AIMS Environmental Science, 8(4): 283–303.
DOI: 10.3934/environsci.2021019
Received: 24 April 2021
Accepted: 01 July 2021
Published: 08 July 2021

http://www.aimspress.com/journal/environmental

Research article

An efficiency dynamic seasonal regression forecasting technique for high variation of water level in Yom River Basin of Thailand

Kitipol Nualtong[1], Ronnason Chinram[1], Piyawan Khwanmuang[1], Sukrit Kirtsaeng[2] and Thammarat Panityakul[1,*]

[1] Division of Computational Science, Faculty of Science, Prince of Songkla University, Hat Yai, 90110, Songkhla, Thailand
[2] Thai Meteorological Department, Bangna, 10260, Bangkok, Thailand

* **Correspondence:** thammarat.p@psu.ac.th.

Abstract: The Yom River Basin is one of 25 river basins in Thailand. The Yom River Basin experiences perennial droughts and floods that heavily impact the agricultural sector. In order to reduce the impact, water management, including water level estimation, must be applied to critical basins like the Yom River Basin. An important task of management is the quantitative prediction of water levels. Four different modeling approaches were applied to forecast the average monthly water level (AMWL) data from four water level measurement stations for the wet season (May–October) and dry season (November–April) from 2007 to 2020. The forecast patterns obtained from the four approaches were similar to the observed historical values, except the upstream in wet season and downstream in dry season. Furthermore, the artificial neural network (ANN) approach overestimated forecasts for almost every station in both seasons. All four approaches were more accurate in the dry season than the wet season. This study proposed a forecasting method called dynamic seasonal regression (DSR), which was obtained by combining multiple linear regression (MLR) and the autoregressive integrated moving average (ARIMA) model of the random error from MLR. DSR was more efficient than ANN, seasonal-ARIMA (SARIMA) and a hybridized SARIMA and ANN approach (SARIMANN). On average, for all stations in wet and dry seasons, DSR reduced RMSE by over 40.86%, 9.10% and 23.07% with respect to ANN, SARIMA and SARIMANN, and MAPE by over 35.01%, 13.02% and 15.96% with respect to ANN, SARIMA and SARIMANN. The RMSE of upstream was higher than the RMSE downstream in the wet season for all methods, and the MAPE of upstream was lower than the downstream in both seasons for all methods. Moreover, the RMSE of upstream was lower than the downstream in the dry season for all methods except the ANN method.

Keywords: dynamic seasonal regression; ANN; SARIMA; water level; high variation of hydrological cycle basin; Yom River basin

1. Introduction

The Yom River Basin is one of the 25 main basins of Thailand. Located in Northern Thailand, the Yom River Basin is important because of the variety of tropical wet and dry, and even savanna, climates that occur throughout the year. This river basin impacts a huge area. The Land Development Department reported in 2013 that the Yom River Basin comprises an area of 907.75 km^2 which experiences fewer than 4 droughts every ten years, an area of 701.3 km^2 which experiences 4–5 droughts every ten years, and an area of 229.34 km^2 which experiences over 5 droughts every ten years. These areas respectively accounted for 49.38, 38.15, and 12.34% of the total area which experiences repeated drought conditions in the basin. The same area represents 3.77, 2.92, and 0.95% of the total area of the Yom River Basin. The Yom River Basin has an area of 2454.06 km^2 which experiences fewer than 4 floods every ten years, an area of 1334.25 km^2 which experiences 4–7 floods every ten years, and an area of 541.90 km^2 which experiences 8 or more floods every ten years. These areas respectively accounted for 56.67, 30.81, and 12.51% of the total area of the basin which experiences repeated flooding, and represent 10.21, 5.55, and 2.25% of the total area of the Yom River Basin [1–4].

There are several factors that affect the streamflow of a river; precipitation, climate change and disturbance from human activities. These factors complicate hydrological modeling [4–7]. Sopipan (2014) [8] studied historical monthly rainfall data from April 2005 to March 2013 in Nakhon Ratchasima Province, Thailand. From the data, he developed forecasts using the autoregressive integrated moving average (ARIMA) model and the multiplicative Holt–Winters method. The performances of the forecasts were measured using the mean absolute percentage error (MAPE), mean squared error (MSE) and mean absolute error (MAE). Forecasts developed from both methods were found to be acceptable but ARIMA gave a better result. Fashae et al. (2018) [9] used the artificial neural network (ANN) and ARIMA models to analyse the discharge of the River Opeki in Oyo State, Nigeria from 1982 to 2010 and used the best model to forecast the discharge of the river from 2010 to 2020. The performance of the two models was subjected to an evaluation based on the root mean square error (RMSE) and correlation coefficient. The result showed that ARIMA performed better, considering the level of RMSE and higher correlation coefficient. Pongsiri (2007) [10] compared the accuracy of forecasts derived from the ARIMA model, the ANN model, and a hybrid ARIMA-ANN model for the daily closing price of the Trans-Pacific Partnership (TPP) time series since 2005 to 2007. The assessment of the accuracy of the three models was based on the MSE, MAE, and MAPE. The ARIMA model concept was suitable for a time series with a linear structure, while the ANN model was able to completely describe the non-linear structure. Therefore, the hybrid ARIMA-ANN model would be completely descriptive of time series data with a linear and non-linear structure. The result showed that the hybrid ARIMA-ANN model was able to provide more accurate short-term (7-day) forecasts than the ARIMA and ANN models, but for a long-term (30-day) forecast, the ANN model provided the most accurate forecast.

This study proposes appropriate forecasting models for time series of the average monthly water level (AMWL) of the Yom River Basin in Northern Thailand. The approach modified the Box-Jenkins method into a seasonal regression time series model. The proposed approach is called the Dynamic

Seasonal Regression (DSR) model, which has been developed from previous works [11–13]. The efficiency of this forecasting model was determined from comparisons to three different approaches. Firstly, a machine learning approach, the ANN model; secondly, a stochastic approach, the seasonal autoregressive integrated moving average (SARIMA) model; and finally, a hybridized approach, seasonal autoregressive integrated moving average and artificial neural network or SARIMANN. The study period was over thirteen hydrological years from April 2007 to March 2020.

2. Materials and methods

2.1. Study region and dataset

The Yom River Basin in Northern Thailand (Figure 1) covers a surface area of approximately 24046.89 km^2, between latitudes 14°50′ N and 18°25′ N and longitudes 99°16′ E and 100°40′ E. The Yom River Basin comprises 11 major sub-river basins and covers 11 provincial administrations, 45 districts, 286 sub-districts, and 2028 villages [4,14].

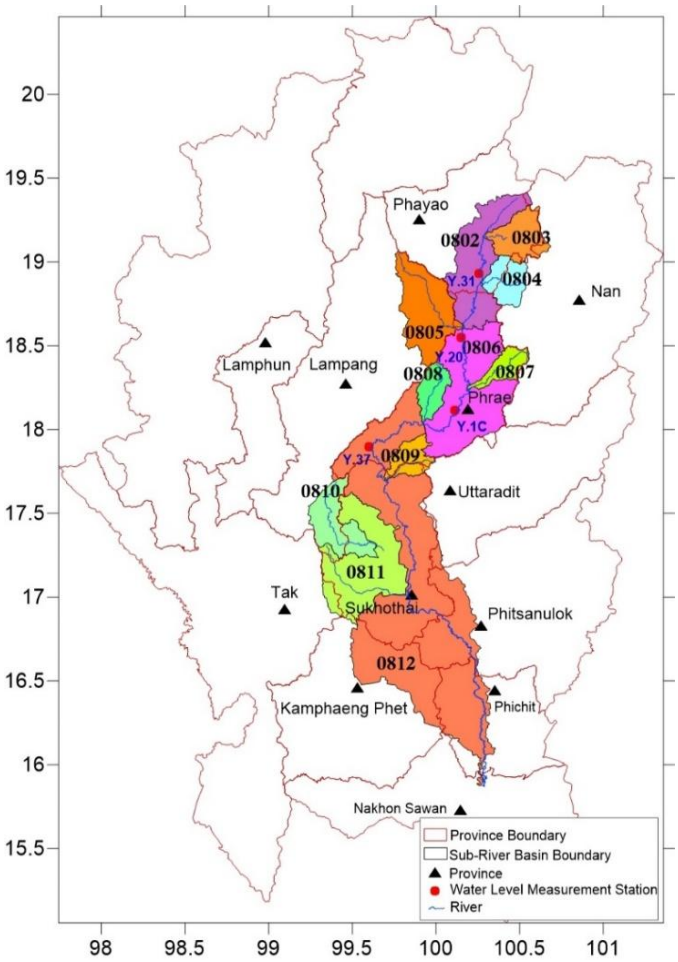

Figure 1. Locations of water level measurement station in Yom River Basin.

In 2014, the Yom River Basin received average annual precipitation of about 1179 mm, average annual runoff of about 5261 million m^3 and an average annual runoff of fewer than 2500 m^3 per year per person, which is less than the average annual runoff in Thailand per year per person (3496 m^3). In

2019, the Yom River Basin contained one large-sized reservoir and five medium-sized reservoirs with a total storage capacity of 295.62 million m^3 [1,4,14,15].

The daily water level data in m (MSL) were selected from four water level measurement stations over the length of the Yom River: Ban Thung Nong [Y.31] station, Ban Huai Sak [Y.20] station, Ban Nam Khong [Y.1C] station, and Ban Wang Chin [Y.37] station (Figure 1). Data were collected from the Upper Northern Region Irrigation Hydrology Center (Royal Irrigation Department) over thirteen hydrological years or 4649 days [16–20], from April 1, 2007 to March 31, 2020**.** The average of monthly data was adopted as the observation of any time series, so the monthly period represented by t. The AMWL data at the four listed water level measurement stations for the wet and dry seasons of Thailand were calculated, the wet season was from May to October and the dry season was from November to April.

2.2. Methodology

2.2.1. Seasonal Autoregressive Integrated Moving Average Model (SARIMA Model)

The ARIMA model is a universal model widely used in the area of time series analysis. An ARIMA model is a combination of autoregressive (AR) and moving average (MA) parts with differencing. Generally, the ARIMA model is denoted by ARIMA(p,d,q) where p,d and q are non-negative integers. In this notation, the p-parameter refers to the autoregressive (AR) part, the d-parameter refers to the order of regular differencing part, and the q-parameter refers to the moving average (MA) part [21–23]. The back-shift operator (B), where $B^j y_t = y_{t-j}$ for j, is a positive integer. The multiplicative SARIMA$(p,d,q) \times (P,D,Q)S$ model is represented by:

$$\Phi_P(B^S)\varphi_p(B)(1-B^S)^D(1-B)^d y_t = \Theta_Q(B^S)\theta_q(B)\varepsilon_t \quad (1)$$

where $\varphi(B)$ is the regular Autoregressive polynomial of order p, i.e., $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$

$\Phi(B^S)$ is the seasonal Autoregressive polynomial of order P, i.e., $\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}$

$\theta(B)$ is the regular Moving Average polynomial of order q, i.e., $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$

$\Theta(B^S)$ is the seasonal Moving Average polynomial of order Q, i.e., $\Theta(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots - \Theta_Q B^{QS}$

$(1-B)^d$ is the d^{th} order differencing process y_t

$(1-B^S)^D$ is the D^{th} order seasonal differencing process y_t and ε_t is independent random variable that represents the error term at time t.

The multiplicative SARIMA model is a form of the Box-Jenkins methodology, It requires stationary time series data of y_t and the independent random variable (ε_t) must qualify as white noise

2.2.2. Artificial Neural Network Model (ANN Model)

Artificial neural networks are the foundation of artificial intelligence (AI). ANNs have self-learning capabilities and two properties: learning and recall. The multi-layer perceptron neural network

or feed-forward back-propagation neural network is the learning function of the neural network. It utilizes a supervised learning technique that requires an input set and a target output as a training pair. During network teaching multiple training sets are normally used which generate actual output. If the actual output is different from the target output, it causes an error value. The network learns both datasets by adjusting the weight to reduce the difference value, or the error value, between actual output and target output to a minimum. The weight adjustment will be adjusted in small increments by the process of repeating the data one by one until the weights in the network converge. Its greatest strength is in non-linear solutions to ill-defined problems [23,24]. In the representation of a multi-layer perceptron neural network shown in Figure 2,

y_n is the input data of input layer at node n^{th}; $n = 1, 2, \dots, N$
x_m is the output data of hidden layer after weight adjustment using sigmoid-activation function at node m^{th}; $m = 1, 2, \dots, M$
z_j is the actual output data of output layer after weight adjustment using linear - activation function at node j^{th}; $j = 1, 2, \dots, J$
t_j is the target output data of output layer at node j^{th}; $j = 1, 2, \dots, J$
e_j is the error of output layer at node j^{th}; $j = 1, 2, \dots, J$
w_{nm} is the weight of network between input layer and hidden layer
w_{mj} is the weight of network between hidden layer and output layer

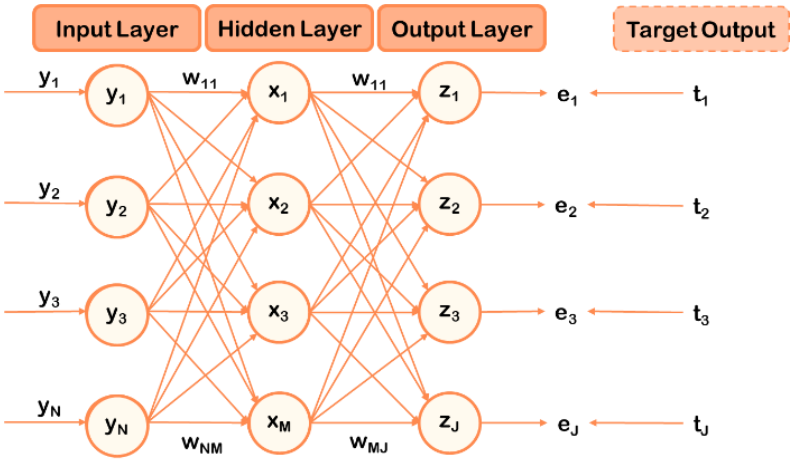

Figure 2. The multi-layer perceptron neural network.

2.2.3. Hybrid between Seasonal Autoregressive Integrated Moving Average and Artificial Neural Network Model (SARIMANN Model)

SARIMA models use a stochastic approach, which is suitable for linear variation with linear seasonal and trends data but is unable to manage non-linear data, such as hydrological data. ANN models, on the other hand, employ a machine learning approach which is more suitable for non-linear situations. Therefore, we hybridized the SARIMA model and the ANN model to deal with linear and non-linear situations simultaneously. The result is called the SARIMANN model. The creation of the SARIMANN model uses the concept of taking the input variables obtained from the best-fit SARIMA model and applying them to input data in the ANN model [23]. In the SARIMANN model shown in Figure 3, φ_p is the parameter of the autoregressive (AR) of order p; Φ_P is the parameter of the

seasonal autoregressive (SAR) of order P; θ_q is the parameter of the moving average (MA) of order q; Θ_Q is the parameter of the seasonal moving average (SMA) of order Q; $(1-B)^d$ is the differencing (a.k.a. Integrated, denoted by I) process y_t of order d; and $(1-B^S)^D$ is the seasonal differencing process (a.k.a. Seasonal Integrated, denoted by SI) y_t of order D.

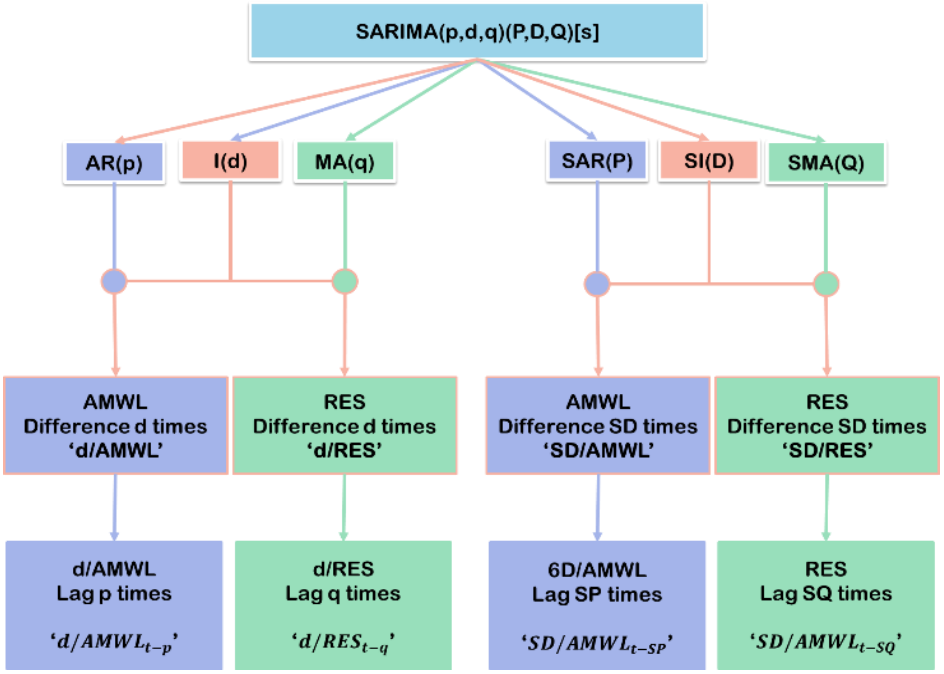

Figure 3. The concept of input variables of SARIMANN models.

2.2.4. The dynamic seasonal regression technique

As mentioned above, the technique combines three approaches: MLR, ARIMA and DSR. Modeling was performed in three stages

Step 1: To find the best multiple linear regression model (MLR model) for each AMWL series, time series data of the AMWL were input with seasonal lag $j =$ 3, 6, 9, and 12. Then the time series with significance lags can be imported to the MLR full model as the predictors, where the significance lags are obtained by testing any autocorrelation function (ACF) value by its 95% confidence limits. After that, the best subset be justified by stepwise selection. The MLR model is represented by:

$$y_t = \beta_0 + \beta_1 y_{t-3} + \beta_2 y_{t-6} + \cdots + \beta_j y_{t-j} + \vartheta_t \quad (2)$$

where y_t is the AMWL at time t

y_{t-j} is the AMWL at time $t-j$

β_j is the coefficients of the MLR method of j^{th}

ϑ_t is the independent random error term of MLR model at time t

t is time period (month)

Step 2: The best ARIMA model for the random error component from above MLR above was found by considering the random error component from the best subset fitting. The $ARIMA(p,d,q)$

model can be written in terms of the back-shift operator (B), where $B^j\vartheta_t = \vartheta_{t-j}$ for seasonal lag j is a positive integer represented by

$$\left(1-\sum_{j=1}^{p}\varphi_j B^j\right)(1-B)^d\vartheta_t = \left(1-\sum_{j=1}^{q}\theta_j B^j\right)\varepsilon_t \quad (3)$$

in which ϑ_t is independent random error term of MLR model at time t

φ_j is the autoregressive parameter of j^{th}

θ_j is the moving average parameter of j^{th}

$(1-B)^d$ is the d^{th} order differencing process ϑ_t

ε_t is the residual of ARIMA model at time t

Step 3: The dynamic seasonal regression model (DSR model) was derived by integrating the MLR and ARIMA components. Therefore, the DSR model is represented by the equation:

$$y_t = \beta_0 + \beta_1 y_{t-3} + \beta_2 y_{t-6} + \cdots + \beta_j y_{t-j} + \left[\frac{\left(1-\sum_{j=1}^{q}\theta_j B^j\right)\varepsilon_t}{\left(1-\sum_{j=1}^{p}\varphi_j B^j\right)(1-B)^d}\right] \quad (4)$$

where the second part of the right-hand side of the equation is from Eq 3.

2.2.5. Accuracy and efficiency

If the RMSE and MAPE values are small, the forecast value of AMWL is highly accurate.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - y_t)^2} \quad (5)$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|x_t - y_t|}{x_t} \quad (6)$$

where y_t is the forecast value of AMWL data of t. Since $t = 1, 2, \ldots, n$; x_t is the observations value of AMWL data at month t and n is the total number of observations of AMWL data.

3. Results

The time series plots of AMWL data from four water level measurement stations show relatively clear seasonal variations for the wet and dry seasons from 2007 to 2019; the wet season is from May to October and the dry season is from November to April. There is a peak in the AMWL every hydrological year. The peak and the dispersion of data are smaller in the dry season than in the wet season, as shown in Figures 4 and 5.

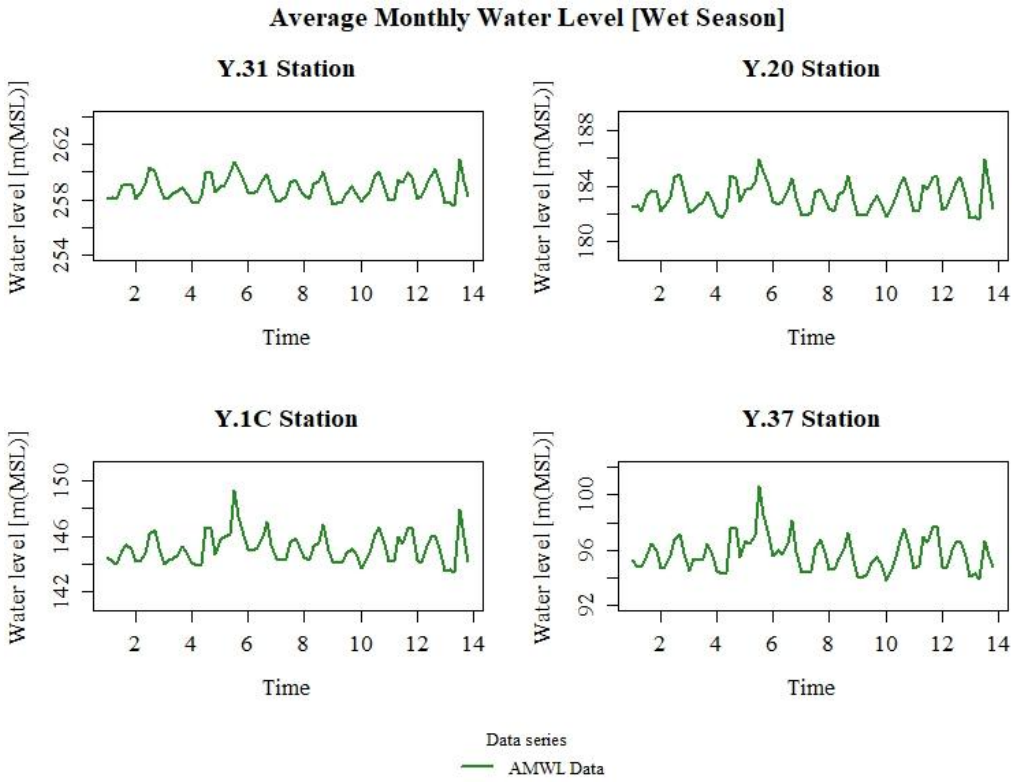

Figure 4. Time series plots of AMWL data from all four stations in the wet season from 2007 to 2019 (12 hydrological years).

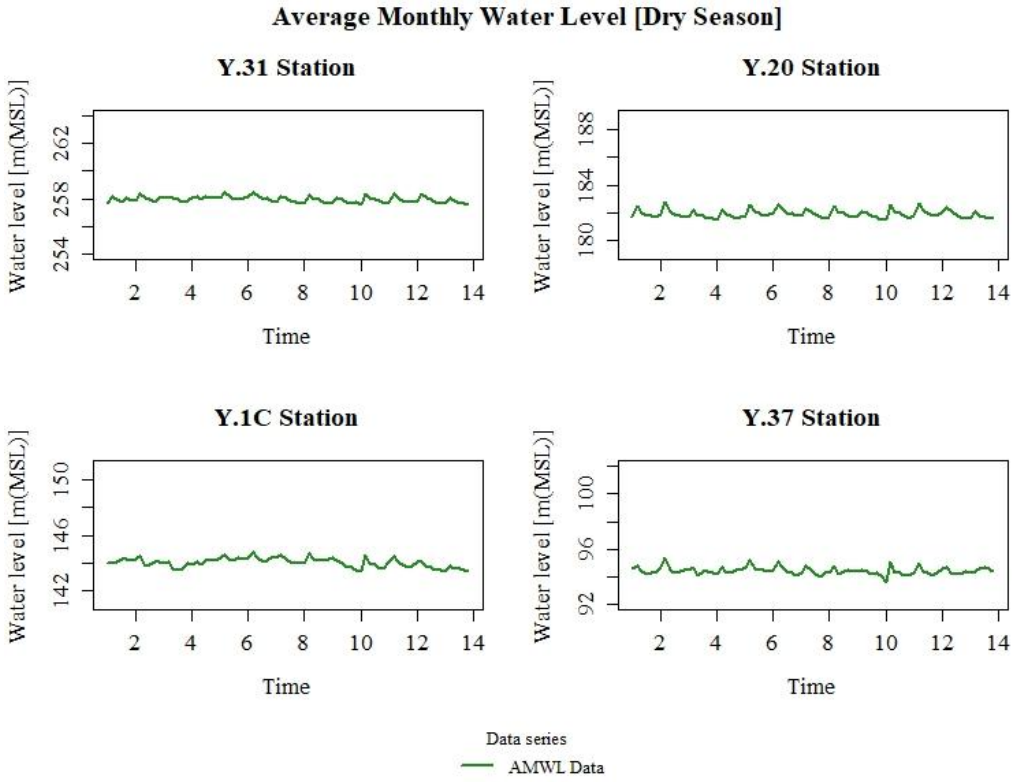

Figure 5. Time series plots of AMWL data from all four stations in the dry season from 2007 to 2019 (12 hydrological years).

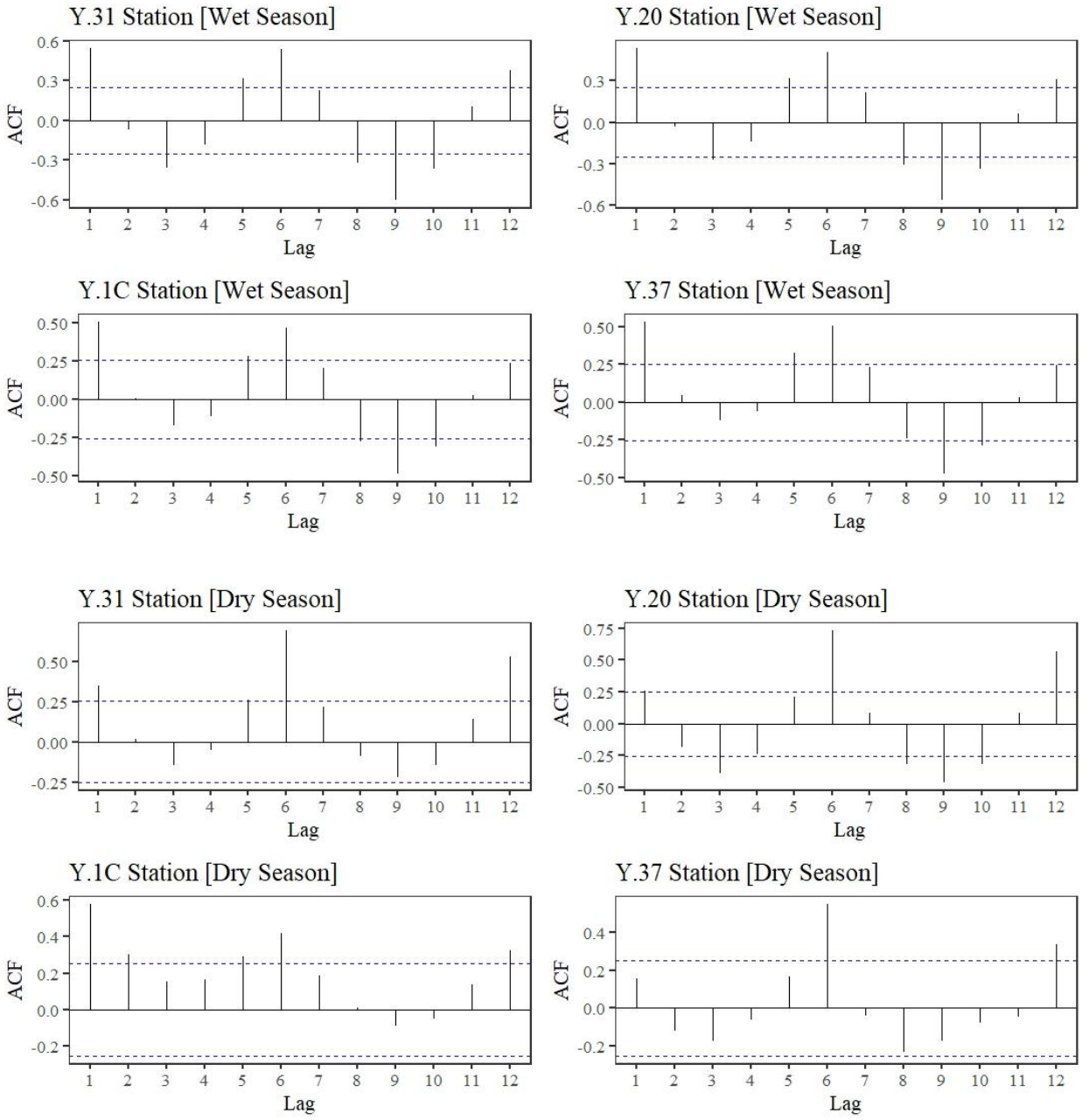

Figure 6. The ACF plots with 95% confidence limits of AMWL data from all four stations in wet and dry seasons from 2009 to 2019 (10 hydrological years).

The AMWL data from all four water level measurement stations for the wet and dry seasons from 2007 to 2019 were prepared with seasonal: Lag 3, Lag 6, Lag 9, and Lag 12. The ACF values were plotted at 95% confidence limits from AMWL data from the four water level measurement stations for the wet and dry seasons from 2009 to 2019 (Figure 6). Series that lag outside the 95% confidence limit were selected as independent variables of the full MLR model, as shown in Table 1.

Table 1 Predictors for the full MLR models of AMWL at all four stations in wet and dry seasons.

Series Data	Series Data
Y.31 Station [Wet season]	$y_{WET\ Y.31_{t-3}}, y_{WET\ Y.31_{t-6}}, y_{WET\ Y.31_{t-9}}, y_{WET\ Y.31_{t-12}}$
Y.20 Station [Wet season]	$y_{WET\ Y.20_{t-3}}, y_{WET\ Y.20_{t-6}}, y_{WET\ Y.20_{t-9}}, y_{WET\ Y.20_{t-12}}$
Y.1C Station [Wet season]	$y_{WET\ Y.1C_{t-6}}, y_{WET\ Y.1C_{t-9}}$
Y.37 Station [Wet season]	$y_{WET\ Y.37_{t-6}}, y_{WET\ Y.37_{t-9}}$
Y.31 Station [Dry season]	$y_{DRY\ Y.31_{t-6}}, y_{DRY\ Y.31_{t-12}}$
Y.20 Station [Dry season]	$y_{DRY\ Y.20_{t-3}}, y_{DRY\ Y.20_{t-6}}, y_{DRY\ Y.20_{t-9}}, y_{DRY\ Y.20_{t-12}}$
Y.1C Station [Dry season]	$y_{DRY\ Y.1C_{t-6}}, y_{DRY\ Y.1C_{t-12}}$
Y.37 Station [Dry season]	$y_{DRY\ Y.37_{t-6}}, y_{DRY\ Y.37_{t-12}}$

The best-fitted models and coefficients of the MLR model of the AMWL data from all four water level measurement stations for wet and dry seasons, are shown in Table 2.

Now take the random error values from the best-fitted models of the MLR to create the ARIMA model. Figures 7 and 8 show the time series plots of the random error series of the MLR model that best fitted the data from the four water level measurement stations for the wet and dry seasons from 2009 to 2019.

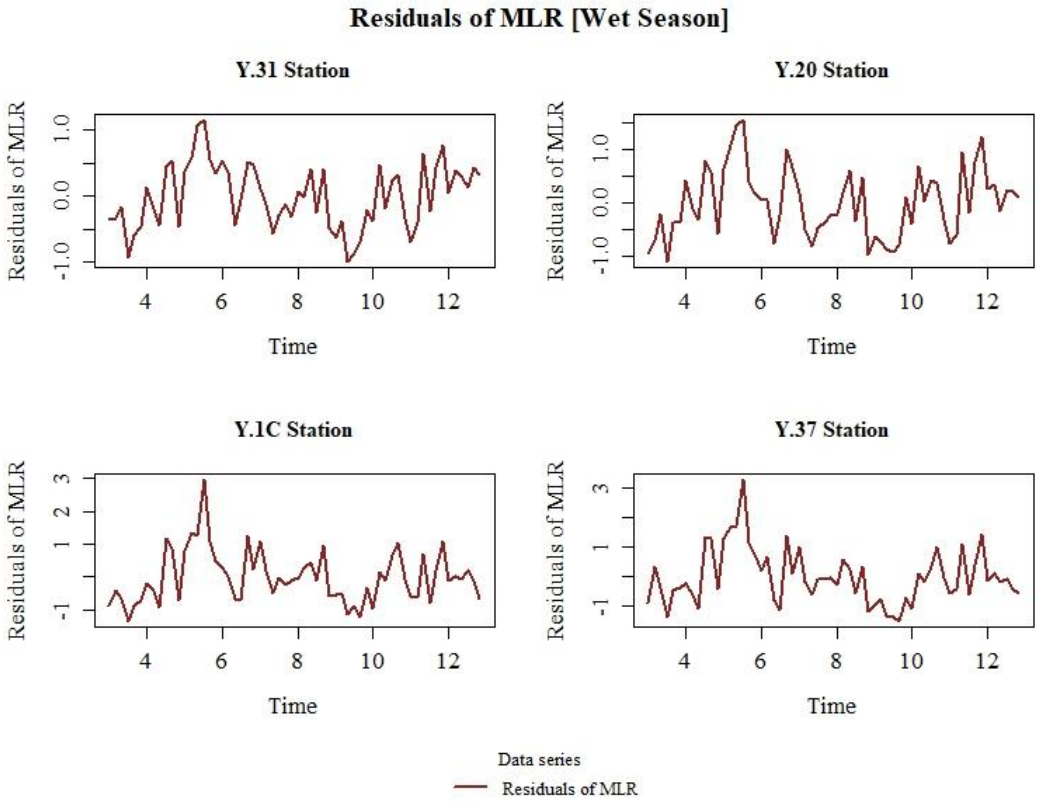

Figure 7. Time series plots of the random error of MLR for all four stations in the wet season from 2009 to 2019 (10 hydrological years).

Table 2 Coefficients of the MLR model of AMWL data from all four stations in wet and dry seasons.

Series Data	Estimate	Std. Error	t value	p-value	
Y.31 Station [Wet season]					
Intercept	284.5455	46.3542	6.1390	9.04×10^{-8}	***
$y_{WET\ Y.31_{t-6}}$	0.2261	0.0990	2.2840	0.0262	*
$y_{WET\ Y.31_{t-9}}$	-0.5342	0.0970	-5.5060	0.0000	***
$y_{WET\ Y.31_{t-12}}$	0.2088	0.1023	2.0400	0.0461	*
Y.20 Station [Wet season]					
Intercept	143.2314	41.0168	3.4920	9.53×10^{-4}	***
$y_{WET\ Y.20_{t-3}}$	0.2756	0.1233	2.2360	0.0294	*
$y_{WET\ Y.20_{t-6}}$	0.2558	0.0998	2.5620	0.0132	*
$y_{WET\ Y.20_{t-9}}$	-0.6264	0.1039	-6.0280	1.45×10^{-7}	***
$y_{WET\ Y.20_{t-12}}$	0.3132	0.1287	2.4340	0.0182	*
Y.1C Station [Wet season]					
Intercept	150.0174	22.7725	6.5880	1.54×10^{-8}	***
$y_{WET\ Y.1C_{t-6}}$	0.4135	0.1005	4.1150	1.26×10^{-4}	***
$y_{WET\ Y.1C_{t-9}}$	-0.4466	0.1032	-4.3260	6.19×10^{-5}	***
Y.37 Station [Wet season]					
Intercept	97.7530	13.9911	6.9870	3.35×10^{-9}	***
$y_{WET\ Y.37_{t-6}}$	0.4597	0.0953	4.8230	1.09×10^{-5}	***
$y_{WET\ Y.37C_{t-9}}$	-0.4800	0.0988	-4.8570	9.67×10^{-6}	***
Y.31 Station [Dry season]					
Intercept	67.6116	23.3034	2.9010	0.0052	**
$y_{DRY\ Y.31_{t-6}}$	0.7379	0.0903	8.1690	3.2×10^{-11}	***
Y.20 Station [Dry season]					
Intercept	109.0405	24.6359	4.4260	4.39×10^{-5}	***
$y_{DRY\ Y.20_{t-6}}$	0.6449	0.0818	7.8830	1.08×10^{-10}	***
$y_{DRY\ Y.20_{t-9}}$	-0.2443	0.0815	-2.9960	0.0040	**
Y.1C Station [Dry season]					
Intercept	72.2529	17.7215	4.0770	1.41×10^{-4}	***
$y_{DRY\ Y.1C_{t-6}}$	0.4983	0.1230	4.0520	1.53×10^{-4}	***
Y.37 Station [Dry season]					
Intercept	46.6447	9.4506	4.9360	7.09×10^{-6}	***
$y_{DRY\ Y.37_{t-6}}$	0.5058	0.1001	5.0550	4.61×10^{-6}	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

The random error of the MLR models that best fitted the data from the four water level measurement stations for the wet and dry seasons can be tested with the augmented Dickey-Fuller method. A p-value higher than the significance level, 0.05, is evidence of stationarity. The wet season time series for the Y.31 station, the Y.1C station, and the Y.37 station, returned p-values of 0.2794, 0.0645, and 0.2594, respectively. The dry season time series for the Y.31 station, the Y.20 station, the Y.1C station, and the Y.37 station, returned p-values of 0.2459, 0.1847, 0.1333, and 0.0706,

respectively. Therefore, all the time series are stationary except the wet season series for the Y.20 station, which gave a p-value of 0.0484 (Table 3).

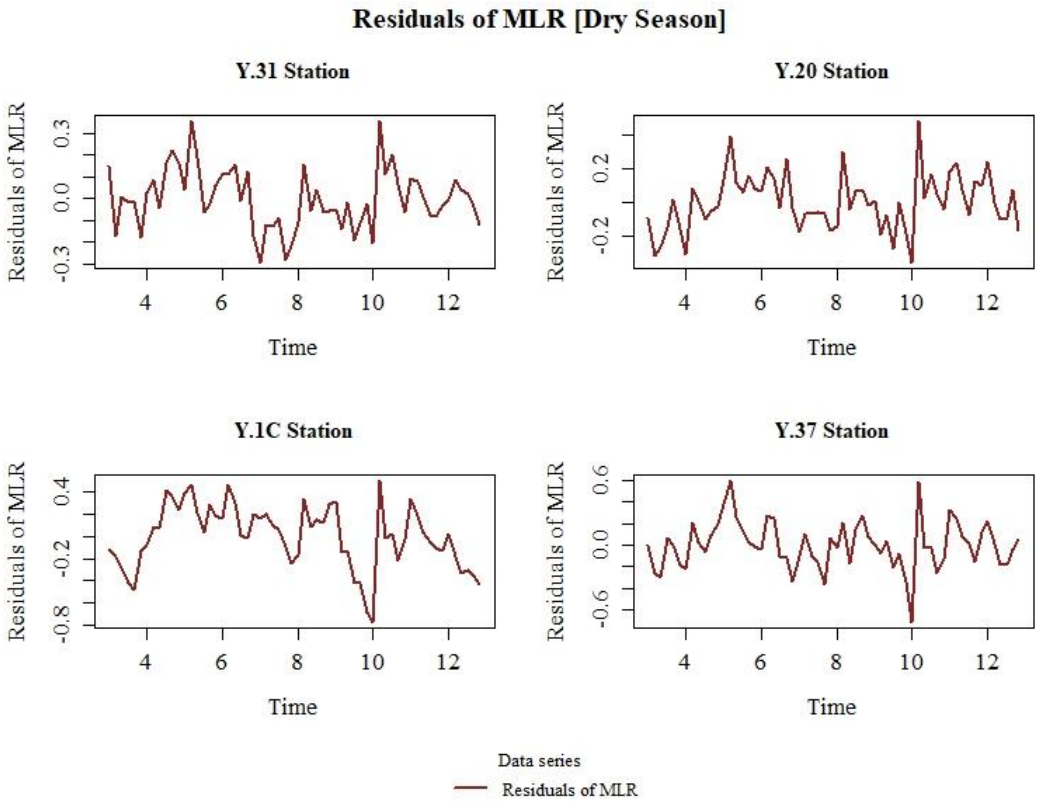

Figure 8. Time series plots of the random error of MLR for all four stations in the dry season from 2009 to 2019 (10 hydrological years).

Table 3 Stationarity consideration of the random error from the MLR models for all four stations in wet and dry seasons.

Series Data	Augmented Dickey-Fuller Test		
	Dickey-Fuller	Lag order	p-value
Y.31 Station [Wet season]	-2.7308	3	0.2794
Y.20 Station [Wet season]	-3.5096	3	0.0484 *
Y.1C Station [Wet season]	-3.3992	3	0.0645
Y.37 Station [Wet season]	-2.7802	3	0.2594
Y.31 Station [Dry season]	-2.8136	3	0.2459
Y.20 Station [Dry season]	-2.965	3	0.1847
Y.1C Station [Dry season]	-3.0923	3	0.1333
Y.37 Station [Dry season]	-3.3604	3	0.0706

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

The ACF plots with 95% confidence limits of the random error of best fitted MLR models for all stations in wet and dry seasons are shown in Figure 9.

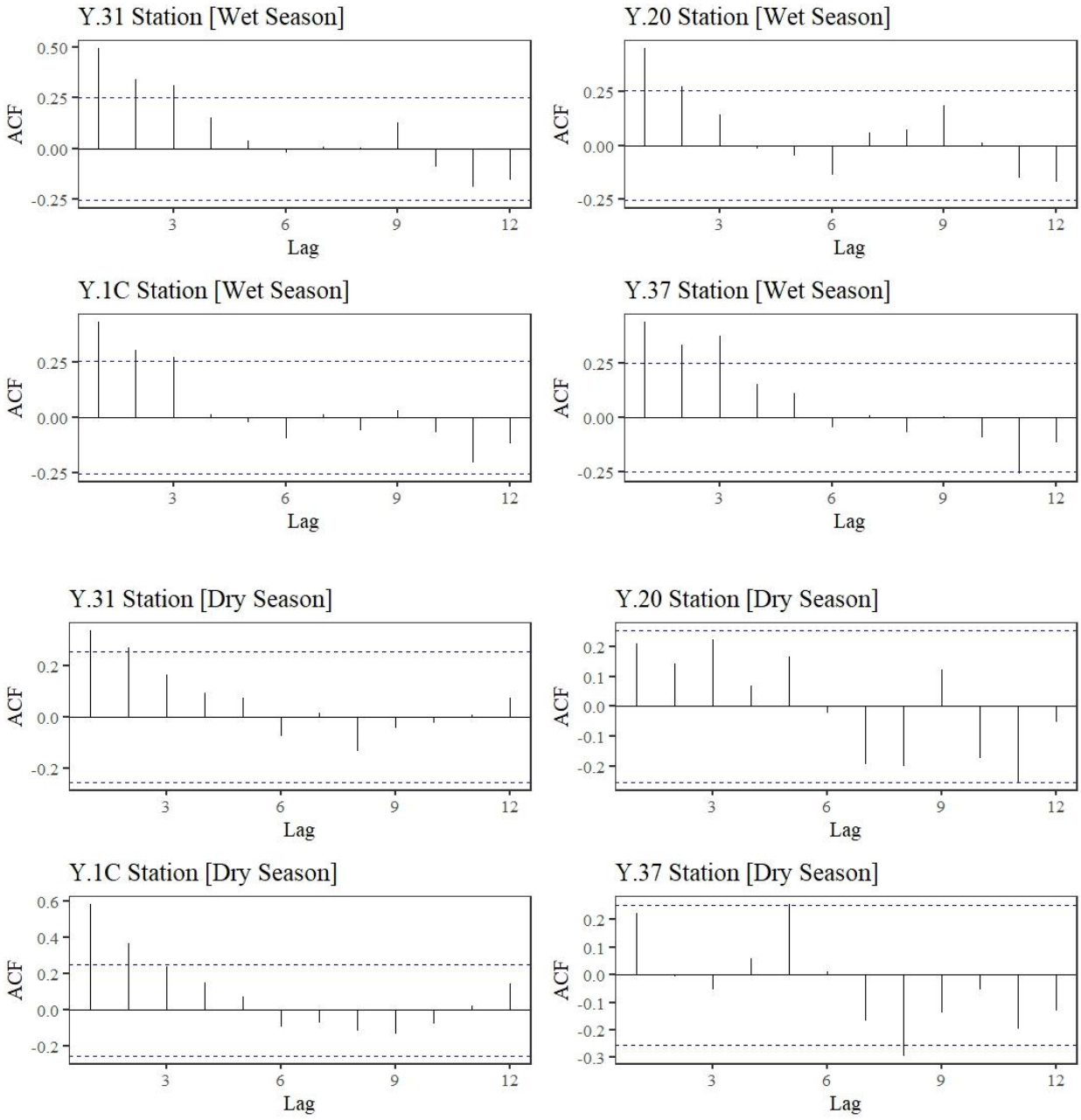

Figure 9. The ACF plots of random error with 95% confidence limits of the MLR models for all four stations in wet and dry seasons.

The best ARIMA models of the random errors from the best fitted MLR models were then determined. The coefficients of the best ARIMA models are shown in Table 4. For the Y.31 Station [wet season], Y.1C Station [wet season], and Y.31 Station [dry season] the best ARIMA model was ARIMA (1,0,0) with zero mean. For the Y.20 Station [wet season], the best ARIMA model was ARIMA (1,1,0). For the Y.37 Station [wet season], the best ARIMA model was ARIMA(1,0,1) with zero mean. For the Y.1C Station [dry season] the best ARIMA model was ARIMA(1,1,1). For the Y.20 Station [dry season] and Y.37 Station [dry season], the best ARIMA model was ARIMA(0,0,0) with zero mean, or random walk.

Table 4 Coefficients of the ARIMA models for all four stations in wet and dry seasons.

Series Data		Estimate	Std. Error	z value	p-value	
Y.31 Station [Wet season]						
ARIMA(1,0,0) with zero mean	ar1	0.4946	0.1112	4.448	8.67×10^{-6}	***
Y.20 Station [Wet season]						
ARIMA(1,1,0)	ar1	-0.3354	0.1213	-2.7646	0.0057	**
Y.1C Station [Wet season]						
ARIMA(1,0,0) with zero mean	ar1	0.4377	0.1163	3.762	1.69×10^{-4}	***
Y.37 Station [Wet season]						
ARIMA(1,0,1) with zero mean	ar1	0.7664	0.1352	5.6698	1.43×10^{-8}	***
	ma1	-0.4147	0.1826	-2.2715	0.0231	*
Y.31 Station [Dry season]						
ARIMA(1,0,0) with zero mean	ar1	0.3409	0.1221	2.7912	0.0053	**
Y.20 Station [Dry season]						
ARIMA(0,0,0) with zero mean						
Y.1C Station [Dry season]						
ARIMA(1,1,1)	ar1	0.5968	0.1451	4.1122	3.92×10^{-5}	***
	ma1	-0.9553	0.0964	-9.9093	$<2.2\times10^{-16}$	***
Y.37 Station [Dry season]						
ARIMA(0,0,0) with zero mean						

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Table 5 Consider the assumption of white noise of the residuals of ARIMA models of all four stations in wet and dry seasons.

Series Data	Box-Ljung Test		
	Chi-squared	df	p-value
Y.31 Station [Wet season]	0.2516	1	0.6159
Y.20 Station [Wet season]	0.2387	1	0.6252
Y.1C Station [Wet season]	0.2618	1	0.6089
Y.37 Station [Wet season]	0.0117	1	0.9140
Y.31 Station [Dry season]	0.2275	1	0.6334
Y.20 Station [Dry season]	2.7767	1	0.0957
Y.1C Station [Dry season]	0.0905	1	0.7636
Y.37 Station [Dry season]	3.1613	1	0.0754

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

To verify the assumptions of white noise, the Box-Ljung test was applied to the residuals of the ARIMA models for the wet and dry seasons at the four stations. The p-values were all higher than significance level of 0.05 and therefore non-significant. The Y.31 station, the Y.20 station, the Y1C station, and the Y.37 station for the wet season show p-values of 0.6159, 0.6252, 0.6089, and 0.9140,

respectively. The Y.31 station, the Y.20 station, the Y1C station and the Y.37 station for the dry season show p-values of 0.6334, 0.0957, 0.7636, and 0.0754, respectively (**Table 5**). Therefore, all residuals are white noise.

The equations of the DSR models of AMWL data from all four stations for wet and dry seasons are shown in Table 6.

Table 6 The equation of the DSR models of AMWL data from all four stations in wet and dry seasons.

Series Data	Equation of the best-fitted SARIMA model
Y.31 Station [Wet season]	$y_{WET\ Y.31_t} = 284.54550 + 0.22611\, y_{WET\ Y.31_{t-6}} - 0.53416\, y_{WET\ Y.31_{t-9}} + 0.20879\, y_{WET\ Y.31_{t-12}} + 0.4946\, \vartheta_{WET\ Y.31_{t-1}}$
Y.20 Station [Wet season]	$y_{WET\ Y.20_t} = 143.23144 + 0.27558\, y_{WET\ Y.20_{t-3}} + 0.25581\, y_{WET\ Y.20_{t-6}} - 0.62644\, y_{WET\ Y.20_{t-9}} + 0.31316\, y_{WET\ Y.20_{t-12}} - 1.3354\, \vartheta_{WET\ Y.20_{t-1}} + 0.3354\, \vartheta_{WET\ Y.20_{t-2}}$
Y.1C Station [Wet season]	$y_{WET\ Y.1C_t} = 150.0174 + 0.4135\, y_{WET\ Y.1C_{t-6}} - 0.4466\, y_{WET\ Y.1C_{t-9}} + 0.4377\, \vartheta_{WET\ Y.1C_{t-1}}$
Y.37 Station [Wet season]	$y_{WET\ Y.37_t} = 97.75304 + 0.45969\, y_{WET\ Y.37_{t-6}} - 0.47995\, y_{WET\ Y.37_{t-9}} + 0.7664\, \vartheta_{WET\ Y.37_{t-1}} + 0.4147\, \varepsilon_{WET\ Y.1C_{t-1}}$
Y.31 Station [Dry season]	$y_{DRY\ Y.31_t} = 67.61155 + 0.73788\, y_{DRY\ Y.31_{t-6}} + 0.3409\, \vartheta_{DRY\ Y.31_{t-1}}$
Y.20 Station [Dry season]	$y_{DRY\ Y.20_t} = 109.04052 + 0.64489\, y_{DRY\ Y.20_{t-6}} - 0.24431\, y_{DRY\ Y.20_{t-9}}$
Y.1C Station [Dry season]	$y_{DRY\ Y.1C_t} = 72.2529 + 0.4983\, y_{DRY\ Y.1C_{t-6}} - 0.4032\, \vartheta_{DRY\ Y.1C_{t-1}} - 0.5968\, \vartheta_{DRY\ Y.1C_{t-2}} + 0.9553\, \varepsilon_{DRY\ Y.1C_{t-1}}$
Y.37 Station [Dry season]	$y_{DRY\ Y.37_t} = 46.6447 + 0.5058\, y_{DRY\ Y.37_{t-6}}$

The forecast results obtained from the four models applied to AMWL data from all four water level measurement stations for the wet (six months: May 2019 – October 2019) and dry (six months: November 2019 – April 2020) seasons of one hydrological year showed that the methods yielded similar forecast patterns to previous recorded values, except upstream in the wet season and downstream in the dry season. Furthermore, ANN overestimated forecasts at almost every station in both seasons, as shown in Figures 10 and 11.

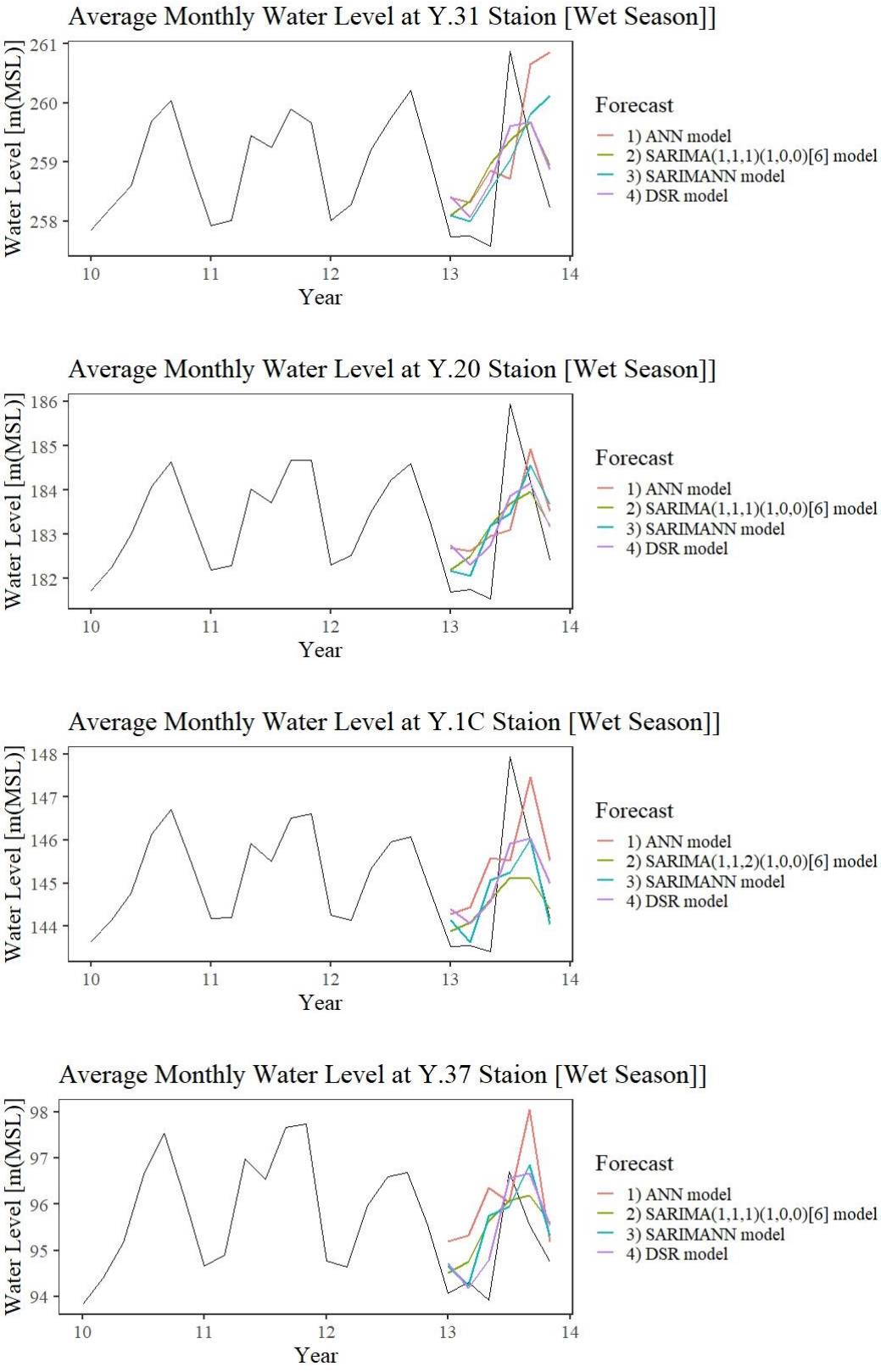

Figure 10. Forecast for one hydrological year applying the four models to AMWL data from all four stations for the wet season.

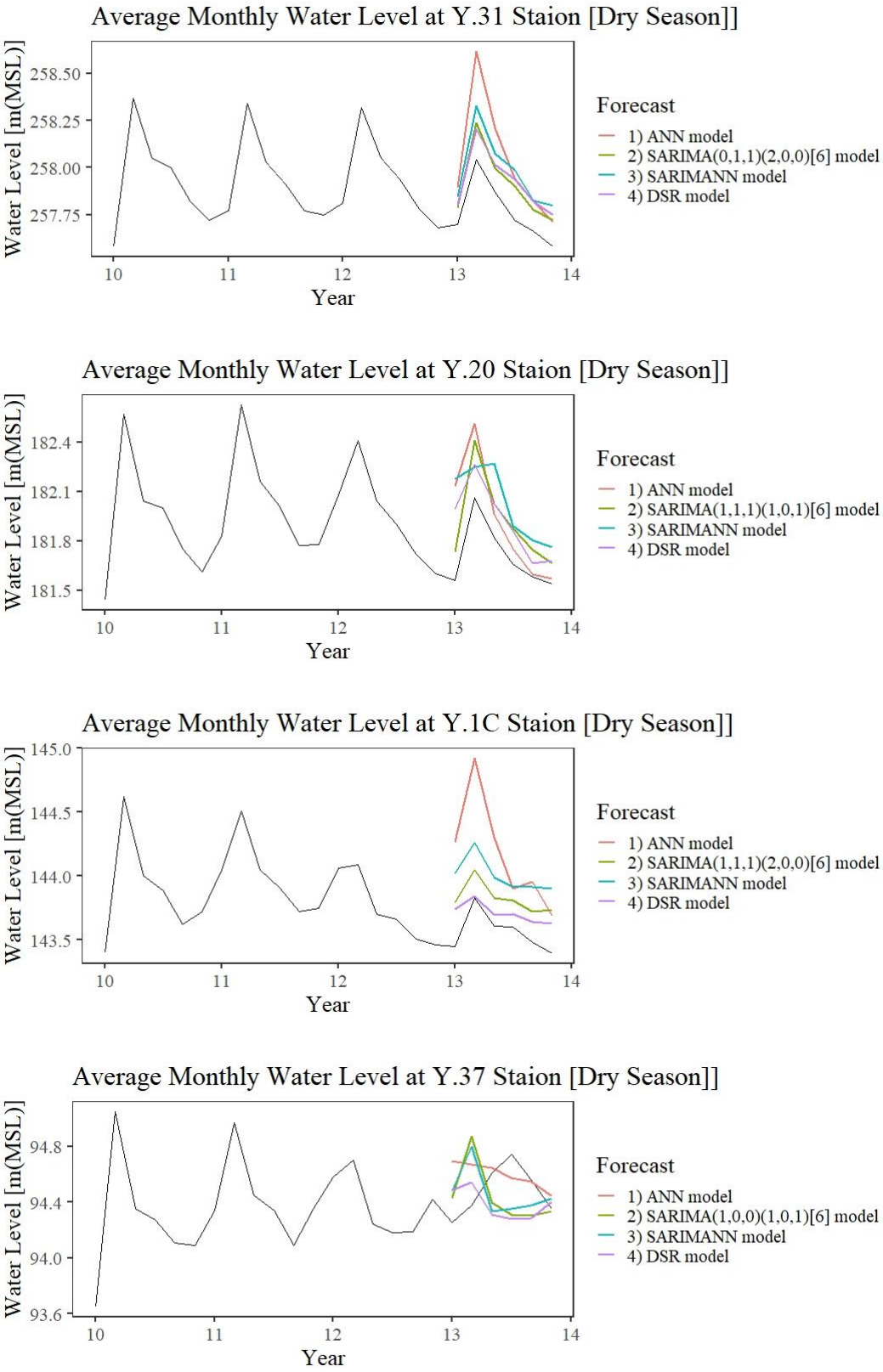

Figure 11. Forecast for one hydrological year applying the four models to AMWL data from all four stations for the dry season.

Table 7 The accuracy (RMSE, MAPE) and efficiency (percentage reduction) between all approaches of the four stations for wet and dry seasons.

		Wet Season					Dry Season				
	Model	Y.31 Station	Y.20 Station	Y.1C Station	Y.37 Station	Average	Y.31 Station	Y.20 Station	Y.1C Station	Y.37 Station	Average
RMSE	ANN	1.62	1.51	1.63	1.58	1.58	0.31	0.31	0.68	0.23	0.38
		50.14%	24.72%	33.34%	54.56%	40.86%	46.98%	22.40%	73.85%	-20.39%	43.87%
	SARIMA	0.94	1.25	1.32	0.90	1.10	0.15	0.22	0.27	0.31	0.24
		14.62%	8.79%	17.94%	19.81%	15.02%	-11.35%	-10.41%	34.01%	10.86%	9.10%
	SARIMANN	1.18	1.36	1.32	1.02	1.22	0.22	0.36	0.45	0.29	0.33
		31.45%	16.18%	17.60%	29.60%	23.07%	25.50%	33.42%	60.38%	3.30%	34.74%
	DSR	0.81	1.14	1.09	0.72	0.94	0.16	0.24	0.18	0.28	0.21
MAPE	ANN	0.55	0.73	1.04	1.43	0.94	0.10	0.12	0.43	0.18	0.21
		49.40%	28.93%	40.03%	54.23%	44.67%	40.32%	2.67%	75.69%	-42.24%	35.01%
	SARIMA	0.32	0.56	0.68	0.82	0.60	0.06	0.11	0.18	0.28	0.16
		11.51%	8.34%	8.33%	20.66%	13.02%	-12.59%	-3.55%	43.17%	7.69%	14.17%
	SARIMANN	0.37	0.60	0.60	0.90	0.62	0.08	0.18	0.31	0.27	0.21
		24.62%	13.71%	-3.49%	26.94%	15.96%	25.03%	34.09%	66.09%	5.57%	35.55%
	DSR	0.28	0.52	0.62	0.65	0.52	0.06	0.12	0.10	0.26	0.14

The accuracy of the four approaches was interpreted from the RMSE and MAPE values (Table 7). The efficiency of DSR was represented in the percentage reduction of RMSE and MAPE with respect to the ANN, SARIMA and SARIMANN models. On average, for wet season forecasts, DSR reduced the RMSE by 40.86%, 15.02%, and 23.07%, with respect to ANN, SARIMA, and SARIMANN, respectively, and reduced MAPE by 44.67%, 13.02%, and 15.96%, with respect to ANN, SARIMA, and SARIMANN, respectively. For dry season forecasts, DSR reduced RMSE by 43.87%, 9.10%, and 34.74%, with respect to ANN, SARIMA, and SARIMANN, respectively, and reduced MAPE by 35.01%, 14.17%, and 35.55%, with respect to ANN, SARIMA, and SARIMANN, respectively.

All approaches were more accurate for the dry season than the wet season but it is difficult to make a conclusion about the accuracy of upstream and downstream forecasts because of the incongruence between the RMSE and MAPE values for these forecasts. The updtream RMSEs (1.62, 0.94, 1.18, and 0.81) were slightly greater than the downstream (1.58, 0.90, 1.02, and 0.72) for the wet season for all methods, while the upstream MAPE scores (0.55, 0.32, 0.37, and 0.28) were slightly lower than the downstream scores (1.43, 0.82, 0.90 , and 0.65) for the wet season for all methods. For the dry season, the upstream MAPE scores (0.10, 0.06, 0.08, and 0.06) were slightly lower than the downstream scores (0.18, 0.28, 0.27 , and 0.26) for all methods, but the upstream RMSEs (0.31, 0.15, 0.22, and 0.16) were slightly smaller than the downstream RMSEs (0.23, 0.31, 0.29, and 0.28) for all methods except ANN.

4. Discussion and conclusions

All approaches were performed to forecast the average monthly water level (AMWL) data of all four water level measurement stations for wet seasons (May–October) and dry seasons (November–

April) from 2007 to 2020. The forecast patterns of the four approaches were similar to the historical patterns, except the upstream in wet season and downstream in dry season. ANN overestimated forecasts for almost every station in both seasons. All approaches were more accurate in the dry season than the wet season. This study proposed an efficient forecasting method called DSR, which by combining the MLR model and the ARIMA model of random error from MLR was more efficient than the ANN, SARIMA and SARIMANN approaches. On average, DSR was more efficient than the other approaches for all stations in wet and dry seasons, reducing RMASs by over 40.86%, 9.10% and 23.07% with respect to ANN, SARIMA and SARIMANN, and reducing the MAPEs by over 35.01%, 13.02% and 15.96% with respect to ANN, SARIMA and SARIMANN. The RMSEs of upstream were higher than the downstream MAPE scores in the wet season for all methods, and the upstream MAPE scores were lower than the downstream MAPE scores in both seasons for all methods. The upstream RMSEs were smaller than the downstream RMSEs in the dry season for all methods except ANN.

The forecasts of the four models of AMWL data from all four water level measurement stations for the wet season (six months: May 2019 – October 2019) and dry season (six months: November 2019 – April 2020) of one hydrological year gave similar forecast patterns to the previously observed values, except the upstream in wet season and downstream in dry season. Furthermore, ANN over estimated forecasts at almost every station in both seasons. All approaches were more accurate in the dry season than the wet season but it is difficult to draw a conclusion about the accuracy of upstream and downstream forcasts because of incongruence between RMSE and MAPE results. The RMSEs of upstream was slightly greater than the downstream RMSEs in the wet season for all methods, while the MAPE scores upstream was slightly lower than the downstream MAPE scores in wet and dry seasons for all methods. Also, the RMSEs of upstream were slightly smaller than the downstream RMSEs in the dry season for all methods except ANN. The proposed forecasting approach in this study, DSR, was obtained by combining MLR and ARIMA of the random errors from MLR and was more efficent than the ANN, SARIMA and SARIMANN approaches.

Acknowledgments

The authors would like to thank the Meteorological Department and the Royal Irrigation Department of Thailand for the valuable data of this paper.

Conflict of interest

The authors declare no conflicts of interest.

References

1. Office of the National Water Resources, The National Water Resources Management Strategies (2015–2026), 2019. Available from: http://www.onwr.go.th/?page id=3684
2. Department of Water Resources, Water Situation: Drought area, 2019. Available from: http://data.dwr.go.th/dwr/watersituation/drought/index
3. Department of Water Resources, Water Situation: Flood Area, 2019. Available from: http://data.dwr.go.th/dwr/watersituation/deluge/index
4. P. Khwanmuang, R. Chinram, T. Panityakul. (2020) The Hydrological and Water Level Data in Yom River Basin of Thailand. *J Math Comput Sci* 10: 3026–3047.

5. Zhang Q, Gu X, David CY, et al (2009) Abrupt behaviors of the streamflow of the Pearl River basin and implications for hydrological alterations across the Pearl River Delta, China. *J Hydrol* 377: 274–283.
6. Zhang Q, Gu X, Singh V P, et al. (2015) Evaluation of ecological instream flow using multiple ecological indicators with consideration of hydrological alterations. *J Hydrol* 529: 711–722.
7. Zhang Q, Gu X, Singh V P, et al. (2015) Homogenization of precipitation and flow regimes across China: changing properties, causes and implications. *J Hydrol* 530: 462–475.
8. Sopipan N (2014) Forecasting Rainfall in Thailand: A Case Study of Nakhon Ratchasima Province, *Int J Environ Ecol Geol Mar Eng* 8: 712–716.
9. Fashae O A, Olusola A O, Ndubuisi I, et al. (2018) Comparing ANN and ARIMA model in predicting the discharge of River Opeki from 2010 to 2020. *River Res Appl* 35: 169–177.
10. Siripanich P. Time series forecasting using a combined ARIMA and artificial neural network model, 2019. Available from: http://www.sure.su.ac.th/xmlui/bitstream/handle/123456789/11717/Fulltext.pdf?sequence=1&isAllowed=y
11. Hyndman R J, Athanasopoulos G (2018) Forecasting: Principles and Practice, 2018. Available from: https://otexts.com/fpp2/
12. Box GEP, Jenkins GM, Reinsel GC, et al. (2015) Time series analysis: Forecasting and control, 5 Eds., Hoboken, New Jersey: John Wiley & Sons.
13. Fu C, Ding F, Li Y, et al. (2021) Learning dynamic regression with automatic distractor repression for real-time UAV tracking, Engineering Applications of Artificial Intelligence. 98: 104116.
14. Hydro - Informatics Institute (Public Organization), Implementation of data collection and analysis of data on the 25 river basin data warehouse of the system development project, and modeling flood and drought (Yom River basin), 2019. Available from: http://www.thaiwater.net/web/attachments/25basins/08-yom.pdf
15. Bureau of Water Management and Hydrology (Royal Irrigation Department), Water allocation and cultivation plan for irrigation area during the dry season in 2019 - 2020, 2019. Available from: http://water.rid.go.th/hwm/wmoc/planing/dry/manage water2562-63.pdf
16. Upper Northern Region Irrigation Hydrology Center (Royal Irrigation Department), Average daily of water level - water content data, 2019. Available from: http://hydro-1.rid.go.th/
17. Upper Northern Region Irrigation Hydrology Center (Royal Irrigation Department), Station history: Hydrological station of Yom River (Y.1C), Ban Nam Khong, Pa Maet sub-district, Mueang Phrae district, Phrae province, 2019. Available from: https://www.hydro-1.net/Data/STATION/History/Y1c.pdf
18. Upper Northern Region Irrigation Hydrology Center (Royal Irrigation Department), Station history: Hydrological station of Yom River (Y.20), Ban Huai Sak, Tao Pun sub-district, Song district, Phrae province, 2019. Available from: https://www.hydro-1.net/Data/STATION/History/Y20.pdf
19. Upper Northern Region Irrigation Hydrology Center (Royal Irrigation Department), Station history: Hydrological station of Yom River (Y.31), Ban Thung Nong, Sa sub-district, Chiang Muan district, Phayao province, 2019. Available from: https://www.hydro-1.net/Data/STATION/History/Y31.pdf

20. Upper Northern Region Irrigation Hydrology Center (Royal Irrigation Department), Station history: Hydrological station of Yom River (Y.37), Ban Wang Chin, Wang Chin sub-district, Wang Chin district, Phrae province, 2019. Available from: https://www.hydro-1.net/Data/STATION/History/Y37.
21. Fouli H, Fouli R, Bashir B, et al. (2018) Seasonal Forecasting of Rainfall and Runoff Volumes in Riyadh Region, KSA, KSCE Journal of Civil Engineering. 22: 2637–2647.
22. R. Nau, Statistical forecasting: notes on regression and time series analysis. Available from https://people.duke.edu/~rnau/411home.htm
23. Nualtong K, Panityakul T, Khwanmuang P, et al. (2021) A Hybrid Seasonal Box Jenkins-ANN Approach for Water Level Forecasting in Thailand, Environment and Ecology Research. 9: 93 – 106.
24. TAN PN, Steinbach M, Kumar V (2014) Introduction to data mining, Pearson Education Limited, U.S.A.

AIMS Press