape of its probability density function, explicitly quantifying the degree of peaked Ness or flatness relative to the standard bell curve.

Kurtosis of normal distribution: 2.909151499229055

Skewness of normal distribution: 1.015269801276841

The value indicated above is 2.9. Therefore, the statement above lacks sufficient strength. The observed distribution in this context is 1.01, representing the analogue distribution, specifically the aerobic graph, rather than the analogue distribution. A number within the range of (-0.5 and 0.5) can be classified as deviating from the norm. Figure 5 is a view of our data, made by the evaporation mass over time, and this is the presentation of the data form that is close to the noise view because the data changes from day to day as evidence of the distribution of evaporation values between 0 and 20 or slightly more appears in the drawing and for the time period (1993-2006). There are also yellow tails and values in red that match these zero values for the same period of data in the CSV file that we obtained with those data in coordination with the (*Centre for Research on Dams and Water Resources/University of Mosul/Iraq).*

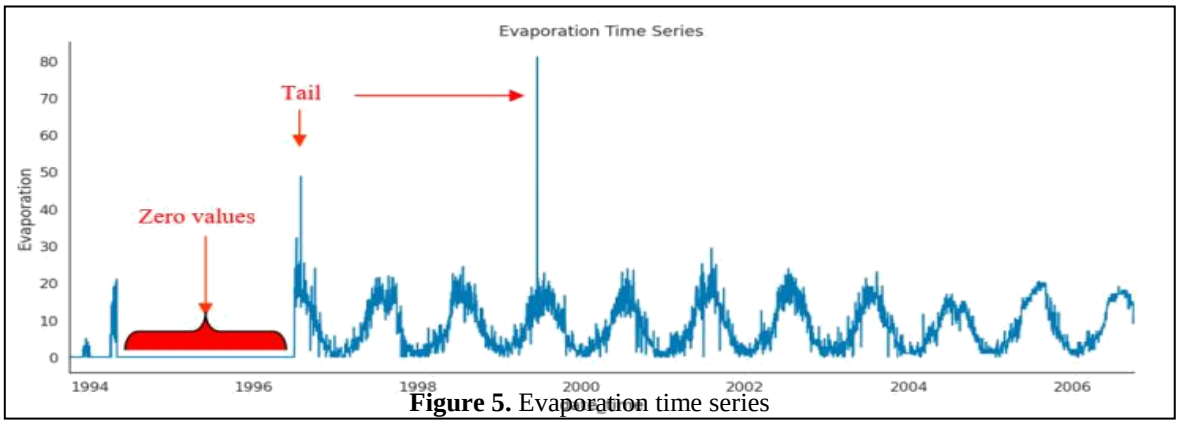


Figure 5. Evaporation time series

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

Figure 6. Represents the Conspiracy Fund for Evaporation variable and is divided into two parts (Yearly and quarterly) the box plot is for each year from the beginning of 1993 to 2006, which is the end of our data. Years and all values go from 245 to 256 with a few exceptions like tails and some simple values, and this was clear from the histogram above Figure 6.

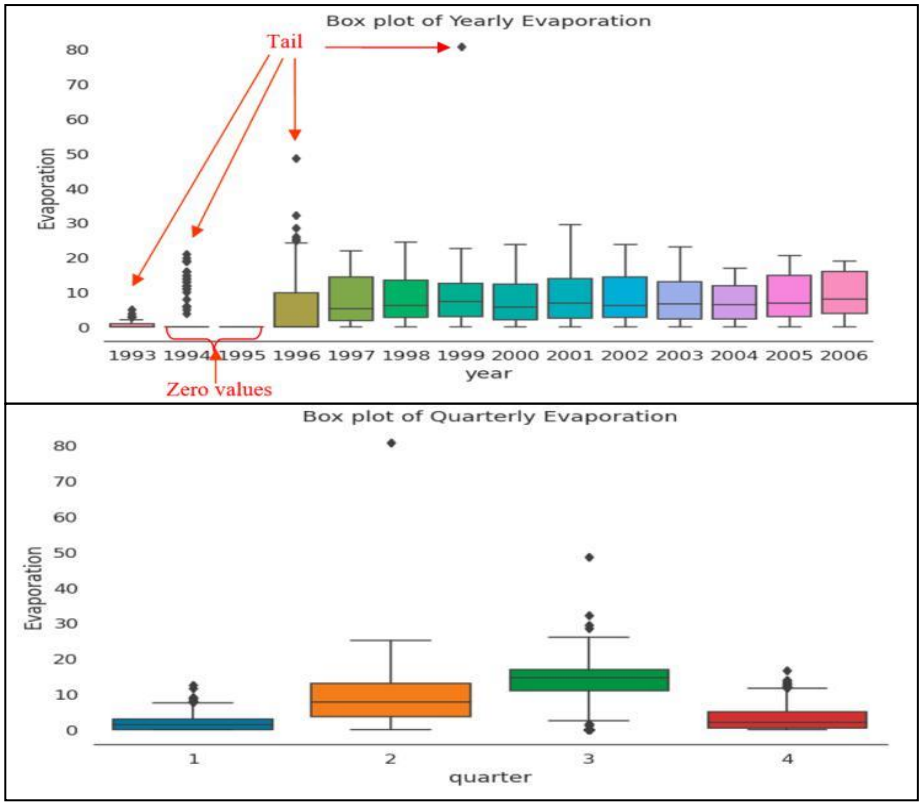


Figure 6. Evaporation Box Plot

The graph, denoted as Figure. 7. Illustrates that the values depicted therein fall within the range of 0 to 20, with a little tendency towards the latter. This range encompasses all values under consideration. The technique above is probabilistic. The red line represents the inherent distribution, whereas the blue line corresponds to our collected data. However, it is observed that the blue line does not accurately correspond to the red line in instances where the numbers are negative. Consequently, the red line does not represent the inherent distribution.

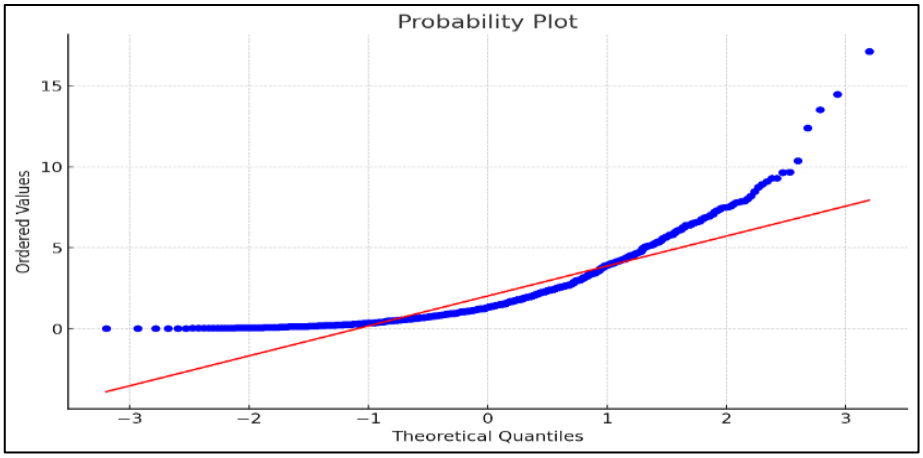


Figure 7. Probability plot

Start resampling the data during a different year and quarterly. To better visualize the data, we generated an average for each year and plotted it beside the point we drew for that year in Figure 8

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

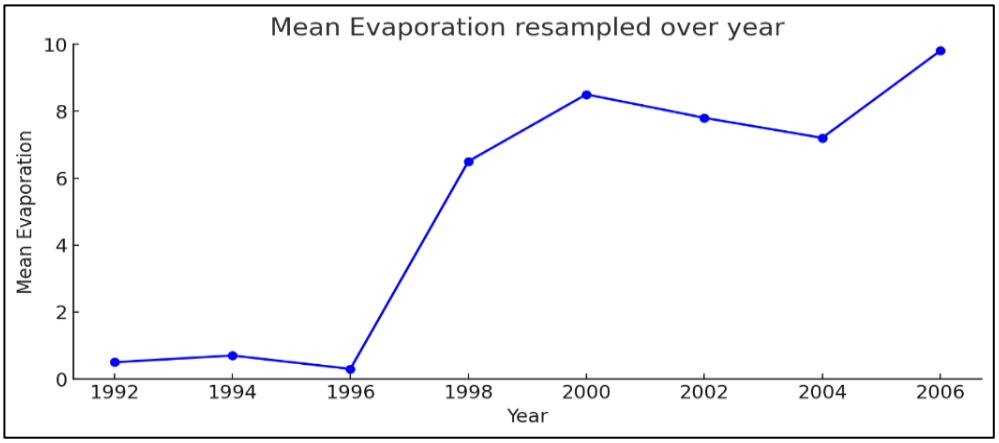


Figure 8. Mean evaporation re-sampling over year

Figure. 9. Illustrates that the sampling procedure remains consistent across all years. Shall comprehend the monthly or daily data as a vision and objects. Why are we proceeding in this manner? Because specific models operate on enormous datasets, the available RAM and computing power may prove insufficient to execute the mathematical and statistical operations on all this data. In such cases, it will be necessary to consolidate these models into a single unit, namely the automatic regression model. It will not utilize the entirety of this module; instead, we will apply deep learning techniques within that module. To fully comprehend the data behavior, which is the central focus of the present study, it is imperative to ascertain and comprehend the correlation between the variables that provide insight into that behavior. To this end, I determined and drew the spread for each column (variable) as yearly in the study's data table.

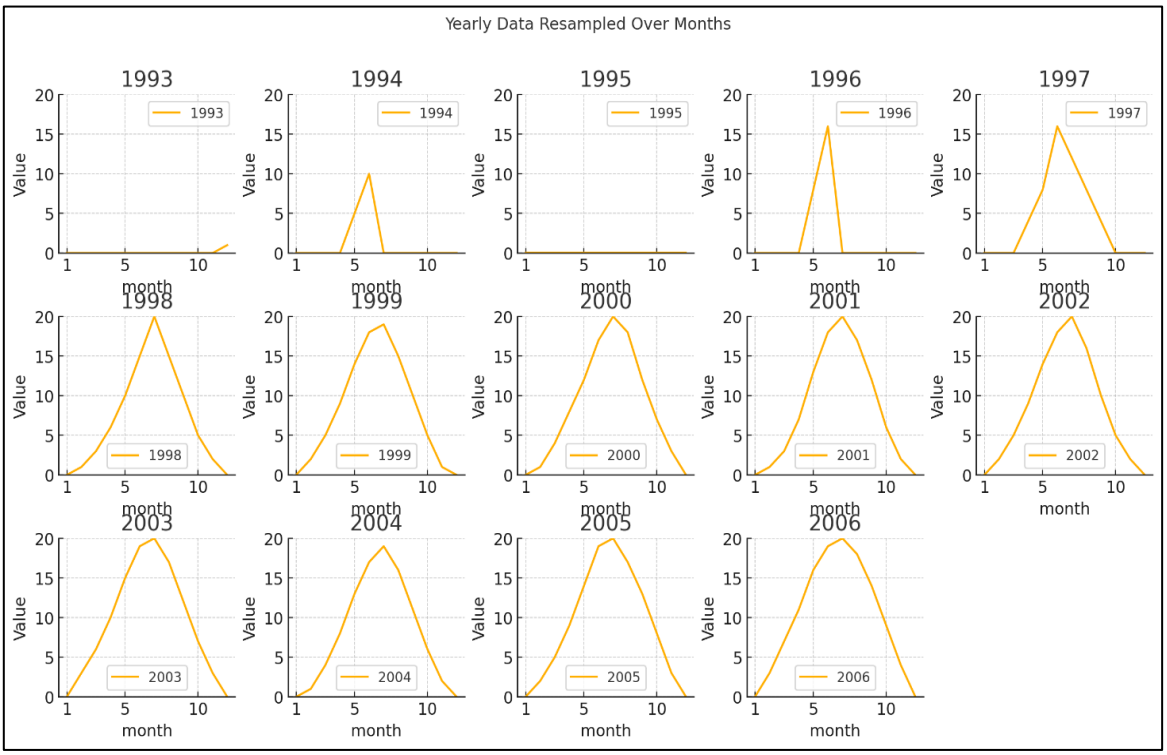


Figure 9. Evaporation resampling each year

Time series data can be analyzed using the main average and standard deviation. The average rotor is a form of moving average calculated by calculating the average values in a specific time window. The variability of results across time can be quantified by calculating the standard deviation. Finding trends and patterns in time series data could be made easier with the help of the main average and standard deviation. High or low data volatility could be identified with the help of a standard deviation. When analyzing time series data, it is common practice to use

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

the average deviation in conjunction with the two primary rotor characteristics. Daily closing rates for stocks over a year are displayed in Figure.10. You can see the original data in the chart's blue line. The red line represents the 12-year average loss of moisture. The standard deviation of evaporation is represented by the black line. During its 12 years in circulation, the average rate of evaporation has fluctuated between (20) and (0). Since no readings (zero data) were taken between the beginning of 1995 and the beginning of 1997, the standard deviation reveals that evaporation was more variable at the end of 1994 and 1999 and at the beginning of 1995 and 1997 than at any other time. This information could be used to produce projections of the future value of evaporation. When analyzing time series data, the main average and standard deviation are two of the most useful statistics. It is possible to get insight into the behavior of time series data and to forecast its future values by familiarizing oneself with the mechanics of these measurements.

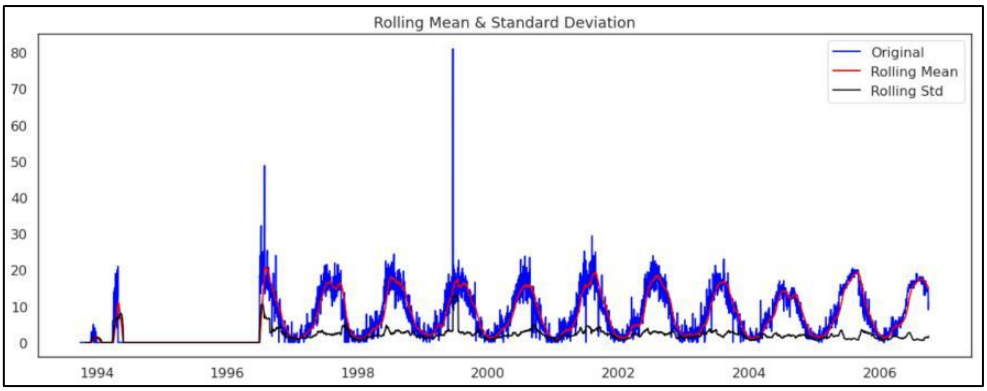


Figure 10. Rolling mean and standard deviation

The autoregressive integrated moving average (ARIMA) model has significantly altered how evaporation is predicted by utilizing time series data. This robust statistical analysis tool effectively utilizes historical evaporation data to accurately forecast future evaporation rates. Through the effective utilization of the seasonal ARIMA model, researchers have accurately modelled weekly evaporation data for long durations, such as the period from 1993 to 2006. This has allowed for precise predictions and insightful trend analysis [10]. It is widely used in fields such as environmental science, medicine, economics, etc. [19][20]. The present research incorporates the ARIMA model, a popular statistical model for time series forecasting and data analysis. After performing stepwise search to minimize and calculate a measure AIC, obtained the following:

$$model = pm.auto\_arima(final\_df['Evaporation']), \quad (1)$$

Through running Function 1, the result is the:

Best model: ARIMA(1,0,0)(2,0,1)[12] intercept

Total fit time: 31.019 seconds

a) The model is the ARIMA model with 1 automatically retrograde term, 0 inte grated terms, and 0 moving averages.
b) The seasonal component of the model contains 2 automatic terms, 0 integrated terms, and 1 medium term mover.
c) The model is installed using 12 seasonal cycles.
d) The term objection is included in the model.
e) The overall time for the model is 31.019 seconds.

$$final\_df.index.min(), final\_df.index.max(), \quad (2)$$

The range of the 'final' data framework indication is determined by Function 2.
The only thing that determines which row is which in the data framework is the indication. In this scenario, the 'final' data framework index is likely to be a date or time, given the data are time series data. In the final data structure, the aforementioned function can be utilized to establish the time range covered by the data, as shown by the following output:

(Timestamp('1993-10-31 00:00:00', freq='M'),

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

Timestamp('2006-09-30 00:00:00', freq='M'))

The outputs have shown the validity of the work of the analysis used in this study, showing the true length of time of the data set adopted in the present study, namely, the period of the data set used in the present study. (1993-2006). First, the data will be divided into three parts (training, Validation and Testing) and the division ratio for each part will be sequential (80,10,10). The output is:

Table 4: Multivariate datasets.

Date Shape	Divide
Train Data Shape	(3398, 1)
Val Data Shape	(850, 1)
Test Data Shape	(500, 1)

In Table 4 the shape of the data passed to the model for training is (3398, 1). This means that the data consists of 3398 samples, each of which is a one-dimensional array, and the shape of the data passed to the model for validation is (850, 1). This means that the data consists of 850 samples, each of which is a one-dimensional array, and the shape of the data passed to the model for testing is (500, 1). This means that the data consists of 500 samples, each of which is a one-dimensional array.

train=final-df[(final-df.index.get-level-values(0) >= '1993-10-31') and
(final- df.index.get-level-values(0) <= '2003-07-31')]

That Python function generates a new data frame for 'train' using the dates '1993-10-31' to '2003-07-31' from the original data frame 'final dist'. Can retrieve the date from the first level of the index by using this function method. and this function can sort the data. The produced data frame, 'train', is limited to the original data frame's entries from '1993-10-31' to '2003-07-31'. A model can subsequently be trained using this data frame. As shown in Table 5.

Table 5: Training dataset

Date	Evaporation
1993-10-31	0.0000
1993-11-30	0.0000
1993-12-31	1.6613
1994-01-31	0.0000
1994-02-28	0.0000

test=final-df[(final-df.index.get-level-values(0) > '2003-07-31')]

That Python function creates a new data frame called 'test' that contains only the data from the original data frame that is after the date '2003-07-31'. The 'index.get-level-values(0)' method is used to get the values of the first level of the index, which in this case is the date. The resulting data frame, 'test', contains only the data from the original data frame that is after the date '2003-07-31'. This data frame can then be used for testing a model. As shown Table 6.

Table 6: Testing dataset

Date	Evaporation
2003-08-31	15.4645
2003-09-30	11.5267
2003-10-31	7.6032
2003-11-30	3.2333
2003-12-31	1.4839

5. ARIMA Model Predictions

When analyzing and predicting time series data that display trends, seasonality, and/or non-stationarity, the ARIMA model is especially beneficial. Among the several seasonal ARIMA models, one stands out:

ARIMA(1,0,0)(2,0,1)[12].

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

Table 7: Predictive Data by ARIMA Model

Month	Date	Actual data	Predication	Difference
1	2003-08-31	10.7	14.47	3,77
2	2003-09-30	12	12.11	0.11
3	2003-10-31	4.6	7.60	3.00
4	2003-11-30	2	4.69	2.69
5	2003-12-31	1	3.11	2.11
6	2004-01-31	1.5	3.17	1.67
7	2004-02-29	2.3	3.11	0.81
8	2004-03-31	5.6	3.72	1.88
9	2004-04-30	6.1	5.37	0.73
10	2004-05-31	13	8.50	4.50
11	2004-06-30	15	11.34	3.66
12	2004-07-31	9	12.81	3.81
13	2004-08-31	14	11.59	2.41
14	2004-09-30	10	9.80	0.20
15	2004-10-31	6	6.61	0.61
16	2004-11-30	2	4.51	2.51
17	2004-12-31	1	3.47	2.47
18	2005-01-31	4.6	3.43	1.17
19	2005-02-28	3	3.50	0.50
20	2005-03-31	9	4.02	4.98
21	2005-04-30	8	5.46	2.54
22	2005-05-31	15	7.45	7.55
23	2005-06-30	17	9.66	7.34
24	2005-07-31	20	10.93	9.07
25	2005-08-31	20	10.09	9.91
26	2005-09-30	10	8.61	1 .39
27	2005-10-31	11	6.10	4 .90
28	2005-11-30	3	4.42	1 .42
29	2005-12-31	2	3.66	1.66
30	2006-01-31	2	3.57	1.57
31	2006-02-28	3	3.71	0.71
32	2006-03-31	6	4.18	1.82
33	2006-04-30	7	5.50	1.50
34	2006-05-31	10	6.91	3.09
35	2006-06-30	17	8.80	8.20
36	2006-07-31	17	9.96	7.05
37	2006-08-31	17	9.31	7.69
38	2006-09-30	11	7.99	3.01

That is, the model accounts for the fact that the time series is seasonal. The 12-month seasonality is indicated by the (12) in the model specification. The intercept element is unnecessary in seasonal ARIMA models since the model already accounts for the time series' seasonality. Since the time series' seasonality is already accounted for, the intercept term is unnecessary; it is employed to account for the overall level of the time series. Table 7. Shows the actual results of the predicting methodology adopted in this realistic study, as the future values of the evaporation variable were predicted for three years in advance and divided for 12 months each year. Since the last year (2006) in the data collection has only 9 months, this brings the total forecast to 33 months over 3 years (2004, 2005 and 2006).

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

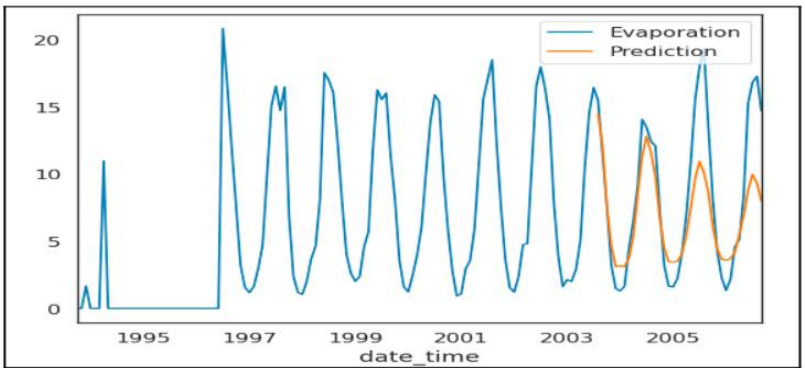


Figure 11. Prediction of evaporation values

Figure 11. Illustrates the extent to which evaporation forecast values correspond to actual evaporation coefficients. The x-axis is the date, the y-axis is the Evaporation value, The orange line shows the forecast, and the blue line shows the actual evaporation values in the Figure. While noting the compatibility of the two lines (actually and prediction), the congruence is scientifically good with a marked variation due to some zero values in the data set, indicating the actual performance of the ARIMA model applied in this study. It also notes that the forecasting methodology worked correctly for the period (2004-2006) and that the outcome of future values was very close to reality, as noted in Table 7. Table 8 shows the results of the prediction methodology adopted in this realistic study. The future values of the evaporation variable were predicted for 2006-2009, subsequent years after the last year 2006 (in months) in the dataset used in this study. It is divided into 12 months each year. Since the dataset's previous year, 2006, contains only 9 months, as in the prediction results, it contains only 3 months. So, the total forecast is 39 months over 4 years (2006, 2007, 2008, and 2009).

Table 8: Future predictive data

Month	Date	Predication
1	2006-10-30	5.8443
2	2006-11-30	4.3852
3	2006-12-30	3.7699
4	2007-01-30	3.6532
5	2007-02-28	3.8217
6	2007-03-30	4.2655
7	2007-04-30	5.526
8	2007-05-30	6.6293
9	2007-06-30	8.3475
10	2007-07-30	9.443
11	2007-08-30	8.9036
12	2007-09-30	7.6674
13	2007-10-30	5.7144
14	2007-11-30	4.3743
15	2007-12-31	3.8349
16	2008-01-30	3.7067
17	2008-02-28	3.8902
18	2008-03-30	4.3184
19	2008-04-30	5.5394
20	2008-05-30	6.4852
21	2008-06-30	8.1064
22	2008-07-30	9.1656
23	2008-08-30	8.6805
24	2008-09-30	7.4951
25	2008-10-31	5.6495
26	2008-11-30	4.3764
27	2008-12-31	3.8783

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

28	2009-01-31	3.745
29	2009-02-28	3.9349
30	2009-03-31	4.3529
31	2009-04-30	5.5474
32	2009-05-31	6.4088
33	2009-06-30	7.9723
34	2009-07-31	9.0075
35	2009-08-31	8.5524
36	2009-09-30	7.3991
37	2009-10-31	5.6175
38	2009-11-30	4.3852
39	2009-12-31	3.9104

6. Accuracy of Results

One common statistic to evaluate the accuracy of time series predicting models is MAE. A single numerical score indicating the average absolute error size between the predicted and actual values is calculated to evaluate the desired model performance. This is why it is a good idea to use the MAE metric to measure the accuracy of the ARIMA model. After applying the procedure, the following result was:

Test MAE: 2.459

It indicates that, in average absolute error, the model predictions are approximately 2.459 units away from the actual values. This is a relatively moderate and expected error, due to the presence of (1028) zero values out of (4748) actual values in the dataset adopted in this study, as well as the presence of 4 missing values from the evaporation values, as in Table 2. A popular statistic for assessing the precision of models for time series forecasting is (RMSE). It calculates a single numerical score that indicates the average magnitude of the mistakes between predicted and actual values to compare models. That's why it's good to use the RMSE scale to measure the accuracy of the ARIMA model. Following the application of the measure, the following result was:

Test RMSE: 1.487.

Depending on the characteristics of the data that may affect the nature of the time series data predicted in interpreting and analyzing the management system, and given the noise in the data collection used in this study despite the fact that they have been processed and that zero values have emerged in an influential way, the test value of monitoring and *RMSE (1.487)* is very good, however, as they are close to number 1 .

7. Conclusion

After the introduction of the water evaporation element in the Mosul dam as an important factor and its application in this study, and through a series of mathematical and statistical actions using many of the relevant python functions, we explored and analyzed the data contained in this study. The ARIMA model was one of the best operating models, with time series data reflecting seasonal and temporal accreditation processes. The missing values were then identified and isolated, and the data interface and interaction between them were developed; high quality data with actual values were obtained that demonstrated their realistic course and measured the accuracy of the model performed by the integrated system standard and the evaluation system and the efficient performance of future forecasting. The study contributed to products designed to provide sound and reliable inputs to innovative learning models, in particular forecasting models. This allows data decision makers to adopt such data in decision making processes in risk and crisis management. This is necessary to address the state of the Mosul dam, the problem of evaporation, drought, and Iraq low water share. The study is therefore a scientific starting point for future researchers to support the development of their scientific research in data analysis and forecasting.

References

[1] B. Yadav and S. Mathur, "River discharge simulation using variable parameter mccarthy-muskingum and wavelet-support vector machine methods," Neural Computing and Applications, vol. 32, pp. 2457–2470, 2020.

[2] E. Moges, Y. Demissie, L. Larsen, and F. Yassin, "Review: Sources of hydrological model uncertainties and advances in their analysis," Water, vol. 13, no. 1, p. 28, 2020.

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024

[3] V. K. Undavalli and B. Khandelwal, "Impact of alternative fuels and fuel properties on PM emissions," in Aviation Fuels, B. Khandelwal, Ed. Academic Press, 2021, pp. 71–111. [Online]. Available: https://doi.org/10.1016/B978-0-12-818314-4.00012-1.

[4] S. S. Akay, O. Ozcan, and F. B. Sanli, "Quantification and visualization of flood-induced morphological changes in meander structures by UAV-based monitoring," Engineering Science and Technology, an International Journal, vol. 27, p. 101016, 2022.

[5] M. Anbarasan, B. Muthu, C. Sivaparthipan, R. Sundarasekar, S. Kadry, S. Krishnamoorthy, A. A. Dasel, et al., "Detection of flood disaster system based on IoT, big data and convolutional deep neural network," Computer Communications, vol. 150, pp. 150–157, 2020.

[6] X. Lei, W. Chen, M. Panahi, F. Falah, O. Rahmati, E. Uuemaa, Z. Kalantari, C. S. S. Ferreira, F. Rezaie, J. P. Tiefenbacher, et al., "Urban flood modeling using deep-learning approaches in Seoul, South Korea," Journal of Hydrology, vol. 601, p. 126684, 2021.

[7] P. A. Garsole, S. Bokil, V. Kumar, A. Pandey, and N. S. Topare, "A review of artificial intelligence methods for predicting gravity dam seepage, challenges and way-out," AQUA - Water Infrastructure, Ecosystems and Society, vol. 72, no. 7, pp. 1228–1248, 2023. [Online]. Available: https://doi.org/10.2166/aqua.2023.042.

[8] M. F. Allawi, F. B. Othman, H. A. Afan, A. N. Ahmed, M. S. Hossain, C. M. Fai, and A. El-Shafie, "Reservoir evaporation prediction modeling based on artificial intelligence methods," Water, vol. 11, no. 6, 2019. [Online]. Available: https://doi.org/10.3390/w11061226.

[9] A. Geetha and G. Nasira, "Time-series modelling and forecasting: Modelling of rainfall prediction using ARIMA model," International Journal of Society Systems Science, vol. 8, no. 4, pp. 361–372, 2016.

[10] A. Hayes, "Autoregressive integrated moving average (ARIMA) prediction model," Investopedia, 2022. [Online]. Available: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp. [Accessed: Apr. 30, 2023].

[11] T.-T.-H. Phan and X. H. Nguyen, "Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red River," Advances in Water Resources, vol. 142, p. 103656, 2020. [Online]. Available: https://doi.org/10.1016/j.advwatres.2020.103656.

[12] M.-C. Wu, G.-F. Lin, and H.-Y. Lin, "The effect of data quality on model performance with application to daily evaporation estimation," Stochastic Environmental Research and Risk Assessment, vol. 27, 2013. [Online]. Available: https://doi.org/10.1007/s00477-013-0703-4.

[13] J. Frost, "Root mean square error (RMSE)," Statistics by Jim, 2024. [Online]. Available: https://statisticsbyjim.com/regression/root-mean-square-error-rmse/. [Accessed: May 24, 2024].

[14] Z. Kayhomayoon, F. Naghizadeh, M. Malekpoor, N. Arya Azar, J. Ball, and S. Ghordoyee Milan, "Prediction of evaporation from dam reservoirs under climate change using soft computing techniques," Environmental Science and Pollution Research, vol. 30, pp. 1–24, 2022. [Online]. Available: https://doi.org/10.1007/s11356-022-23899-5.

[15] H. S. Munawar, A. W. Hammad, S. T. Waller, M. J. Thaheem, and A. Shrestha, "An integrated approach for post-disaster flood management via the use of cutting-edge technologies and UAVs: A review," Sustainability, vol. 13, no. 14, p. 7925, 2021.

[16] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," Econometric Reviews, vol. 29, no. 5-6, pp. 594–621, 2010.

[17] B. Aljoumani, J. A. Sanchez-Espigares, B. Kluge, G. Wessolek, and B. Kleinschmit, "Analyzing temporal trends of urban evaporation using generalized additive models," Land, vol. 11, no. 4, p. 508, 2022.

[18] H. Bazzi, H. Ebrahimi, and B. Aminnejad, "A comprehensive statistical analysis of evaporation rates under climate change in southern Iran using WEAP (case study: Chahnimeh reservoirs of Sistan plain)," Ain Shams Engineering Journal, vol. 12, no. 2, pp. 1339–1352, 2021.

[19] M. Farsi, D. Hosahalli, B. Manjunatha, I. Gad, E.-S. Atlam, A. Ahmed, G. Elmarhomy, M. Elmarhoumy, and O. A. Ghoneim, "Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data," Alexandria Engineering Journal, vol. 60, no. 1, pp. 1299–1316, 2021.

[20] K.-R. Kim, J.-E. Park, and I.-T. Jang, "Outpatient forecasting model in spine hospital using ARIMA and SARIMA methods," Journal of Hospital Management and Health Policy, vol. 4, 2020.

DOI: https://doi.org/10.54216/FPA.170225
Received: February 15, 2024 Revised: May 17, 2024 Accepted: October 14, 2024