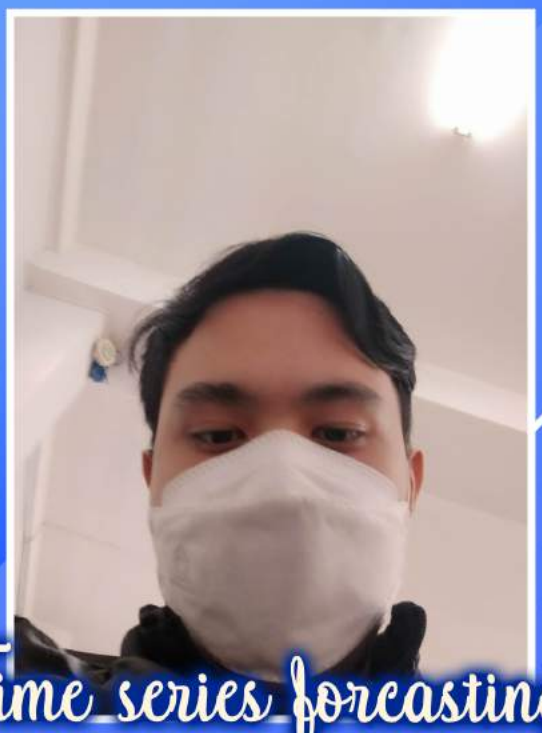
UNIVERSITY OF INFORMATION TECHNOLOGY
FACULTY OF INFORMATION SYSTEMS



Time series forcasting Using statistical methods

• NGUYỄN VĂN TRƯỜNG KHOA - 20521472

Time series forcasting using statistical methods

A case study of predicting daily rainfall in Thua Thien Hue province.

Nguyen Van Truong Khoa

University of Information Technology - VNUHCM

Ho Chi Minh City, Vietnam

20521472@gm.uit.edu.vn

***Abstract*—Presenting proposed algorithms such as: ARIMA, Linear regression, Long Short-Term Memory (LSTM) to solve the problem of predicting rainfall in the province losing to thien hue in central Vietnam, where there are only 2 dry and rainy seasons and frequent occurrences. The occurrence of droughts and floods are of great interest to the Vietnamese government and scientific community. The dataset used for testing is the statistical data of daily rainfall collected from the weather station (Station location: Latitude 16,433 Longitude 107,583 Altitude 9m) in Thua Thien Hue province in website freemeteo[12] from January 1, 2020 to December 29, 2022. Summarying preliminary knowledge of the application models, then build the application models proposed above. Divide the dataset into a train set used to train the model and a test set to test the results of the trained model. Evaluate the results through standard error such as RMSE and MAE.Predict rainfall of 30 days from 01/01/2023 to 30/01/2023. Give my own opinion with the study.**

Keywords—Linear regression model, ARIMA ,LSTM , Data Forecasting, Rainfall , Time Series.

I. INTRODUCTION.

In the tropics, where several countries only have two seasons per year (dry season and rainy season), many countries, particularly those that rely heavily on agricultural commodities, will need to forecast rainfall in order to determine the best time to begin planting their products and maximize their harvest. Another example is forecasting, which may be used by a business to estimate raw material price swings and devise the optimal plan to maximize profit. This study focuses on Rainfall forecasting in Hue province located in Vietnam from 01/01/2020 to 29/12/2022.

Climate forecast information, especially rainfall forecast, has great significance for socio-economic development activities such as planning for agricultural production, tourism, fishing and aquaculture, and management. management, efficient exploitation of water resources... The level of confidence in rainfall forecasts is often lower than that of other predictors, due to the spatial distribution and variation with The duration of precipitation depends on many other factors. Therefore, although the study of rainfall forecasting is not new, it is still of great interest in many countries around the world, including Vietnam.

Thua Thien Hue province is located in the tropical monsoon climate of Southeast Asia with high temperature, influenced by EL Nino and La Nina phenomena, so it is subject to extreme weather such as storms, floods, thunderstorms, and landslides. landslides, droughts, saltwater intrusion, forest fires... In recent years, climate change has negatively impacted the agricultural sector of Thua Thien Hue, making people's lives difficult [1]. Precipitation will be done by agro-meteorological models through historical data. These results will be the scientific basis for planning water storage for reservoirs in the region, reasonable arrangement of time periods for planting agricultural crops, crops, and guiding decisions. about crop structure. This is a very important problem in directing agricultural production in the same rural areas of Thua Thien Hue province, whose success depends heavily on the results of the total rainfall forecast in this period.

Currently, in Vietnam, seasonal drought forecasts have partly met the needs of socio-economic development and natural disaster prevention and mitigation, however, to solve the above problem, it

will be difficult. Due to the difficulty of input data sources, the rain forecast results are still qualitative, not yet quantified and moreover, the forecast period is 3 consecutive months, not really suitable for the requirements of the agricultural problem in this region.

Currently, there are two approaches to study the forecast of rainfall: statistical method and dynamic model method. In general, the traditional statistical method has achieved certain results, many statistical models have the main contribution in making the forecast of total rainfall. The dynamic modeling method is a research direction that is receiving great attention to develop, it has outstanding advantages in terms of providing predictive products, however, solving numerical models is very complicated and expensive. requires a highly configurable computer tool and moreover, the rain forecast results have not yet achieved the desired accuracy.

A linear regression model is a model that analyzes the relationship between a dependent variable and one or more independent variables. A linear regression trendline uses the least squares method to plot a straight line through values so as to minimize the distances between the values and the resulting trendline. This linear regression indicator plots the trendline value for each data point.

The ARIMA model is a time series analysis model, it not only considers the self-moving cycles of the forecast data series, the interactions in the self-moving process of other influencing factors but also evaluate the error rules in the simulation process to improve the accuracy of the forecast. Although this model has been applied in many countries around the world, in Vietnam so far there are still very few studies applied in seasonal climate forecasting.

Long Short-Term Memory (LSTM) is an artificial neural network widely used in the field of artificial intelligence and deep learning. As there can be unknown delay between significant events in time series, LSTM network is well suited for classification, processing and making predictions based on time series data . To verify this, it is included in this report.

II. RELATED WORK.

There are a large number of researches and studies in regard to the prediction of rainfall, for which remains a popular topic in both literature and industry.

M Sidiq (2018) [2] used monthly rainfall of 48 data got from Badan Meteorologi dan Geofisika (BMG) Bandung from January 2011 to December 2013 to built ARIMA model for forcasting rainfall of Bandung in Indonesia. By pointing out the limitation of ARIMA model, it can just forecast immediate future.

Pazvakawambwa G.T. and Ogunmokun A. A. (2010) [3] used monthly Windhoek rainfall data from 1891 to 2011 to built ARIMA model for predicting the Windhoek rainfall in Namibia.

D.A Attah, G.M.Bankole (2011) [4] used rainfall history data of Kaduna South meteorological station within the Lower Kaduna catchment for a period of 47 years (1960 –2006) to built ARMA model.

Ebenezer Afrifa-Yamoah , Bashiru I. I. Saeed, Azumah Karim (2016) [5] forecast monthly rainfall in the Brong Ahafo Region of Ghana by using SARIMA model

Yashon O. Ouma, Rodrick Cheruyot & Alice N. Wachera (2022) [6] compares LSTM neural network and wavelet neural network (WNN) for spatio-temporal prediction of rainfall and runoff time-series trends in scarcely gauged hydrologic basins.

Mohammad Mahmudur Rahman Khan, Md. Abu Bakr Siddique, Shadman Sakib, Anas Aziz, Ihtyaz Kader Tasawar, Ziad Hossain [7] predicted Temperature and Rainfall in Bangladesh using Long Short Term Memory and Recurrent Neural Networks

Agalya, Annapoorani., Arundhati, C. Geetha (2019) [12] predict rainfall by using linear regression

S. Prabakaran, P. Naveen Kumar and P. Sai Mani Tarun (2017) [8] predict rainfall in by using linear regression

Bahareh Karimi, Mir Jafar Sadegh Safari, Ali Danandeh Mehr, Mirali Mohammadi (2019) [9] predicted monthly rainfall using ARIMA and Gene Expression

Charity Oseiwah Adjei, Wei Tian, Bernard-Marie Onzo, Emmanuel Adu Gyamfi Kedjanyi, Oscar Famous Darteh (2021) [10] forecated rainfall in Sub-Sahara Africa-Ghana using LSTM Deep Learning Approach

MAI Navid, NH Niloy (2017) [11] using multiple linear regressions for predicting rainfall for Bangladesh

III. DATA USED.

A. Data overview.

This study predicts daily rainfall of Thua Thien Hue province. The data are collected from the weather station (Station location: Latitude 16,433 Longitude 107,583 Altitude 9m) in Thua Thien Hue province in website freemeteo[12], consisting of 1094 observations of daily rainfall from 01/01/2020 to

29/12/2022 and the raifall measurement is the total rainfall depth during a given period, expressed in millimeters (mm).

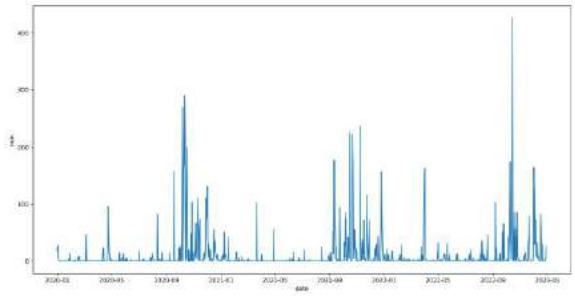


Fig 1.Plot of data of daily rainfall in Thua Thien Hue province from 01/01/2020 to 29/12/2022.

TABLE 1. DATA ABSTRACT

Rainfall	
Mean	10.60027422
Standard Error	1.032655851
Median	0
Mode	0
Standard Deviation	34.15578499
Sample Variance	1166.617649
Kurtosis	40.13953808
Skewness	5.590052022
Range	425.9
Minimum	0
Maximum	425.9
Sum	11596.7
Count	1094

B. Train and test data.

To train and evaluate the trained model, the data set is divided by me in 4 cases:

First case i call it is *6:4 case*, 60% for train and 40% for test as shown in the following image

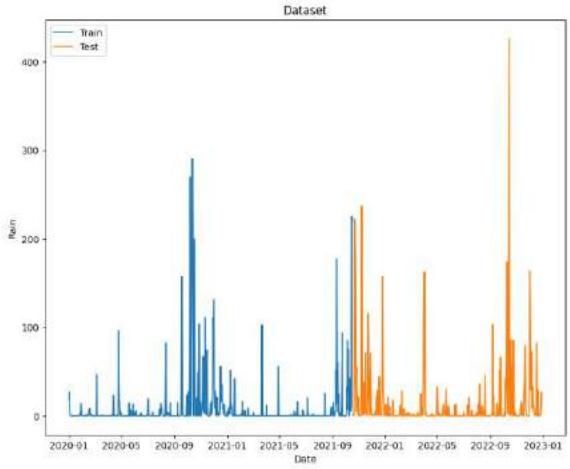


Fig 2.Plot of train and test data in case 6:4.

Second case i call it is *7:3 case*, 70% for train and 30% for test as shown in the following image

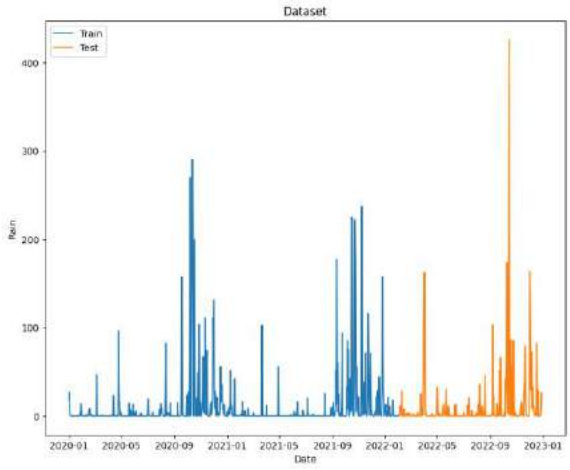


Fig 3.Plot of train and test data in case 7:3.

Third case i call it is *8:2 case*, 80% for train and 20% for test as shown in the following image

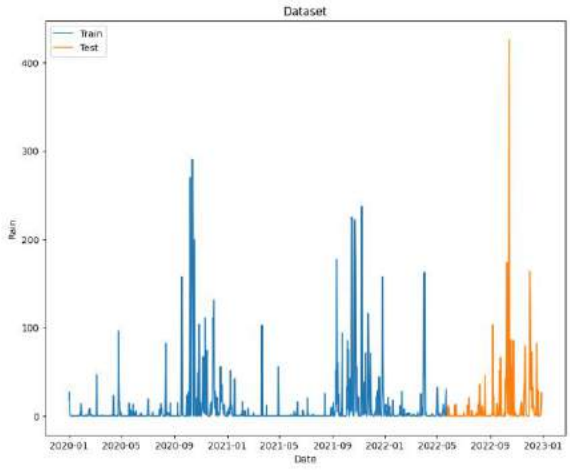


Fig 4.Plot of train and test data. in case 8:2.

Fourth case i call it is *9:1 case*, 90% for train and 10% for test as shown in the following image

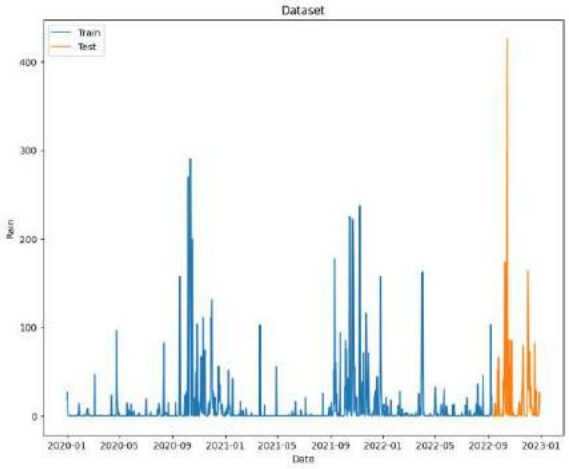


Fig 5.Plot of train and test data in case 9:1.

IV. APPLICATION OF LINEAR REGRESSION.

Linear regression is a widely used method that fits a data set to a model in which the forecasted variable y depends linearly on a number of predictor variable x. This multiple linear regression model can be expressed as,

$$Y_t = \beta_0 + \beta_1 * X_t + \varepsilon_t \quad (1)$$

The coefficients β_0 and β_1 denote the intercept and the slope of the line respectively. ε denote a deviation from the underlying straight line model.

After performing the formulas of linear regression by using tool in python program, the results show the parameters of intercept and confienct variables as shown below:

TABLE 2. LINEAR REGRESSION PARAMETER

Case	Intercept	Slope
6:4	5.53649488	0.01060026
7:3	4.66210413	0.01511306
8:2	6.79706849	0.00719206
9:1	8.55464343	0.00140794

Thus, it can be deduced that the equation of linear regression is

- Case 6:4:

  $y = 5.53649488 + 0.01060026x$

- Case 7:3:

  $y = 4.66210413 + 0.01511306x$

- Case 8:2:

  $y = 6.79706849 + 0.00719206x$

- Case 9:1:

  $y = 8.55464343 + 0.00140794x$

V. APPLICATION OF ARIMA

A.Methods used .

A time series is a series of data observed over time. ARIMA models are a type of model that can describe both stationary and non-stationary time series and make reliable forecasts based on a description of historical data for a single variable. This approach differs from previous forecasting models in that it does not presume any specific pattern in the past data of the time series to be forecasted. The BoxJenkins methodology for building ARIMA models is based on the following steps: (1) Model Identification, (2) Parameter Estimation and Selection, (3) Diagnostic Checking (or Modal Validation); and (4) Model's use.

Step 1: Model identification

Model identification involves determining the orders (p, d, and q) of the AR and MA components of the model. Basically it seeks the answers for whether data is stationary or nonstationary? What is the order of differentiation (d), which makes the time stationary?

Step 2: Parameter Estimation and Selection

The parameters of the model will estimated by the Least squares method

Step 3: Modal Validation

After determining the parameters of the for the ARIMA process, the next thing to do is to proceed test whether the error term εt of the model is white noise?

This is a condition of a good model (Wang & Lim, 2005 [9])

Step 4: Model's use

Based on the equation obtained from the model, determine the price forecast value and the confidence interval of the forecast.

B. Stationary test.

In mathematics, stationarity is used as a tool in time series analysis. To form a complete model of statistical significance, the time series of data first needs to check the stationarity of the series. A stationary process has the property that themean, variance and autocorrelation structure do not change over time. In fact, most economic data series (base series) are non-stationary. This means that those time series have a time-varying sample mean and variance.

But when we take the difference, the time series often become stationary series.

The data series used in the ARIMA model is assumed to be stationary. Because reliably, to predict the rainfall by using the ARIMA model, we need to consider whether the data series are stationary series or not. To confirm this, it is possible to first rely on direct observation of the graph of the data series, and then test it. One of the two most popular testing methods is the Augmented Dickey-Fuller (ADF) which econometricians call the unit root test for original data series and difference series.

The formula for the ADF test is represented as follows:

$$\text{ADF} = \Delta y_t - \beta_1 \Delta y_{t-1} - \beta_2 \Delta y_{t-2} \text{-} \ldots \text{-} \beta_p \Delta y_{t-p} \quad (2)$$

Inside:

- y_t is the value of the time series at time t
- Δy_t is the difference between Yt value and Yt-1 value
- $\beta_1, \beta_2, \ldots \beta_p$ are the parameters of the model.

The ADF test uses the above formula to calculate the ADF value and compares this value with a threshold critical value. If the ADF is less than the critical value, the time series is considered to be a static series. Conversely, if the ADF is greater than the critical value, then the time series is not a static series.

Figure 1 shows that the daily rainfall variation (in milliliters) (01/01/2020-29/12/2022) is stable and does not tend to increase or decrease, specifically, its average tends to not increase or decrease over time. Thus, it can be speculated that the rainfall value chain is stationary.

TABLE 3. ADF TEST RESULT FOR THE ORIGINAL SERIES.

Data	Case	ADF value	p-value
Original series	**6:4**	-5.844	0.000
	7:3	-7.137	0.000
	8:2	-7.913	0.000
	9:1	-8.378	0.000

Note: The critical values at 1%, 5% and 10% statistical significance are -3.44, -2.87 and -2.57 respectively.

For the original series, the above test values all exceed the critical value at the 1%, 5% and 10% significance levels. Therefore, the hypothesis H0 (the original series is a stationary series) is rejected, i.e. the Original series is a stationary series and d parameter in ARIMA is 0 in all case.

C. ARIMA model.

ARIMA is an algorithm consisting of 3 algorithms including: Autoregressive (AR), Integrated (I), Moving Average (MA).

Integrated (I) is the process of co-integration or differentiation. Due to the requirements of the ARIMA algorithm, the sequence must be stationary. The simplest way to form a stationary series is to differentiate until there is a stationary series. The order of the generation difference to create a stationary series is d, the s-difference process d is shown as follows:

Difference of order 1:

$$\text{I(1)} = \Delta(x_t) = x_t - x_{t-1} \quad (3)$$

Difference of order d:

$$\text{I(d)} = \Delta^d(x_t) = \underbrace{\Delta(\Delta(\ldots\Delta(x_t)))}_{\text{d time}} \quad (4)$$

Autoregressive (AR) indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. Thus, an autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t \quad (5)$$

where ε_t is white noise. This is like a multiple regression but with *lagged values* of as y_t predictors. We refer to this as an AR**(**p**)** model, an autoregressive model of order p.

Moving Average (MA) indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. An Moving Average model of order p can be written as

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \quad (6)$$

where is white noise. We refer to this as an MA(q) model, a moving average model of order q. Of course, we do not *observe* the values of , so it is not really a regression in the usual sense.

As we observe from Figure 7,8,9, the maximum horizontal value of both Autocorrelation and Partial Autocorrelation is less than 35. So p and q can be in the set of integer values [0.35] To estimate the coefficients of the ARIMA(p,0,q) models as identified above, the Python programming language was used.

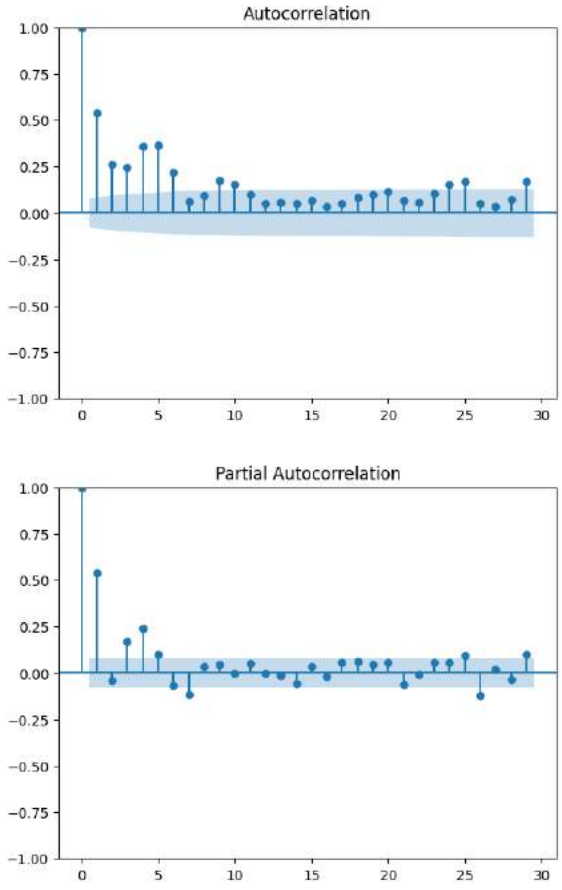
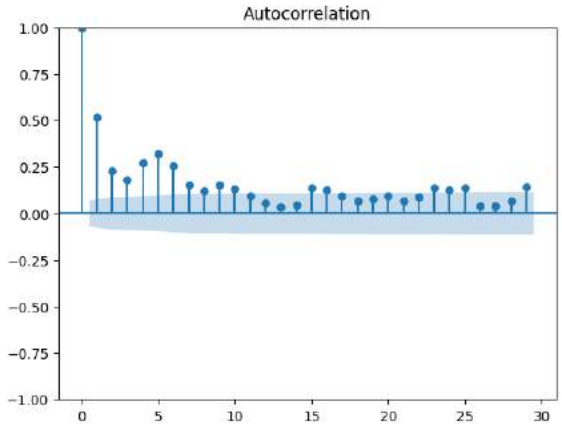


Fig 6.Correlation and Partital Autocorrelation of the chain of fluctuations in rainfall in case 6:4.



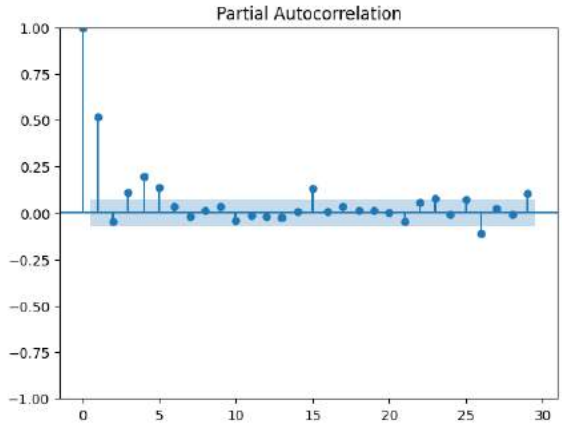


Fig 7.Correlation and Partital Autocorrelation of the chain of fluctuations in rainfall in case 7:3.

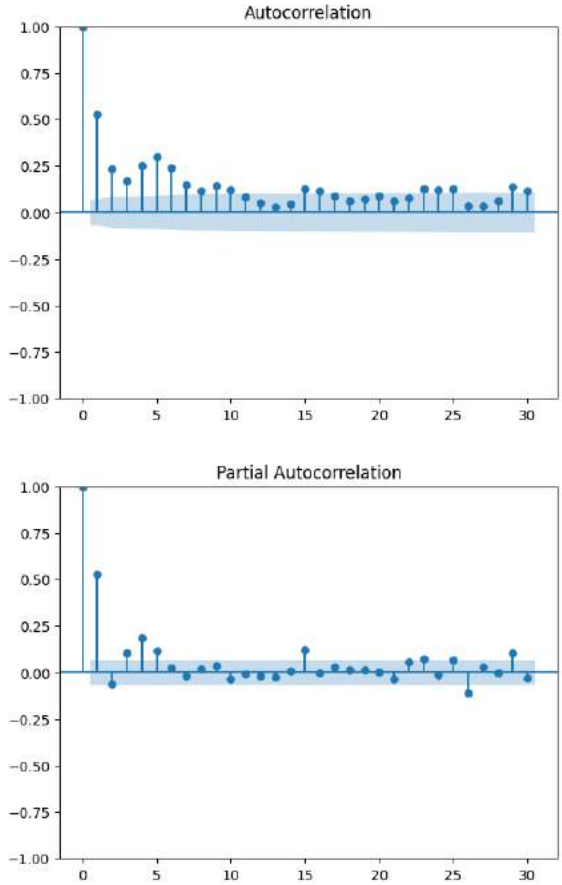


Fig 8.Correlation and Partital Autocorrelation of the chain of fluctuations in rainfall in case 8:2.

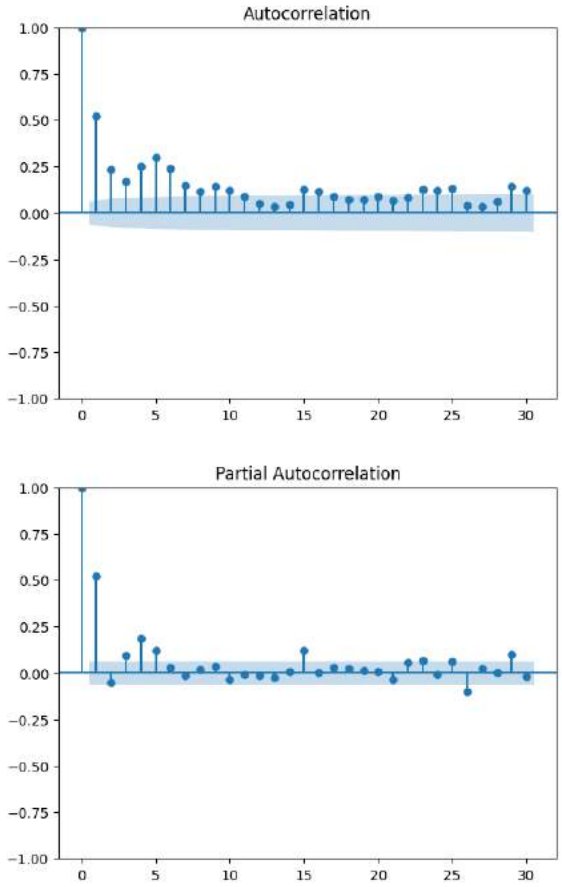


Fig 9.Correlation and Partital Autocorrelation of the chain of fluctuations in rainfall in case 9:1.

To check the suitability of the models, we rely on the Akaike Information Criteria (AIC) as small as possible. By using python program tool which is autoARIMA, I have AIC result of all unerror ARIMA model as following table:

TABLE 4. AIC VALUE OF UNERROR ARIMA MODEL IN CASE 6:4

Model	**AIC value**
ARIMA(0,0,0)	6443.245
ARIMA(1,0,0)	6164.956
ARIMA(0,0,1)	6228.954
ARIMA(2,0,0)	6166.914
ARIMA(1,0,1)	6166.846
ARIMA(2,0,1)	6168.619
ARIMA(1,0,0)	6152.806
ARIMA(0,0,0)	6393.813
ARIMA(2,0,0)	6154.675
ARIMA(1,0,1)	6154.539
ARIMA(0,0,1)	6198.951
ARIMA(2,0,1)	6155.836

TABLE 5. AIC VALUE OF UNERROR ARIMA MODEL IN CASE 7:3

Model	**AIC value**
ARIMA(0,0,0)	7610.242
ARIMA(1,0,0)	7326.498
ARIMA(0,0,1)	7379.271
ARIMA(2,0,0)	7328.207
ARIMA(1,0,1)	7327.975
ARIMA(2,0,1)	7328.988
ARIMA(1,0,0)	7306.869
ARIMA(0,0,0)	7540.787
ARIMA(2,0,0)	7307.072
ARIMA(1,0,1)	7306.22
ARIMA(0,0,1)	7337.755
ARIMA(2,0,1)	7307.693
ARIMA(1,0,2)	7278.931
ARIMA(0,0,2)	7310.312
ARIMA(2,0,2)	inf
ARIMA(1,0,3)	7279.097
ARIMA(0,0,3)	7311.384
ARIMA(2,0,3)	7278.962
ARIMA(1,0,2)	7282.456

TABLE 6. AIC VALUE OF UNERROR ARIMA MODEL IN CASE 8:2

Model	**AIC value**
ARIMA(0,0,0)	8644.926
ARIMA(1,0,0)	8303.616
ARIMA(0,0,1)	8365.798
ARIMA(2,0,0)	8304.464
ARIMA(1,0,1)	8303.771
ARIMA(2,0,1)	8304.712
ARIMA(1,0,0)	8282.481
ARIMA(0,0,0)	8567.681
ARIMA(2,0,0)	8280.919
ARIMA(3,0,0)	8273.667
ARIMA(4,0,0)	8245.929
ARIMA(5,0,0)	8235.756
ARIMA(6,0,0)	8237.178
ARIMA(5,0,1)	8237.29
ARIMA(4,0,1)	8236.775
ARIMA(6,0,1)	8238.983
ARIMA(5,0,0)	8244.629

TABLE 7. AIC VALUE OF UNERROR ARIMA MODEL IN CASE 9:1

Model	**AIC value**
ARIMA(2,0,0)	9251.223
ARIMA(1,0,1)	9250.809
ARIMA(2,0,1)	9251.803
ARIMA(1,0,0)	9226.566
ARIMA(0,0,0)	9539.059
ARIMA(2,0,0)	9225.827
ARIMA(3,0,0)	9218.599
ARIMA(4,0,0)	9186.839
ARIMA(5,0,0)	9174.96
ARIMA(6,0,0)	9176.22
ARIMA(5,0,1)	9176.361
ARIMA(4,0,1)	9175.997
ARIMA(6,0,1)	9177.988
ARIMA(5,0,0)	9184.862
ARIMA(2,0,0)	9251.223
ARIMA(1,0,1)	9250.809
ARIMA(2,0,1)	9251.803
ARIMA(1,0,0)	9226.566

After estimating the ARIMA models and have the results in table. We choose the model ARIMA(6,0,0) for case 6:4, 8:2, 9:1 and model ARIMA(7,0,1) for case 7:3 as the best model best fit to the dataset with minial AIC value.

SARIMAX Results

Dep. Variable:	y	No. Observations:	656
Model:	SARIMAX(1, 0, 0)	Log Likelihood	-3073.403
Date:	Sun, 01 Jan 2023	AIC	6152.806
Time:	17:35:51	BIC	6166.264
Sample:	01-01-2020	HQIC	6158.024
	- 10-17-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	4.0026	2.044	1.958	0.050	-0.003	8.008
ar.L1	0.5760	0.015	38.370	0.000	0.547	0.605
sigma2	686.2758	15.199	45.153	0.000	656.487	716.065

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	24285.22
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	7.68	Skew:	4.00
Prob(H) (two-sided):	0.00	Kurtosis:	31.72

The summary of Best model ARIMA(1,0,0) model with minimize AIC in case 6:4.

SARIMAX Results

Dep. Variable:	y	No. Observations:	765
Model:	SARIMAX(1, 0, 2)	Log Likelihood	-3634.465
Date:	Sun, 01 Jan 2023	AIC	7278.931
Time:	17:44:44	BIC	7302.130
Sample:	01-01-2020	HQIC	7287.862
	- 02-03-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.7156	0.615	1.164	0.244	-0.489	1.921
ar.L1	0.9305	0.020	45.836	0.000	0.891	0.970
ma.L1	-0.4408	0.029	-15.458	0.000	-0.497	-0.385
ma.L2	-0.3050	0.020	-15.111	0.000	-0.345	-0.265
sigma2	783.1812	17.430	44.934	0.000	749.020	817.343

Ljung-Box (L1) (Q):	0.11	Jarque-Bera (JB):	23190.77
Prob(Q):	0.74	Prob(JB):	0.00
Heteroskedasticity (H):	11.92	Skew:	3.95
Prob(H) (two-sided):	0.00	Kurtosis:	28.79

The summary of Best model ARIMA(1,0,2) model with minimize AIC in case 7:3.

SARIMAX Results

Dep. Variable:	y	No. Observations:	875
Model:	SARIMAX(5, 0, 0)	Log Likelihood	-4110.878
Date:	Sun, 01 Jan 2023	AIC	8235.756
Time:	17:54:47	BIC	8269.176
Sample:	01-01-2020	HQIC	8248.540
	- 05-24-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	3.2251	1.914	1.685	0.092	-0.526	6.977
ar.L1	0.5286	0.017	31.448	0.000	0.496	0.562
ar.L2	-0.0990	0.019	-5.297	0.000	-0.136	-0.062
ar.L3	0.0100	0.020	0.493	0.622	-0.030	0.050
ar.L4	0.1179	0.023	5.087	0.000	0.072	0.163
ar.L5	0.1174	0.019	6.305	0.000	0.081	0.154
sigma2	704.6709	14.598	48.273	0.000	676.060	733.282

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	29200.44
Prob(Q):	0.94	Prob(JB):	0.00
Heteroskedasticity (H):	1.57	Skew:	3.99
Prob(H) (two-sided):	0.00	Kurtosis:	30.15

The summary of Best model ARIMA(5,0,0) model with minimize AIC in case 8:2.

SARIMAX Results

Dep. Variable:		y	No. Observations:			984
Model:	SARIMAX(5, 0, 0)		Log Likelihood			-4580.480
Date:	Sun, 01 Jan 2023		AIC			9174.960
Time:	18:03:14		BIC			9209.201
Sample:	01-01-2020		HQIC			9187.984
	- 09-10-2022					
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
intercept	3.0357	1.699	1.787	0.074	-0.293	6.365
ar.L1	0.5156	0.015	34.149	0.000	0.486	0.545
ar.L2	-0.0857	0.017	-5.015	0.000	-0.119	-0.052
ar.L3	0.0048	0.018	0.258	0.796	-0.031	0.041
ar.L4	0.1191	0.021	5.640	0.000	0.078	0.161
ar.L5	0.1194	0.017	7.020	0.000	0.086	0.153
sigma2	646.4082	12.083	53.498	0.000	622.726	670.090
Ljung-Box (L1) (Q):	0.01		Jarque-Bera (JB):		38085.99	
Prob(Q):	0.93		Prob(JB):		0.00	
Heteroskedasticity (H):	0.92		Skew:		4.14	
Prob(H) (two-sided):	0.44		Kurtosis:		32.33	

The summary of Best model ARIMA(5,0,0) model with minimize AIC in case 9:1.

From the summary of model ARIMA in all case, The corresponding expression of the model is

- Case 6:4:

$$y_t = 4.0026 + 0.5760y_{t-1}$$

- Case 7:3:

$$y_t = 0.7156 + 0.9305y_{t-1} - 0.4408\varepsilon_{t-1} - 0.3050\varepsilon_{t-2}$$

- Case 8:2:

$$y_t = 3.2251 + 0.5286y_{t-1} - 0.0990y_{t-2} + 0.0100y_{t-3} + 0.1179y_{t-4} + 0.1174y_{t-5}$$

- Case 9:1:

$$y_t = 3.0357 + 0.5156y_{t-1} - 0.0857y_{t-2} + 0.0048y_{t-3} + 0.1191y_{t-4} + 0.1194y_{t-5}$$

VI. APPLICATION OF LSTM

Long short-term memory networks (LSTM) are a type of recurrent neural network (RNN), which is a general name for a class of neural networks that can handle sequential input. LSTM is a network structure that consists of three "gate" components (shown in Fig. 10). An LSTM unit contains three gates: an input gate, a forgetting gate, and an output gate. Rules can be applied to information as it enters the LSTM network. Only information that adheres to the algorithm will be retained, while information that does not conform will be erased via the forgetting gate.

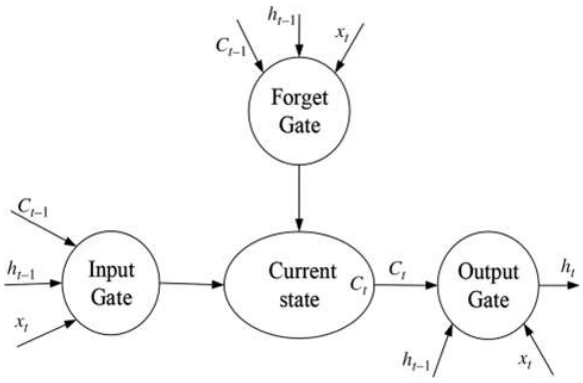


Fig 10. LSTM unit structure.

The gate permits input to be transferred selectively, while (7) depicts the sigmoid activation function of the LSTM network. Through the gating unit, the LSTM may add and delete information for neurons. It comprises of a Sigmoid neural network layer and a pair multiplication operation to assess whether information is sent or not. Each Sigmoid layer element is a real integer between [0, 1], denoting the weight via which the associated information goes. There is also a layer in the LSTM neural network that contains the tanh activation function, as stated in (8). It is utilized to maintain the condition of neurons.

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (7)$$

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (8)$$

The LSTM neural network's forgetting gate selects what information should be deleted, which reads h_{t-1} and x_t and provides the neuron state C_{t-1} a value of 0-1. The calculating technique for forgetting probability is shown in (9).

$$f_t = \sigma(w_f \cdot [h_{t-1}, x_t] + b_f) \quad (9)$$

where h_{t-1} represents the output of the previous neuron and x_t is the input of the current neuron. is the sigmoid function.

The amount of fresh information added to the neuron state is determined by the input gate. As demonstrated in (9), the input layer containing the sigmoid activation function first identifies which information needs to be updated, and then a tanh layer provides candidate vectors $\hat{c}_t$, an update to the neuron's state is made, as shown in (10).

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \quad (10)$$

where the calculation methods of i_t and $\hat{c}_t$ are shown in (11) and (12)

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_i) \quad (11)$$

$$\hat{C}_t = tanh\,(w_c \cdot [h_{t-1}, x_t] + b_c) \qquad (12)$$

The output gate controls how many current neural units states and how many controlling units states are filtered, as illustrated in (13) and (14).

$$\sigma_t = \sigma(w_o \cdot [h_{t-1}, x_t] + b_o) \qquad (13)$$

$$h_t = O_t{}^{*} tanh\,(C_t) \qquad (14)$$

Regarding the input, my idea is to get the data of the previous day to predict for the next day, i.e. $x_t = y_{t-1}$. Since we get the data first to begin with, I decided to remove the first element of the train and test datasets.

After testing many cases, I decided to choose the hyperparameter to apply to the tool available in the python program as follows.

TANLE 8.LSTM PARAMETER

Parameter	**Values**
Number of cell	365
Number of input	1
Activation funtion	Rectified Linear Activation Function (relu)
Layer output	1
Loss value	'mean_squared_error'
Epochs	100
Optimizer	'adam'

VII. Methods of evaluating results.

A. Evaluating prediction results in test dataset.

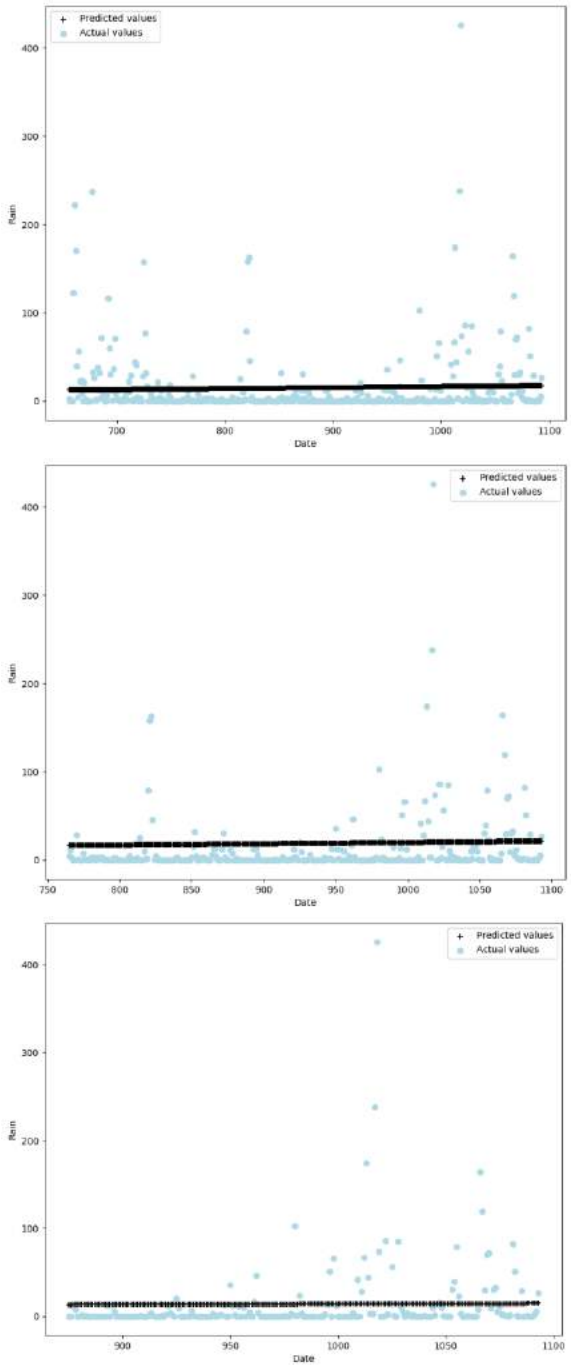
To evaluate the salinity prediction efficiency of the model, this study uses using evaluation indicators: NSE (Nash–Sutcliffe efficiency coefficient) and mean error root squared error (RMSE–Root Mean Squared Error). The NSE index represents the degree of association connection between real measured and simulated values, ranging from –∞ to 1, the closer the value is to 1, the greater the degree. The accuracy of the model is higher. The RMSE index represents the difference between the predicted and observed values, the lower the value, the better the model (ranging from 0 to ∞)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \qquad (15)$$

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \qquad (16)$$

Where: y_j is the jth sample real value, $\hat{y}_j$ is the jth sample prediction value, n is the number of samples used for evaluation.

Visually, we can see the accuracy of the prediction results that the models give compared to the actual results through the fig 11,12,13.



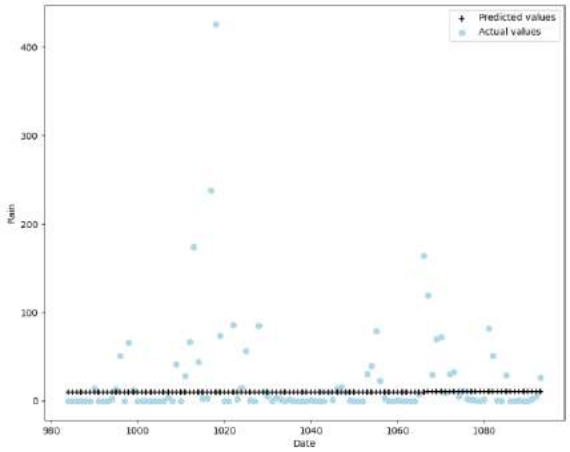
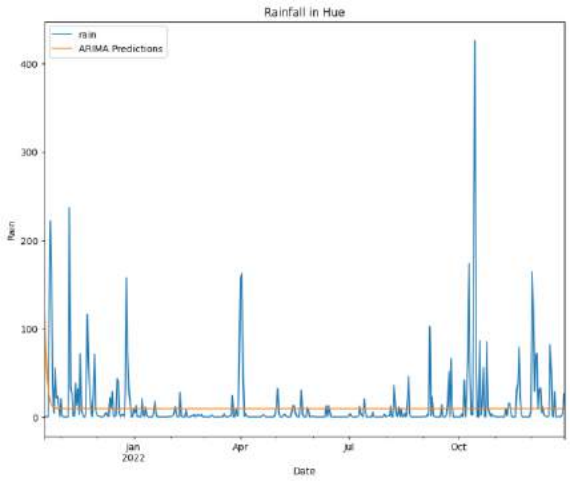
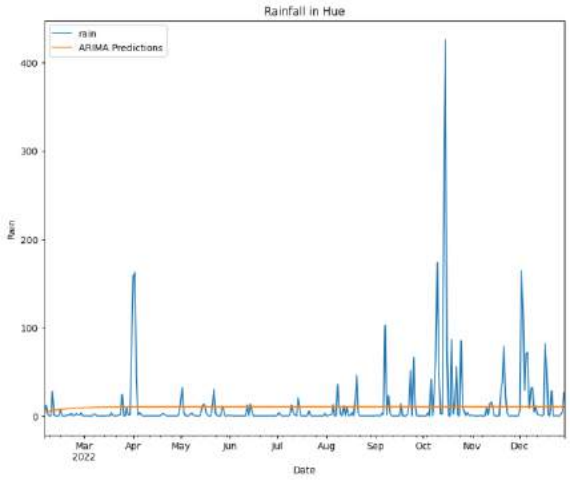
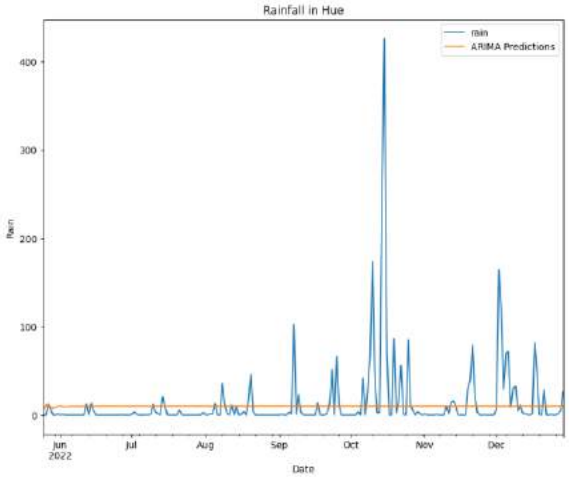


Fig 11. Linear regression prediction and actual plot of test dataset with cases in top-down order of case 6:4, 7:3, 8:2, 9:1.







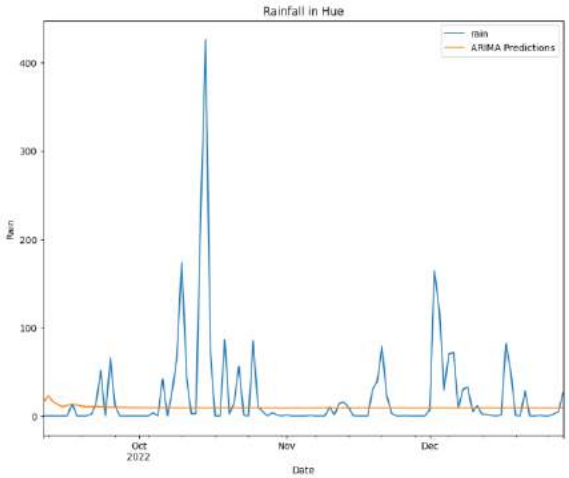
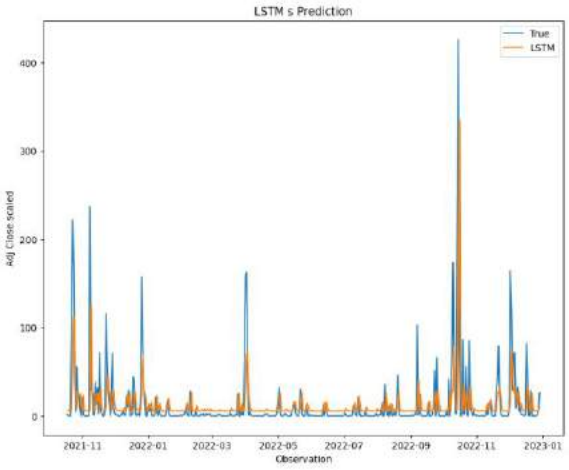


Fig 12. ARIMA prediction and actual plot of test dataset with cases in top-down order of case 6:4, 7:3, 8:2, 9:1.



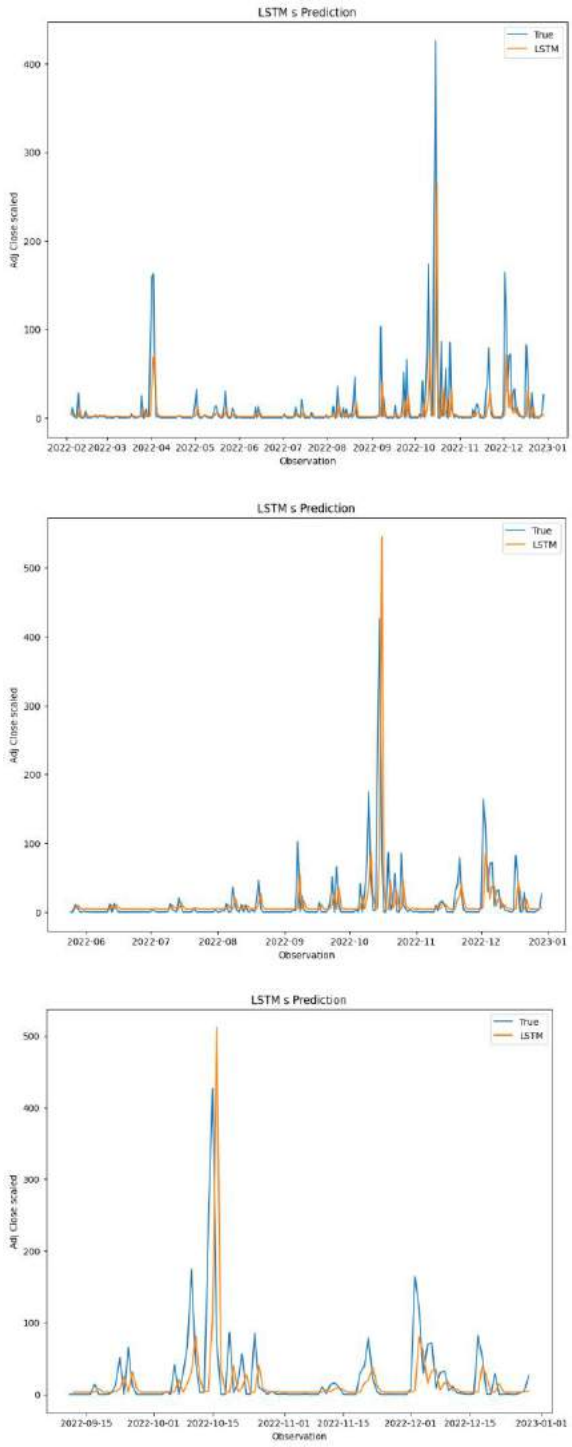


Fig 13. LSTM prediction and actual plot of test dataset with cases in top-down order of case 6:4 in scale[0,1], 7:3 in scale[-1,0], 8:2 in scale[0,1], 9:1 in scale[0,1].

Alternatively, the accuracy can be assessed through standard indicators such as RMSE or MAE shown in Table 9.

TABLE 9.EVALUATING RESULTS BY RMSE, MAE

Method	**Case**		**RMSE**	**MAE**
LR	6:4		37.630	19.916
	7:3		36.503	21.659
	8:2		40.525	19.577
	9:1		55.928	24.663
ARIMA	6:4		38.286	17.168
	7:3		35.873	15.774
	8:2		40.734	17.247
	9:1		56.174	24.752
LSTM	6:4	Scale[0,1]	0.065	0.031
		Scale[-1,0]	0.157	0.15
		Scale[-1,1]	0.147	0.112
	7:3	Scale[0,1]	0.058	0.036
		Scale[-1,0]	0.065	0.023
		Scale[-1,1]	0.159	0.051
	8:2	Scale[0,1]	0.061	0.028
		Scale[-1,0]	0.07	0.027
		Scale[-1,1]	0.176	0.055
	9:1	Scale[0,1]	0.084	0.04
		Scale[-1,0]	0.095	0.06
		Scale[-1,1]	0.248	0.099

As we can see from the table, I decided to choose linear model in case 7:3, ARIMA model in case 7:3, LSTM model in case 7:3 with scale[-1,0]

B. Predicting real data in next 30 day.

TABLE 10. PREDICTING RAINFALL IN HUE 30 DAYS AGO

Date	LR	ARIMA	LSTM
1/1/2023	21.22602	10.41809	8.23361266
1/2/2023	21.24113	10.41809	5.5218051
1/3/2023	21.25625	10.41809	2.56468204
1/4/2023	21.27136	10.41809	1.03207135
1/5/2023	21.28647	10.41809	2.10692063
1/6/2023	21.30159	10.41809	3.53152583
1/7/2023	21.3167	10.41809	2.41882738
1/8/2023	21.33181	10.41809	2.05378536
1/9/2023	21.34693	10.41809	4.74510673
1/10/2023	21.36204	10.41809	9.41597322
1/11/2023	21.37715	10.41809	9.73880054
1/12/2023	21.39226	10.41809	6.27443474
1/13/2023	21.40738	10.41809	0.28364462
1/14/2023	21.42249	10.41809	-2.38337795
1/15/2023	21.4376	10.41809	0.85547153
1/16/2023	21.45272	10.41809	4.48820507
1/17/2023	21.46783	10.41809	4.0078822
1/18/2023	21.48294	10.41809	3.09224119
1/19/2023	21.49806	10.41809	8.38217644
1/20/2023	21.51317	10.41809	17.2624901
1/21/2023	21.52828	10.41809	11.47367272
1/22/2023	21.5434	10.41809	5.64286436
1/23/2023	21.55851	10.41809	-1.53360231
1/24/2023	21.57362	10.41809	-4.47691719
1/25/2023	21.58873	10.41809	0.64547059
1/26/2023	21.60385	10.41809	5.72444897
1/27/2023	21.61896	10.41809	4.46583002
1/28/2023	21.63407	10.41809	2.83987086
1/29/2023	21.64919	10.41809	9.64991597
1/30/2023	21.6643	10.41809	25.98040007

V. CONCLUSION.

In this paper, three models including: ARIMA, Linear regression, LSTM are proposed to predict daily precipitation. Model structure, algorithm framework and experimental design are presented. The feasibility and accuracy in applying the LSTM to the problem of solving the problem of precipitation forecasting in the lowland province of Hue was verified by comparing the model with the arima model and the linear regression model through the analysis methods. error standards such as RMSE, MAE. Experiments show that the average accuracy of the LSTM is not only better than that of the other two models. Furthermore, the average accuracy of each prediction value is more than 95%. Although the model has relatively good performance, the above model can only predict with high accuracy with rainfall below 50 mm, and for heavy rains, the accuracy is less than 50%. It can be due to the effects of erratic climate changes or man-made factors such as increased emissions, etc. However, I firmly believe that being able to explore, learn and collect data of many other factors can improve the accuracy of the models.

VI. Reference.

[1] Văn Dinh ,"Thừa Thiên Huế: Ngành nông nghiệp chủ động ứng phó với biến đổi khí hậu",23/12/2019, https://baotainguyenmoitruong.vn/thua-thien-hue-nganh-nong-nghiep-chu-dong-ung-pho-voi-bien-doi-khi-hau-297164.html

[2] M Sidiq, (2018) "Forecasting Rainfall with Time Series Model" in IOP Conference Series: Materials Science and Engineering.

[3] Pazvakawambwa G.T. and Ogunmokun A. A.,(2017) "A time-series forecasting model for Windhoek Rainfall, Namibia." in collection of open access research paper (CORE).

[4] D.A Attah, G.M.Bankole, (2011) "Time Series Analysis Model for Annual Rainfall Data in Lower Kaduna Catchment Kaduna, Nigeria" in global journal of researches in engineering civil and structural engineering, Volume 11 Issue 6 Version 1.0 November 2011

[5] Ebenezer Afrifa-Yamoah, Bashiru I. I. Saeed, Azumah Karim, (2016) "Sarima Modelling and Forecasting of Monthly Rainfall in the Brong Ahafo Region of Ghana", DOI:10.5923/j.env.20160601.01,in website ResearchGate

[6] Yashon O. Ouma, Rodrick Cheruyot & Alice N. Wachera, (2022) "Rainfall and runoff time-series trend analysis using LSTM recurrent neural network and wavelet neural network with satellite-based meteorological data: case study of Nzoia hydrologic basin" in website SpringerLink

[7] Mohammad Mahmudur Rahman Khan, Md. Abu Bakr Siddique, Shadman Sakib, Anas Aziz, Ihtyaz Kader Tasawar, Ziad Hossain, "Prediction of Temperature and Rainfall in Bangladesh using Long Short Term Memory ,Recurrent Neural Networks" in website science publishing group.

[8] S. Prabakaran, P. Naveen Kumar and P. Sai Mani Tarun, (2017) "Rainfall Prediction Using Linear Regression Model" in ARPN Journal of Engineering and Applied Sciences, VOL. 12, NO. 12, JUNE 2017, ISSN 1819-6608

[9] Bahareh Karimi, Mir Jafar Sadegh Safari, Ali Danandeh Mehr, Mirali Mohammadi, (2019) "RAINFALL PREDICTION USING MODIFIED LINEAR REGRESSION" in ARPN Journal of Engineering and Applied Sciences

[10] Charity Oseiwah Adjei, Wei Tian, Bernard-Marie Onzo, Emmanuel Adu Gyamfi Kedjanyi, Oscar Famous Darteh, "Monthly Rainfall Prediction Using ARIMA and Gene Expression" in Online Journal of Engineering Sciences and Technologies (OJEST), Vol. 2, No. 3; Summer 2019; Pages 8-14; DOI: 10.21859/ojest-02032.

[11] MAI Navid, NH Niloy, (2017) "Rainfall Forecasting in Sub-Sahara Africa-Ghana using LSTM Deep Learning Approach" in International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 10 Issue 03, March-2021

[12] Freemeteo website, https://freemeteo.vn/thoi-tiet/hue/current-weather/location/?gid=1580240&language=vietnamese&country=vietnam