PENERBIT UMT
Menjana Khazanah Ummah

Journal of Mathematical Sciences and Informatics
Journal Homepage: https://journal.umt.edu.my/index.php/jmsi
eISSN: 2948-3697
DOI: http://doi.org/10.46754/jmsi.2022.06.003

JMSI
Journal of Mathematical Sciences and Informatics

MODELLING INDIAN OCEAN AIR TEMPERATURE USING ADDITIVE MODEL

MIFTAHUDDIN[1], ANANDA PRATAMA SITANGGANG[1], NORIZAN MOHAMED[2*] AND MAHARANI A. BAKAR[2]

[1]Department of Statistics, FMIPA, Syiah Kuala University, Banda Aceh Indonesia; miftah.unsyiah.ac.id. [2]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Malaysia; norizan@umt.edu.my

Corresponding author: norizan@umt.edu.my

ARTICLE INFO

Article History:
Received 17 JUNE 2021
Accepted 29 MARCH 2022
Available online 29 SEPTEMBER 2022

Section Editor: Muhammad Safiih Lola

Keywords:
Air temperature;
Linear model;
Generalized Linear model;
Generalized additive model;
Gaussian process

ABSTRACT

In this study, we used the fluctuating air temperature dataset. The change is caused by data fluctuations, trend, seasonality, cyclicity and irregularities. The generalized additive model (GAM) data approach is used to describe these phenomena. The aim of this research is to find out the factors that affect the air temperature in the Indian Ocean, find a suitable model, and obtain the best model from three approximate methods, namely the Linear Model (LM), the Generalized Linear Model (GLM), and the GAM models, which use a dataset of factors that affect the temperature of the Indian Ocean (close to Aceh region). For the air temperature of $\alpha = 0.05$, the significant effects are precipitation, relative humidity, sea surface temperature, and the wind speed. The LM, GLM and GAM models are quite feasible because they all meet and pass the classical hypothesis tests, namely the normality test, multicollinearity test, the heteroscedasticity test, and the autocorrelation test. The appropriate model is GAM model based on adaptive smoothers. Compared to the LM, GLM and GAM models, GAM model with the adaptive smoothers base gave smallest AIC values of 4552.890 and 2392.396 where modeling was without and with time variable respectively. Therefore, it can be said that the correct model used at air temperature is the GAM model for adaptive smoothers base.

2020 Mathematics Subject Classification:
2020 ACM Computing Classification Codes:

INTRODUCTION

Climate change is a global phenomenon, and there has been a tremendous interest to look at its impact. Climate change can change the structure and function of coastal and marine resources [12, 13].

Temperature is the ability level of an object to transfer and receive heat, and it can be expressed as a measure of the average kinetic energy of the object's molecular motion. Temperature is a state of hot air or cold air. The highest temperature on the Earth's surface based on latitude appears in tropical regions (near the equator) and temperatures are at its lowest in polar regions [7,8].

Haghbin *et al.* [14] provided a comprehensive review of the soft computing (SC) model for estimating sea surface temperature over the last two decades, and reviewed more than 50 papers to assess the trend of SC model application for SST estimation. They presented the frequency of application (%) of SC where 88.24% was ANN-based models, and 11.76% were other SC models.

Cheng *et al.* [15] applied the back-propagation neural network (BPNN) method to determine the subsurface temperature of the North Pacific Ocean (NPO) at 16 different depths (30, 50, 75, 100, 125, 150, 200, 250, 300, 400, 500, 600, 700, 800, 900 and 1000 m) by selecting the optimum input combination of sea surface parameters obtained from satellite measurements. They used sea surface height (SSH), sea surface temperature (SST), sea surface salinity (SSS) and sea surface wind (SSW) as the input or independent parameters for the BPNN model, and the output or dependent parameters were the temperature at 16 depths. In addition, they also included the sea surface velocity (SSV) as a new component in their study. They found that the BPNN model can accurately estimate the subsurface (upper 1000 m) temperature of the North Pacific Ocean. The corresponding mean square errors were 0.868 and 0.802 using four (SSH, SST, SSS and SSW) and five (SSH, SST, SSS, SSW and SSV) input parameters and the average coefficients of determination were 0.952 and 0.967, respectively. They concluded that, input of the SSV in addition to the SSH, SST, SSS and SSW therefore gave a positive impact on the BPNN model and helps to improve the accuracy of the estimation.

Air quality on earth shows a downward trend year after year. The increase in world population, the growth of urban and rural industries, the accumulated emissions of gases from transport and other technologies and electronic products, and human activities on natural resources, such as forest fires, land for houses, office buildings and toll roads, have caused the change in the quality of air temperature on the earth's surface. In addition to the above factors, changes in the earth's surface temperature are also considered to be caused by natural factors, such as rain, humidity, solar radiation, sea surface temperature and wind speed [7,8].

The Indian Ocean is the third largest ocean in the world and is believed to account for 20% of the earth's total surface water. The Indian Ocean directly borders Indonesia, in particular Aceh province. The Indian Ocean region directly bordering Aceh province is in the tropics and the temperature is relatively high, so it is interesting to study the temperature in this region. The purpose of this research is to find out the factors that affect the temperature of the Indian Ocean, and find a suitable model to obtain the best model of three approximation methods, namely the linear model (LM), Generalized Linear Model (GLM) and Generalized Additive Model (GAM).

METHODOLOGY

Linear Model (LM)

Consider the linear assumption of the training dataset $D = (x_i, y_i)$, $i = 1, 2, \ldots n$ where x_i is the input variable (as a predictor variable) and y_i represents the output variable (as a response). The relationship between the variables X and Y can be written as a linear model in matrix form as:

$$Y = X\beta + \varepsilon \tag{1}$$

where $Y = (y_1, \ldots, y_n)^T \in R^n$ is the response variables, $\beta = (\beta_0, \ldots, \beta_p)T \in R^{p+1}$ is the unknown parameter, $X \in R^{nx(p+1)}$ is a matrix of n rows and $p + 1$ columns of a set of p input $X_0, X_1, \ldots, X_p$ of length n including an intercept, and the element of ε is assumed independent and identically distributed (i.i.d), i.e. normal random variables. The linear model in the form of equation (1) is as follows [3,4,5,6]:

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \; + \beta_k X_{1m} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \; + \beta_k X_{2m} + \varepsilon_2 \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
Y_i &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \; + \beta_k X_{nm} + \varepsilon_n
\end{aligned}$$

Hypothesis testing is a type of multiple regression analysis to see if the data used can be analysed. Below are some assumptions that must be met [1]:

Normality Test

The purpose of the normality test is to test whether the dependent variable and the independent variable have a normal distribution in the regression model. The normality test can be performed using the histogram test, the plot normality test, the chi-square test, the skewness and kurtosis test, or the Kolmogorov Smirnov test.

Heteroscedasticity Test

The purpose of the heteroscedasticity test is to test whether the regression model has unequal residual variance from one observation to another. If the residuals from one observation to another remain unchanged, it is called homoscedasticity, and if they are different, it is called heteroscedasticity. A good regression model is homoscedastic or where heteroscedasticity does not occur.

Multicollinearity Test

The multicollinearity test is designed to test whether the regression model finds a correlation between independent (mutually independent) variables. A good regression model should not produce correlation between independent variables. If the independent variables are related to each other, these variables are not orthogonal. Orthogonal variables are independent variables whose correlation value between the independent variables is equal to zero.

Autocorrelation Test

The autocorrelation test is designed to test if the linear regression model correlates between the error of use of the T period and the error T-1 (front). If there is a correlation, it is called autocorrelation problems. A good regression model is a regression of autocorrelation. The Durbin-Watson test (DW test) is used to find out if there is a self-correlation in a regression model. It requires sections in regression models, and there are no more variables among the independent variables.

Generalized linear Model (GLM)

GLM is the development of a "classical" linear model, especially in overcoming the limitations of non-normal dependent variables. However, the response variable in GLM is assumed to have a distribution that belongs to the family of exponential distributions [9,10.11].

There are three main components in GLM [5], including:

1) Random component, namely the dependent variables $Y_1, Y_2, ..., Y_n$ which are random variables where $Y_i \sim (\mu_i, \sigma^2)$ belongs to the family of exponential distribution.
2) Systematic components which are the functions of the predictor variables:

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

3). The link function that connects a function from the mean value of the random component to the systematic component: $g(\mu_i) = \eta_i$.

If Y is a random variable, both continuous and discrete, and belongs to the family of exponential distribution, then the probability density function of *Y* can be modelled as follows:

$$f_y(y, \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \quad (2)$$

where θ is called the natural parameter and ϕ is the dispersion parameter, where a, b, and c are specific functions derived based on the probability function or the probability density function of Y.

Generalized Additive Model (GAM)

GAM is an extension of linear model and generalized linear model, with the form of equations developed from additive models as well as the development of linear models into compressed linear models. The GAM model is as follows:

$$\eta = g(E(y|x_1, x_2, \ldots, x_k)) = \alpha + \sum_{j=1}^{p} \quad f_j(x_{ij}); i = 1,2,3,\ldots,n \quad (3)$$

where g is the link function and ff is a function which is assumed to be non-parametric by the smoothing method. One of the uses of GAM is to overcome the non-linear influence of independent variables that are difficult to do in the linear regression method. In addition, another reason for the use of GAM is that there is no need to make strong assumptions about the distribution of residuals as in the linear regression model [2].

The dataset used is air temperature, which is influenced by several factors. Secondary data is obtained from the Global Tropical Moored Buoy Array website, from 2000 to 2019, comprising 2321 observations. Dataset on air temperature and influencing factors were taken in the Indian Ocean. Data types is based on the time unit of collection, air temperature dataset and factors that affect air temperature in the Indian Ocean as time series data. In this study, five variables are used: air temperature (Y), rainfall (X_1), relative humidity (X_2), sea surface temperature (X_3), and wind speed (X_4).

RESULTS AND DISCUSSION

Descriptive Analysis

The data patterns of each variable used in the study are as follows:

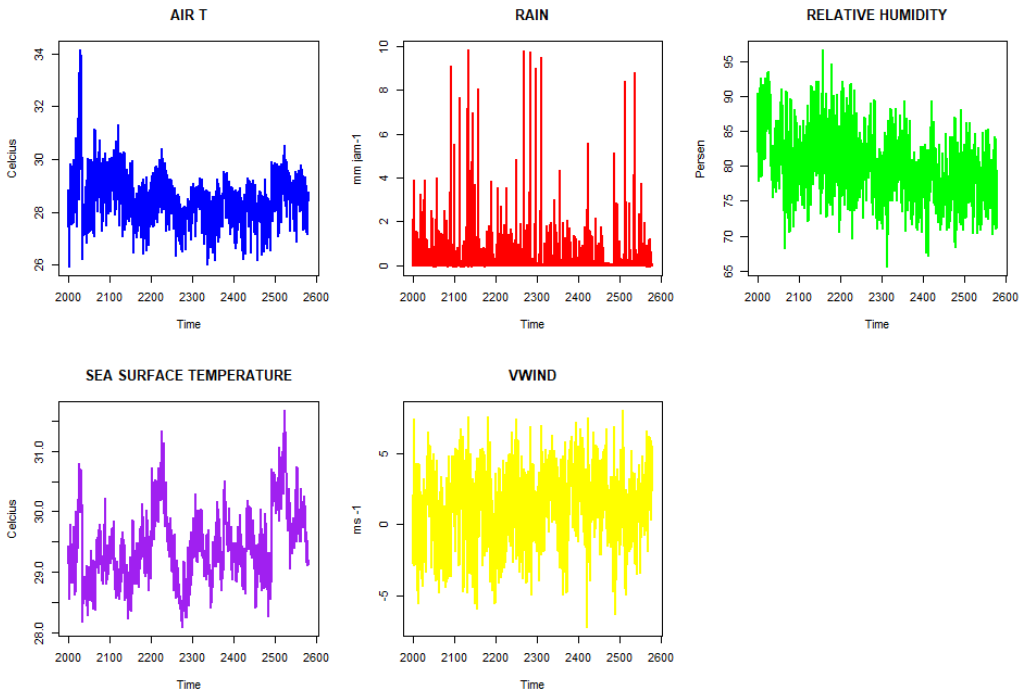

Figure 1: Air temperature dataset pattern for each variable

We can see that the air temperature in the Indian Ocean appears in the range between 26°C and 31°C as shown in the data pattern in Figure 1 (Figure Air T). The temperature only rose for a few days. Precipitation in the Indian Ocean can be seen in Figure 1 (Figure Rain), showing that the volume of precipitation is mostly low. It can be seen that the relative humidity of the Indian Ocean is between 75% and 90% (Figure 1, Figure Relative Humidity). The sea surface temperature of the Indian Ocean looks like the seasonal density shown in Figure 1 (Figure Sea Space Temperature). The wind speed in the Indian Ocean appears in the range between -5 m/s and 5 m/s.

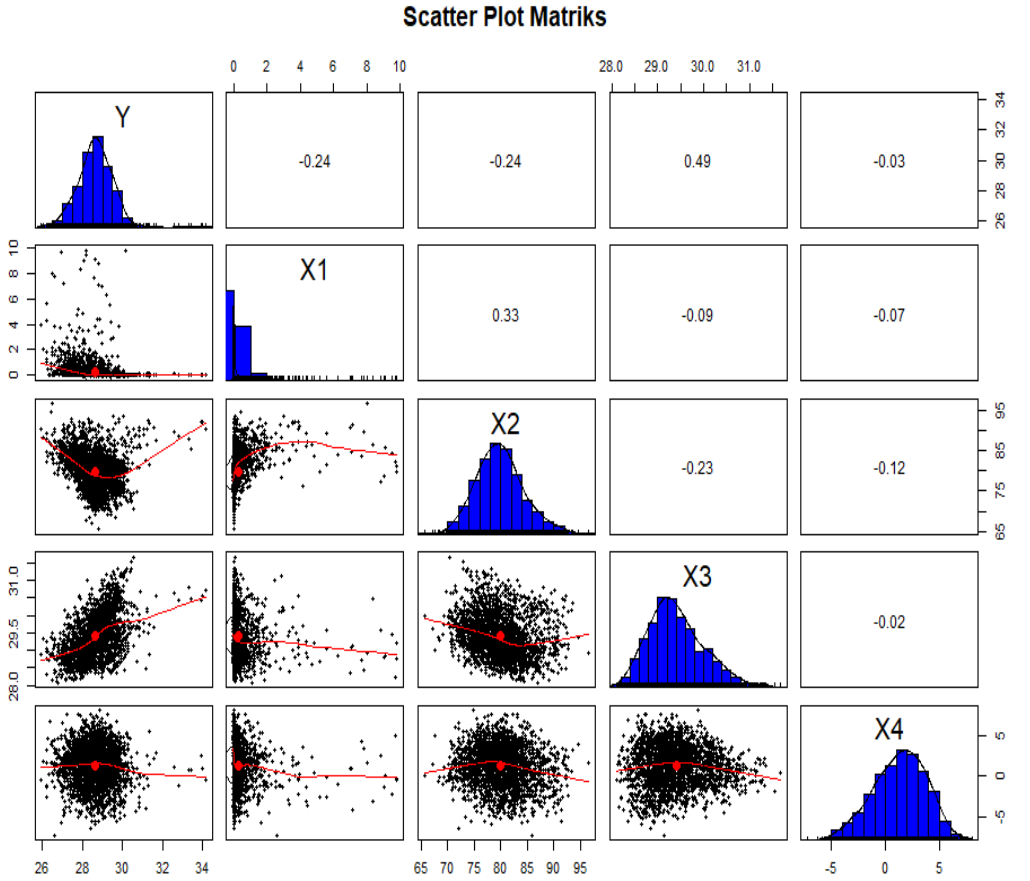

Figure 2: Scatterplot matrix of factors affecting air temperature on 2000-2019

According to the scatter diagram in Figure 2, the correlation value between temperature and rainfall is -0.24, that is, temperature and rainfall are negatively correlated. Therefore, if the temperature rises, the rainfall will decrease, and vice versa. The correlation value between air temperature and humidity is -0.24, that is, air temperature and humidity are negatively correlated. Therefore, if the air temperature increases, the humidity will decrease, and vice versa. The correlation value between air temperature and sea surface temperature is 0.49, indicating that air temperature is positively correlated with sea surface temperature. So the higher the temperature, the higher the sea surface temperature. The correlation value between air temperature and wind speed is -0.03, indicating that air temperature and wind speed are negatively correlated. Therefore, if the wind speed increases, the air temperature will decrease.

According to summary statistics, the minimum temperature variable is 25.94°C, Q1 is 28.11°C, the median is 28.64°C, the mean is 28.64°C, Q3 is 23.19°C and the maximum is 34.14°C. For precipitation, the minimum value is -0.0600 mm/hour, Q1 is 0 mm mm/hour, the median is 0 mm/hour, the average is 0.2644 mm/hour and Q3 is 0.1500 mm/hour. The maximum value is 9800 mm/hour. For humidity, the minimum value is 65.60%, Q1 is 76.80%, the median value is 79.60%, the average value is 79.87%, Q3 is 82.50% and the maximum value is 96.60%. Similarly, the minimum value of the

sea surface temperature variable is 28.09°C, Q1 is 28.99°C, the median is 29.32°C, the average is 29.40°C, Q3 is 29.75°C, and the maximum is 31.67°C. For wind speed, the minimum value is -7.200 m/s, Q1 is -0.300 m/s, the median value is 1.500 m/s, the average value is 1.314 m/s, Q3 is 3.100 m/s and the maximum value is 8.000 m/s. The minus sign of wind speed values refers to the direction of wind speed from high pressure areas to low pressure areas [16].

LM Results

We develop the initial model (M1) by using the LM approach with air temperature as the response variable and four predictor variables, which are rainfall, relative humidity, SST and wind speed.

Table 1: The results of initial model

Coefficients	Estimate	Std. Error	t value
Intercept	10.285317	0.889836	11.559
Rainfall	-0.181188	0.018626	-9.728
Relative Humidity	-0.014393	0.003743	-3.846
SST	0.665728	0.026337	25.277
Wind speed	-0.014969	0.006255	-2.393

Table 1 shows that, rainfall, relative humidity and SST have significant effects, however, wind speed gave less significant effect. The adjusted R^2, R^2 and standard error for initial model are 28.56%, 28.69% and 0.7305 respectively. To improve the result of the LM model, we then proposed the second model by adding time variables as such Nrdays (annual effect) and Doy (seasonal effect). Table 2 shows the results of second model.

Table 2: The results of second model

Coefficients	Estimate	Std. Error	t value
Intercept	1.086e+01	8.400e-01	12.923
Rainfall	-1.366e-01	1.641e-02	-8.321
Relative Humidity	-4.652e-02	3.559e-03	-13.071
SST	7.670e-01	2.633e-02	29.126
Wind speed	9.793e-03	5.568e-03	1.759
Nrdays	-3.283e-04	1.348e-05	-24.361
Doy	-1.262e-03	1.419e-04	-8.891

The adjusted R^2, R^2 and standard error of second model are 31.31%, 31.49% and 0.7163 respectively. The AIC of second model is 4515.289 which is lower than initial model of 5133.574. Time variables are Nrdays (number of days) and Doy (day of year).

GLM Results

The third model (M3) was constructed without using time variables, however we applied the Gamma distribution family. The results as follows:

Table 3: The results of third model

Coefficients	Estimate	Std. Error	t value
Intercept	5.713e-02	1.080e-03	52.888
Rainfall	2.345e-04	2.383e-05	9.841
Relative Humidity	1.722e-05	4.575e-06	3.763
SST	-8.049e-04	3.192e-05	-25.216
Wind speed	1.791e-05	7.670e-06	2.335

By comparing LM without time variables with AIC, which was 5133.574, and GLM approach with Gaussian family and function identity as link function with AIC which was 5110.9, we showed that GLM approach was better.

We then constructed the fourth model by adding time variables such as day, month and year, like in the second model. Table 4 shows the results of the fourth model. By adding the time variables we found the AIC values decreased from 5110.9 to 4489.6.

Table 4: The results of fourth model

Coefficients	Estimate	Std. Error	t value
Intercept	5.635e-02	1.015e-03	55.513
Rainfall	1.765e-04	2.082e-05	8.481
Relative Humidity	5.657e-05	4.322e-06	13.086
SST	-9.258e-04	3.182e-05	-29.100
Wind speed	-1.294e-05	6.816e-06	-1.899
Nrdays	3.974e-07	1.640e-08	24.227
Doy	1.550e-06	1.731e-07	8.953

GAM with base application

The types of bases used for testing that will be precisely used for the implementation of bases are seen from the smallest Akaike's Information Criterion (AIC) value.

Table 5: Base Determination for GAM models without and with time variable

Basis	AIC of GAM model without time variable	AIC of GAM model with time variable	
P-spline	4644.301	3166.606	R-sq.(adj) = 0.699 Deviance explained = 70.5%
Duchon splines	4638.951	22154.67	R-sq.(adj) = 0.000431 Deviance explained = -1.1e+05%
Cubic regression splines	4634.350	3132.775	R-sq.(adj) = 0.704 Deviance explained = 71.1%
Cubic splines	4634.349	3132.775	R-sq.(adj) = 0.704 Deviance explained = 71.1%
Cyclic cubic spline	4778.541	3513.964	R-sq.(adj) = 0.65 Deviance explained = 65.7%

Factor smooth interaction	4645.899	3176.978	R-sq.(adj) = 0.699 Deviance explained = 70.5%
Random effect	5133.309	4514.915	R-sq.(adj) = 0.453 Deviance explained = 45.5%
Adaptive smoothers	4552.890	2392.396	R-sq.(adj) = 0.791 Deviance explained = 80.2%
Gaussian process smooth	4631.038	3089.376	R-sq.(adj) = 0.71 Deviance explained = 71.7%
Splines regression with thin plate	4645.899	3176.978	R-sq.(adj) = 0.699 Deviance explained = 70.5%

Based on the Table 5, it can be seen that the smallest AIC value is owned by the adaptive smoothers base with an AIC value of 4552.890 and 2392.396, where the model is without and with time variable respectively. Referring to Table 5, we found that, the smallest AIC is a model with Duchon splines base, however the R-sq (adj) is so small, hence this model is not appropriate. Therefore, the model with the basis of adaptive smoothers with small AIC and the largest R-sq (adj) value of 79.10% (Deviance explained 80.2%) is selected. We also present the GAM model for independent variables and time variables in Figures 3-9.

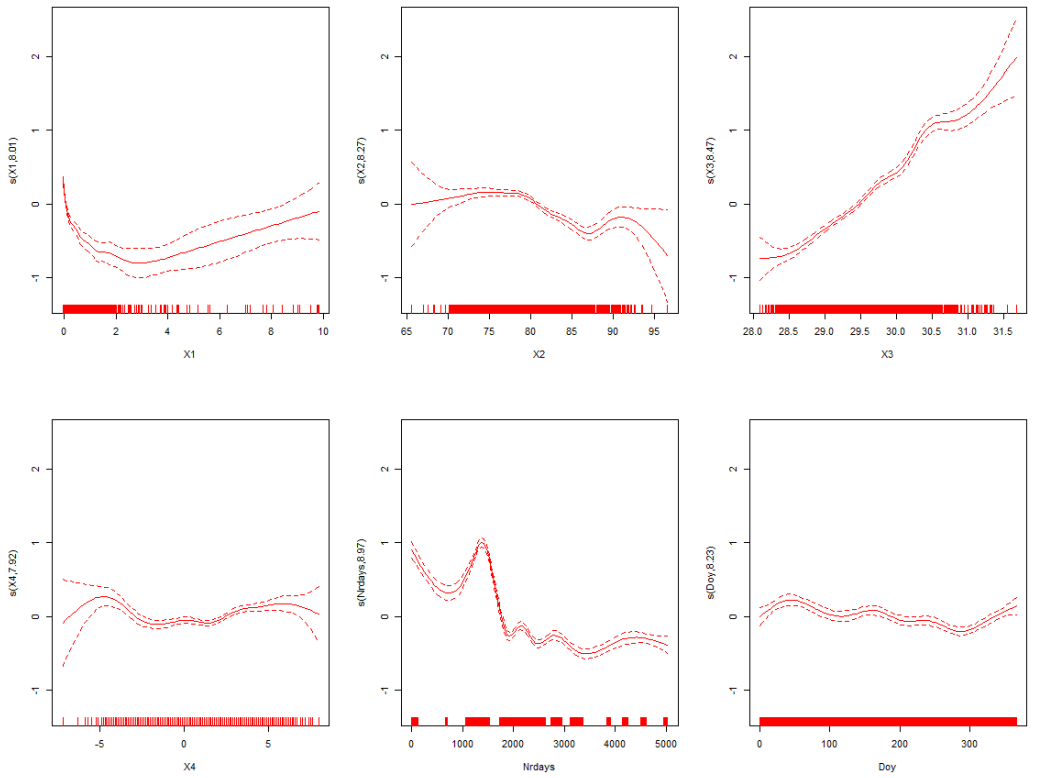

Figure 3: GAM model of independent variables and time variables with P-spline basis

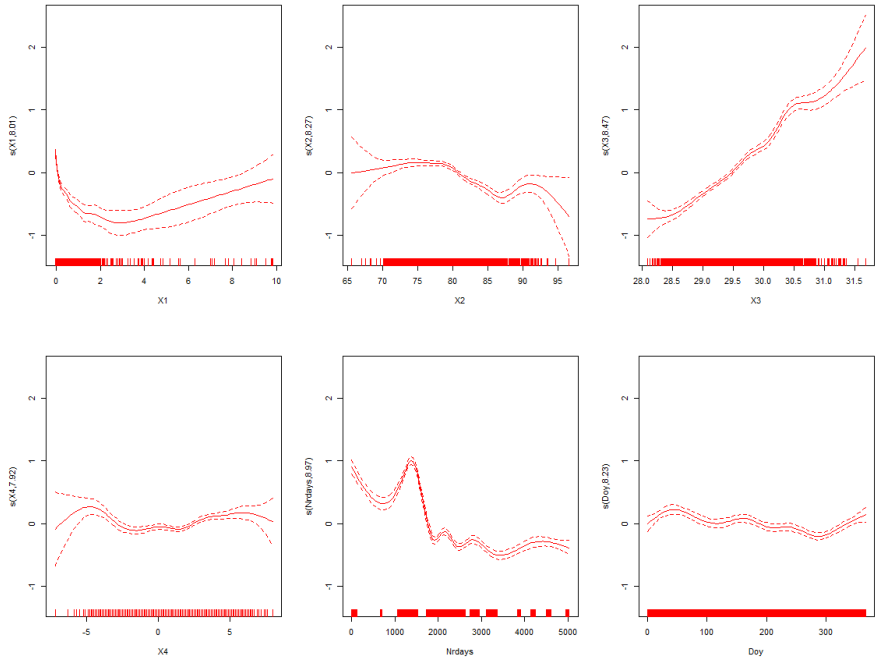

Figure 4: GAM model of independent variables and time variables with cubic regression splines

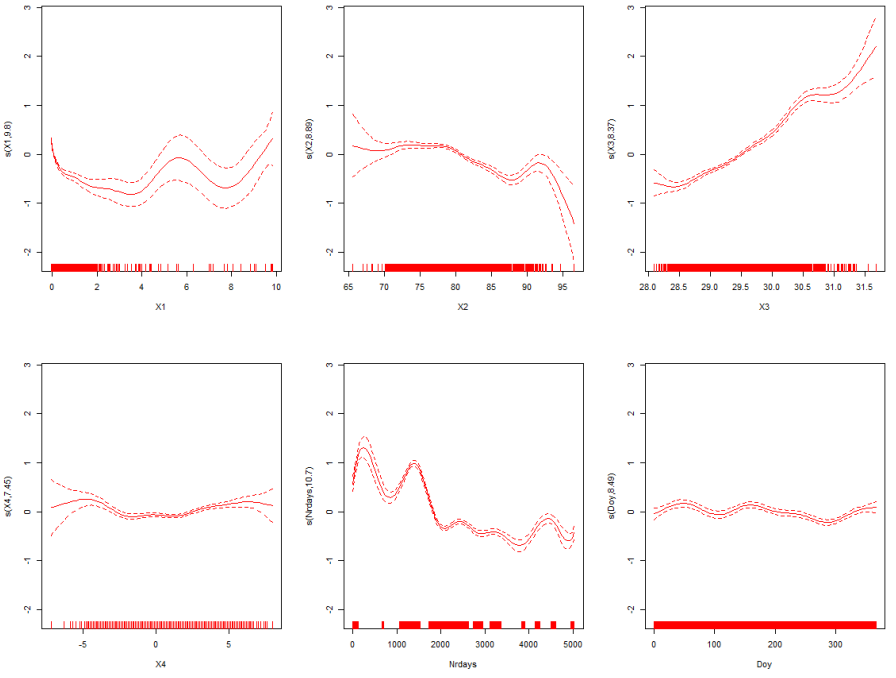

Figure 5: GAM model of independent variables and time variables with cubic splines basis

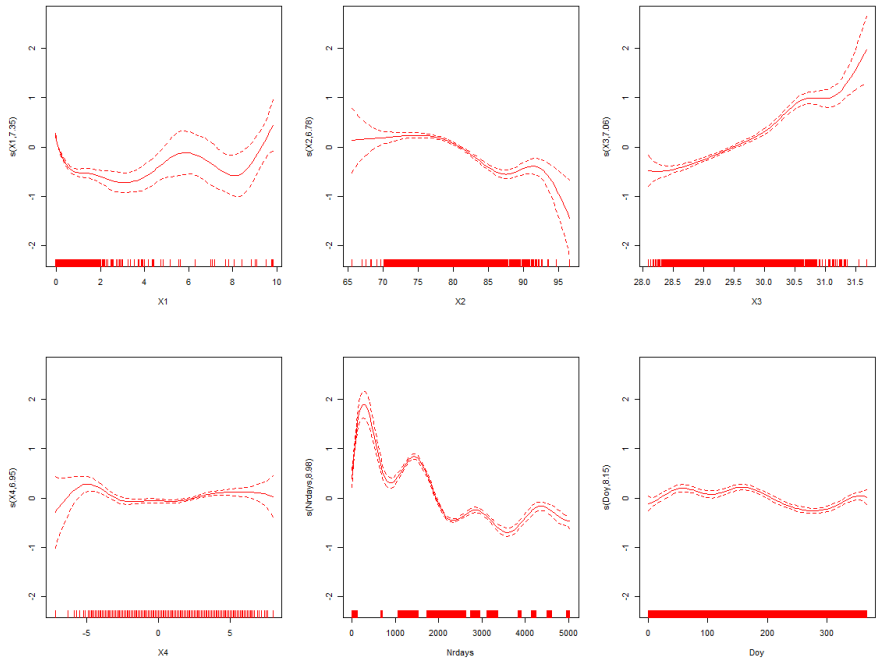

Figure 6: GAM model of independent variables and time variables with factor smooth interaction

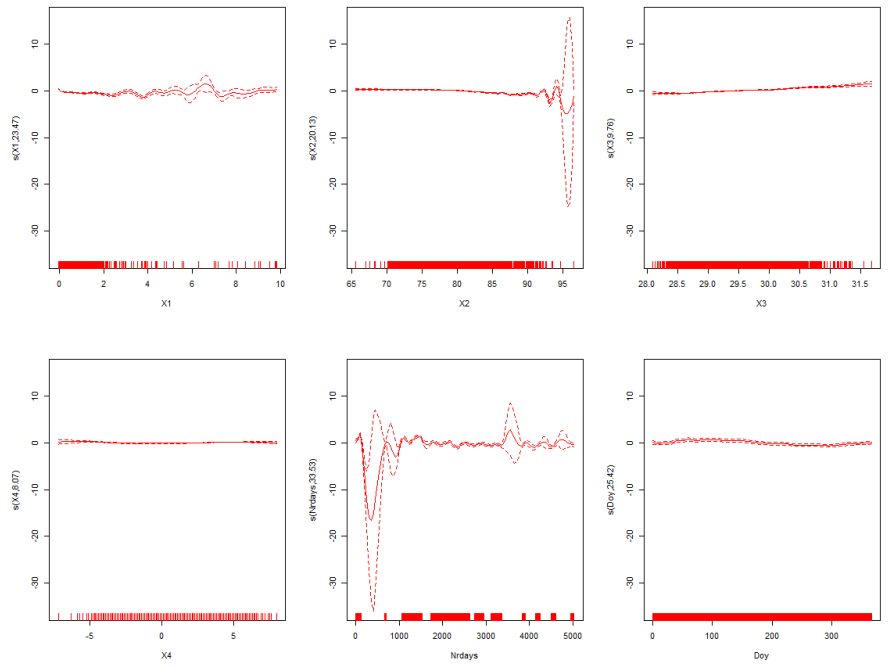

Figure 7: GAM model of independent variables and time variables with adaptive smoothers basis

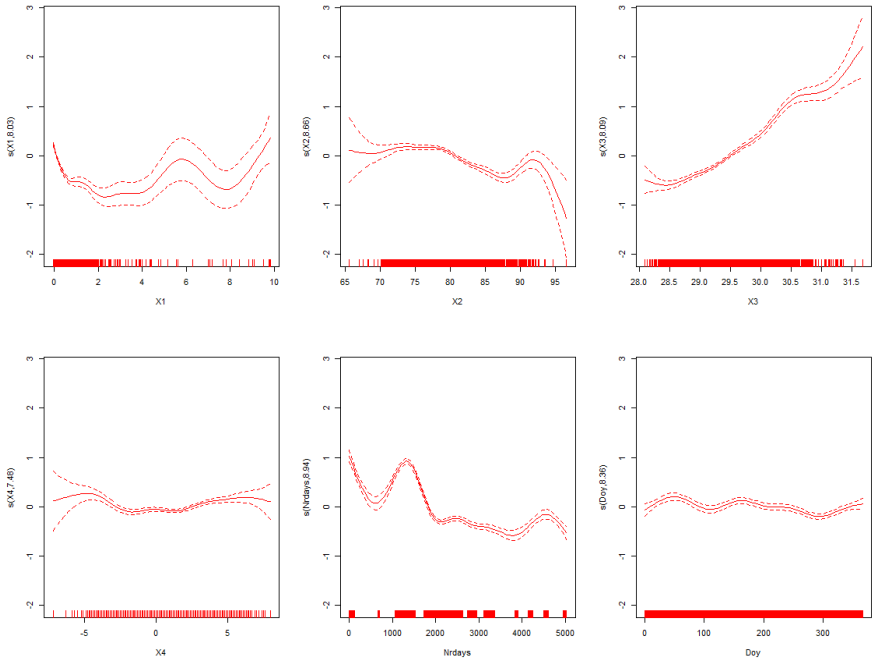

Figure 8: GAM model of independent variables and time variables with Gaussian process smooth

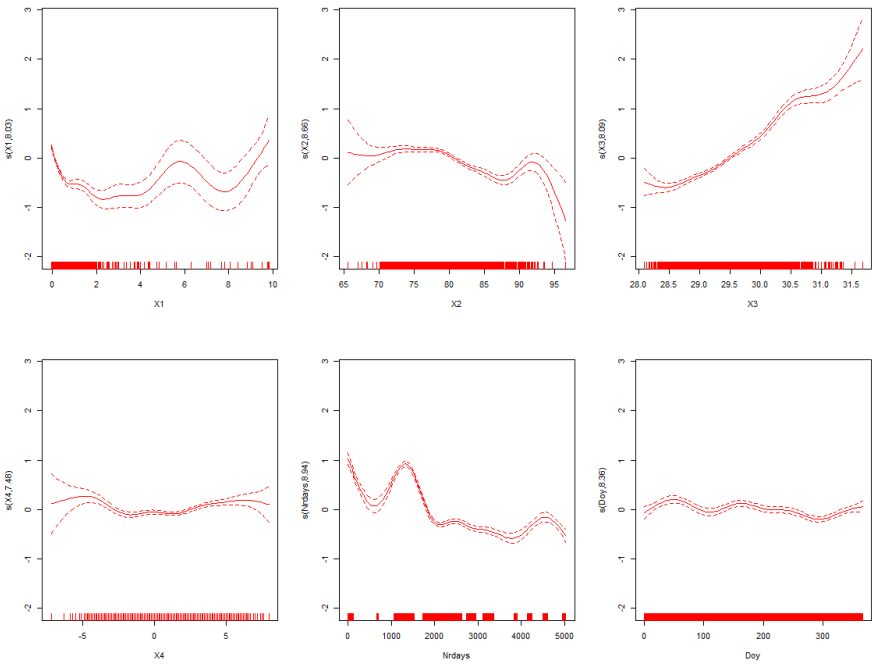

Figure 9: GAM model of independent variables and time variables with splines regression with thin plate

The selection model aims to obtain which model is suitable for use at air temperature. The best model can be seen based on the small AIC value and appropriate model visualisation. Following are the AIC values from the model.

Table 6: AIC values of the best model for air temperature dataset

MODEL	AIC of GAM model without time variable	AIC of GAM model with time variable
LM	5133.574	4515.289
GLM	5110.900	4489.600
GAM with adaptive smoothers &	4552.890	2392.396
P-spline basis	4644.301	3166.606

GAM model based on adaptive smoothers has the smallest AIC value compared to LM and GLM models, which is 2392.396. So it can be said that the right model used at air temperature is the GAM model with the application of the adaptive smoothers basis. However, the appropriate model visualisation is GAM model with P-spline basis. This can also be proven based on the following model plots,

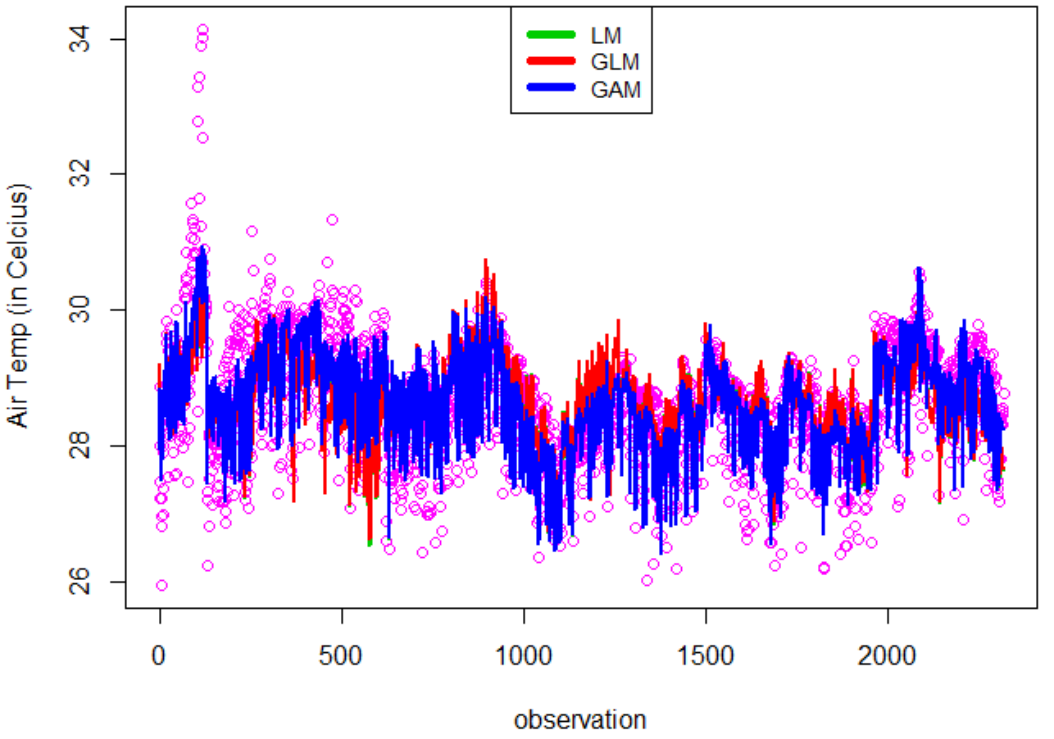

Figure 10: Comparisons of LM, GLM and GAM models with time variables (circle = air temperature)

Figure 10 shows that the distribution of data using the GAM model with the application of the P-spline basis has followed the distribution of the actual data. Therefore, the right model used in the air temperature data is the GAM model with the application of the P-spline basis.

CONCLUSION

Based on the analysis results, it can be concluded that factors influencing the temperature of the Indian ocean (close to Aceh region), which are significant to the air temperature at $\alpha = 0.05$ are rainfall (X_1), relative humidity (X_2), sea surface temperature (X_3), and wind speed (X_4). Furthermore the LM, GLM and GAM models are quite feasible because they have met and passed the classic assumption test, namely the normality test, multicollinearity test, heteroscedasticity test, and autocorrelation test. Lastly the appropriate model is GAM model based on adaptive smoothers basis, which has the smallest AIC value compared to the LM and GLM models, which is 2392.396. However, the appropriate model for visualisation is GAM with P-spline basis. So, it can be said that the right model used at air temperature is the GAM Model with the application of the P-spline basis.

CONFLICTS OF INTEREST

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Statistics Faculty of Mathematics and Sciences, Syiah Kuala University, Institute for Research and Community Service Syiah Kuala University, and Directorate of Research and Community Service (or DPRM) Ristekdikti Jakarta.

REFERENCES

[1] I. Ghozali. (2009). *Aplikasi Analisis Multivariate dengan Program SPSS*. Semarang: UNDIP.

[2] I. Komang. (2001). *Penerapan gam untuk pendugaan model produksi*. Bogor: IPB

[3] M. H., Kutner, C. J., Nachtsheim & J. Neter, W. Li. (2005). *Applied linear statistical models* (5th ed.). New York: Mc Graw-Hill.

[4] S. Weisberg. (2014). *Applied Linear Regression* (4th ed.)., Hoboken, New Jersey: John Wiley & Sons, Inc.

[5] T. Z. Keith. (2014). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (2nd ed.). New York: Routledge.

[6] P. l. Roback & J. Legler. (2021). *Beyond multiple linear regression: Applied generalized linear models and multilevel models in R*. United Kingdom: Taylor and Francis Group.

[7] M. Miftahuddin. (2016). Fundamental fitting of the SST data using linear regression models. *Proceedings 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA)*, (pp. 128-133). IEEE. Xplore, ISBN: 978-1-5090-3385-0.

[8] M. Miftahuddin & Y. Ilhamsyah. (2018). Modeling of sea surface temperature using linear models with autocorrelation Indian Ocean. *IOP Conference Series: Earth and Environmental Science*, *176*, 1755-1315, 176 (2018) 012038.

[9] P. McCullagh & J. A. Nelder. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

[10] P. K. Dunn & G. K. Smyth. (2018). *Generalized linear models with examples in R*. New York: Springer.

[11] A. G. Barnett & A. J. Dobson. (2018). *An introduction to generalized linear models* (4th ed.). United Kingdom: Taylor and Francis Group.

[12] Y. N. Kamaruzzam, M. A. Muzzneena, A. Mustapha & M. Abd Ghaffar. (2021). Determination of fishing grounds distribution of the Indian mackerel in Malaysia's Exclusive Economic Zone off South China Sea using boosted regression trees model. *Thalassas: An International Journal of Marine Sciences*, *27*. https://doi.org/10.1007/s41208-020-00282-0

[13] K. H. D. Tang. (2019). Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations. *Science of the Total Environment*, *650*, 1858-1871. doi: 10.1016/j.scitotenv.2018.09.316.

[14] M. Haghbin, A. Sharafati, D. Motta, N. Al-Ansari & M. H. M. Noghani. (2021). Applications of soft computing models for predicting sea surface temperature: A comprehensive review and assessment. *Progress in Earth and Planetary Science*, *8*(4), 1-19.

[15] H. Cheng, L. Sun & J. Li. (2021). Neural network approach to retrieving ocean subsurface temperatures from surface parameters observed by satellites. *Water*, *13*(3), 1-20.

[16] S. Prawirowardoyo. (1996). *Meteorologi*. Bandung: ITB.