Rainfall Data Fitting based on An Improved Mixture Cosine Model with Markov Chain

THITIPONG KANCHAI[1], NAHATAI TEPKASETKUL[1], TIPPATAI PONGSART[2], WATCHARIN KLONGDEE[1]
[1]Department of Mathematics, Faculty of Science, Khon Kaen University, THAILAND
[2]Department of Statistics, Faculty of Science, Khon Kaen University, THAILAND

Abstract: - This article proposes a model that uses the adjusted mixture cosine model of two components with Markov chain (MC_2MC) for predicting the monthly rainfall with actual data from Khon Kaen meteorological station (381201) in Khon Kaen province, Thailand. The data considers 31 years of historical data from January 1991 to December 2021. The evaluation is measured by the root mean square error ($RMSE$) and the R^2 values. We found that the mixture cosine model has $RMSE$ and R^2 values of 70.72 and 52.49%, respectively, and the MC_2MC model has $RMSE$ and R^2 values of 42.43 and 82.53%, respectively.

Key-Words: - Markov chain, mixture cosine, rainfall, imputation, missing data

Received: April 15, 2022. Revised: November 17, 2022. Accepted: December 19, 2022. Published: January 26, 2023.

1 Introduction

The rainfall data is essential for meteorological parameters. It significantly impacts our daily lives, causing issues with flooding and drought. One particularly has an impact on farming. We are currently dealing with climate change, which impacts rainfall. Because of these, agricultural yields are less specific, and crop insurance is made to reduce the risk of loss in unexpected events. As a result, analysis of rainfall is frequently carried out for a variety of applications, including the impact of rainfall on agricultural yields, [1], in addition, for building crop insurance as a weather index insurance, [2]. Crop insurance is challenging due to the presence of missing rainfall data. For this reason, data imputation has attracted a lot of attention from researchers to fill in the missing values with estimation. The traditional prediction approaches include regression, [3], [4], [5], [6], [7], machine learning, [8], [9], [10], and neural networks, [11], [12], [13].

Each year, the rainfall records significantly increase, especially in rainy periods. This behavior repeats on a yearly basis. Therefore, the overall characteristic of the rainfall data can be said to be a time series with a seasonal pattern. The characteristic of the seasonality was captured using either sine, cosine, or mixture cosine functions. Researchers often choose sine and cosine functions to estimate the data that have seasonal components, [11], [14], [15]. Moreover, the parameter vector of the mixture cosine model can be obtained using the differential evolution algorithm, [16].

In 1906, Markov chain was named after Andrei A. Markov, who first published his result, [17]. Markov chain is a stochastic process of a mathematical model in probability behavior. Many authors have used Markov chain to improve the model for fitting data. In 2014, Sous et al., [18], improved Grey model (1,1) using Markov chain and middle points matrix for forecasting gold prices. In 2019, Azizah et al., [19], proposed an application of Markov chain for predicting rainfall data at West Java using data mining approach. In 2021, Yutong, [20], proposed applications of Markov chain in weather and market share forecasts.

In this research, we propose a model that uses the adjusted mixture cosine model of two components with Markov chain (MC_2MC) for predicting the monthly rainfall. The rainfall data from Khon Kaen meteorological station (381201) in Khon Kaen province, Thailand, are chosen for illustration. Khon Kaen is located in northeastern Thailand, as shown in the red line in Fig. 1. The data considers 31 years of historical data from 1991 to 2021.

Fig. 1: Location of Khon Kaen province, Thailand.

2 Materials and Method

2.1 Data

According to the historical data series of the monthly rainfall, we have data from January 1991 to December 2021 that has complete data for 346 months and missing data for 26 months. Fig. 2 shows the arrangement of the data.

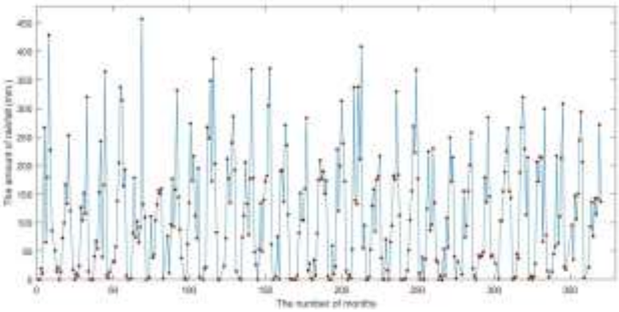
Fig. 2: The arrangement of the monthly rainfall data.

Next, Table 1 shows the summary of statistical information of monthly rainfall data from January 1991 to December 2021.

Table 1. Statistic analysis of monthly rainfall data.

Variable	Details	Min	Median	Mean	Max
Rainfall (mm.)	Monthly rainfall data from January 1991 to December 2021.	0	91.9	109.7	457.1

2.2 Mixture Cosine Model

From Fig. 2, we found that the monthly rainfall data can be represented as a time series. Moreover, it has a behavior like seasonal. Therefore, we shall consider our data as a periodic function and choose the mixture cosine model of m components formulated by

$$\hat{x}_r = \sum_{b=1}^{m} a_b \cos\left(\frac{r-\alpha_b}{G}\right) + a_{b+1}$$

where $a_1, a_2, \dots, a_{m+1}$ are real numbers, and $\alpha_1, \alpha_2, \dots, \alpha_m \in \{1,2,\dots,12\}$ represent the months with the peaks of each year. Since the cosine function has a period equal to 12 months, we choose $G = \frac{12}{2\pi} = \frac{6}{\pi}$. In this article, we shall estimate the parameter vector $(\alpha_1, \alpha_2, \dots, \alpha_m, a_1, a_2, \dots, a_{m+1})$ as the following procedure.

1. Consider $\alpha_1 = 1,2,\dots,12$, $\alpha_2 = \alpha_1, \alpha_1 + 1, \dots, 12$, and $\alpha_m = \alpha_{m-1}, \alpha_{m-1} + 1, \dots, 12$.
2. For each $(\alpha_1, \alpha_2, \dots, \alpha_m)$, use the differential evolution (DE) algorithm without crossover (population size $= 100$ and differential weight $= 0.8$) to estimate the parameters, $a_1, a_2, \dots, a_{m+1}$ with minimizing the root mean square error $E(a_1, a_2, \dots, a_{m+1} | \alpha_1, \alpha_2, \dots, \alpha_m)$ given by
$$E(a_1, a_2, \dots, a_{m+1} | \alpha_1, \alpha_2, \dots, \alpha_m) = \sum_{all\ r} (x_r - \hat{x}_r)^2$$
where
$$\hat{x}_r = \sum_{b=1}^{m} a_b \cos\left(\frac{r-\alpha_b}{G}\right) + a_{b+1}.$$
3. Choose $(\alpha_1^*, \dots, \alpha_m^*, a_1^*, \dots, a_{m+1}^*) = \underset{(\alpha_1,\dots,\alpha_m,a_1,\dots,a_{m+1})}{\operatorname{argmin}} E(a_1, a_2, \dots, a_{m+1} | \alpha_1, \alpha_2, \dots, \alpha_m)$

2.3 Adjust Mixture Cosine Model with Markov Chain

We adjust the mixture cosine model with Markov chain to fit the monthly rainfall data. Firstly, we construct the transition probability matrix by the residual error (e_r) of actual data (x_r) and predicted data $(\hat{x}_r)$ of mixture cosine model, i.e.,

$$e_r = x_r - \hat{x}_r$$

where $r = 1,2,3, \dots, n$, and n is the amount of data.

We separate the residual errors into k states. Define $L = 25^{\text{th}}$ percentile of $\{e_r\}$ and $U = 75^{\text{th}}$ percentile of $\{e_r\}$. The length of the interval (I) is calculated by

$$I = \frac{U-L}{k-2}.$$

Each interval of the state is calculated as follows:
State 1 (S_1): $x_r \in S_1$, if $e_r - L \leq 0$.
State 2 (S_2): $x_r \in S_2$, if $0 < e_r - L \leq I$.
State 3 (S_3): $x_r \in S_3$, if $I < e_r - L \leq 2I$.
$\vdots$
State $k-1$ (S_{k-1}): $x_r \in S_{k-1}$, if $(k-3)I < e_r - L \leq (k-2)I$.
State k (S_k): $x_r \in S_k$, if $e_r - L > (k-2)I = U - L$.

Let $F = [m_{ij}]_{k\times k}$ be a matrix given by m_{ij}, which is the number of x_r in state i and $x_{r+1}$is in state j where x_r, x_{r+1} are not missing data and $r = 1,2,3,\dots,n$. Next, let M_i be the number of data belonging to the state i such that

$$M_i = \sum_{j=1}^{k} m_{ij}\,, i = 1,2,\dots,k.$$

Therefore, the transition probability of moving one step from the i^{th} state to the j^{th} state is given by

$$p_{ij} = \frac{m_{ij}}{M_i},$$

where $i, j = 1,2,\dots,k$. Thus, the transition probability matrix is denoted by $T = [p_{ij}]_{k\times k}$.

Let $\Delta = [\delta_1 \quad \delta_2 \quad \cdots \quad \delta_k]'$ where δ_i is the represented value of state i^{th} given by

$$\delta_i = L + I\,\frac{2(i-1)-1}{2}$$

for all $i = 1,2,\dots,k$.

Therefore, we can adjust the mixture cosine model of m components with Markov chain, shortly called an MC$_m$MC model, which is formulated by

$$x_r^* = \hat{x}_r + T_i\Delta$$

where $x_r \in S_i$, $T_i = [p_{i1}\, p_{i2} \dots p_{ik}]$, and $\hat{x}_r$ is predicted value of the mixture cosine model.

3 Results

The mixture cosine model experiment uses 346 months of rainfall data. We determine the mixture cosine's parameters and function using the smallest sum square error based on the actual data. The mixture cosine model is fitted via differential evolution— the root mean square error value as displayed in Table 2.

We obtain the best mixture cosine model for fitting the monthly rainfall data when $m = 2, \alpha_1 = 6$, and $\alpha_2 = 10$.

It follows that:
$a_1 = 120.8039, a_2 = 79.0787$ and $a_3 = 103.4161$.

We then have

$$\hat{x}_r = 120.8039\cos\left(\frac{\pi}{6}(r-6)\right) + 79.0787\cos\left(\frac{\pi}{6}(r-10)\right) + 103.4161.$$

A comparison of the mixture cosine model with actual data is presented in Fig. 3.

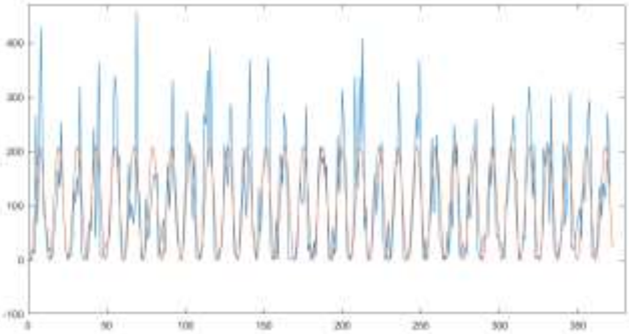

Fig. 3: Graph of mixture cosine model.

The residual error of the mixture cosine model and actual data is shown in Fig. 4.

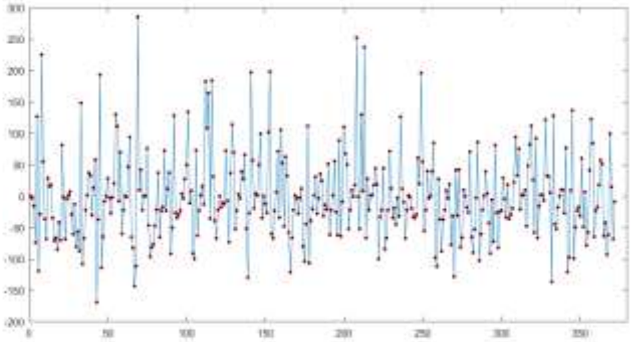

Fig. 4: The residual for mixture cosine model.

As mentioned in section 2.3, we separate the residual errors into 8 states. We have
$L = -47.2472, U = 35.2396$, and $I = 13.7478$.

Each interval of the state is calculated as follows:

State 1 (S_1): $x_r \in S_1$, if $e_r \le -47.2472$.
State 2 (S_2): $x_r \in S_2$, if $-47.2472 < e_r \le -33.4994$.
State 3 (S_3): $x_r \in S_3$, if $-33.4994 < e_r \le -19.7516$.
State 4 (S_4): $x_r \in S_4$, if $-19.7516 < e_r \le -6.0038$.
State 5 (S_5): $x_r \in S_5$, if $-6.0038 < e_r \le 7.7440$.
State 6 (S_6): $x_r \in S_6$, if $7.7440 < e_r \le 21.4918$.
State 7 (S_7): $x_r \in S_7$, if $21.4918 < e_r \le 35.2396$.
State 8 (S_8): $x_r \in S_8$, if $e_r > 35.2396$.

The matrix of represented value for each state is obtained by:

$$\Delta = \begin{bmatrix} -54.1211 \\ -40.3733 \\ -26.6255 \\ -12.8777 \\ 0.8701 \\ 14.6179 \\ 28.3657 \\ 42.1135 \end{bmatrix}.$$

Table 2. The fitting results in terms of the root mean square error of each α and β.

α \ β	1	2	3	4	5	6	7	8	9	10	11	12
1	230.31	225.80	212.81	187.55	157.46	125.80	107.11	115.77	143.10	174.81	202.41	220.32
2	-	229.72	221.54	202.06	173.90	141.31	113.87	108.16	125.08	157.73	187.33	212.21
3	-	-	216.91	204.32	179.97	149.81	118.17	99.42	106.05	132.48	162.29	190.54
4	-	-	-	194.95	176.49	149.75	117.67	91.88	85.31	103.33	132.08	160.77
5	-	-	-	-	161.95	141.14	114.77	89.23	73.16	77.64	99.24	125.06
6	-	-	-	-	-	125.14	107.76	90.20	76.62	**70.72**	76.66	92.84
7	-	-	-	-	-	-	102.22	96.40	92.93	89.01	86.15	84.83
8	-	-	-	-	-	-	-	107.06	117.03	118.11	116.04	109.74
9	-	-	-	-	-	-	-	-	137.30	147.73	150.36	145.18
10	-	-	-	-	-	-	-	-	-	170.14	178.51	177.06
11	-	-	-	-	-	-	-	-	-	-	196.00	203.47
12	-	-	-	-	-	-	-	-	-	-	-	215.72

Therefore, we have the matrix (F) as mentioned in Section 2.3 shown below:

$$F = \begin{bmatrix} 99 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 38 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 19 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 45 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 26 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 14 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 64 \end{bmatrix}.$$

The transition probability matrix is obtained by:

$$T = \begin{bmatrix} 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0256 & 0.9744 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0500 & 0.9500 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0217 & 0.9783 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0370 & 0.9630 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0588 & 0.9412 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0667 & 0.9333 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0154 & 0.9846 \end{bmatrix}.$$

Therefore, we obtain the MC_2MC model illustrated in the equation below:

$$x_r^* = \max\{x_r^\circ, 0\}$$

where $x_r \in S_i$ and

$$x_r^\circ = 120.8039\cos\left(\frac{2\pi}{12}(r-6)\right) + 79.0787\cos\left(\frac{2\pi}{12}(r-10)\right) + 103.4161 + T_i\Delta.$$

Fig. 5 shows the graphs of the actual data, the mixture cosine model, and the MC_2MC model. The lines were derived from actual data, the mixture cosine model, and the MC_2MC model using blue, red, and purple, respectively. The x-axis and y-axis of each graph in Fig. 5 are the number of the month and amount of rainfall (mm.), respectively.

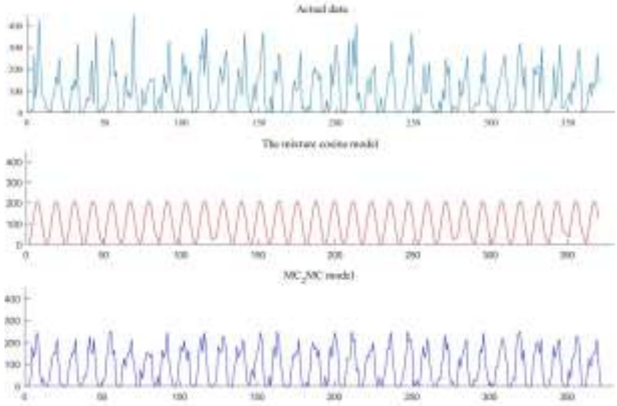

Fig. 5: Comparison of the actual data, the mixture cosine model, and the MC_2MC model.

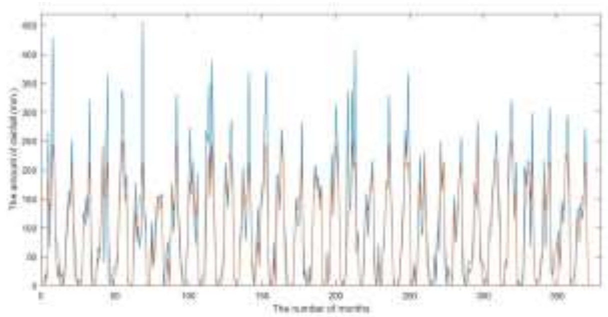

Fig. 6: Actual and simulated rainfall for MC_2MC model.

Fig. 6 shows the actual and generated monthly rainfall data in Khon Kaen province, Thailand. The red and blue lines are based on actual data and the MC_2 MC model, respectively. The number of months and amount of rainfall (mm.) represent the x-axis and y-axis of Fig. 6.

4 Measuring the Quality of Fitting

To evaluate the performance of a statistical learning method on a given data set, we would like to measure how well its predictions match the actual data. The evaluation is measured by the root mean square error and the R-square is how well the regression model explains observed data.

The root mean square error ($RMSE$) is defined by

$$RMSE = \sqrt{\frac{1}{n}\sum_{r=1}^{n}(x_r - q_r)^2},$$

and the R-square (R^2) is defined by:

$$R^2 = 1 - \frac{\sum_{r=1}^{n}(x_r - q_r)^2}{\sum_{r=1}^{n}(x_r - \bar{x}_r)^2},$$

where $x_r, q_r, \bar{x}_r$ are the actual value, predicted value of the model, and mean of actual value, respectively, and $r = 1,2,\ldots,n$. The $RMSE$ and R^2 of the models are illustrated in Table 3.

Table 3. Evaluation value of the models

Model	$RMSE$	R^2
Mixture cosine	70.72	52.49%
MC_2MC	42.43	82.53%

Table 3 shows the performance of the mixture cosine model and the MC_2MC model for fitting the actual data.

5 Conclusion

The proposed model uses the adjusted mixture cosine model of two components with Markov chain (MC_2MC) for predicting the monthly rainfall data from Khon Kaen meteorological station (381201) in Khon Kaen province, Thailand. The data considers 31 years of historical data from January 1991 to December 2021. We found that the mixture cosine model has $RMSE$ and R^2 values of 70.72 and 52.49%, respectively, and the MC_2MC model has $RMSE$ and R^2 values of 42.43 and 82.53%, respectively. According to these findings, the MC_2MC model has a 40.00% better $RMSE$ than the mixture cosine model. The MC_2MC model can describe the monthly rainfall data since it has an acceptance rate of $R^2 = 82.53\%$.

The application of this work can be utilized to anticipate the missing variables or to predict the value of the periodic data such as annual rainfall, daily temperature, or the number of tourists visiting the famous place.

Acknowledgement:
The first author would like to express gratitude to the Science Achievement Scholarship of Thailand (SAST) for financial assistance for this paper.

References:

[1] Verón, Santiago R., Diego de Abelleyra, and David B. Lobell, Impacts of Precipitation and Temperature on Crop Yields in the Pampas, *Climatic Change*, Vol.130, 2015, pp. 235–245.

[2] Kath, Jarrod, Shahbaz Mushtaq, Ross Henry, Adewuyi Adeyinka, and Roger Stone, Index Insurance Benefits Agricultural Producers Exposed to Excessive Rainfall Risk, *Weather and Climate Extremes*, Vol.22, 2018, pp. 1–9.

[3] S. Prabakaran, P. N. Kumar, and P. S. M. Tarun, RAINFALL PREDICTION USING MODIFIED LINEAR REGRESSION, *ARPN Journal of Engineering and Applied Sciences*, Vol. 12, No.12, 2017, pp. 3715-3718.

[4] J.Refonaa, M. Lakshmi, Raza Abbas, and Mohammad Raziullha, Rainfall Prediction using Regression Model, *ijrte*, Vol.8, No.2S3, 2019, pp. 543–546.

[5] R. E. Chandler and H. S. Wheater, Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland: GENERALIZED LINEAR MODELING OF DAILY RAINFALL, *Water Resour. Res.*, Vol.38, No.10, 2002, pp. 10-1-10–11.

[6] R. Coe and R. D. Stern, Fitting Models to Daily Rainfall Data, *J. Appl. Meteor.*, Vol.21, No.7, 1982, pp. 1024–1031.

[7] N. Sethi and K. Garg, Exploiting Data Mining Technique for Rainfall Prediction, *IJCSIT*, Vol.5, No.3, 2014, pp. 3982–3984.

[8] C. M. Liyew and H. A. Melese, Machine learning techniques to predict daily rainfall amount, *J Big Data*, Vol.8, No.153, 2021, pp. 1-11.

[9] N. Oswal, Predicting Rainfall using Machine Learning Techniques, *Atmospheric and Oceanic Physics*, 2021, pp. 1-23.

[10] N. Salaeh *et al.*, Long-Short Term Memory Technique for Monthly Rainfall Prediction in Thale Sap Songkhla River Basin, Thailand, *Symmetry*, Vol.14, No.8, 2022, pp. 1-24.

[11] P. Chan Chiu, A. Selamat, O. Krejcar, K. Kuok Kuok, E. Herrera-Viedma, and G. Fenza, Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network, *International Journal of Interactive*

Multimedia and Artificial Intelligence, Vol.6, No.7, 2021, pp. 39-48.
[12] Shakib Badarpura, Abhishek Jain, Aniket Gupta, Deepali Patil, and SHREE L.R TIWARI COLLEGE OF ENGINEERING, Rainfall Prediction using Linear approach Neural Networks and Crop Recommendation based on Decision Tree, *IJERT*, Vol.9, No.4, 2020, pp. 394-399.
[13] R. Venkata Ramana, B. Krishna, S. R. Kumar, and N. G. Pandey, Monthly Rainfall Prediction Using Wavelet Neural Network Analysis, *Water Resour Manage*, Vol.27, No.10, 2013, pp. 3697–3711.
[14] H. Jin, Q. Shao, and S. Crimp, Daily rainfall data infilling with a stochastic model, 23rd International Congress on Modelling and Simulation, Canberra, ACT, Australia, 2019.
[15] K. Mammas and D. Lekkas, Rainfall Generation Using Markov Chain Models; Case Study: Central Aegean Sea, *Water*, Vol.10, No.7, 2018, pp. 856-866.
[16] J. Ilonen, J.-K. Kamarainen, and J. Lampinen, Differential Evolution Training Algorithm for Feed-Forward Neural Networks, *Neural Processing Letters*, Vol.17, 2003, pp. 93-105.
[17] W. K. Ching and M. K. Ng, *Markov chains: models, algorithms and applications*. New York, N.Y: Springer, 2006.
[18] S. Sous, T. Thongjunthug, and W. Klongdee, Gold Price Forecasting Based on the Improved GM (1,1) Model with Markov Chain by Average of Middle Point, *KKU Sci. J.*, Vol.42, No.3, 2014, pp. 693-699.
[19] A. Azizah, R. Welastika, A. N. Falah, B. N. Ruchjana, and A. S. Abdullah, An Application of Markov Chain for Predicting Rainfall Data at West Java using Data Mining Approach, *IOP Conf. Ser.: Earth Environ. Sci.* 303, 2019, pp. 1-10.
[20] X. Yutong, Applications of Markov Chain in Forecast, *J. Phys.: Journal of Physics: Conference Series*, 2021.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

-Thitipong Kanchai carried out the conceptualization, investigation, methodology, software, writing-original draft, and writing-review & editing.
-Nahatai Tepkasetkul guides the differential evolution in Matlab and writing-review & editing.
-Tippatai Pongsart carried out the conceptualization, investigation, methodology, and writing-review & editing.
-Watcharin Klongdee carried out the conceptualization, investigation, methodology, writing-original draft, and writing-review & editing.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The first author would like to express gratitude to the Science Achievement Scholarship of Thailand (SAST) for financial assistance for this paper.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.