Generalized additive models fitting with autocorrelation for sea surface temperature anomaly data

IOP Conf. Series: Materials Science and Engineering **523** (2019) 012002 doi:10.1088/1757-899X/523/1/012002

Generalized additive models fitting with autocorrelation for sea surface temperature anomaly data

S Ananda[1,*], **Miftahuddin**[1]
[1]Department of Statistics, Faculty of Mathematics and Sciences, Syiah Kuala University, Banda Aceh 23111, Indonesia

E-mail: shafia.ananda@students.stat.unsyiah.ac.id

Abstract. Climatic conditions in Sumatra Island are affected by Sea Surface Temperature Anomaly (SSTA) in the Indian Ocean or more commonly referred to as Indian Ocean Dipole (IOD). Extreme climate events also closely related to SSTA. Several climate features that affect SSTA such as air temperature, precipitation rain, relative humidity, wind speed, and solar radiation. SSTA is an increase or decrease in the mean of Sea Surface Temperature so required the analysis to assess extreme climatic events to the risk due to the occurrence of anomalies. Generalized Additive Models (GAM) with autocorrelation can be used to model this phenomenon. GAM method accommodates the nonlinear influence between response variables and predictor variables. The data used in this research is the time series data, daily data from 2006-2017 and there are gaps in it, where there is an autocorrelation value. The purpose of this research is to get a representative model and to know the factors that influence to SSTA. The results show that GAM's best model with autocorrelation is GAM model by including month and year variables with the monthly autocorrelation structure. Factors that affecting SSTA are the air temperature, month and year as time covariates.

1. Introduction

The Earth's climate system is affected by many parameters. Sea Surface Temperature (SST) is one of those parameters. The SST parameter is very useful in obtaining indications of Earth's climate and variability, such as tropical climate variability. SST prediction data is used as an indicator to detect many marine phenomena, such as the Indian Ocean Dipole (IOD), monsoon, and El Nino Southern Oscillation (ENSO) consisting of El Nino and La Nina phenomena [1]. 71% of the Earth's surface is an ocean, and 97% of all water in the earth with a total volume of more than 1 billion km^3 is sea water, so it is inevitable that the oceans greatly affect the movement and circulation of the atmosphere and weather in any region of the earth [2] .

Indonesia's climate conditions are strongly influenced by Indonesia's position and its surrounding atmospheric and surface status, such as the Sea Surface Temperatures Anomaly (SSTA) in the Indian Ocean region. SSTA in the Indian Ocean region is often referred to as Indian Ocean Dipole (IOD) or Dipole Mode Index (DMI) [3]. The SSTA is also closely related to extreme climatic events. The province of Aceh is a province located on the westernmost part of the island of Sumatra and faces directly with the Indian Ocean so that it has a considerable impact due to climate change caused by the incident SSTA. Key weather and climate features and also affect SSTA such as air temperature, rainfall, humidity, wind, and solar radiation [4].

*Corresponding author: shafia.ananda@students.stat.unsyiah.ac.id

An analysis that examines extreme climatic events is needed to minimize the adverse effects of the SSTA event. One method for estimating models in extreme studies is Generalized Additive Models (GAM) [5]. GAM methods can accommodate well with non-linear influence between response variables and predictor variables. The modelling started from model fitting using GAM whereas previously it has been shown that linear regression model as fundamental fitting for SST data [6]. The data used in this research comes from the National Oceanic and Atmospheric Administration (NOAA) with 8N90E position, which is a point close to Aceh Province.

2. Literature Review

2.1. Sea Surface Temperature Anomaly (SSTA)

SSTA in western Sumatra and eastern Africa is often referred to as the Indian Ocean Dipole (IOD) or Dipole Mode Index (DMI), an sea surface temperature anomaly in the Indian Ocean characterized by a zonal pattern. SSTA is an increase or decrease in the mean of Sea Surface Temperature (SST). SSTA is calculated based on the result of the reduction between the actual SST value and the average SST value of the place in question [7].

If the actual SST value is higher than the average value, then the SSTA value will be positive. Otherwise, if the actual SST value is lower than the average value, then the SSTA value will be negative. Indian Ocean Dipole (IOD) has its own classification. The classification of IOD is distinguished:

1) Normal IOD, if the IOD value is in the range of -0.4 to 0.4.
2) IOD is strong, if the IOD value is in less than -0.4 and more than 0.4.

If the IOD value is < -0.4 then there is an increase in water vapor which in turn increases rainfall. Conversely, if the value of IOD is > 0.4 then there is a decrease in the number and incidence of rain so causing the incidence of drought [8].

2.2. Generalized Additive Models (GAM)

Generalized Additive Models (GAM) is the extended of Generalized Linear Models (GLM) by replacing linear functions $\sum_{j=1}^{p} \beta_j X_j$ with additive functions $\sum_{j=1}^{p} f_j (X_j)$ [9]. GAM generalizes the additive model into the form of exponential family distribution. The general form of the GAM model can be formulated as follows:

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j (X_j) \quad (1)$$

where:

$g(\mu)$ = *link function* f_j = *smoothing function* of the predictor variable
β_0 = constant coefficient X_j = predictor variable to *j*

2.3. Autocorrelation

GAM with autocorrelation structure refers to the Autoregressive Moving Average (ARMA) model. ARMA is a family of models for analyzing time series. The notation ARMA(p,q) refers to a model with p autoregressive terms and q moving-average terms [10]. The GAM model that has been formed is then formulated into a GAM model with autocorrelation in which the residual model has an autocorrelation value. The basic idea is to modelling year and month with autocorrelation structure with different order into GAM model [11].

3. Result and discussion

3.1. Testing Distribution and Link Function Determination

Based on the tests performed, SSTA histogram forms a normal distribution. But to prove it needed Kolmogorov Smirnov test and obtained P-value is 0.4325 so it can be concluded that the SSTA data is

normal distribution. There is one link function that is suitable for normal distribution data that is identity. Therefore the link function used in this research is link function identity.

3.2. Modeling

3.2.1. Linear Model. The first modeling done before forming a GAM model in this research is to make annual and seasonal pattern using Linear Model to know the pattern of SSTA occurrence based on annual effect and seasonal effect. The results obtained can be seen in Figure 1.a and 1.b below:

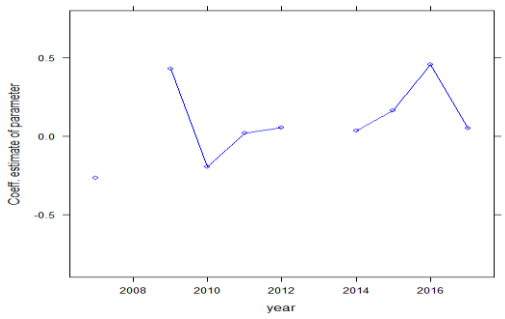

Figure 1.a. Annual pattern from annual effects

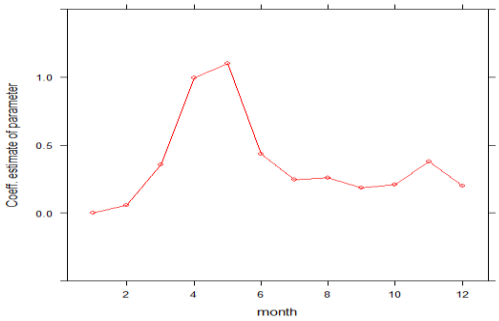

Figure 1.b. Seasonal Pattern from seasonal effects

Figure 1.a for the annual effect shows a strong SSTA decline from 2009 to 2010. While for 2011-2016 the annual effect increased strongly and declined again in 2017. Figure 1.b for seasonal effects indicates that an increase of SSTA from January to February, and a strong increase from February to May (peak season), as [12] has done for SST data. Further declines from May to October and an increase again in November.

3.2.2. Outlier Examination

a) *Studentized Residuals*

An observation is said to be the outlier data if the absolute value of studentized residuals is greater than 2 [13]. Figure 2 show a visualization of studentized residuals:

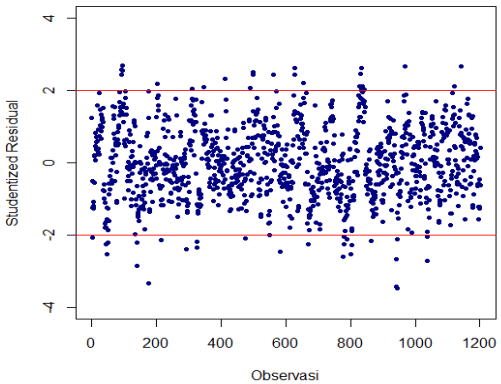

Figure 4. Studentized Residual plot

Based on figure 4 above, it can be seen that there is an observations that an outlier because has an absolute value of studentized residuals greater than 2. The picture above also shows there is 53 outlier data

3.2.3. Influential Observation Analysis

a) *Cook's Distance* (Di)

An observation is said the influential observation if the value of Di is greater than 4/(n-k-1) [13]. Table 1 show the value Di based of data from studentized residual value:

Table 1. *Cook's Distance* (Di)

Obs	*Studentized Residual*	Di	Obs	*Studentized Residual*	Di	Obs	*Studentized Residual*	Di	Obs	*Studentized Residual*	Di
2	-2.0725	0.0065	293	-2.3872	0.0030	627	2.6205	0.0053	834	2.6124	0.0066
46	-2.2668	0.0063	312	2.0576	0.0022	628	2.4447	0.0042	835	2.0047	0.0036
47	-2.5327	0.0084	326	-2.1998	0.0045	656	2.2091	0.0033	838	2.1075	0.0045
51	-2.2168	0.0058	327	-2.3427	0.0048	669	-2.2566	0.0043	844	2.0306	0.0036
93	2.4372	0.0040	347	2.0901	0.0032	778	-2.5990	0.0066	865	-2.1666	0.0033
94	2.5839	0.0049	414	2.3140	0.0065	779	-2.2610	0.0050	942	-2.6719	0.0055
95	2.5541	0.0044	474	-2.0879	0.0025	780	-2.0555	0.0043	943	-3.4377	0.0066
96	2.6898	0.0065	489	2.0648	0.0031	789	-2.1205	0.0036	944	-3.4843	0.0082
141	-2.2050	0.0052	498	2.5010	0.0036	802	-2.5304	0.0054	946	-2.1184	0.0024
142	-2.8581	0.0092	500	2.4487	0.0043	803	-2.2764	0.0041	968	2.6747	0.0049
176	-3.3452	0.3951	549	-2.0011	0.0028	828	2.1064	0.0037	1037	-2.0497	0.0033
204	2.1761	0.0043	562	2.4340	0.0050	831	2.4649	0.0060	1038	-2.7172	0.0052
214	-2.1419	0.0052	583	-2.4631	0.0152	833	2.0622	0.0045	1120	2.1232	0.0030
									1141	2.6702	0.0049

By using the value 4/(nk-1) = 4/(1202-5-1) = 0.00334, then based on table 1 it can be seen that observations 293, 312, 347, 474, 489, 549, 656, 865, 946, 1037, and 1120 are the outlier data and also have the value of Cook's Distance < 0,00334. So that it can be said that the-11 data is data the outlier data that is not the influential observations. So the data can be excluded from observation. The next model is a model by excluded the outlier data that is not the influential observations.

3.2.4. Generalized Additive Models (GAM). Before built a GAM model with autocorrelation, the first GAM model that built in this research is a GAM model with 5 predictor variables (climate features). The ANOVA table of the GAM model is as follows:

Table 2. GAM Model

Source of Variation	Edf	Ref.df	F	P-value
Air Temperature	7,766	8,557	541,72	2×10^{-16}
Precipitation Rain	8,565	8,933	22,40	2×10^{-16}
Relative Humidity	3,574	4,561	12,46	$9{,}78\times10^{-11}$
Wind Speed	8,881	8,995	93,29	2×10^{-16}
Shortwave Solar Radiation	4,355	5,404	11,01	$8{,}74\times10^{-11}$

Table 2 shows that all predictor variables are significant because they have P-value < 0.05. The visualization between actual data with fitted GAM model results for SSTA is show in figure 5:

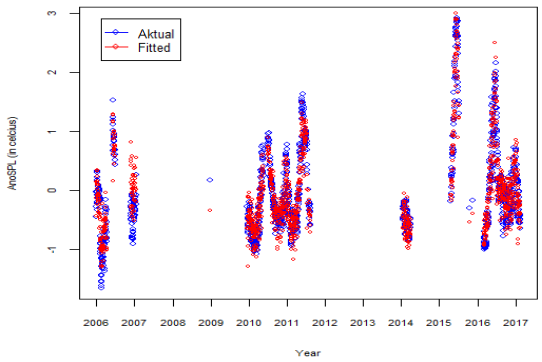

Figure 5. GAM model visualization

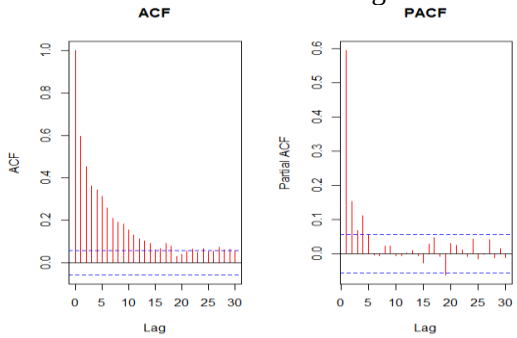

Figure 6. ACF and PACF plot GAM model

Figure 5 shows that the model is not compatible with the data because the fitted model has not approached the actual data and is still spreading. In addition there is a gap on visualization because data containing NA is excluded from the study. After that, checking of GAM's residual model was

performed using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to build GAM models with autocorrelation. Figure 6 shows that the plot of the ACF decreases exponentially and the PACF plot drops dramatically after the first lag so the GAM model with autocorrelation that suitable is the AR (1) model.

3.2.5. GAM with Autocorrelation. After build a GAM model with 5 predictors variable, then reseracher built several GAM models with autocorrelation that show in tabel 3 below:

Table 3. Summary of GAM Models with Autocorrelation Structures

Model	Time Variable	Autocorrelation	R^2 (adjusted)	AIC
GAM_1	Month and Year	Month	73,40%	-1022,501
GAM_2	Month and Year	Year	60,70%	-1512,852
GAM_3	Month	Month	72,10%	-1014,341
GAM_4	Month	Year	60,70%	-1514,852
GAM_5	Year	Month	50,80%	-1006,123
GAM_6	Year	Year	37,90%	-1413,718

Based on table 3 it can be concluded that the best model is GAM_1 model because it has the highest R^2 (adjusted) value and low AIC value.

3.2.6. Visualization of the Best Model. Visualization between actual data with fitted GAM_1 model can be seen in figure 7 below:

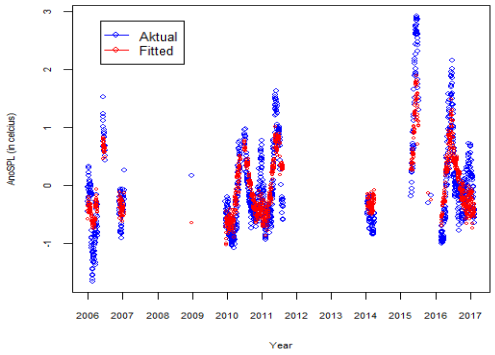

Figure 7. Visualization the best model (GAM_1 model)

Based on the visualization in figure 7 can be seen that the model has been compatible with the data because the fitted model has approached the actual data and not spread. The picture also shows that IOD turn up to be maximum around 2015 and 2016 with SSTA $\geq 2^oC$. While the negative IOD's occurred around 2007, 2010, and 2014.

3.3. Parameter Significance Testing

The parameter significance test is performed to find out whether the predictor variable used in this research has significant effect on the response variable. The significance testing of the best GAM model show in table 4:

Table 4. Parameter Significance Testing

Coefficient	Estimation	P-value	Coefficient	Estimation	P-value
Intercept	-0,0386	0,4709	Wind Speed	-0,0056	0,6956
Air Temperature	-0,1154	0,0138*	Shortwave Solar Radiation	$7{,}95\times10^{-10}$	0,9999
Precipitation Rain	-0,1004	0,2111	Month	0,6147	$5{,}02\times10^{-9}$***
Relative Humidity	0,0226	0,4336	Year	-0,1298	$1{,}68\times10^{-4}$***

Description: Significant on α 0,1% (***), 1% (**), 5% (*) and 10% (.)

Based on Table 4 can be seen that the variables that significantly affect SSTA is air temperature, month, and year with a significance level of 5%.

4. Conclusion

The best model in this research is GAM_1 model, that is GAM model which includes month and year factor with month autocorrelation structure. Measurements of R^2 adjusted and AIC GAM model are 73.40% and -1022,501, respectively. The GAM_1 model has also been compatible with the data because the result of fitted GAM_1 model approached the actual SSTA data. Factors that affect the SSTA is air temperature, month, and year, for period 2006-2017 at 8N90E.

5. Acknowledge

The authors thank the mentor, Dr. Miftahuddin for the help so that this research can be done.

6. References

[1] Miftahuddin. 2016. *Semirata PTN Barat Bidang Ilmu MIPA*. Sriwijaya, 0:732-741.
[2] Reid, C., Marshall, J. Logan, D., Kleine, D. 2009. *Coral Reefs and Climate Change*. The University of Queensland, Australia.
[3] Pramudia, A., Estiningtyas, W., Susanti, E., Sucianti. 2015. *Fenomena dan Perubahan Iklim Indonesia serta Pemanfaatan Informasi Iklim untuk Kalender Tanam*. Agency for Agricultural Research and Development, Ministry of Agriculture.
[4] Tjasjono, B. 1999. *Klimatologi Umum*. ITB, Bandung.
[5] Handayani, L. 2014. *Statistical Downscaling* dengan Model Aditif Terampat untuk Pendugaan Curah Hujan Ekstrim. *Tesis*. Bogor Agricultural Institute, Bogor.
[6] Miftahuddin. 2016. Fundamental fitting of the SST Data using Linear Regression Models. Program Book 12th ICSMA. ISBN 978-1-5090-3385-0.
[7] Swarinoto, Y. Makmur E.E.S. 2009. *Bulletin of Meteorology Climatology and Geophysics*, **5**(3): 55-56.
[8] BMKG. 2018. *Analisis Dinamika Atmosfer*.
[9] Hastie, T. dan Tibshirani, R. 1986. Statistical Science, **1**(3):297-318.
[10] Hamilton J.D. 1994. *Time Series Analysis*. Princeton University Press, Princeton.
[11] Wood, S. N. 2006. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Press, London.
[12] Miftahuddin. 2017. Assessment of Sea Surface Temperature in the Indian Ocean using Generalized Additive Models. *Proceeding SEMIRATA*. Jambi, 0:225-237
[13] Fox, John. 2009. *Regression Diagnostics*. McMaster University, Canada.