Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Using Hybrid Regression Tree and ARIMA Model with Wavelet Transforms for Evaporation Time Series Forecasting

Zinah Mudher ALbazzaz[1] and Naam Salem Fadhil[2]

[1,2]Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Article information

Article history:
Received June 14,2024
Accepted July 25 ,2024
Available online 1 December, 2024

Keywords:
ARIMA model
Regression Tree
Wavelet Transforms
Forecasting
Evaporation

Correspondence:
Naam Salem Fadhil
naamsalem@uomosul.edu.iq

Abstract

Forecasting accuracy of evaporation time series is an importance to control environmental impacts, damages, and risks affecting especially plant life and growth, and thus that impact on human and animal life. Evaporation data are considered from climate time series, which are characterized by its nature a non-linearity data, as they suffer from the problem of heterogeneity because they contain many seasonal and periodic components, and necessarily that complexity may lead to inaccurate forecasts. The time stratified method will be used in this study with the proposed forecasting methods to achieve greater homogeneity and less complex temporal behavior. Two forecasting methods will be used, represented by the regression tree (RT) method and the integrated autoregressive and moving average (ARIMA)model, and it is proposed to hybridize them with a method that combines both within the hybrid ARIMA-RT model as a way to improve forecasting results by dealing more accurately with the non-linearity data. The effect of wavelet transformations (WT) will also be tested with both the ARIMA model and the hybrid ARIMA-RT model, and whether it will have a role in improving forecasting results. A time series modeling structure will be adopted to determine the input structure of the RT model within the proposed hybrid approach by using multiplicative seasonal ARIMA. Also, the use of WT will be limited to filtering a random errors series (residuals), which the rest of its time lags depended on, represented by the moving average variables process. The forecasting results of the proposed methods might comparisons with the traditional forecasting method. This study was concerned with investigating various methods for forecasting evaporation time series for an agricultural meteorological station in the city of Mosul, Iraq for hot and cold seasons. The results of this study reflected the superiority of the hybrid method compared to the traditional ARIMA model. The results also included that forecasts were clearly affected by the use of WT. it can be concluded that the ARIMA-RT hybrid model has a clear role in improving the accuracy of forecast results through this study. Using WT leads to a slight improvement in the accuracy of forecasts, and it may vary according to the data and its nature and homogeneity.

DOI: 10.33899/iqjoss.2024.185246

Introduction

The diversity of patterns and components within the behavior of climate time series data in general and evaporation data in particular suffers from the problem of non-linearity as is most climate and environmental data because of its complexity in terms of seasonal and periodic fluctuations that affect the homogeneity of the data, which will lead to additional complications and obstacles in analyzing as time series and forecasting it, and this may negatively affect the

accuracy of the forecast results, as the results may be unsound. The accuracy of forecast results for evaporation may depend on the selection of the appropriate method used for analysis and forecasting mainly.

The problem of data heterogeneity due to the effects of seasonal and periodic patterns, the evaporation time series data will be divided into two seasons; the first is the hot season, which includes data for the months (May-September), and the cold season, which includes data for the months (November-March). The autoregressive and integrated moving average (ARIMA) model is one of the most famous time series models that is commonly used for univariate time series analysis as a traditional statistical model for forecasting. The ARIMA model is a linear model that may not be suitable for dealing with data that suffer from non-linearity problems. However, the ARIMA model is often used with climate data as with climate variables, including variable evaporation time series (1, 2). Two forecast methods, represented by the regression tree (RT) and the ARIMA model will be proposed and then hybridized by reconciling them together within the hybrid ARIMA-RT model as a developed method used to improve forecasting results by dealing well with the problem of non-linearity data. Wavelet transformations (WT) will be used also with both the ARIMA model and the hybrid ARIMA-RT model to test the role of their effect in improving forecasting results in two developed models, namely ARIMA-Wavelet and RT-Wavelet, based on the ARIMA model.

In previous studies, many researchers used the ARIMA model in the presence of the WT effect with many time series data, whether climate or other, as in (4 ,3). In light of the presence of the WT effect, the RT model was used, and it was also used after reconciling it in the hybrid ARIMA-RT model to forecast many series data. The results of this study reflected the superiority of the hybrid method compared to the traditional ARIMA model. The results included that forecasts were unaffected by the use of WT clearly. There was a clear improvement in the forecasting accuracy of the ARIMA-Wavelet model compared to the traditional ARIMA model. In the case of comparing the two hybrid models, ARIMA-RT and RT-Wavelet, there was no clear effect and role for using WT. This is because the basic development and improvement was accomplished after proposing RT after reconciling it with ARIMA to deal well with the problem of nonlinearity. From this study, it can be concluded that the ARIMA-RT hybrid model has a clear role in improving the accuracy of forecast results. As for using WT, it leads to an improvement in the accuracy of forecasts in the absence of introducing the effect of RT and reconciling it with traditional models, and this effect may vary depending on the data, its nature, and its homogeneity.

Materials and methods

The framework for this study will include:

a. Data initialization.

b. Modeling using ARIMA models.

c. Using the WT method with the residuals series resulting from the ARIMA model and forecasting using the ARIMA-Wavelet hybrid method.

d. Modeling using the RT method based on the ARIMA model structure, referred to as the hybrid ARIMA-RT model.

e. Modeling RT method based on hybrid ARIMA-Wavelet method structure and referred to as the hybrid RT-Wavelet model.

f. Comparison of forecast results using ARIMA, ARIMA-Wavelet, ARIMA-RT, and RT-Wavelet methods. This framework can be illustrated as in Figure 1 below.

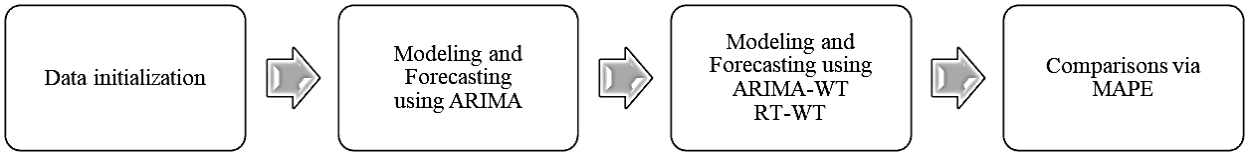


Figure 1: The general framework of the study

Autoregressive integrated moving average (ARIMA) model

The reason of unstable the time series is the variance or the series average, or both, change with time (unstable). Then some manipulations can be performed on the time series for the purpose of achieving stability in it. If the variance varies by time, then we can perform some transformation formulas on the data to achieve stability in it. This type of transformations is called power transformations. The condition of stability in the time series can also be achieved through the method of differences for unstable series, which is called (first-order differences), as it represents the first difference between two consecutive observations values ($W_t = Z_t - Z_{t-1}$), and when stability is not achieved, one resorts to differences from Higher rank or seasonal rank. The ARIMA model is one of the traditional models for univariate time series forecasting. The Box-Jenkins method will be adopted in its four stages: identification, parameter estimation, diagnostic examination, and forecasting. As an optimal method for analyzing time series, the multiplicative seasonal ARIMA model [ARIMA(p,d,q)(P,D,Q)s] was used, which takes the following general mathematical formula.

$$\phi(B)\Phi(B)(1-B^S)^D(1-B)^d Z_t = \theta(B)\Theta(B)a_t \quad (1)$$

$$\phi(B)\Phi(B)W_t = \theta(B)\Theta(B)a_t \quad (2)$$

$$W_t = (1-B^S)^D(1-B)^d Z_t$$

$$\phi(B) = (1-\phi_1 B - \phi_2 B^2 - \mathrm{L} - \phi_p B^p)$$

$$\Phi(B) = (1-\Phi_1 B^s - \Phi_2 B^{2s} - \mathrm{L} - \Phi_P B^{Ps})$$

$$\theta(B) = (1-\theta_1 B - \theta_2 B^2 - \mathrm{L} - \theta_q B^q)$$

$$\Theta(B) = (1-\Theta_1 B^s - \Theta_2 B^{2s} - \mathrm{L} - \Theta_Q B^{Qs})$$

ϕ_k and θ_k are the non-seasonal parameters of the auto regression and the moving averages at step k respectively, and reflect the effect of changing the time series variable (Z_{t-k}) and the random variable (a_{t-k}) at step k respectively, $a_t = y_t - \hat{y}_t$, Θ_k and Φ_k are the seasonal parameters of the auto regression and the moving averages respectively, ***S*** represents the seasonal period, p indicates the rank of the autoregressive model, d represents the number of differences necessary to achieve stability, q symbolizes the rank of the moving averages, P indicates the rank of the seasonal autoregressive model, Q symbolizes the rank of the seasonal moving averages, D represents the number of seasonal differences necessary to achieve stability, B is the backshift operator, a_t is the white noise with zero arithmetic mean and variance σ_a^2, and we can write it as follows $a_t \sim i.i.d.N(0, \sigma_a^2)$ (6 ,5).

It is necessary, when analyzing a time series, to draw the time series graphically in order to identify many of its features, especially determining whether the time series is stable or unstable, in addition to other features. The autocorrelation and partial autocorrelation functions are useful tools for demonstrating the stability of a time series. The stability condition of an ARIMA model require that that the mean and variance are constants (independent of time t) . In many cases, time series are unstable and the reason for this is due to a change in the mean over time, for example it has a general trend, or due to a change in the variance of the series over time. If the time series is unstable, we can achieve weak stability in it, or sometimes we call it second-order stability (8 ,7).

Beginning to identify the features of the series and determine the ranks of the polynomials in the (ARIMA) model and the number of parameters (p, q, P, Q) after achieving the stability of the series. The table below shows a simplified methodology for determining the ranks of the polynomials in (ARIMA) and the number of parameters in the model through the (ACF) (PACF) functions. This is a summary of the recognition phase of the Box and Jenkins methodology. After completing the first stage or step, which is identifying the hypothetical (ARIMA) model using the Box-Jenkins method, the second step is estimating the model parameters through the model's potential function method. Such estimates are referred to as maximum potential estimates.

From the statistical aspect, the non-significance of the estimated parameters is one of the reasons that impair the accuracy of the model. Therefore, the null hypothesis will be tested, which states that the parameter estimates are not significantly different from zero, that is, equal to zero, as the critical value of the (t) test is the tabular value multiplied by the estimated standard error of the parameter. If the absolute value of the value calculated for the t-test for each estimator is at least equal to the critical value, then the null hypothesis, meaning that the estimator is significant, will be rejected. The significance of all model parameters is one of the most important conditions that must be achieved as one of the most important diagnostic examination procedures. Also, the autocorrelation state of the residual's series can be used to test whether the residuals series is identical to the white noise process $a_t \sim \mathrm{i.i.d}\ N(0, {\sigma_a}^2)$. Therefore, the autocorrelation function of the residuals must not contain significant correlation coefficients as an enhancement measure to pass the diagnostic tests. There are other tests to diagnose the quality of the model, but the significance of the parameters and the ACF status of the residuals are among the most prominent and most important for the model to be qualified to be used to forecast future values with good performance (9, 10).

From the statistical aspect, the non-significance of the estimated parameters is one of the reasons that impair the accuracy of the model. Therefore, the null hypothesis will be tested, which states that the parameter estimates are not significantly different from zero, that is, equal to zero, as the critical value of the (t) test is the tabular value multiplied by the estimated standard error of the parameter. If the absolute value of the value calculated for the t-test for each estimator is at least equal to the critical value, then the null hypothesis, meaning that the estimator is significant, will be rejected. The significance of all model parameters is one of the most important conditions that must be achieved as one of the most important diagnostic examination procedures. Also, the SACF of residuals series can be used to test whether the residuals series is identical to the white noise process $[a_t \sim \mathrm{i.i.d}\ N(0, {\sigma_a}^2)$. Therefore, the autocorrelation function of the residuals must not contain significant correlation coefficients as an enhancement measure to pass the diagnostic

examination. There are other tests to diagnose the quality of the model, but the significance of the parameters and the ACF status of the residuals are the most prominent and most important for the model to be qualified to be used to forecast future values with good performance.

Regression tree (RT) model

Regression trees are a case of decision trees. Decision trees are defined as a graph that shows the actions that can be taken from natural states and their probabilities, and they are one of the forecasting models used in statistics, data mining, and machine learning. Decision trees are also important in analyzing decision issues that contain a series of decisions or a series of sequentially occurring states of nature.

Decision trees are created by dividing the original data set to form the root node of the tree and then dividing it into subgroups until the leaves represent the final decisions. Regression trees differ from classification trees which are the second type of trees, whose are forecasting results are categorical, just as the dependent variable in them is originally of the categorical type, while regression trees differ in terms of their outputs and forecasts, which are real values, because the dependent variable in them is a continuous variable. This method is based on the consideration that all elements of the studied society form one group and are then divided into two or more groups, subdivided, and so on. The process of branching and decision-making takes place through breaks represented as logical sentences and mathematical conditions placed on the joints of the tree in the form of nodes that branch into two or more branches, leading to the leaves as purposeful endings that represent the final decisions when the stopping order determined by the researcher is achieved (11).

The design of the classification and regression tree goes through several stages: (13 ,12)

1. The construction or building stage. This stage consists of several steps, which are:

a. Determine the dependent variable y .

b. Choosing explanatory variables $x_1, x_2, x_3, \mathrm{L}$.

c. Choosing the origin node so that it is suitable for the research goal.

d. Determining the rules of branching (fission).

2. The process of division or fragmentation.
3. Stopping process.
4. Pruning process.
5. Assembly process.
6. Draw the tree.

On each node t , there are two subjects. The first one that expresses the left node resulting from the branching at the point t , and the subject $r(t)$ expresses the right node resulting from the branching at the point t . The following figure 2 shows the regression tree.

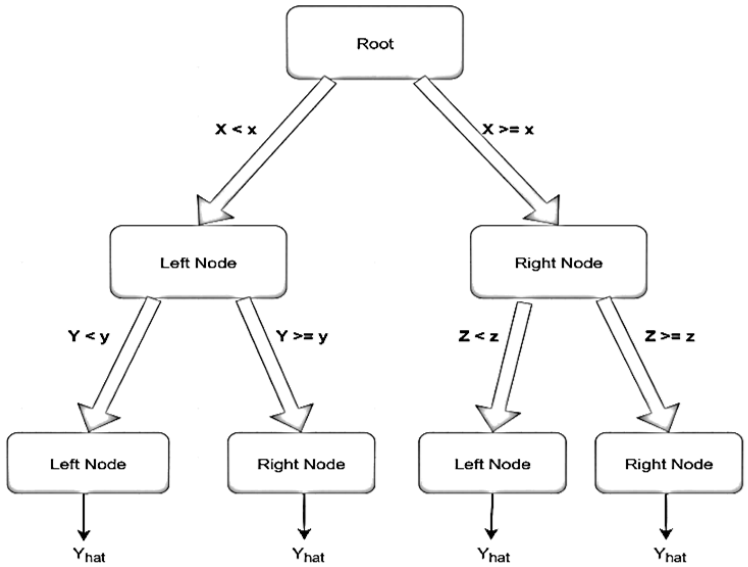


Figure 2: General structure of the regression tree

The regression tree is designed using the characteristics of the data sets taken from the elements of sample size n, which include the observations values of the variables x $x_1, x_2, x_3, ...$ and the observation values of the corresponding groups in the dependent variable y.

$N(t)$ is the number of sample elements that belong to the region $R(t)$ corresponding to the node t, which is distributed among a number of groups that make up the dependent y according to their properties. $N_j(t)$ represents the number of sample elements that belong to the region $R(t)$ and to the group G_j, as the following relationship is satisfied.

$$\sum_{j=1}^{g} N_j(t) = N(t) \quad (3)$$

The probabilities of belonging and distribution to groups can be calculated as follows (14).

The probability that any element i of the samples belongs to the region $R(t)$ corresponding to the node t can be calculated as follows.

$$P(i \in R(t)) = \frac{N(t)}{n} \quad (4)$$

The probability that any element i of the samples belonging to the region $R(t)$ corresponding to the node t belongs to the group G_j can be calculated as follows.

$$P(G_j / i \in R(t)) = \frac{N_j(t)}{N(t)} \quad (5)$$

Also, the probability that the sample elements corresponding to the node t are distributed as a node to the left and a node to the right can be calculated, respectively, as follows.

$$P[l(t)] = \frac{N[l(t)]}{n} \quad (6)$$

$$P[r(t)] = \frac{N[r(t)]}{n} \quad (7)$$

The probability that any element of the sample corresponding to the node t will go to the left node or the right node can be calculated, respectively, as follows.

$$P_l = \frac{P[l(t)]}{P(t)} = \frac{N[l(t)]}{N(t)} \quad (8)$$

$$P_r = \frac{P[r(t)]}{P(t)} = \frac{N[r(t)]}{N(t)} \quad (9)$$

Thus, the features of each group corresponding to the node t can be determined according to what is proportional to it in terms of the number of sample elements to which it belongs from the region $R(t)$, according to their probabilities. The probabilities corresponding to the groups are compared and the group that corresponds to the greatest probabilities is chosen (15).

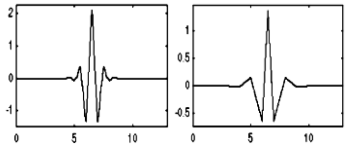
$$P(G_K / t) = \max P(G_j / t) \quad (10)$$

Wavelet Transform (WT) (17 ,16)

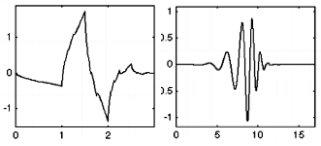
Wavelet Transform or Wavelet Analysis is a mathematical analytical method used to process signals for many practical applications. The basis of this theory is based on Fourier's theory, which is a method used to represent periodic signals in the form of a sine and cosine series. This conversion transfers the signal from the time domain to the frequency domain and vice versa. The wavelet transform is considered a complement to spectral analysis, which depends on the conversion from the time domain to the frequency domain based on the rules of complex trigonometric or exponential functions that lead to the detection of periodicity in the wavelet movement by relying on the conversion to the field of measurement and displacement using the rules of orthogonal functions that lead to detection. About the regularity or irregularity of wave movement.

This short-time Fourier transform represents the signal using a specific window depending on its temporal and frequency resolution. The main problem in this conversion is the loss in time and frequency. High accuracy is obtained for signals that change quickly when using a small window, but this accuracy is not high for signals that change slowly and when using a large window, the exact opposite happens. For this reason, the theory of wavelet transformation is developed, so the wavelet transform is divided into three types: continuous wavelet transform, discrete wavelet

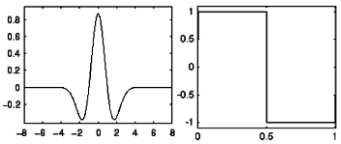
transform, and wavelet packet transform. Perhaps the most famous types of wavelet functions can be represented graphically as follows:



Bi-orthogonal functions bior 2.6



Daubchies functions db2 and db9



Haar function and Mexican hat function

Figure 3: Some types of wavelet transform functions

Wavelets emerged an interest statistical in 1989 when Mallat introduced a multiple analysis method for processing discrete data (18). And (19) also showed that the contracted wavelet has ideal statistical properties that are desirable in problems associated with noise reduction. There are several useful applications for wavelets, such as data compression, image feature detection, and noise reduction from time series.

The measurements of forecasting Error (20)

One of the most important criteria used to express the accuracy of forecast data is as follows:

The mean absolute percentage errors is:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|\times 100 \qquad (11)$$

(12)

n represents the number of observations, y_t is the variable of the original time series at time t, while $\hat{y}_t$ represents the forecast variable, and e_t represents the residuals series at time t.

Results and conclusions

Data used in this study

The ARIMA model will be studied using the Box-Jenkins method methodology and the RT method with the wavelet transform effect to analyze evaporation time series data in the city of Mosul/Iraq after separating it into hot and cold seasons. Data were taken from the Agricultural Meteorological Observatory at the location specified by longitude E = 43.16 and latitude N = 36.33. Because of the data it contains, it is due to the heterogeneity between the different seasons and the diversity of the seasonal seasons and their fluctuations. To achieve greater consistency, the data was separated into two seasons, the first being the cold months (November-December-January-February-March). The second season is hot for months (June - July - August - September). The study data included 372 hot season observations and 303 cold season observations as daily data. About 20% of the end-of-season time series data will be allocated as test period data and used as hypothetical future data used to test models built on the training period data. Therefore, the training data for the hot season will include (292 observations), while it will be (80 observations) at the end of the hot season as a testing period. As for the training period for the cold season, it will contain (243 observations) and (60 observations) for the cold time series as data for the test period.

Steps for the general framework of the practical aspect

1. Alignment of evaporation time series data for homogeneity with data set for the training and testing periods for each season.
2. Building an ARIMA model based on the Box-Jenkins methodology.
3. Taking the right side of the ARIMA models as input for the RT method and setting it as an input matrix whose number of columns is the number of terms of the right side and the number of rows represents the number of observations of the training period. The target variable y_t is the one on the left side of the ARIMA model.
4. Using the WT method to analyze the components of the residuals series resulting from the ARIMA model and obtain forecasts from the ARIMA-Wavelet hybrid method for the training and testing periods for the two seasons.
5. Modeling the training data for each season using the RT method, relying on the structure of the ARIMA models for each season and obtaining forecasts from the hybrid ARIMA-RT model for the training and testing periods.
6. Modeling the training data for each season using the RT method, relying on the structure and results of the hybrid ARIMA-Wavelet method for each season, and obtaining forecasts from the hybrid RT-Wavelet model for the training and testing periods.
7. Comparing the forecast results for the training and testing periods and for all seasons for the evaporation time series using the ARIMA, ARIMA-Wavelet, ARIMA-RT, and RT-Wavelet methods through the forecast error criterion.

ARIMA Model

The Box-Jenkins methodology is applied to analyze time series and obtain the best models that express the behavior of the data. The first step is to identify and diagnose the state of stability and the necessity of achieving it, while determining the ranks of the model. To reveal the stability of the data, the series will be plotted with the ACF and PACF functions. Figures 4 and 5 below show the time series plotting of evaporation and the ACF and PACF functions for the hot and cold seasons.

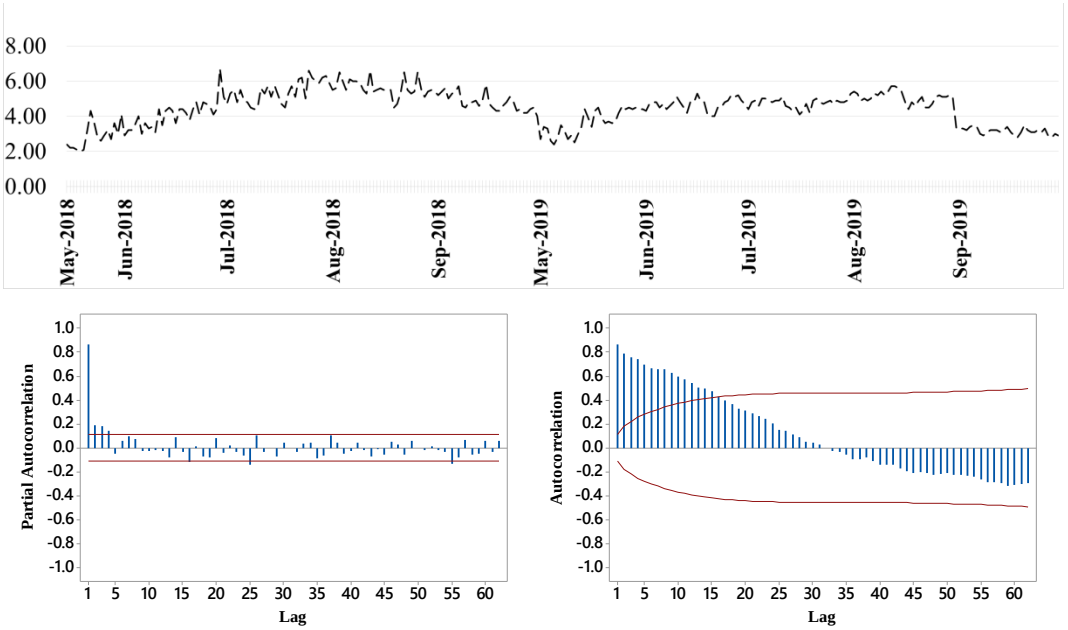


Figure 4. Time series plotting of evaporation, ACF and PACF for the hot season

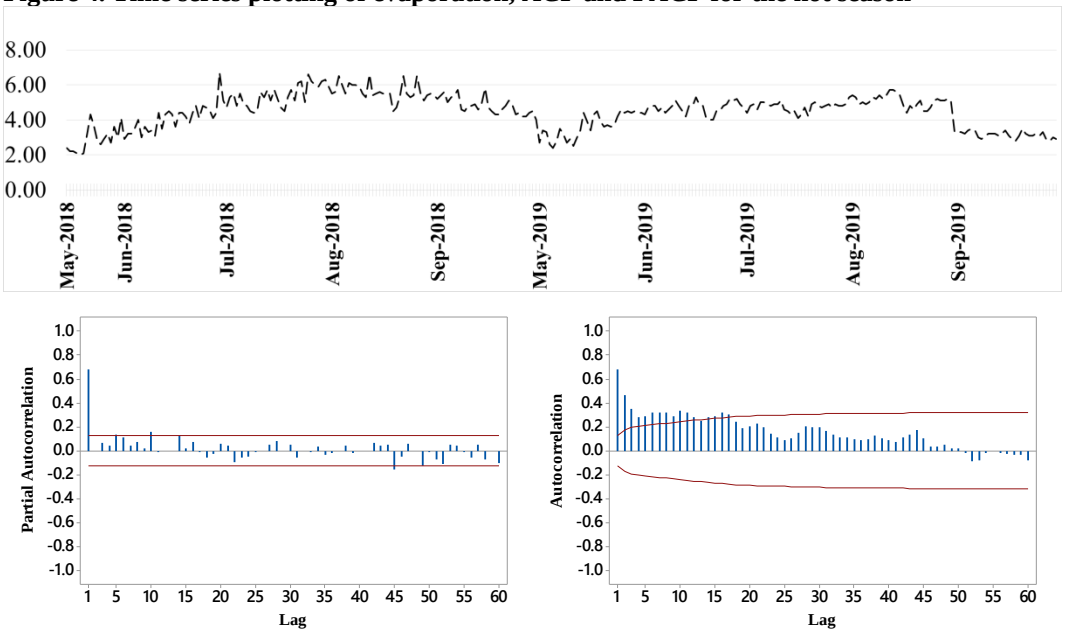


Figure 5. Time series plotting of evaporation, ACF, and PACF for the cold season

Figures 4 and 5 above show the unitability (since more than 5% of correlation coefficients out of confidence interval)of the evaporation time series for the hot and cold seasons. Therefore, stability will be achieved by taking successive and seasonal differences. Through experimentation and testing, two decisions were made in parallel to achieve stability. In the first, one regular team and one seasonal team were taken, respectively. In the second, only one regular team was taken, for both seasons. It turned out that both decisions were correct, and each of them gave a different advantage than the other, while it turned out that the data was completely stable after making both decisions. To

determine the ranks of the models, the ACF and PACF functions are drawn. Figures 6 and 7 show the ACF and PACF after achieving stability with both of the above decisions for the hot and cold evaporation seasons, respectively.

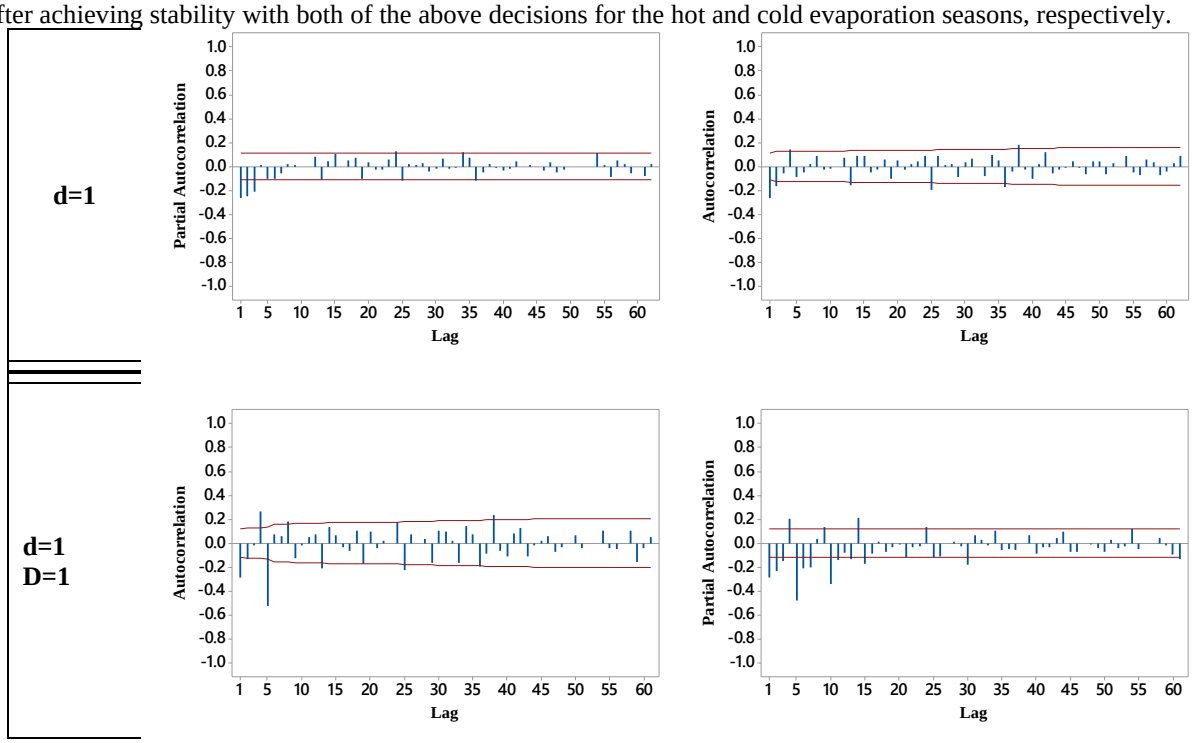


Figure 6. Time series plotting of evaporation, ACF and PACF for the hot season

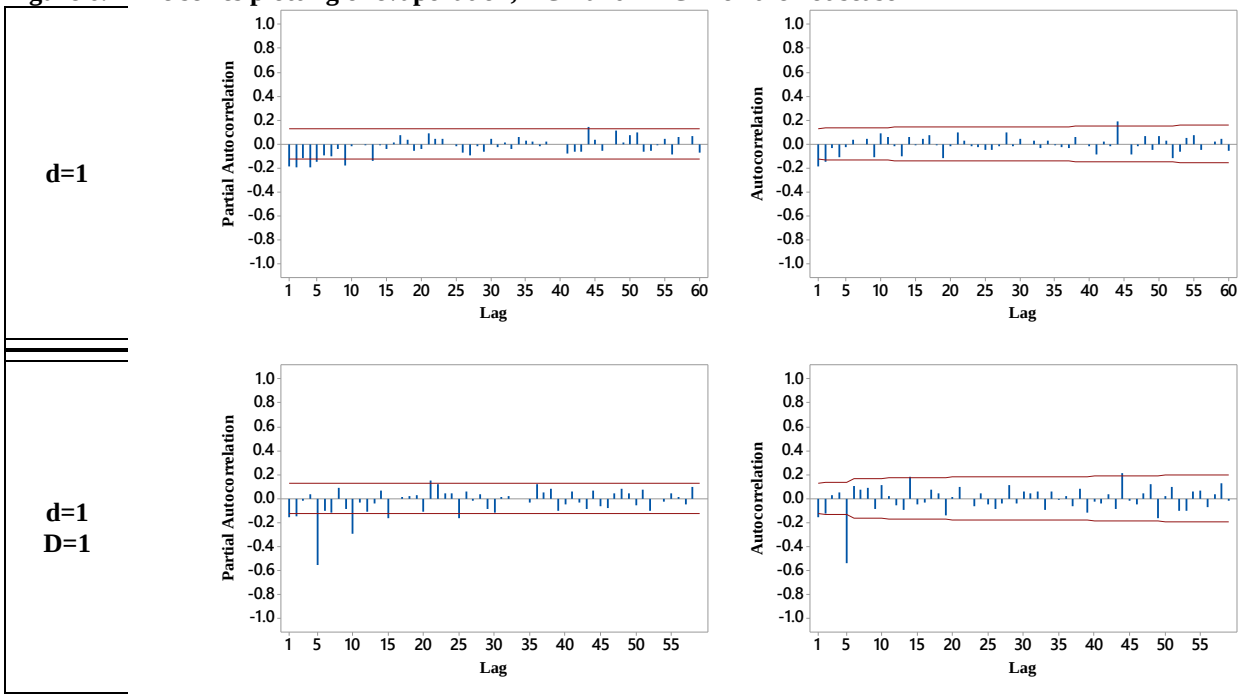


Figure 7. Time series plotting of evaporation, ACF, and PACF for the cold season

From Figures 6 and 7 above, and according to the similar behavior of the ACF and PACF functions in both the hot and cold seasons, the best ARIMA models will be as follows:

1. The first model: ARIMA(0,1,2) for both seasons and the results of significance are as in Table 1 below.
2. The second model: ARIMA(1,1,1)(0,1,1)5 for both seasons and the results of significance are as in Table 2 below.

Table 1. Parameter values and their significance for the ARIMA(0,1,2) model for both hot and cold seasons.

Hot Season				Cold Season			
Type	**Parameter**	**Calc. t**	**p-value**	**Type**	**Parameter**	**Calc. t**	**p-value**
θ_1	0.4128	7.14	0.000	θ_1	0.4216	7.12	0.000
θ_2	0.1837	3.18	0.002	θ_2	0.3881	6.55	0.000

Table 2. Parameter values and their significance for the ARIMA(1,1,1)(0,1,1)5 model for both hot and cold seasons.

Hot Season				Cold Season			
Type	**Parameter**	**Calc. t**	**p-value**	**Type**	**Parameter**	**Calc. t**	**p-value**
ϕ_1	0.2886	2.75	0.006	ϕ_1	0.5228	7.89	0.000
θ_1	0.7182	9.32	0.000	θ_1	0.9354	34.06	0.000
Θ_1	0.9726	49.33	0.000	Θ_1	0.9446	30.79	0.000

From Tables 1 and 2 above, for both seasons, it is clear that all estimated parameter values for the ARIMA(0,1,2) and ARIMA(1,1,1)(0,1,1)5 models are significant. The ACF function was also tested for the residuals of the above models and it turned out that it conforms to the conditions of good models. Thus, the above two models have successfully passed the diagnostic tests. Table 3 below shows the MAPE error criterion values for the training and testing periods for both hot and cold seasons for the above two models.

Table 3. MAPE error criterion values for ARIMA models for the training and testing periods for hot and cold seasons.

Models	**Hot Season**		**Cold Season**	
	Training	**Testing**	**Training**	**Testing**
ARIMA(0,1,2)	7.48748	59.96663	17.49708	48.69273
ARIMA(1,1,1)(0,1,1)$_5$	7.56499	60.77533	16.68917	56.71475

From Table 3 above, it is noticeable that preference varies from one model to another and it is not possible to judge the better predictive performance of one model over the other.

RT model

The structure of the models that were concluded as the best ARIMA models and referred to above was relied upon to be used in building the structure of the input regression tree models, whose foundations, equations, and methods of use in the theoretical aspect are indicated, and it is conventionally referred to as the hybrid ARIMA-RT model. Table 4 below shows the MAPE error criterion values for the training and testing periods for both hot and cold seasons for the hybrid ARIMA-RT model based on the above ARIMA models.

Table 4 the MAPE error criterion values for the hybrid ARIMA-RT model for the training and testing periods for both hot and cold seasons

Models	**Hot Season**		**Cold Season**	
	Training	**Testing**	**Training**	**Testing**
ARIMA(0,1,2)	4.16565	32.97756	8.174444	63.02950
ARIMA(1,1,1)(0,1,1)$_5$	5.04347	32.41332	10.77235	66.52910

Table 4 above, it is noticeable that the preference varies from one RT model to another, and it is not possible to judge the better forecasting performance of one model over the other. Comparing the results to table 3 and 4 above, the hybrid ARIMA-RT model most likely outperformed the ARIMA model, unlike the results of the testing period for the cold season only.

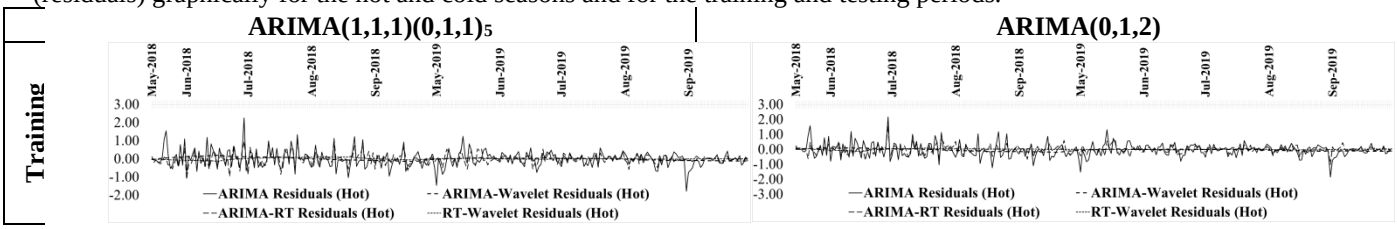
Wavelet transform

The wavelet transform WT was used in this study to purify the variable of the residuals series of the ARIMA model, which is referred to as the hybrid ARIMA-Wavelet model, then returned it to the original ARIMA model from which it was taken. The residuals series variate of the hybrid ARIMA-RT model was also purified based on WT in the same previous manner and is referred to as the hybrid RT-Wavelet model. This purification is carried out through the analysis of the components and periodic influences that control the non- linear and heterogeneous behavior of the original series and the resulting uncertainty in the modeling, which negatively affects the forecast performance and the forecasting residuals affected by all these influences and obstacles, using rules and foundations that lead to detecting regularity or irregularity of wave movement. Using the ready-made tool in the MATLAB system regarding stochastic modeling with one-dimensional regression estimators, it therefore requires determining the type and level of the wave function, in addition to the importance of structuring the inputs and organizing their numbers and dimensions in a scientifically correct manner. The Haar wave functions are used, as well as db1, db2, and db3 at the fifth level, after obtaining the results, the priority forecasting performance was concluded after using the db2 function for the hot season and the db3 function for the cold season. The results were as in table 5 below, which shows the values of the MAPE the error criterion for the training and testing periods for both the hot and cold seasons for the ARIMA-Wavelet and the hybrid RT-Wavelet model, according on the previously mentioned ARIMA model.

Table 5. values of the MAPE for the hybrid ARIMA-Wavelet and RT-Wavelet models, based on ARIMA for the training and testing periods for both the hot and cold seasons.

Models		Hot Season		Cold Season	
		Training	Testing	Training	Testing
ARIMA-Wavelet	ARIMA(0,1,2)	1.67931	61.99826	2.833533	49.71198
	$ARIMA(1,1,1)(0,1,1)_5$	1.61756	62.81019	3.405706	57.24044
RT-Wavelet	ARIMA(0,1,2)	4.07420	32.95010	8.176733	63.02666
	$ARIMA(1,1,1)(0,1,1)_5$	5.05101	32.40999	10.79895	66.52363

By comparison with results of the previous tables and from table 5 above, the forecasting performance of the hybrid ARIMA-Wavelet model was superior in the training periods and its failure to perform well in the testing periods, which may be explained by the lack of observations in it when compared to the testing period. As for the hybrid RT-Wavelet model, it didn't add anything new to the accuracy of forecast performance expect for a slight improvement compered to the original hybrid ARIMA-RT model, which lead us to the increase in the number of techniques and methods used to reach good forecasts may be negatively reflected in the complexity of the methodology of the proposed method, with the absence of noticeable improvement in forecast accuracy. figures 8 and 9 below illustrate and compare forecast errors (residuals) graphically for the hot and cold seasons and for the training and testing periods.



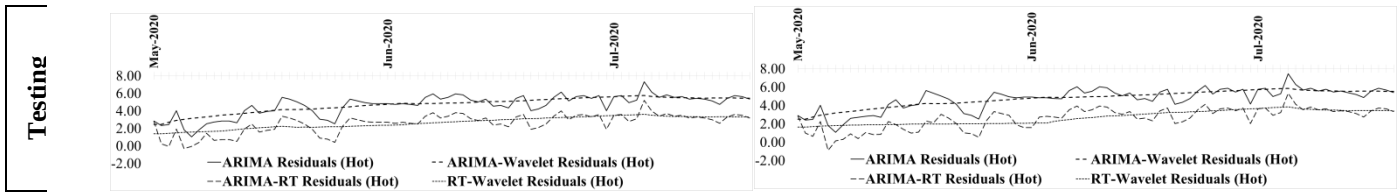


Figure 8. comparison of evaporation time series forecast errors for models used for the hot seasons

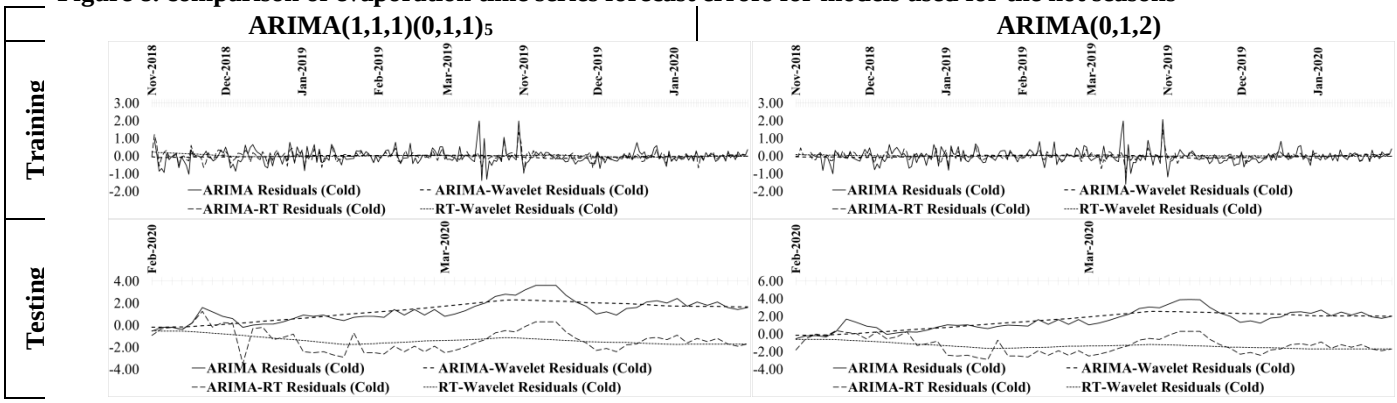


Figure 9. comparison of evaporation time series forecast errors for models used for the cold seasons

Figures 8 and 9 above give a clearer picture than the results of the forecast performance of the ARIMA, ARIMA-Wavelet, ARIMA-RT, and RT-Wavelet methods reported in the previous tables. From the above figures, the superiority of the forecasting performance of the hybrid ARIMA-Wavelet and RT-Wavelet models in the training and testing periods is evident through the smoothness that characterizes the residuals of the two models, which indicates a high behavioral match of the original evaporation series variable with the corresponding forecast series variable for the hot and cold seasons periods and for the two seasons, which reflects the high quality of forecasting performance of the hybrid ARIMA-Wavelet and RT-Wavelet methods.

Conclusions

In this research, the traditional method represented by the ARIMA models and the machine learning method represented by RT model were used, in addition to the use of wavelet transform WT as a specialized method in analyzing complex seasonal and cyclical components and patterns, and dismantling and purifying them from obstacles and heterogeneous compounds in order to forecast evaporation time series data after dividing It into the hot and cold seasons, and two periods for the training and testing. From the results and discussion obtained, it is possible to conclude the advantage of using both of the hybrid ARIMA-Wavelet and RT-Wavelet models in forecasting for the training and testing periods for evaporation data in particular or other climate data with similar behavior, due to their superiority in forecasting compared to the two ARIMA and ARIMA-RT models. Also the model and ARIMA-RT model have a really effective in improving forecasting when compared to the forecasting performance of the traditional ARIMA model. The traditional ARIMA model doesn't have the ability to deal with non-linearity data and forecast it. The methodology of dividing the data into hot and cold seasons addressed the problem of heterogeneity in the data due to the diversity of seasonal and cyclical components in the data.

References

1. Chaudhuri, S. and D. Dutta, *Mann–Kendall trend of pollutants, temperature and humidity over an urban station of India with forecast verification using different ARIMA models.* Environmental monitoring and assessment, 2014. **186**: p. 4719-4742.
2. Eymen, A. and Ü. Köylü, *Seasonal trend analysis and ARIMA modeling of relative humidity and wind speed time series around Yamula Dam.* Meteorology and Atmospheric Physics, 2019. **131**: p. 601-612.

3. Gupta, A. and A. Kumar. *Mid term daily load forecasting using ARIMA, wavelet-ARIMA and machine learning*. in *2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*. 2020. IEEE.
4. Aladağ, E., *Forecasting of particulate matter with a hybrid ARIMA model based on wavelet transformation and seasonal adjustment.* Urban Climate, 2021. **39**: p. 100930.
5. Wei, W.W.S., *Time series analysis : univariate and multivariate methods*. 2nd ed ed. 2006: Pearson Addison Wesley.
6. Liu, L.-M., *Time Series Analysis and Forecasting*. 2nd ed. 2006. 578.
7. Chan, N.H., *Time Series Applications to Finance*. 2002, John Wiley & Sons, Inc.
8. Palma, W., *Long-memory time series: theory and methods*. Vol. 662. 2007: John Wiley & Sons.
9. Shukur, O.B. and M.H. Lee, *Daily wind speed forecasting through hybrid KF-ANN model based on ARIMA.* Renewable Energy, 2015. **76**: p. 637-647.
10. Das, R.C., *Forecasting incidences of COVID-19 using Box-Jenkins method for the period July 12-Septembert 11, 2020: A study on highly affected countries.* Chaos, Solitons & Fractals, 2020. **140**: p. 110248.
11. Pham, B.T., I. Prakash, and D.T. Bui, *Spatial prediction of landslides using a hybrid machine learning approach based on random subspace and classification and regression trees.* Geomorphology, 2018. **303**: p. 256-270.
12. Loh, W.-Y., *Classification and regression tree methods.* Encyclopedia of statistics in quality and reliability, 2008. **1**: p. 315-323.
13. Choubin, B., et al., *Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches.* Environmental earth sciences, 2018. **77**(8): p. 1-13.
14. Rutkowski, L., et al., *The CART decision tree for mining data streams.* Information Sciences, 2014. **266**: p. 1-15.
15. Loh, W.Y., *Classification and regression trees.* Wiley interdisciplinary reviews: data mining and knowledge discovery, 2011. **1**(1): p. 14-23.
16. Zhang, J., et al., *A hybrid approach of wavelet transform, ARIMA and LSTM model for the share price index futures forecasting.* The North American Journal of Economics and Finance, 2024. **69**: p. 102022.
17. Jaishi, H.P., et al., *Comparing wavelet-based artificial neural network, multiple linear regression, and ARIMA models for detecting genuine radon anomalies associated with seismic events.* Proceedings of the Indian National Science Academy, 2024: p. 1-15.
18. Mallat, S.G., *A theory for multiresolution signal decomposition: the wavelet representation.* IEEE transactions on pattern analysis and machine intelligence, 1989. **11**(7): p. 674-693.
19. Donoho, D.L. and I.M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage.* Journal of the american statistical association, 1995. **90**(432): p. 1200-1224.
20. Hyndman, R.J. and A.B. Koehler, *Another look at measures of forecast accuracy.* International journal of forecasting, 2006. **22**(4): p. 679-688.

إستخدام النموذج الهجين لشجرة الانحدار و ARIMA مع التحويلات المويجية للتنبؤ بالسلسلة الزمنية للتبخر

زينة مضر البزاز[1] و نعم سالم فاضل[2]

[1,2]قسم الاحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق.

الخلاصة: إن دقة التنبؤ بالسلاسل الزمنية للتبخر أمر من الاهمية بمكان للتحكم في التأثيرات والأضرار والمخاطر البيئية المؤثرة خصوصا على حياة النبات ونموه وبالتالي تأثير ذلك على حياة الانسان والحيوان. تعتبر بيانات التبخر من السلاسل الزمنية المناخية والتي تتميز بطبيعتها غير الخطية البيانات من خلال ما تعانيه من مشكلة عدم التجانس لاحتوائها على العديد من المركبات الموسمية والدورية، وقد يؤدي ذلك التعقيد بالضرورة إلى تنبؤات قليلة الدقة. لتحقيق تجانس اكبر وسلوك زمني اقل تعقيدا سيتم استخدام اسلوب التراصف الزمني في هذه الدراسة مع اساليب التنبؤ المقترحة. سيتم استخدام اسلوبين تنبؤيين متمثلان باسلوب شجرة الانحدار regression tree (RT) ونموذج الانحدار الذاتي والمتوسطات المتحركة التكاملية integrated autoregressive and moving average (ARIMA) ومنهما يتم اقتراح تهجينهما باسلوب يجمع كليهما ضمن النموذج ARIMA-RT الهجين كأسلوب لتحسين نتائج التنبؤ من خلال التعامل بدقة اكبر مع عدم خطية البيانات. كذلك سيتم اختبار تأثير التحويلات المويجية Wavelet transformations (WT) مع كل من نموذج ARIMA والنموذج ARIMA-RT الهجين وهل سيكون له دور في تحسين النتائج التنبؤية. سيتم اعتماد هيكلية نمذجة السلاسل الزمنية باستخدام ARIMA الموسمي المضاعف لتحديد هيكلية مدخلات نموذج RT ضمن الأسلوب المقترح الهجين. كذلك فان استخدام WT سيقتصر على فلترة سلسلة الاخطاء العشوائية (البواقي) والتي تعتمد عليها باقي تخلفاتها الزمنية المتمثلة بمتغيرات المتوسطات المتحركة. سيتم إجراء مقارنات لنتائج التنبؤ للأساليب المقترحة ومقارنتها بالطريقة التقليدية للتنبؤ. اهتمت هذه الدراسة بالبحث في الاساليب المتنوعة للتنبؤ بالسلاسل الزمنية للتبخر لإحدى محطات الأرصاد الجوية الزراعية في مدينة الموصل، العراق. عكست نتائج هذه الدراسة تفوق الاسلوب الهجين مقارنة بنموذج ARIMA التقليدي. كما تضمنت النتائج عدم تأثر التنبؤات بصورة واضحة في ظل استخدام WT. من خلال هذه الدراسة يمكن استنتاج ان النموذج الهجين ARIMA-RT له دور واضح في تحسين دقة نتائج التنبؤ. اما استخدام WT فانه يؤدي الى تحسن بسيط في دقة التنبؤات وقد تختلف باختلاف البيانات وحسب طبيعتها وتجانسها.

الكلمات المفتاحية: نموذج ARIMA؛ شجرة الانحدار؛ التحويل المويجي؛ التنبؤ؛ التبخر.