Thai Journal of **Math**ematics
Special Issue: The 17^{th} IMT-GT ICMSA 2021
Pages 135–145 (2022)

http://thaijmath.in.cmu.ac.th

ISSN 1686-0209

An Investigation of the Predicting Model for Daily Temperature in the Capital of Laos

Rusma Khaleng[1], Rattikan Saelim[1,*], Salang Musikasuwan[1], Santhana Chaimontree[1] and Kriangsak Damchoom[1]

[1] *Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani, Thailand 94000*
e-mail: 6420320006@email.psu.ac.th (R. Khaleng); rattikan.s@psu.ac.th (R. Saelim); salang.m@psu.ac.th (S. Musikasuwan); santhana.c@psu.ac.th (S. Chaimontree); kriangsak.d@psu.ac.th (K. Damchoom)

Abstract Laos is one of those countries whose people are engaged in farming. Farming in developing countries strongly relies on climate. Being one of the key factors of climate change, the temperature has become a fundamental topic of recent research in this area. The aims of this study were two folds, namely (i) to study how temperature in Laos changed in the last decades, and (ii) to propose an appropriate model for temperature prediction in Laos. In this work, the daily temperatures in Laos in the last 26 years (1995-2020) were analyzed. Suggested by the ARIMA model that the current temperature in Laos could depend on the temperatures up to 8 previous days. However, taking the Akaike Information Criterion (AIC) and Root Mean Square Error (RMSE) into account, the ARIMA (4,0,0) seems to be a good enough model. While the multiple linear regression (MLR) of time-lag variables suggested that the last 1, 4, 6, and 8 days of temperatures have contributed significantly to the current day temperature. In a way, the two models agreed on the number of the predictors which is 4, even though they have different time lags. Furthermore, the random forest model with 4 predictors (lags 1, 4, 6, and 8) even slightly outperformed the random forest model with predictors of 8 previous consecutive days in terms of RMSE. Besides, by fitting a simple linear regression to the seasonally adjusted temperature, it showed that the temperature in the capital of Laos has not significantly changed in the last 26 years.

MSC: 49K35; 47H10; 20M12

Keywords: temperature pattern and trend; climate change; autoregressive integrated moving average; multiple linear regression; random forest

Submission date: 31.01.2022 / Acceptance date: 31.03.2022

1. Introduction

Temperature change is one of the major concerns in this developing world. A small amount of average temperature change could have a greater effect on the environment and all creatures on Earth. Especially, when it comes to food production and sustainability, it is the issue that drew everyones attention. Temperature could have effects in any field such as environment, society, economic, etc. One way or another, inevitably, the effects

*Corresponding author.

Published by The Mathematical Association of Thailand.

will come to human beings. Several research have confirmed that temperature has effects on crop yields [1–4] and cattle production [5]. Even though crops need temperature to survive, extreme high or low temperatures can have a negative effect on plants growth and yields [6]. In fact, a study on temperature thresholds [1] for several food crops such as soybean, wheat, tomato, peanut, has reported that extreme temperature has effects on crop production. Moreover, an increasing temperature and precipitation variability also increases the risk to crop production [3].

Laos is a developing country in Southeast Asia with a primarily agricultural economy. Vientiane is Laos' capital and the largest city, located on the Mekong River near Thailand's border. According to the 2020 Census, the city has a population of 948,477 people. Vientiane is a 3,920-square-kilometer metropolis with an average elevation of 174 meters above sea level. The climate in Vientiane is tropical savanna, with a distinct wet and dry season. The dry season in Vientiane lasts from November to March. The wet season begins in April and lasts for around seven months in Vientiane. Throughout the year, Vientiane is extremely hot and humid, while temperatures in the city are slightly milder during the dry season than during the wet season [7]. According to the United Nations' Food and Agriculture Organization (FAO) [8], biological resources contribute more than 66% percent of Laos' GDP. Understanding how the temperature in Laos would change is therefore quite beneficial.

The goal of this study was to predict daily temperatures in Laos from 1995 to 2020 based on daily temperatures. The seasonal pattern was first explored on the assumption that the season repeats itself every year. The trend of the daily temperature was then calculated using simple linear regression on the seasonally adjusted daily temperature over a 26-year period. Following that, the ARIMA model was used to make a prediction. Finally, the time-lag variables were employed as predictors for the current daily temperature in multiple linear regression (MLR) and random forest models.

2. DATA AND METHODS

Following the collection of daily temperatures, preprocessing for missing value imputation was carried out, as shown in Figure 1. The data was then fitted with ARIMA, MLR, and random forest models, as well as seasonal patterns and trends were studied. The methodologies were addressed in depth in the following three subsections.

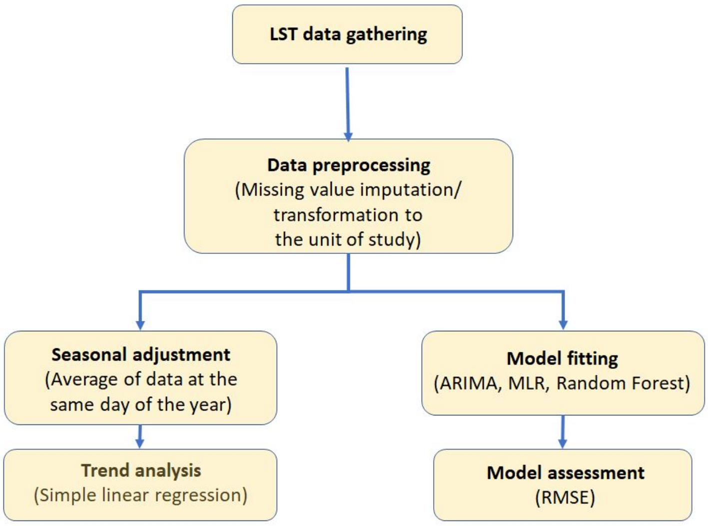

FIGURE 1. The conceptual framework of this study

2.1. Study area and data collection

In this study, daily temperature in Vientiane city of Laos has been considered. The location of Vientiane city is presented in Figure 2.

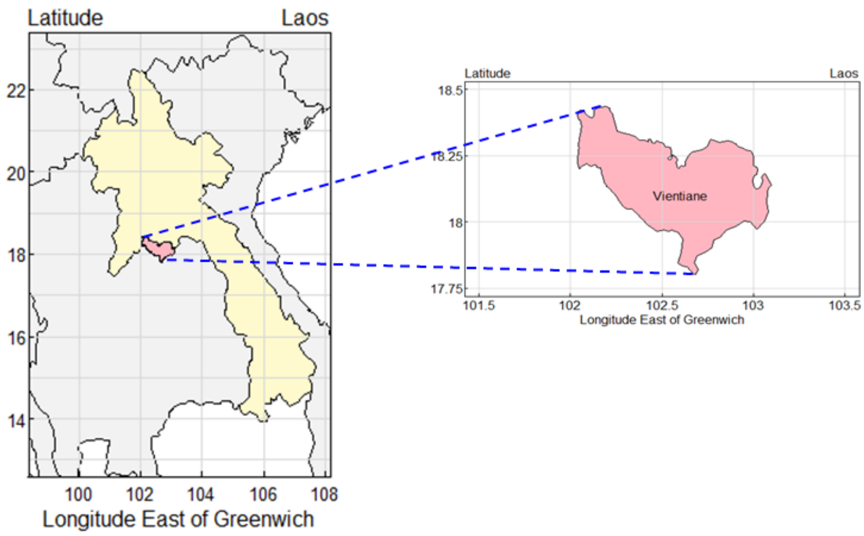

FIGURE 2. The location of Vientiane city in Laos map

With a total of 9266 measurements, the daily average temperature of Vientiane, Laos was obtained from https://academic.udayton.edu/kissock/http/Weather/ [9]. The National Climatic Data Center contributed this information, which is free to use for research and non-commercial reasons. It was discovered that there were 56 missing values in the original data. Each missing value was replaced with the 26-year average of temperatures on the corresponding day of the year. Figure 3 shows the time-series plots of daily temperature in Vientiane from 1995 to 2020 after all missing values were imputed.

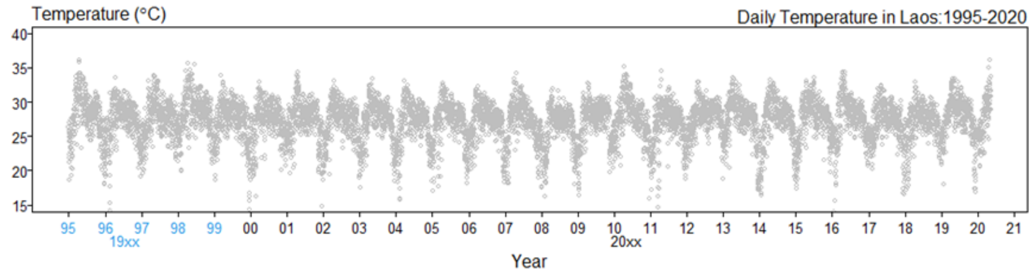

FIGURE 3. The plots of daily average temperature in Vientiane city during 1995-2020

The data was separated into two portions for the prediction models, with 70 percent and 30 percent for the training and testing datasets, respectively. The predictive model was built using the training dataset, and the model was evaluated using the testing dataset. Figure 4 shows time-series graphs of observation data for the training and testing sets.

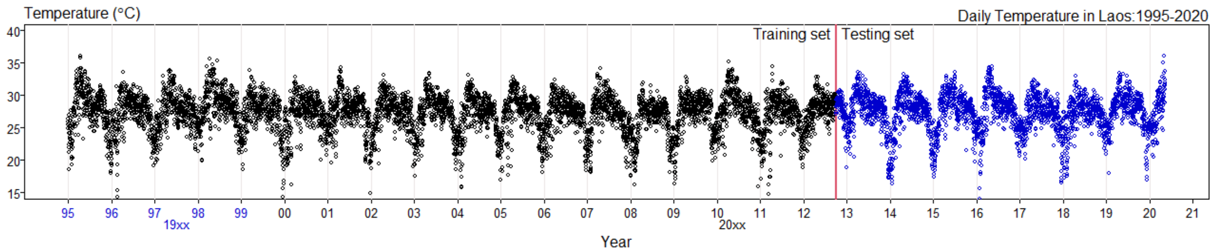

FIGURE 4. Time-series plots of daily average temperature for training dataset (black) and testing dataset (blue)

2.2. METHODS

In this study, simple linear regression was used to examine the trend of observation data. The following is the formula for a simple linear regression model (Equation 2.1):

$$\hat{y} = \alpha_0 + \alpha_1 X + \epsilon, \tag{2.1}$$

where $\hat{y}$ is the outcome variable, α_0 is the intercept, α_1 is the slope, X is the independent variable and ϵ is the error term.

Time series models can be constructed using stationary variables, which have statistical qualities that remain constant across time, such as mean and variance. The prediction model was fitted using the autoregressive integrated moving average model. According to the literature, the ARIMA model is one of the best models for predicting a time series. The process of obtaining the model [10] involved fitting an appropriate model, estimating the parameters, and evaluating the model. As related to the result of this paper, the ARIMA(4,0,0) is given by (Equation 2.2):

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + +\phi_3 y_{t-3} + \phi_4 y_{t-4} + e_t \tag{2.2}$$

where, $\hat{y}_t$ is the predicted average daily temperature at observation t; ϕ_1, ϕ_2, ϕ_3 and ϕ_4 are coefficients of the lag variable y_{t-1}, y_{t-2}, y_{t-3} and y_{t-4}, respectively, and e_t is the value not explained by the lag variables.

Multiple linear regression analysis is a statistical method for developing a suitable regression model for predicting the value of the outcome variable, or the variable under investigation. The multiple linear regression model takes the following form based on the values of independent variables or predictors (Equation 2.3):

$$\hat{y}_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + ... + b_k y_{t-k} \tag{2.3}$$

where $\hat{y}_t$ is the outcome variable (predicted temperature at time t)
y_{t-i} is the temperature at time $t - i$, $i = 1, 2, ..., k$
b_0 is the intercept (constant term)
b_i is the coefficients for temperature at time $t - i$, $i = 1, 2, ..., k$
k is the number of predictors (lag variables)

In general, time-series data is always faced with seasonal effects. To investigate the effect of season in the observation data, the seasonal patterns have been analyzed. It is natural to assume that at the same time of the year the season would approximately be the same. Therefore, for simplicity, the season has been taken from the average values of temperatures at the same day of the year throughout the whole 26 years. By removing (subtracting) the seasonal effect from the original observations and adjusting by adding back the average value of the original series. Once the daily average temperature was seasonally adjusted, its trend has been investigated by fitting simple linear regression [11, 12]. As the seasonal patterns give the picture of how temperature repeatedly vary during a particular time of the period, its trend provided a long-term direction of changes. The linear change could be either increasing, decreasing, or not going anywhere.

ARIMA and regression models has been used for prediction in many studies [13–16]. For example, Hassan (2014) used ARIMA and regression models to predict the daily and monthly clearness index for several Canadian cities [13]. The two models were both linear models. The main difference was that ARIMA model takes only the time-lag variables as its predictors, while the predictors of regression model can be any variables including time-lag variables. To avoid ambiguity, an ARIMA(p,d,q) with different choices of the parameters p, d, and q will be considered as a different model. Moreover, the appropriated model will be selected by measuring the values of the Akaike Information Criterion (AIC) and the Root Mean Squared Error (RMSE). To determine the parameters q and p parameters of ARIMA model, the graph of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) as shown in Figure 5 will be considered.

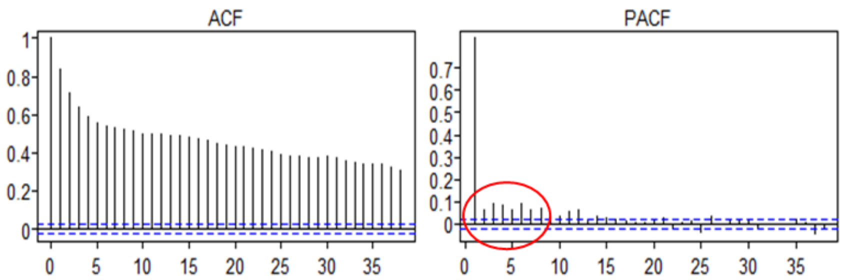

FIGURE 5. ACF and PACF of the training dataset

Once the parameter p, d, and q were fixed for an ARIMA model, the MLR and random forest will join in and play the role of confirmation. Notice that the integer p also indicates the number of factors. Namely, all the time-lag variables starting from p days earlier including all days later until yesterday (cannot be skipped) will be taken as the predictors. For example, if $p = 8$, it means that the ARIMA model includes all time-tag variables of the previous 8 consecutive days as its predictors. But this is not the case for multiple regression since any predictors can be included or excluded from the model. In fact, fitting MLR with, say, 8 factors of 8 consecutive days, the most insignificant factor is removed. Repeatedly fitting with the remaining factors and removing one most insignificant factor at a time until only significant factors are left. Finally, the random forest model [17] is applied to confirm again the result by the ARIMA and MLR models.

3. RESULTS

3.1. Seasonal Pattern and Trend Analysis

The seasonal pattern (red curve in Figure 6) is a repetition, for each year, of those 366 average values derived earlier. The seasonal pattern was subtracted from the original temperature and adjusted by adding its overall average to achieve the seasonally adjusted temperature (Figure 7).

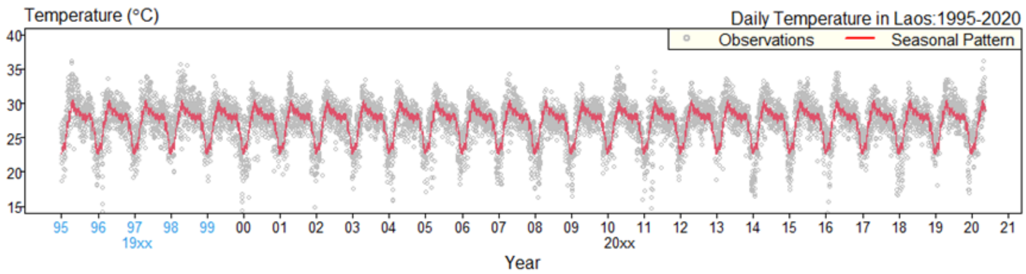

Figure 6. The plots of original daily temperature and its seasonal patterns in Vientiane city during 1995-2020

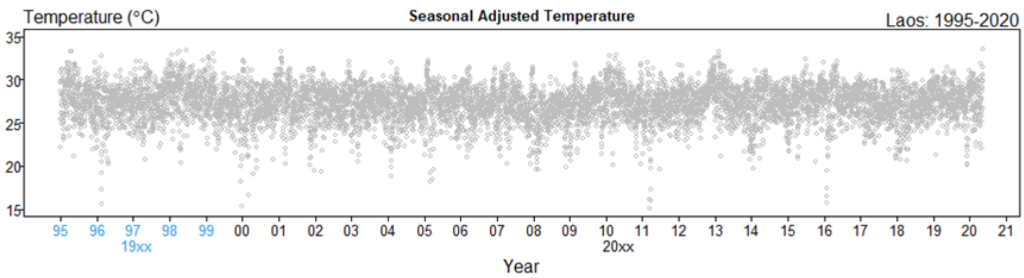

Figure 7. The plots of seasonal adjusted daily temperature in Vientiane city from 1995 to 2020

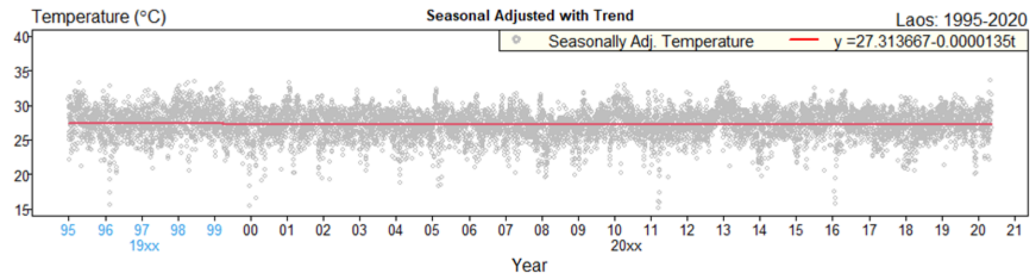

FIGURE 8. The plots of the trend (red line) and seasonal adjusted daily temperature (grey dots) in Vientiane city during 1995-2020

In this study, the trend (red line in Figure 8) revealed changes of -0.0000139 (decreases of 0.0000139) degree Celsius per day with p-value 0.088. Moreover, by simple calculation, the temperature in Vientiane city has a tendency of decreasing around 0.0049275 degree Celsius per year or 0.4927500 degree Celsius per century. However, notice that the p-value is greater than 0.05. Therefore, the change was not statistically significant. In other words, it is statistically saying that the change was not different from zero.

3.2. Predictive Models

In case of daily temperature in Laos, the ACF and PACF in Figure 9 suggested that there could be up to 8 time-lag variables to be considered as factors in the ARIMA model. In other words, the temperatures of the 8 previous days predict the current temperature. Therefore, the ARIMA$(p, 0, 0)$ for $p = 1, 2, , 8$ has been created. It turned out that for the increasing number of p from 1 to 8, the AIC and RMSE of the training dataset were strictly decreasing. However, when applying the model to the training dataset, the RMSE were strictly increasing, even though the increase and decrease were small. In fact, the AIC and RMSE of the training set were decreasing by the proportion of less than 1% and only at the 2^{nd} place decimals, respectively.

The ARIMA(4,0,0) seems to be a reasonable model for making economic sense so that the result may be acceptable with a less difficult equation. The resulting model was illustrated and showed how well it fit the training and testing datasets in Figure 9 with their RMSE were 1.5756 and 2.0028, respectively. Moreover, the equation of ARIMA (4,0,0) is given by

$$\hat{y}_t = 3.5184 + 0.7700y_{t-1} - 0.0099y_{t-2} + 0.0224y_{t-3} + 0.0882y_{t-4} \tag{3.1}$$

where $y_{t-k},\ k = 1, 2, 3, 4$ are the time-lag temperatures of k days earlier, and $\hat{y}_t$ is the predicted current temperature.

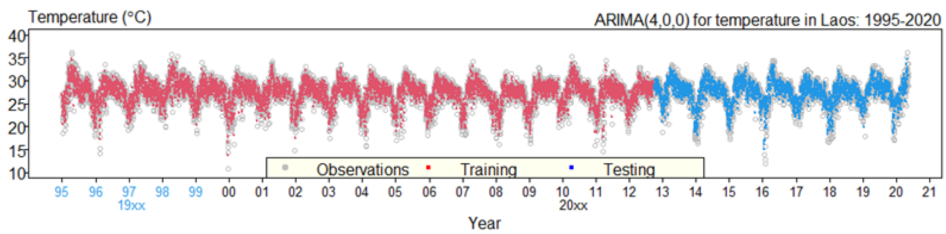

FIGURE 9. The time-series plots of ARIMA(4,0,0) predicted values for training (red dots) and testing (blue dots) datasets compared with original observations

Now, taking the first suggestion by the PACF for the ARIMA model, it was reasonable to start multiple linear regression with 8 time-lag variables. Then, following the general process of removing one less significant variable at a time, namely the variable with highest p-value (greater than 0.05) given by the regression model. It turns out that there are 4 time-lag variables left. Indeed, only the lags 1, 4, 6, and 8 days earlier that significantly had contribution to the current temperature. The result by the regression model fitting to the training (RMSE=2.4286) and testing (RMSE=2.1132) datasets is shown in Figure 10 and its equation is given by

$$\hat{y}_t = 2.6293 + 0.7248y_{t-1} + 0.0399y_{t-4} + 0.0464y_{t-6} + 0.0744y_{t-8} \tag{3.2}$$

where $y_{t-k}, k = 1, 4, 6,$ and 8 are the time-lag temperatures of k days earlier, and $\hat{y}_t$ is the predicted temperature at time t.

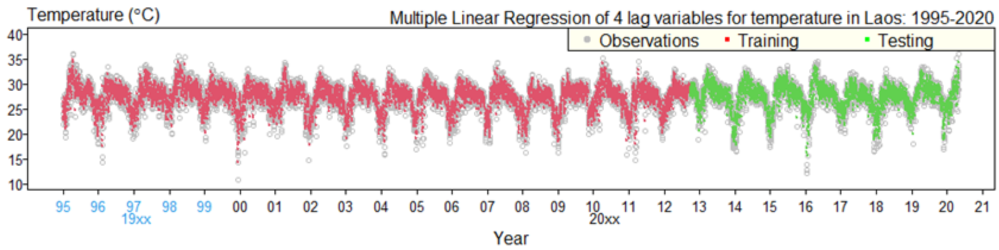

Figure 10. The time-series plots of MLR model predicted values for training (red dots) and testing (green dots) datasets against original observations

Two random forest models have been created with 4 and 8 time-lag variables as suggested by MLR, and PACF, respectively. The random model with 4 time-lag variables included the temperature at day 1, 4, 6 and 8 days earlier. Both models have been treated with 500 trees and randomly selected 3 independent variables for each tree. The final predictive value has been calculated by taking arithmetic mean of all predicted value from 500 trees.

The result by the both of random forest models (treated with 4 and 8 time-lag variables) fitting to the training (RMSE=1.6247 and 1.6026) and testing (RMSE=1.5171 and 1.6026) datasets are shown in Figures 11 and 12, respectively.

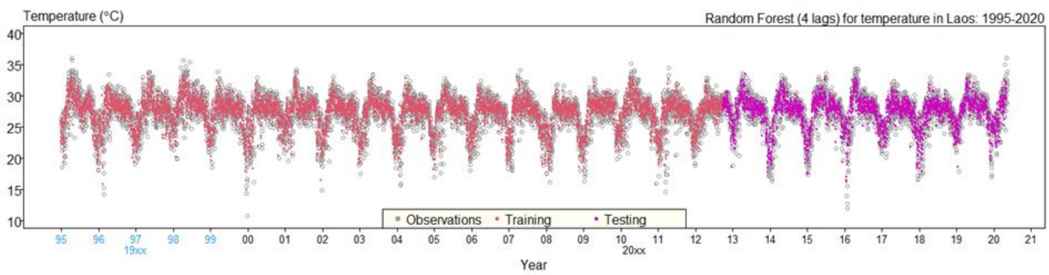

Figure 11. The time-series plots of random forest model (4 time-lag variables) predicted values for training (red dots) and testing (purple dots) datasets against original observations

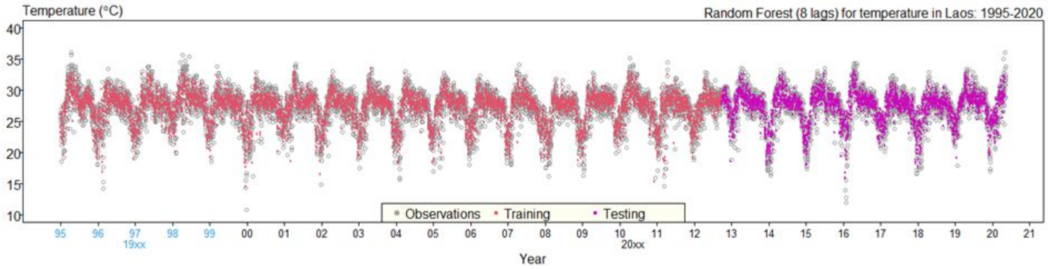

FIGURE 12. The time-series plots of random forest model (8 time-lag variables) predicted values for training (red dots) and testing (purple dots) datasets against original observations

Finally, the results by the random forest models show that when the number of factors reduced from 8 to 4, following the best MLR obtained, the RMSE slightly change, as shown in Table 1.

TABLE 1. The RMSE of the Training and Testing sets by ARIMA, MLR, and the random forest (RF) models

	RMSE	
	Training set	Testing set
ARIMA (4,0,0)	1.5756	2.0028
MLR with lags 1, 4, 6 and 8	2.4286	2.1132
RF with 8 factors of lags	1.6026	1.6026
RF with lags 1, 4, 6, and 8	1.6247	1.5171

4. DISCUSSION AND CONCLUSION

This research aimed to study how temperature in Laos change in the last decades and propose an appropriate model for temperature prediction in Vientiane city, Laos. By analyzing the trend of daily average temperature by using simple linear regression, it can be observed that trend of the temperature in Vientiane city has not been changed.

From the results, the equations of ARIMA(4,0,0) and the multiple linear regression model were finally obtained as the Equations (3.1) and (3.2), respectively. The first notice is that they had the same number of predictors which is four. However, the fours are different. While the current day temperature by ARIMA(4,0,0) was predicted by the temperatures of the previous four consecutive days up until yesterday, the current day temperature by the multiple regression model was predicted by the temperatures of the previous 8, 6, 4, and 1 days. Furthermore, the result by the random forest model also provides evidence for the lags and the number of lags of the factor. Noticed that the lag predictors could also be suggested by the importance of variables of random forest method. One could argue that the multiple linear regression and the random forest would be better since they could select any lag variables. However, the ARIMA was helpful for giving the idea of the number of time-lag variables to start with.

ACKNOWLEDGEMENTS

We would like to express our thank to the referees for their comments and suggestions on the manuscript. This work was partially supported by Faculty of Science and Technology, and Graduate School, Prince of Songkla University, Pattani Campus, Thailand.

References

[1] U.C. Jha, A. Bohra, R. Jha, Breeding approaches and genomics technologies to increase crop yield under low-temperature stress, Plant Cell Reports 36 (1) (2017) 1-35.

[2] K. Kawasaki, S. Uchida, Quality matters more than quantity: asymmetric temperature effects on crop yield and quality grade, American Journal of Agricultural Economics 98 (4) (2016) 1195-209.

[3] E. Wang, P. Martre, Z. Zhao, F. Ewert, A. Maiorano, R.P. Rtter, B.A. Kimball, M.J. Ottman, G.W. Wall, J.W. White, M.P. Reynolds, The uncertainty of crop yield projections is reduced by improved temperature response functions, Nature Plants 3 (8) (2017) 1-3.

[4] T. Yuliawan, I. Handoko, The effect of temperature rise to rice crop yield in Indonesia uses Shierary Rice model with geographical information system (GIS) feature, Procedia Environmental Sciences 33 (2016) 214-20.

[5] A.J. Zaied, H.M. Geli, J.L. Holechek, A.F. Cibils, M.N. Sawalhah, C.C. Gard, An evaluation of historical trends in New Mexico beef cattle production in relation to climate and energy, Sustainability 11 (23) (2019) 1-17.

[6] C. Zhao, B. Liu, S. Piao, X. Wang, D.B. Lobell, Y. Huang, M. Huang, Y. Yao, S. Bassu, P. Ciais, J.L. Durand, Temperature increase reduces global yields of major crops in four independent estimates, Proceedings of the National Academy of Sciences 114 (35) (2017) 9326-31.

[7] Wikipedia. 2021. Vientiane. Available online: https://en.wikipedia.org/wiki/Vientiane [October 20, 2021]

[8] Food and Agriculture Organization of the United Nations. 2021. FAO in laos. Available online: https://www.fao.org/laos/fao-in-laos/laos-at-a-glance/en/ [October 16, 2021]

[9] University of Dayton. 2020. Environmental protection agency average daily temperature archive. Available online: http://academic.udayton.edu/kissock/http/Weather/default.htm [August 15, 2021]

[10] R. Anokye, E. Acheampong, I. Owusu, I.E. Obeng, Time series analysis of malaria in Kumasi: using ARIMA models to forecast future incidence, Cogent Social Sciences 4 (1) (2018) 1-13.

[11] L.S. Hingane, K.K. Rupa, M.B.V. Ramana, Longterm trends of surface air temperature in India, Journal of Climatology 5 (5) (1985) 521-8.

[12] N.H. Miswan, R.M. Said, S.H. Anuar, ARIMA with regression model in modelling electricity load demand, Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 8 (12) (2016) 113-6.

[13] J. Hassan, ARIMA and regression models for prediction of daily and monthly clearness index, Renewable Energy 68 (2014) 421-7.

[14] J. Mohamed, Time series modeling and forecasting of Somaliland consumer price index: a comparison of ARIMA and regression with ARIMA errors, American Journal of Theoretical and Applied Statistics 9 (4) (2020) 143-53.

[15] M. Murat, I. Malinowska, M. Gos, J. Krzyszczak, Forecasting daily meteorological time series using ARIMA and regression models, International Agrophysics 32 (2) (2018) 253-64.

[16] M. Sabry, H. Abd-El-Latif, N. Badra, Comparison between regression and ARIMA models in forecasting traffic volume, Australian Journal of Basic and Applied Sciences 1 (2) (2007) 126-36.

[17] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, Advances in Neural Information Process System 26 (2013) 4319.