MODELLING SEA SURFACE TEMPERATURE USING GENERALIZED ADDITIVE MODELS FOR LOCATION SCALE AND SHAPE BY BOOSTING WITH AUTOCORRELATION

M. Miftahuddin

A thesis submitted for the degree of

Doctor of Philosophy

at the

Department of Mathematical Sciences

University of Essex

April 2016

Dedicated to

My merciful parents and family.

Friends, colleagues and all who timely pray for goodness.

Acknowledgements

I gratefully acknowledge the support of my supervisor, Prof Berthold Lausen. He has provided and assisted me with his remarkable insights, guidance, and for continually boosting myself to work harder. Due to the support, motivation and patience this research has become a reality. I am thankful to my supervisory board member Dr Alexei Vernitski for his valuable suggestions and encouragement. I would also like to give thanks to Dr Andreas M. and Dr Benjamin H. for their suggestions and inspirations on this research.

I am very thankful to the Ministry of Research, Technology and Higher Education of Indonesia for providing me a scholarship, Prof Lilik Hendrajaya, Prof Mustanir, and Prof Darussman for the support and help they have given, Prof Samsul Rizal, Rector of the Syiah Kuala University, Dr Hizir, vice Rector for Education Affairs, Dean of the Faculty of MIPA and the colleagues of the Department of Mathematics and Statistics.

I am thankful to Prof Abdel Salhi, Head of the Department of Mathematical Sciences, and thankful to the administration of the Department of Mathematical Sciences, University of Essex for their support. I would like to give thanks to Asma, Zardad, Arief, Luqiyan, Hadi, Maharani, Sofia K., Suko, Yacob, Masrizal and all my colleagues, my friends in and outside the department, PPI/KIBAR Essex, with whom I enjoyed my stay here and who helped me a lot. I am thankful to all my family members for their moral support and help.

Abstract

Sea surface temperature (SST) is one of many important parameters that influence the climate system of the earth. Modelling of and prediction from the SST data are challenging due to the fact that gaps in the data lead to incomplete information over time. Generalized additive models by boosting with location scale and shape (gamboostLSS) can be applied to overcome this problem. Moreover, they also deal with sparsity, irregular peaks, and autocorrelation in the data.

We propose in this thesis extended gamboostLSS models by considering time autocorrelation. In our experiments, we initially used 1231 daily observations in the period between November 2006 and September 2012. The data is then further extended from three different moored buoys. The data consisting of the SST as the response from buoys in the Indian Ocean and the air temperature (in Celsius), humidity (in percentage) and rainfall (in millimetre) covariates are considered from land stations in Sumatra Island.

Removing autocorrelation with an AR(1) model has a large impact on global and local model fitting. GamboostLSS-AR(1) models are an advanced technique for removing autocorrelation. We also computed marginal prediction interval with autocorrelation (MPI-AR(1)) of the model. MPI-AR(1) of the gamboostLSS-AR(1) model can be used to predict the missing data in various gaps and to obtain a prediction interval of submodels. The MPI-AR(1) that is applied to different buoys indicated that gamboostLSS-AR(1) model fitting is better than MPI by gamboostLSS model with and without transformation of rainfall. The MPI-AR(1) is more flexible to follow the pattern of the SST data fitting. Our proposed gamboostLSS-AR(1) models are more flexible, interpretable and capable to handle missing data, as well as to deal with high dimensional data and capture complex data structures.

Declaration

The work in this thesis is based on research carried out at the Department of Mathematical Sciences, University of Essex, United Kingdom. I certify that this is all my own work, unless referenced in the text, no part of this thesis has been submitted elsewhere for any other degree or qualification. No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is all my own work, unless referenced, to the contrary, in the text.

Acronyms

AIC : Akaike's Information Criterion

AR(1) : Autocorrelation Lag 1

CV : Cross Validation

CV-risk : Cross Validation-risk

df : degree of freedom

Doy : Day of the year

edf : effective degree of freedom

ER : Empirical Risk

FGD : Functional Gradient Descent

GAIC : Generalized Akaike's Information Criterion

GAM : Generalized Additive Models

gamboost : generalized additive models by boosting

gamboost-AR(1) : generalized additive models by boosting with consider Autocorrelation Lag 1

GAMLSS : Generalized Additive Model for Location, Scale, and Shape

gamboostLSS: generalized additive model for Location, Scale, and Shape by boosting

gamboostLSS-AR(1):generalized additive model for Location, Scale, and Shape by boosting with consider Autocorrelation Lag 1

gMDL : generalized Minimum Description Length

MPI : marginal prediction interval

MPI-AR(1): marginal prediction interval for Autocorrelation Lag 1

m_{stop}: stopping iteration

Pre : without transformation of rainfall

Post : with transformation of rainfall

PLS : Penalized Least Squares

PSSE : Penalized residual sum of square

RSE : Residual Standard Error

RSS : Residual Sum of Squares

Nrdays : Number of the days

v_{slf}: size-length factor

SST : Sea Surface Temperature

SS_{Model}: Sum of Squares of Model

SS_E : Sum of Squares of Error

TAO : Tropical Atmosphere Ocean

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The increase in global temperature has a significant impact on the earth's climate. The earth's climate system is influenced by a large number of parameters. Sea Surface Temperature (SST) is one of them. It affects the regional climate that influences the global and local climate variability, specifically in the tropical Indian Ocean [1,2]. There is no unique definitions of SST, due to its dependence on many factors, such as measurements of SST, instruments, at depth level positions and heterogeneous parameters related to SST data, such as depth of currents and currents velocity, ocean turbulence, salinity, short-long waves radiation, air-sea fluxes of heat, conductivity, moisture, rainfall, relative humidity, winds velocity, sea level pressure, and air temperature. A model fitting of SST dataset can be used to identify the effect of potential relationships among these many variables over time.

SST data is very useful in getting an indication of the earth's climate and its variability, such as the tropical climate variability [1,3–5]. The SST data prediction can be used as an indicator for the detection of many phenomena in the ocean, such as Indian Ocean Dipole (IOD) Mode, monsoon, and El Nino-Southern Oscillation (ENSO) which consist of El Nino

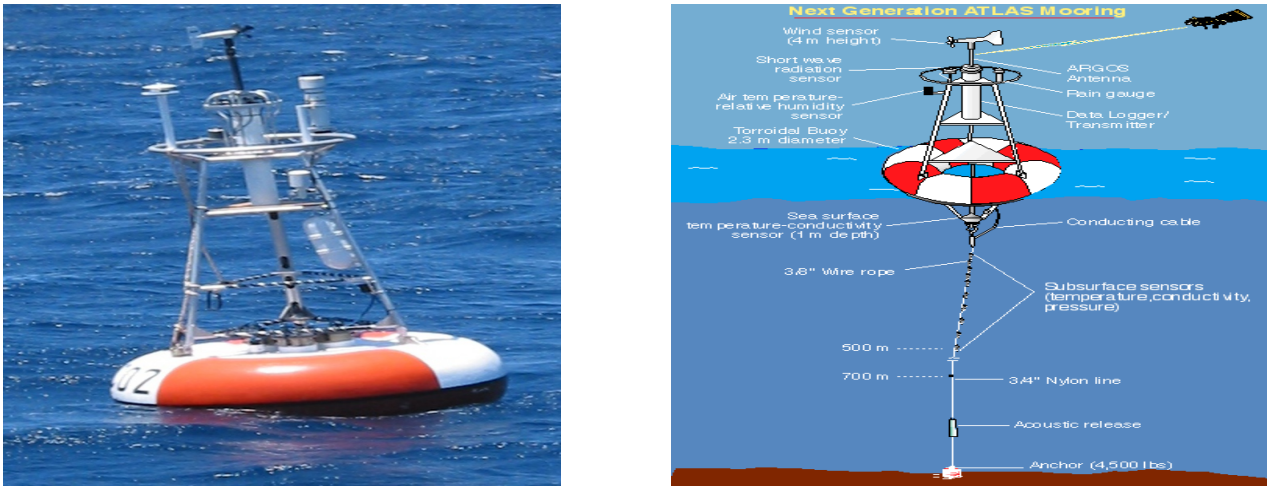
and La Nina. In [5] stated that El Nino phenomenon is close to the equator position, including eastern equatorial Indian Ocean. IOD associates with climate condition in the Indian Ocean (the third largest ocean, about 28 million square miles) and ENSO represents climate condition in the equatorial Pacific Ocean (the largest ocean, about 64 million square miles) where the SST anomalies differentiate between the eastern and western hemisphere. These phenomena can be captured by complex models, having variability and interactions among covariates with a large number of unknown parameters. However, the final model selection of global fitting and variable selection of local fitting are issues of the main concern in overall model fitting over time.

1.1 Characteristics of Sea Surface Temperature Data

The real-time measurements of SST data is usually obtained from different buoys and locations in the sea, as illustrated by moored buoy depicted in Figure 1.1. This data contains missing observations and has a complex structure.

The ocean-atmosphere plays an important role in the climate system. SSTs is a special parameter, it has a key role in circulating climate and its variability [4]. Global warming interacts with SST patterns [6]. Recently in 2012, [7] proposed a climate model to investigate the effects of solar radiation and the greenhouse effect on global warming. Their analysis is based on the data from land stations only and does not consider the relationship between sea and land data, where the surface of the earth consists of 70.9% water. In our study, the SST data is used to reveal the relationship of variables in both. The variables have different measurement scales. The SST data obtained from sea buoys and other climate data from

land station in the Sumatra Island are utilized for modelling and prediction.



(a) (b)

Figure 1.1: *A moored buoy (a), Measurements of SST data of the buoy (b),www.pmel.noaa.gov/tao/.*

The island is considered as climate region B in the regionalization related to Sea Surface Temperature and rainfall variability [8]. This location is influenced by monsoon effects and sensitivity relationship with ENSO [9]. The influence of ENSO over the Indonesian rainfall variability is around 50%, whereas the impact of SST variability over the Indian ocean is around 10-15% [8–10]. Several studies stated that the influence of ENSO in the Indian Ocean [11–13] is not significant (independent). In [11] suggested that the SST variability in this location is influenced by IOD Mode with contribution around 12%. Nevertheless, we can see that there is interaction between SST in the Indian and the Pacific oceans, such as the relationship between SST with period of annual and seasonal effects [14].

The real-time daily SST data used in the preliminary research comes from the Tropical Atmosphere Ocean (TAO) moored ocean buoy positioned at 4^0N90^0E, depth 1 m, from the period between 16 November 2006 to 26 September 2012 which is found in: www.pmel.noaa.gov/tao/. Figure 1.2 shows the SST data which was not observed due to long-term malfunction over the period of 23 July 2008 to 3 July 2010 and several days in

2007, 2011 and 2012.

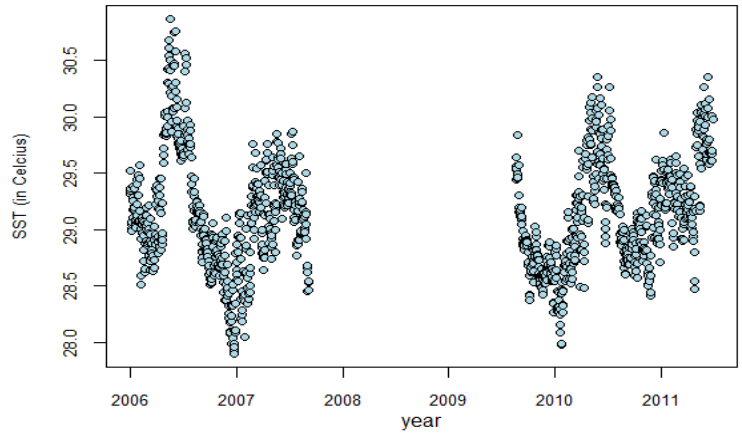


Figure 1.2: *The Sea Surface Temperature (SST) observations in degree celcius (2006 - 2012)*

The complete SST dataset is comprised of 1231 daily observations with a response variable of SST (in ^0C) at 00.00 to 12.00 pm in GMT time records, and given covariates, i.e. air temperature (in ^0C), relative humidity (in %), both covariates have average with the same time records at 07.00 am, 13.00 pm and 18.00 pm, rainfall (in mm) over three-hours period, seasonal and annual factors. Then in our study, the SST dataset is extended by including ocean data from two other buoys. The data was collected at two locations, in the Indian Ocean and Meulaboh land station, during the period of 2006 to 2012. Figures 1.3 (a), (b), and (c) show the irregular patterns of the SST data during the periods of 2006-2015. Figure 1.3 (d) illustrates several buoys position in the Indian Ocean and land stations in Sumatra island. We marked 3 particular buoys with blue circles to clearly understand their positions. Buoy 1 is at the position 4N90E, and buoys 2 and 3 are at positions 1.5N90E and 8N90E, respectively. The observed ranges are; SST (27-31 ^0C), air temperature (23-29 ^0C), relative humidity (70-100 %) and rainfall (0-400 mm). The SST data for the first part of the study comes from one buoy at the Indian Ocean in position 1.5N90E for the period of 2006-2012 with 2263 daily observation. Three climate features have several missing values,

i.e. 4.1 percent of air temperature, 0.044 percent of humidity and 4.286 percent of rainfall covariate.

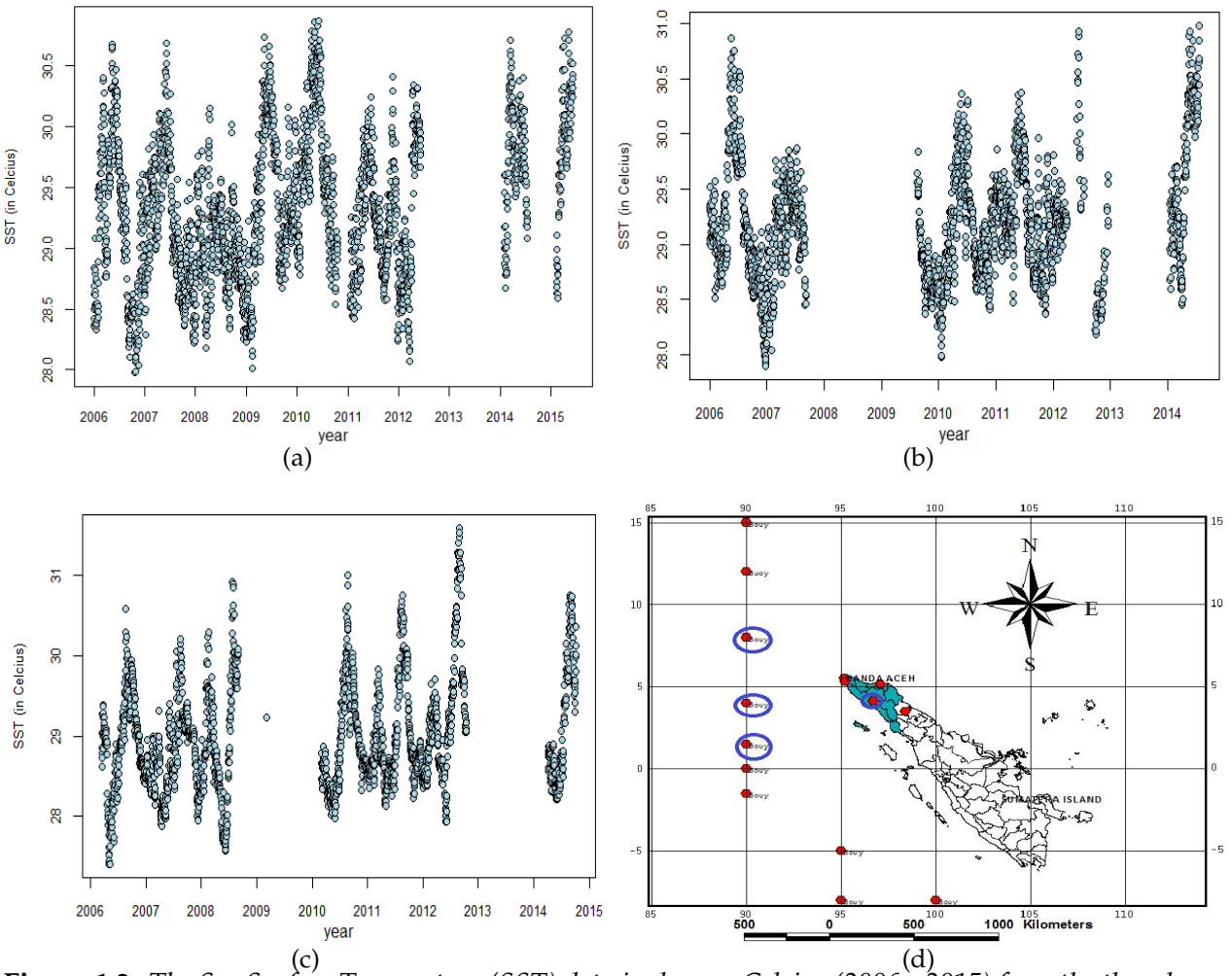


Figure 1.3: *The Sea Surface Temperature (SST) data in degrees Celcius (2006 - 2015) from the three buoys at position 1.5N90E (a), 4N90E (b), and 8N90E (c) in the Indian Ocean. The buoys positions in the Indian Ocean and Meulaboh land station are used in gamboostLSS-AR(1) experiment, in blue circles.*

Missing observations of buoys 1, 2, 3 are as follows: 0.3177 percent (711 observations), 0.0374 percent (88 observations), and 0.1089 percent (244 observations) respectively in the study period.

1.2 Flexible Framework of Additive Models

Generalized Additive Models (GAMs) as semiparametric approaches are considered to model the impact of covariates on response and are capable to capture many effects of covariates. GAM for Location, Scale and Shape (GAMLSS) is introduced by [15–17]. Boosting is one of the most important techniques for fitting regression models with improved accuracy. The boosting technique can be used to improve GAMs. The gamboost model is proposed in [18, 19], that is initially applied to predictions in binary outcome problems. The GAMLSS approach is extended by incorporating boosting, hence gamboostLSS, which is used for variable selection and to deal with high dimensional datasets in [20, 21].

The core issues of the SST data modelling and prediction are the presence of data gaps (or called missing data), sparsity, irregular peaks, and autocorrelation of the available data. Furthermore, we propose modified gamboostLSS models using basis functions to overcome these problems in model fitting and prediction of SST data. In our suggested model the autocorrelation is explicitly considered. The gamboostLSS model considering autocorrelation effect provides many useful insights. The hyper-parameters such as Location, Scale, and Shape (LSS) allow a more detailed interaction. The proposed models have similar properties as the gamboostLSS models, such as flexible structure, smoothness that incorporates many effects of covariates, interpretable, and efficient.

1.3 Structure of this Study

This thesis consists of seven chapters in total. Chapter 1 introduces the background of Sea Surface Temperature data. It also explains the characteristics of SST data, flexible frame-

work of additive models and structure of this study. Chapter 2 provides a methodology and detailed description of the SST data from buoys. In addition, we also given advantages and disadvantages of each model fitting.

In Chapter 3, we focus on these issues and provides a detailed discussion of the additive model concept to deal with the nonlinear influence of covariates. To obtain appropriate model fitting with a basis function, a flexible and precise model is needed. Hence, the additive model can be used to explore information through functional, distributional, and structural terms within the location, scale, and shape (LSS) functions.

In Chapter 4, we presents the model fitting of the SST data. We applied linear models in order to identify the effects of covariates on the SST data. The results indicate significant effects of the annual and seasonal patterns of time covariates on the model. These models are simple to apply, however, they cannot handle Location, Scale, and Shape (LSS) information of covariates and reveal of the SST data phenomena in detail. In this chapter we apply gamboostLSS models, GAM, GAMLSS, and gamboost models to the SST data. All the models result in the smallest error and overcome the fitting problems that are caused by a long gap in observations, sparsity, and nonlinearity. However, gamboostLSS provides more detailed information of the data and can handle complex data better compared to the other three models.

Chapter 5 introduces the gamboost and gamboostLSS models by allowing autocorrelation AR(1) of one buoy. The experimental comparisons show that the gamboostLSS-AR(1) models result in a better fitting. GamboostLSS-AR(1) models can be applied to different datasets collected from several buoys in the Indian Ocean and stations on Sumatra Island. In this chapter, the propose models that are applied to three data sets from different buoys.

Several statistical aspects of model fitting for exploring the data are analyzed and discussed in this chapter.

In Chapter 6, we provides general discussion of the whole results of our experiment and to provide some study insights about the properties of the model fitting related to another methods of machine learning for regression. Finally, Chapter 7 presents the conclusion and a summary of the overall findings, and possible future research using gamboostLSS-AR(1) to model SST data.

Chapter 2

Structure of Statistical Modelling

In this chapter, we present an overview of data description of SST dataset, methodology of model fitting, and objectives of our main applications, firstly, we construct a model using SST data from a buoy and secondly, using SST data from three buoys. In addition, we implement the fitting process from linear models to structured additive models and generalized additive models for location, scale, and shape by boosting with autocorrelation for sea surface temperature data as depicted in Figure 2.2.

2.1 Description of SST Dataset of A Buoy

We use performance on real-time SST data in our study to obtain realistic phenomena and data updating. Besides it is given several statistical characteristics. In this section, firstly, we provide SST dataset to fit linear models and additive models of one buoy as in Chapters 4 and 5. Secondly, we provide SST dataset to fit gamboost-AR(1) and gamboostLSS-AR(1) models of different buoys as in Chapter 6. The patterns of the data, with respect to the

three continuous covariates; air temperature, relative humidity, and rainfall, are displayed in Figure 2.1.

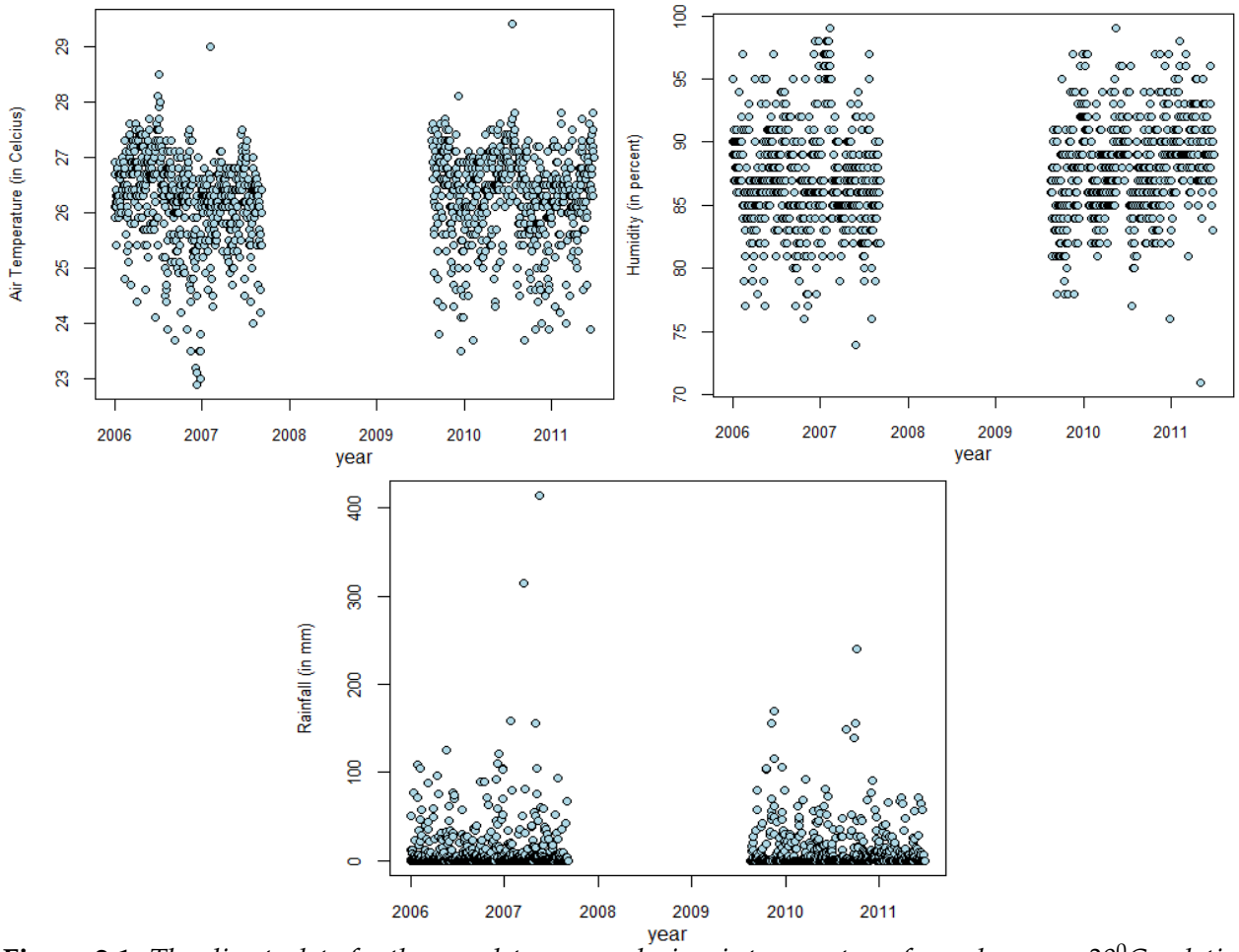


Figure 2.1: *The climate data for the complete case analysis: air temperature, few values over 29^0C; relative humidity tends relatively close to 100 percent and few values under 75 percentages; rainfall, few values over 200 millimeter.*

The bivariate relationship of the covariates are shown in the scatterplot matrix (Figure 2.2). This figure reveals that the relationship between SST and three covariates and within covariates have various degrees of correlation, i.e. the Pearson correlation between the SST and air temperature is 0.27; the SST and relative humidity is 0.02; the SST and rainfall is -0.05, the air temperature and relative humidity is -0.48; the air temperature and rainfall is -0.21; and the relative humidity and rainfall is 0.25.

We observe that the univariate distributions of the variables have one mode. The

variables have different measurement scales. The observed ranges are; SST (27-31 ^0C), air temperature (23-29 ^0C), relative humidity (70-100 %) and rainfall (0-400 mm).

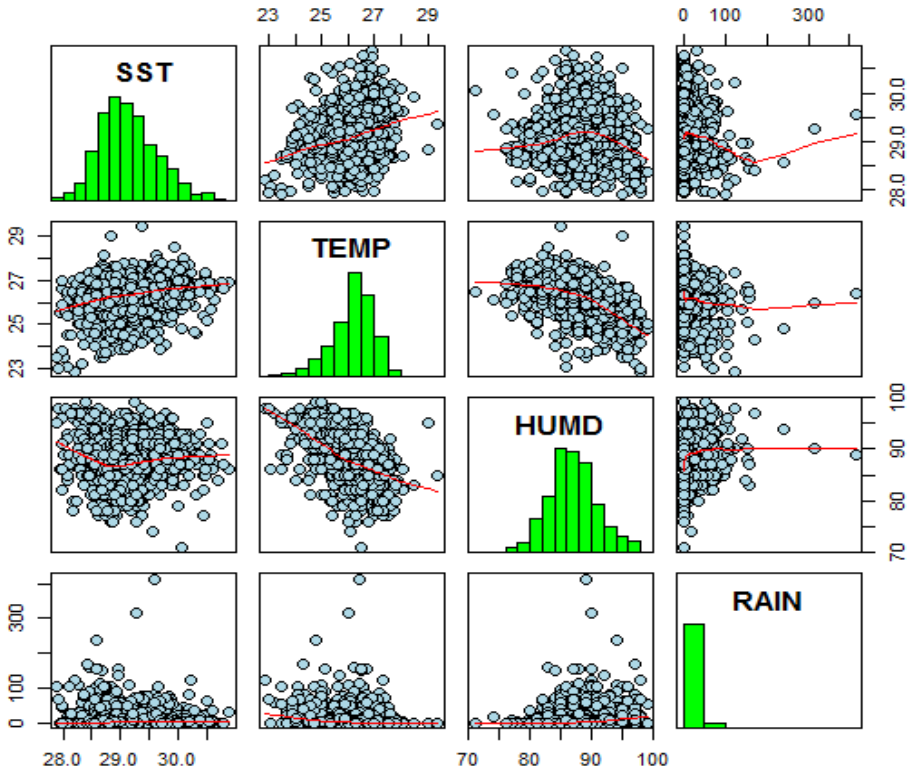


Figure 2.2: *The scatterplot matrix of daily SST, air temperature, humidity and rainfall observations. The relationship between SST response and each covariate shows characteristic changes in direction and transient trends, i.e. the SST with air temperature and the SST with relative humidity. Both patterns have data in the centre scale. The rainfall data has few extreme values. Red color represents smooth lines between a pair of variables.*

Table 2.1: *Univariate description of SST dataset*

Variable	Min	Q1	Median	Mean	Q3	Max
SST	27.90	28.78	29.09	29.13	29.44	30.87
Temp	22.90	25.80	26.30	26.24	26.80	29.40
Humd	71.00	85.00	87.00	87.45	90.00	99.00
Rain	0.00	0.00	0.80	11.75	10.05	414.00

Table 2.1 displays an overview of the numerical SST climate features with summary statistics. By comparing the temperature at sea and land, the minimum value difference is 5^0C. However, the difference at maximum point is smaller than minimum value difference. By looking at the central tendency of SST data in the Table one can see that the mean and

median of all the covariates are almost similar, except for the rainfall covariate. SST dataset have various dispersion, i.e. SST (2.97^0C), air temperature (6.5^0C), relative humidity (28%), and rainfall (414 mm).

2.2 SST Dataset of Different Buoys

We consider data from three buoys positions in the Indian ocean in the same period of time as depicted in Figure 1.3. We summarized all of the SST data of three buoys above mentioned in the following table.

Table 2.2: *Univariate description of SST data set of different buoys during the period of 2006-2012*

Buoy	Variable	Min	Q1	Median	Mean	Q3	Max
1	*SST*	27.98	28.83	29.21	29.26	29.67	30.87
	Temp	23.35	25.90	26.43	26.48	27.00	30.95
	Humd	71.00	85.00	88.00	88.12	91.00	102.23
	Rain	0.00	0.00	0.70	11.28	10.00	414.00
2	*SST*	27.90	28.78	29.09	29.13	29.42	30.87
	Temp	23.57	25.90	26.47	26.45	27.00	30.95
	Humd	74.00	86.00	89.00	89.15	92.27	102.23
	Rain	0.00	0.00	0.90	12.18	11.00	414.00
3	*SST*	27.41	28.49	28.77	28.90	29.23	31.01
	Temp	22.90	25.87	26.40	26.41	26.95	30.95
	Humd	74.00	86.00	89.00	88.97	92.00	102.23
	Rain	0.00	0.00	0.95	11.91	11.50	414.00

Table 2.2 displays the statistical description of SST climate features from three different buoys. The central tendency of SST data in the above table shows almost similar value, except for rainfall covariate. The dispersion of SST data are as follows: SST (2.89^oC, 2.97^oC, and 3.6^oC), air temperature (7.6^oC, 7.38^oC, and 8.05^oC), relative humidity (31.23%, and 28.23%), and rainfall (414mm) for buoys 1, 2, and 3, respectively. We also capture the autocorrelation of the above figures 1.3 as follows.

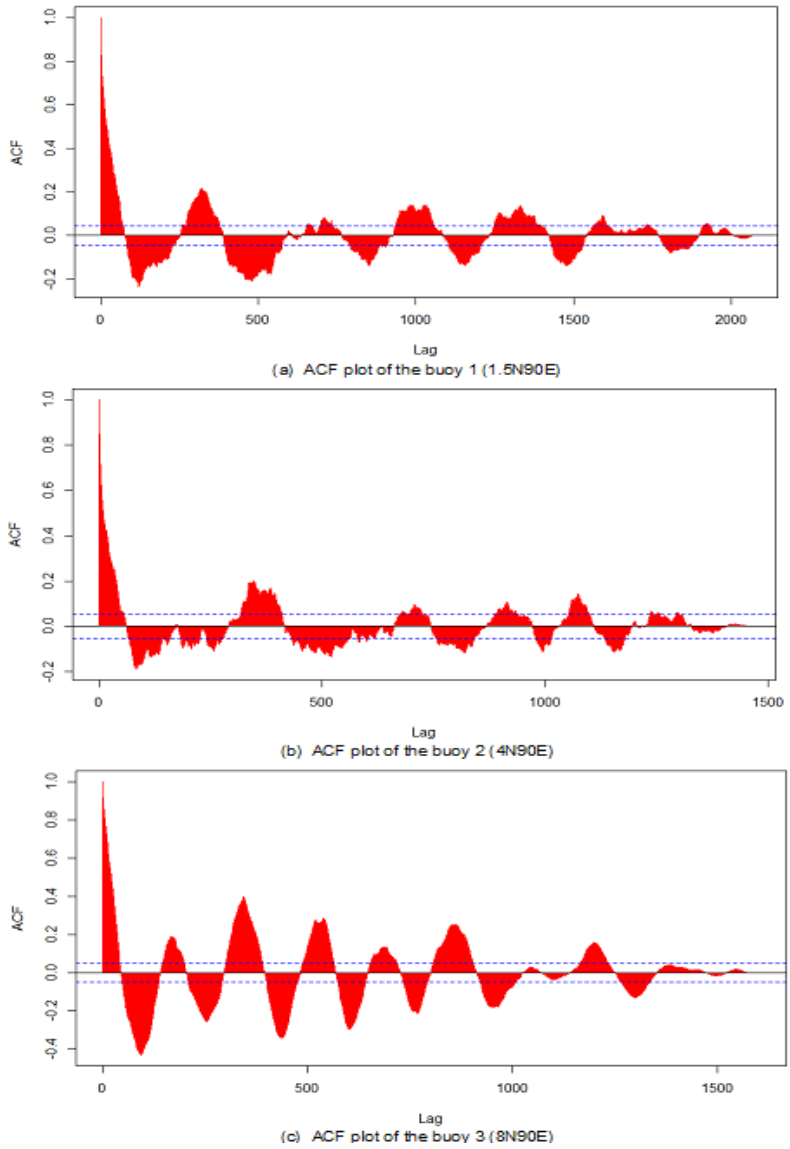


Figure 2.3: *The ACF of the SST data from three buoys using linear models show the similar patterns of buoys 1 and 2. In general, ACF of buoy 3 is bigger than ACF of buoys 1 and 2. For lag=1, ACF model of buoys 1, 2, and 3 is 0.8835944, 0.8477007, 0.9466932, respectively.*

Figure 2.3 shows positive and negative sign values alternately. More specifically, Figures 2.3(a), 2.3(b), and 2.3(c) are the autocorrelations of figures 1.3(a), 1.3(b), and 1.3(c), respectively. The ACF plot can be used to detect the pattern of autocorrelation errors of the SST data. The three figures show a high autocorrelation at lag 1. The change of pattern in peaks and magnitudes is also displayed in various period. All the figures show that ACF tends to zero at the end of the lags.

Figure 2.4: *Taxonomy of the Relationship Model Fitting of the Sea Surface Temperature Data*

2.3 Methodology

As can be seen in Figure 2.2, we begin with an investigation of the effects of time covariates using simple to complex models. Simple linear models can be utilized to get a statistical representation of the SST data.

A. Linear Models

Linear Regression Model (LRM) follows two steps, pre-fitting and fitting SST data.

A.1 Pre-fitting SST data using LRM models are the following:

(1.1) Identification of pattern for each variable by scatterplot data.

(1.2) Identification of the relationship between variables by scatterplot matrix.

(1.3) Determination of assumption for the response distribution.

(1.4) Specification of the structure of the model.

(1.5) Determination of assumption for the expectations of response variables.

A.2 Fitting SST data by a Linear Regression Model (LRM) are:

(2.1) Initially we applied a LRM to the SST data in order to study the effect of time covariates in the model as in Chapter 4. The LRM is an approach to model the conditional function of a continuous variable Y, denoted as response variable, depending on further variables or covariates X.

(2.2) Determination of scenario of model fitting.

We applied LRM in two scenarios. Firstly, we applied the model without considering the time covariates. Secondly, we applied LRM while considering time covariates along with the other covariates. From these experiments, we observed

that there is a significant effect of time variability on the SST data. In the SST data modelling, time covariates are crucially important to be included in the model.

(2.3) Identification of the effects of the covariates in the model

We initially identify the effects of covariates on the SST data. We then propose two models that have the capability to capture these patterns. First sections of the chapter describe the SST data used in detail, and in the latter sections LRMs are applied to the SST data. We visualize and interpret model to analyze effects and patterns of time covariates. Further the diagnostic analysis of the models are reported.

(2.4) While considering time covariates, we investigated the best model for LRM based on the order of covariates in the model with forward selection, backward elimination, and both methods in stepwise regression.

(2.5) Furthermore, we used LRM fitting with and without transformed covariate and also we analyze of variance for both models. Note that rainfall transformation is used for $Rain = log(RAIN + 0.01)$.

Advantages of linear model fitting are simple, easy to apply model fitting, can used without specific pre-fitting process, such as centering and/or scaling data, and capable to detect patterns of submodels included in time covariates. The model provides convenient way to estimate response with given covariates in the dataset and sum up contribution of each covariate with one coefficient. In addition, transformation effect of rainfall covariate in M1 model fitting is decreased by 0.03% and no effect on the time covariates pattern. Linear model fitting of the SST data has the applicability to

identify the data and, therefore, represents an important array of statistical diagnostic tools. However, they have many disadvantages, such as it only provides one pattern of the model, non-smooth curve, large bias, low R^2, and limited linearity.

B. An Overview of Additive Model Fitting

Power of additive modelling does not interpret "additive" in covariates context but concerns with a linear combination of functions (or the estimators are arranged additively) [19,22,23], where this linearity has flexible structure, so that it can be extended to add many functions, LSS functions, boosting, combination of LSS and boosting, etc. Hereafter, we describe four model classes which we perceive as being closely related to additive models for SST data fitting. We introduce GAM, gamboost, GAMLSS, and gamboostLSS models with P-spline basis as in Chapter 3.

C. Applied Additive Models Fitting

Furthermore, we apply the following additive models: GAM, gamboost, GAMLSS, and gamboostLSS with P-spline basis for SST data in Chapter 4. The following sub-sections provide explanation and experiments of additive models.

(1) Experimental Setup of GAM Models

Pre-fitting SST data using GAM models are as follows:

(1a) Data Identification

In our experiments we considered data collected over six years, 2006 to 2012. The data for 2009 is missing and we only have complete data for 2007 and 2011. The remaining four years contain incomplete data with missing observations.

(1b) Implement R-package with extended GAM models

We use the R-package mcgv, for GAM models proposed by Wood [24, 25] in the experiments. This package use Likelihood approach and includes several setups to extend the GAM models.

(1c) Determine measurement of the models by using AIC

We use the Akaike Information Criteria (AIC) as the performance measure for evaluation and comparison of the models.

(1d) Determine assumption of distribution function

In GAM models, as will be discussed in details in Chapter 3, the conditional mean $\mu_i = E(y_i)$ is linked to the additive covariates η_i where $\mu_i = h(\eta_i)$ with h as the response function is assumed to follow Gaussian distribution.

(1e) Specification of the relationship between response and covariates in the model

We consider the SST data observation as $(y, \mathbf{x}^T)$, where y is the response and $\mathbf{x}^T$ is the vector of covariates. The types of covariates are different: continuous and categorical for example. In Chapter 4, we model the relations of the response to the covariates in the recall equation 3.4 as in:

$$\eta(\mathbf{x}) = \beta_\mathbf{0} + f_1(\mathbf{x_1}) + f_2(\mathbf{x_2}) + f_3(\mathbf{x_3}) + f_4(Nrdays) + f_5(Doy), \tag{2.1}$$

where $\mathbf{x_1}$ is the air temperature, $\mathbf{x_2}$ is the relative humidity, $\mathbf{x_3}$ is the rainfall, *Nrdays* is the number of days and *Doy* is the day of the year.

Further by fiting GAM models with P-splines basis to the SST data we can consider the following criteria:

1) Choosing the structure of each covariate in model fitting of the SST data. This includes choice of basis function and consideration of covarites interactions or lack thereof.
2) Choosing the functional term of the covariates and distributional term of the model.
3) Checking the maximum degree or the order of P-spline in the GAM models. It is useful to control the effective degrees of freedom *edf* of each covariate in order to avoid model misfitting.
4) Dividing GAM models with and without transformed covariate, and with and without time covariates in the model fitting by Algorithm 3.
5) The degree of penalization is selected during the model fitting by using the AIC.
6) Trade-off among the AIC, *edf*, and the marginal model in the model fitting.

In GAM models fitting, we fitted SST data in two scenarios, i.e. by using the models without and with transformed rainfall covariate as described in Tables 4.11 and 4.15.

The effect of various compositions of the degrees of freedom *df* for covariates and for model on the AIC without transformation of rainfall covariate is shown in Table 4.11, whereas for with transformation of rainfall covariate is shown in Table 4.15.

Excess of GAM model fitting are flexible model, sophisticated and easy to be applied, without specific pre-fitting process as well, such as centering and/or scaling data,

and capable to detect nonlinearity patterns of submodels included in time covariates. GAM model fitting of the SST data have capability to identify the data better than LRM models and, therefore, represent an important array of statistical modelling. However, they have several shortages, such as the smallest AIC values as model measurement does not guarantee optimal fitting of the data, as well as difficult to control wiggliness in the gaps and limited to achieve irregular peaks.

(2) Experimental Setup of Gamboost Models

Following pre-fitting of SST data by gamboost models are centering of the continuous covariates. The model fitting by gamboost models are as follows:

(2a) In a stage-wise additive modelling by boosting, we use the gamboost model for model fitting, with an mboost package proposed by Buhlmann and Hothorn, *et al.* [19, 26, 27] for our experiments.

(2b) The SST data is modelled through the gamboost model in two setups, the data with its original setup and data with the transformation of rainfall covariate. The general model for gamboost is given in Model 2.3 as follows,

$$
\begin{aligned}
&\text{Model} < -gamboost(SST \sim bols(int, intercept = FALSE)+ \\
&\quad bols(Temperature, intercept = FALSE)+ \\
&\quad bols(Humidity, intercept = FALSE)+ \\
&\quad bols(Rainfall, intercept = FALSE)+ \\
&\quad bbs(Temperature, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
&\quad bbs(Humidity, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
&\quad bbs(Rainfall, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
&\quad bbs(Nrdays, df = 2.5, differences = 2, knots = 100)+ \\
&\quad bbs(Dayofyear, df = 1.5, cyclic = TRUE, boundary.knots = c(1, 365)), \\
&\quad family = Gaussian(), \\
&\quad control = boost_{control}(mstop = 1000, nu = 0.1, trace = TRUE), data = databr)
\end{aligned}
$$

(2c) Determine assumption of family distribution.

(2d) The value of the step-length of factor ν_{slf}= 0.1 in the updated step to avoid model misfitting whereas the number of boosting iterations m_{stop} is selected by cross validation. We take the ν_{slf} value at 0.1 in order to obtain an appropriate value for m_{stop} and to avoid computational time cost [28].

(2e) Evaluate measurement of the models with and without transformation of covariate by using Akaike Information Criteria (GAIC).

Pros of gamboost models fitting of the SST data are pre-fitting process, specification of base-learner for each covariate, faster than GAM model, handle nonlinearity curve, more flexible, and smooth. Whereas cons of the models are trade-off among hyper-parameters such as degrees of freedom, number of knots, stopping iteration, and step-length of factor, which sometimes trade-off for hyper-parameters are not easy practically. In fact, the model fitting have difficulty to achieve peak data. In addition, the smallest AIC value of gamboost model fitting doest not guarantee optimal fitting of the SST data.

(3) Experimental Setup of GAMLSS Models

Pre-fitting SST data by GAMLSS models consists of several steps:

(3a) We considered GAMLSS with 8^5 combinations of degree of P-spline as a starting point in fitting process (initial condition).

(3b) The GAMLSS is an extended form of the work given in R-package GAMLSS for GAMLSS models, proposed by Rigby and Stasinopoulos *et al.*, [16, 29–32].

(3c) Scenario in the algorithm of models fitting with and without transformation of covariate, we use eight as the maximum degree of P-splines smoothing and its

algorithm 4 is given in Appendix.

(3d) Determine assumption of family distribution and LSS function.

(3e) Specification of each covariate based on the initial condition.

For fitting SST data by GAMLSS models consists of:

(3f) Model fitting can be obtained in detail as composition of hyper-parameters, given pre-fitting steps.

(3g) Assess measurement of the models by using Generalized Akaike Information Criteria (GAIC) are reported.

Superiority of GAMLSS model fitting of the SST data provides detail information of the model via LSS functions and significantly decrease AIC, high degrees of freedom, without pre-fitting process such as centering, and interpretability for continuous covariates. Shortcomings of this model fitting include singularity issue, computational time cost in fitting process when determining composition of hyper-parameters for each covariate automatically as in algorithm 4, and difficult to achieve interpretable submodels especially for time covariates.

(4) Experimental Setup of GamboostLSS Models

Pre-fitting SST data by using gamboostLSS models:

(4a) We use the R-package gamboostLSS models proposed by Mayr, A. *et al.* [20, 33] in our experiments.

(4b) The setup in our research is arranged as in the following steps. Firstly, the *Nrdays* covariate with dominant parameter is observed. Secondly, the *Doy* covariate

with dominant parameter by linear and smooth base-learners is investigated.

(4c) In addition, centering of the covariates improves the prediction performance of the model. We scaled the continuous covariates, rainfall, relative humidity and air temperature by centering the data. Moreover, this centering of the covariates as a preprocessing step leads to more smooth model fitting.

(4d) When fitting the gamboostLSS models to the SST data we consider nine base-learners in each model. The setup for model G20 is given in Model 2.3:

$$
\begin{aligned}
G20 = gamboostLSS(&SST \sim bols(int, intercept = FALSE)+ \\
&bols(Temperature, intercept = FALSE)+ \\
&bols(Humidity, intercept = FALSE)+ \\
&bols(Rainfall, intercept = FALSE)+ \\
&bbs(Temperature, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
&bbs(Humidity, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
&bbs(Rainfall, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
&bbs(Dayofyear, df = 1.5, cyclic = TRUE, boundary.knots = c(1, 365))+ \\
&bbs(Nrdays, df = 2.5, degree = 2, knots = 100), \\
&families = GaussianLSS(), \\
&control = boost_{control}(mstop = 1000, nu = 0.1, trace = TRUE), data = databr)
\end{aligned}
$$

Note that all continuous covariates are used as smooth base-learners with true centers, *knots*= 20, *degree*= 3, and *difference*= 2 are fixed.

(4e) Determine assumption of family distribution for location, scale, and shape.

(4f) The value for the hyper-parameter gamboostLSS model is selected through CV-risk.

(4g) Scenario of model fitting by gamboostLSS without and with transformation setup is carried out with different values of the parameters.

Furthermore, we get the general procedure of fitting the SST data by using gamboost-LSS models is given as follows:

1. We construct base-learners through P-splines basis for covariates according to their structures, such as linear, nonlinear, and smooth functions.

2. The continuous covariates with nonlinear base-learners should be centralized (with mean), before the fitting process, then centered to guarantee identifiability [34].

3. For a continuous covariate in smooth function the degrees of freedom (df) used starts from one. A small value of df produces a minimum final risk. Selection bias of base-learners can be reduced by considering a small value for df.

4. Furthermore, continuous covariate can be modeled by smooth function of base-learner and centering by linear function without interception.

5. A time covariate can be modeled using smooth and linear function base-learners.

6. The appropriate number of knots and degrees of freedom should be considered at time covariates with gap observation.

7. Tuning parameters of control boosting (m_{stop} or ν_{slf}) can be selected by evaluating the model for these parameters in different values. The chosen parameter is from coarse to finer values (scale sizes). Initial m_{stop} begins from small to large values. Moderate value of iterations m_{stop} and ν_{slf} should be considered, where default ν_{slf} is 0.1.

8. In order to assess the appropriate number of boosting iterations, we can use a default value of out of sample empirical risk (ER).

9. ER for the hyper-parameters selection is used to Cross Validation (CV) estimations.

10. The precise parameter setting in base-learners specification and control boosting is essential in order to obtain precise model fitting and model prediction. Both of these aspects can be used to avoid misspecified models in the fitting and prediction.

11. The above mentioned steps can be implemented in the gamboostLSS models selection. These steps lead to an appropriate model fitting and prediction, efficiency in the model design and computational time.

12. In addition, checking for model fitting and model meaning (plausible interpretable) is an important step in obtaining the validation of an appropriate model for analysing the results and conclusion.

The major advantages of gamboostLSS models fitting of the SST data are much faster computationally than previous models, interpretability for continuous and time covariates, stable for the *Doy* covariate in μ and σ parameters, and also the *Nrdays* covariate stable in σ parameter. In addition, gamboostLSS models fitting via LSS function and boosting technique have reduced the stopping iteration (m_{stop}) values about 50% to fit SST data compared with gamboost models with the same specification of covariates. Whereas the disadvantages of the models fitting are for the *Nrdays* covariate slightly unstable after the gaps in μ parameter, computational cost in CV-risk when high m_{stop}, and still not an optimal global model fitting.

(5) Experimental Setup Gamboost-AR(1) Models

The following steps capture pre-fitting data using the gamboost-AR(1) model for a buoy as in Chapter 5. They are:

1) Determine Auto-Correlation Function (ACF) of the SST data at lag 1.

2) Setup the centering of the continuous covariates.

3) Determine assumption of family distribution of the model fitting.

4) Specify the hyper-parameters of continuous and time covariates in base-learners.

5) Apply the single autocorrelation coefficient of ρ of step 1) in gamboost-AR(1) model.

6) Setup scenario of model fitting with and without transformation by carrying out different values of the hyper-parameters.

The steps of gamboost-AR(1) model fitting are as follows:

1) Decide the acceptable fitting for gamboost-AR(1) model to obtain the appropriate global model fitting. The global model is related to data response, while submodel, which is called local fitting, is related to the covariates. By tuning hyper-parameters we can fit the time covariates in the model to obtain appropriate global and local models fitting.

2) Model choice to obtain the optimal models fitting by cross-validation of the final risk (CV-risk).

Superiority of gamboost-AR(1) models fitting of the SST data are faster in fitting process and more appropriate model fitting than gamboost models especially for global fitting. Gamboost-AR(1) models fitting produce more stable time covariates than GAM and gamboost models fitting. There are shortcoming of the models fitting, such as trade-off between global and local fitting is not easy to implement and less

number of submodels than gamboost models.

(6) Experimental Setup GamboostLSS-AR(1) Models

To apply gamboostLSS model fitting in autocorrelation model for the data, we use autocorrelation of AR(1) model. Then, we proposed the gamboostLSS-AR(1) model fitting for the SST dataset as in Chapter 5. Suppose a response variable and a set of covariates in $(y_i, \mathbf{x}_i^{'})$, $i = 1, ..., n$, then a procedure to find coefficient of autocorrelation ρ, when ρ is unknown, as follows;

1) Initialize dataset for the response and covariates in n observation and relationship between the variables x and y by using the Linear Model (called the LM) as in Chapter 4.
2) Construct and find the LM 1 and residual of the LM1 model, called e_1.
3) Initialize dataset for the response and covariates in a subset n-1 observation.
4) Construct and find the LM 2 and residual of the LM 2 model, called e_2.
5) Investigate autocorrelation AR(1) between residuals of the LM 1 model and residuals of the LM 2 model. Then find the coefficient of autocorrelation AR(1), called ρ_1. The slope in this model will be an estimator $\hat{\rho_1}$ of ρ_1.
6) Use the differencing method to produce a new dataset with transformed variables. Then construct and find LM with a new dataset, as the LM 3 model.
7) Produce parameter values by using the coefficient of autocorrelation AR(1) based on linear model of step 6 and estimation for initial response y in n.
8) Construct a residual matrix and find the residual from initial estimation y in n and y in n-1.

9) Develop autocorrelation AR(1) of residual of initial estimation y in n and y in n-1.

10) Find coefficient of autocorrelation AR(1), ρ_2 and a new dataset of the step 8.

11) Repeat steps 8 to 10 in order to obtain the next ρ and new dataset.

(7) Application of GamboostLSS-AR(1) Models for Different Buoys

In Chapter 5, we apply gamboostLSS-AR(1) model into two different ways. One has the similar specifications as the three buoys, while the other is different. The idea of generating single model is for the efficiency terms, which means that we save computational time. On the other hand, we build multi models of each buoy in case the data is heterogeneous.

The procedure of fitting SST data initially starts from the pre-fitting procedure which has been explained in the previous section. This procedure includes generating the scatterplot, identifying the statistical description of the data, and calculating the ACF of the SST data by using generalized least squares techniques of detecting autocorrelated errors. We can also use the residual of linear model for this purpose.

The procedure of gamboostLSS-AR(1) model fitting are as follows:

a). Determine the parameter ρ's by using generalized differencing method of AR(1) model. These values of ρ's are important to achieve the minimal residual and optimal submodel of the model fitting.

b). Determine assumption of family distribution.

c). Specify the parameters of continuous and time covariates.

d). Apply the rainfall covariate with and without transformation data.

e). Apply the single autocorrelation coefficient of ρ in gamboostLSS-AR(1) model fitting.

f). Determine the suitable fitting for gamboostLSS-AR(1) model to obtain the appropriate global model fitting, which produces submodels. The global model is related to data response, while submodel, which is called local fitting, is related to the covariates. By tuning hyper-parameters we can fit the time covariates in the model to obtain appropriate global model fitting.

g). Select the appropriate model fitting to obtain the optimal global and local models fitting by cross-validation of the final risk (CV-risk).

The gamboostLSS-AR(1) model fitting for SST data are more robust in with and without transformation of rainfall covariate covariate, sophisticated and faster in fitting process, have flexibility and pre-fitting process, and capable to detect nonlinearity patterns of submodels, it does not change patterns and trends of time covariates in with and without transformation of rainfall. However, the model fitting are still not an optimal fitting performance as indicated by low number of submodels and yet to achieve all irregular peaks.

2.4 Summary

In this chapter, we described the sea surface temperature (SST) dataset from one buoy and different buoys in the Indian Ocean. We also presented methodology of models fitting in our experiment, as well as pros and cons of each model approach from linear regression model (LRM), generalized additive model (GAM), generalized additive model by boosting

(gamboost), generalized additive model for location, scale, and shape (GAMLSS), generalized additive model for location, scale, and shape by boosting (gamboostLSS), generalized additive model by boosting with time-autocorrelation (gamboost-AR(1)), and generalized additive model for location, scale, and shape by boosting with time-autocorrelation (gamboostLSS-AR(1)).

Chapter 3

Additive Model Fitting

3.1 Introduction

So far, we have fitted SST data using linear models with different variants of the time covariates by various criteria (details are given in Chapter 4). We have observed that including time covariates in the model results in an increase in the coefficient of determination, i.e. R^2, F-test values, degrees of freedom, and decrease in residual. The SST model fitting can further be improved by increasing SS_{Model}, adjusting time-group, and reducing (SS_E), through P-spline base functions. There are several reasons for using P-spline basis in our study: this basis can be used to fit short term scale of seasonal effect and long term scale of annual effect with many fluctuations in various gaps. Besides, this basis can control wiggliness of both fluctuation effects, and in addition, the P-spline fit is capable to interpolate and extrapolate SST data fitting in discrete series (e.g. daily observations), periodic, and multi-dimensional smoothing.

We assumed that the domain in $\mathbb{R}^p$ is p-dimensional Euclidean, so that various co-

variates can be accommodated in the model fitting. Linear models are relatively simple in application and interpretation. However, these models have limitations, such as only being concerned with linear relationships, at the mean of the response (or dependent variables), sensitive for outliers, and independency of data. By using linear models (M0 and M1) it shows that there is a nonlinear trend and the extreme values of the covariate (e.g. rainfall) with several outliers. Due to the nonlinearity assumption in annual and seasonal effects, discrepancy could occur in model fitting. To analyze this further we apply a more sophisticated smooth modelling approach in additive models.

Furthermore, there are several advantages to using linear combination in additive models with P-spline basis, including: flexibility fashion for sparseness and irregular peaks, interpretability in high dimensional cases, and trade-off in the specification of basis components, such as degree of freedom, penalty, and knots. Several reasons why these three components are used: the degree of freedom is very important parameter due to related with the number of parameters and smooth estimates directly in model fitting; the penalty can contribute in smoothness supplementary and continuous control in fitting process; and the knots can be used to cover many gaps with equally-spaced grid of the number of knots in model smoothing. Various gaps can be interpolated and extrapolated automatically in smoothing with given the number of knots is proportional. An equally-spaced grid of knots and proportional number of knots have affects in computational speed. For detail we can see in Chapter 6 for general discussion.

Therefore, to avoid misfitting of the SST data, different model types of the P-spline basis can be used. Over-fitting can happen if the number of knots is chosen large so that giving an outcome many fluctuations or otherwise under-fitting. Moreover, the P-spline basis can

be used to capture changes over time of the SST dataset and describes complex effects.

To select an appropriate model fitting with P-spline basis we used various assessment measures, such as the Akaike Information Criteria (AIC), Generalized AIC (GAIC), generalized Minimum Description Length (gMDL) and Cross Validation of empirical risk (CV-risk).

We proposed fitting additive models with P-splines basis mainly to address the issues of nonlinearity, sparsity in the SST data, missing observations (gaps) and autocorrelation. This approach fitting has flexibility to deal with nonlinearity by smoothing models, while sparse matrix design can cover sparsity in the SST data. In addition, the approach has stable basis for large scale spline [35], the properties can accommodate various gaps in the SST data, whereas autocorrelation can handle interpolate fitting in discrete series. For the selection of models we also considered modelling issues, such as variable selection, efficiency in terms of low error and low computational time. We used these selected models for prediction.

The proposed models are GAM, gamboost, GAMLSS, and gamboostLSS with P-spline basis. The chapter is organized as follows: in Section 3.2 GAMs are described in detail, while introduction to basis function with penalized splines is given in Section 3.3, subsection 3.3.1 deals with grouped effect by base-learners, boosting for GAM and GAMLSS models is discussed in the next Section 3.4. In section 3.5, we presents functional gradient-based boosting, and then the formulation of gamboostLSS by considering time covariates in Section 3.6. The chapter closes with the summary.

3.2 Generalized Additive Models

If the Y and $\mathbf{X}$ are random variables representing response (output) and covariates (input) respectively, then the conditional relationship of covariates and response can be written as

$$Y = E[Y|X_1, ..., X_p] + \varepsilon, \text{ where } E[Y|X_1, ..., X_p] = \beta_0 + \sum_{j=1}^{p} f_j(X_j). \tag{3.1}$$

An additive model is defined as:

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \quad i = 1, ..., n, \tag{3.2}$$

where β_0 is an intercept, the f_j are types of models, such as linear, nonlinear, smooth functions, spatial, interaction, etc., incorporating the effects of covariates. The errors ε are independent of the X_j, $E[\varepsilon_i] = 0$, and $\text{var}(\varepsilon_i) = \sigma^2$, $\text{cov}(\varepsilon) = \sigma^2 I_n$.

A GAM is the extension of linear models and Generalized Linear Models (GLM) through a link function $g(.)$ with the assumption that the response variable follows some exponential family distribution. The general GAM structure is given as:

$$g(\mu) = g(E(Y|X_1, X_2, ..., X_p)), \tag{3.3}$$

where $g(.)$ is a known as a link function. In other words, from equation 3.2 it is,

$$f^*(\mathbf{X}) = \beta_0 + \sum_{j=1}^{p} f_j(\mathbf{X_j}). \tag{3.4}$$

where f^* is the expectation of the response by an interpretable additive function.

3.3 Basis Functions

To construct model fitting by structural terms, f_j can be used through a basis function with penalized splines regression based on the basis of the beta spline (B-spline) as discussed by Eilers and Marx [36,37] and by Schmid and Hothorn [27]. In the following definition of B-splines is

$$B_i(x,k) = \frac{x-\pi_i}{\pi_{i+k}-\pi_i}B_i(x,k-1) + \frac{\pi_{i+k+1}-x}{\pi_{i+k+1}-\pi_{i+1}}B_{i-1}(x,k-1),$$

where

$$B_i(x,0) = \begin{cases} 1, & \text{if } x \in [\pi_i, \pi_{i+1}); \\ 0, & \text{if otherwise.} \end{cases}$$

as the ith B-spline basis of degree k recursively, [35,38,39]. B-splines are easier implementing than fitting polynomial regression, e.g. quadratic or cubic polynomial and cubic spline interpolation. Although both fitting techniques mentioned it can used to fit the data in generated functions, in more erratic functions, and able to correlate data which doesn't follow any specific patterns. These techniques have limitation properties, such as it is better for small data sets, high degree affects some distance away, sparse data with various missing observations, and less many control points. Whereas P-splines can corporate with large basis, for example, a smooth curve can be produced by a combination linear with third or fourth degree B-splines. Hereinafter, consider the sum of squares of error (SS_E) for any function f is defined as:

$$SS_E = \sum_{i=1}^{n}(y_i - f(x_i))^2,$$

minimizing SS_E in model fitting of training data leads to infinitely many solutions, as with any function $\hat{f}$ via training data (x_i, y_i) is a solution. This issue becomes complex, where

y over time has various gap observations (missing values), for instance, SST data. To overcome this issue, we can reduce the residual as a penalized residual sum of squares given by minimizing

$$PSS_E(f,\lambda) = \sum_{i=1}^{n} (y_i - f(x_i)^2 + \lambda \int f^{''}(x)^2 dx = \underbrace{SS_E(f)}_{loss\ \ term} + \underbrace{\lambda W(f)}_{penalty\ \ term} \tag{3.5}$$

where $\lambda > 0$ is a smoothing (tuning) parameter, $\lambda = 0$ means no penalty (unpenalized estimator) and if $\lambda = \infty$ then the smoothest curve estimation, a straight line. There is variability in f smoothing, so that trade-off between loss and penalty terms is weighted by λ. λ can be used to control the bias-variance trade-off of the smoothing spline.

Moreover, the smoothing spline has parameters and degrees of freedom (*df*). Whereas, the n parameters are constrained and tend to shrink down in smoothing spline. Effective degrees of freedom (*edf*) become important to keep balancing smoothness via the lower-bias and higher-variance of the wiggliness effect. Furthermore, the penalized smoothness with penalty $W(f)$ can be used to overcome the fitting problem (wiggliness),

$$W(f) = \int f''(x)^2 dx = \int (D_2\beta)^T D_2\beta dx = \beta^T \mathbf{P}\beta, \tag{3.6}$$

where $\mathbf{P} = \mathbf{D_2^T D_2}$ with

$$\mathbf{D_2} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & ... \\ 0 & 1 & -2 & 1 & 0 & ... \\ 0 & 0 & 1 & -2 & 1 & ... \\ .... & ... & ... & ... & ... & ... \\ ... & ... & ... & 1 & -2 & 1 \end{pmatrix}.$$

In basis functions, local basis functions such as P-splines are the most complete for fitting regression models and smoothing techniques. They are also useful for slope (trend) estimation, density smoothing, and mixed models [36, 37]. This approach can be used for fitting polynomial function, density smoothing, give flexible interpolation and extrapolation related to the SST data. In [36] stated low rank smoother, equispaced knots, and a difference penalty can control wiggliness in P-splines basis with a high flexibility on order penalties. This basis has strictly local and benefit range with effective degrees of freedom [40], these properties give excess in various gaps with small aggregate or solitary data like SST.

Thus, P-splines is a standard technique to estimate GAM models and capable to handle hyper-parameters [27]. As in equation 3.5, to obtain smoothness of model fitting, it can use the negative gradient vector $\mathbf{u}$ by penalized least squares regression method,

$$PLS(\boldsymbol{\beta}) = (\mathbf{u} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{u} - \mathbf{B}\boldsymbol{\beta}) + \lambda W(\boldsymbol{\beta}, m), \tag{3.7}$$

where $W(\boldsymbol{\beta}, m) = \boldsymbol{\beta}^T\mathbf{P}\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{D}_{\mathbf{m}}^{\mathbf{T}}\mathbf{D}_{\mathbf{m}}\boldsymbol{\beta}$ as in equation 3.6 where vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ and m is the order value of the difference penalty. The response $\mathbf{u}$ can be defined by the negative gradient. Consider an estimate of function f of equation 3.4 and related to equation 3.7 is

$$\hat{f}_j(\mathbf{x}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{H})^{-1}\mathbf{X}^T\mathbf{u}, \tag{3.8}$$

where $\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{H})^{-1}\mathbf{X}^T\mathbf{u}$ is as a penalized estimate of the coefficient vector β.

Consider penalized beta regression splines where $E(y) = \mu = B\beta$, and B are the values $B_j(x; m)$ at x of the jth B-spline for degree m, it gives equal grid as knots k. In other words, a B-spline is a special polynomial function as defined in [36] and strictly local basis

functions [41]. The (m+1)th order spline can be written as

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\beta_i, \tag{3.9}$$

where β_is are unknown parameters and $B_i(x)$ are known B-spline basis functions, which are defined recursively. Choosing the proper value of k is important in that it may cause misfitting. This basis is very stable for scale spline interpolation [42].

A smooth function produces smooth curves for time covariates in calender time [41]. The choice of knots is of crucial importance in smooth functions as small number of knots can be inflexible to cover the variability of data and a large number of knot can over-fit the data [43]. It also effects the coefficients of covariates in the model. The model becomes more complex because nonlinear effects, random effects, spatial effects, and time effects are involved.

3.3.1 Grouped Effects by Base-Learners

The structural terms of a boosting model for various types of effects of covariates can be specified and identified by base-learner terms. These terms can be obtained for all types of boosting models, for example, gamboost and GAMLSS model by boosting. In gamboostLSS model fitting selection of base-learners for each covariate is essential, as each base-learner represents a type of effect of the covariates. There are several types of base-learners construction for covariates [44,45], such as, linear, ridge penalized, penalized ordinal, P-spline, bivariate P-spline or B-spline, radial basis function, constrain, etc. Details of three of these types are given here.

1. Base-learners with linear effects:

Base-learners with linear effects can be written as, $f_{j,linear}(x_j) = x_j\beta_j$. These type of effects can be used to represent linear effects for groups of covariates, interaction between covariates, and a factor covariate. This can also be used with or without intercept in the model. For a continuous covariate x, then a design matrix $\mathbf{X}$ of autocorrelation by differencing approach is

$$\mathbf{X}_{i-1} = (1, x_i - \rho x_{i-1}),\ i = 2, ..., n.$$

In a group of continuous covariate within one base-learner, i.e.,

$$\mathbf{x}_{i-1} = (x_i^{(1)} - \rho x_{i-1}^{(1)}, x_i^{(2)} - \rho x_{i-1}^{(2)}, ..., x_i^{(q)} - \rho x_{i-1}^{(q)})^T,$$

where q is the number of covariates. The ith row of continuous covariate with intercept is

$$\mathbf{X}_{i-1} = (1, x_i^T - \rho x_{i-1}^T).$$

2. Base-learners with categorical effects:

   These Base-learners are represented as, $f_{j,\text{cat}}(x_j) = \mathbf{z}^{\mathbf{T}}\beta$, where $\mathbf{z}$ are categorical effects of covariates $x_j s$. A categorical type covariate can have ordinal or nominal scale. Through basis function B_i, we can categorize a continuous variable [36, 37, 46],

$$I_k(x) = \begin{cases} 1, & \text{if } x = k; \\ 0, & \text{if } x \neq k; \end{cases}$$

   where $k = 1, ..., K$ are ordered levels. In this case, the basis function is used to

count discrete data with measurements assumption on nominal scale. The base-learners with categorical effects are to specify time-shifts, such as a shift in the level of response and a shift in the process of the submodel. For a categorical covariate $\mathbf{X}$ with p categories (e.g. time covariate) in autocorrelation is,

$$\mathbf{X}_{i-1} = (1, x_i^{(2)} - \rho x_{i-1}^{(2)}, x_i^{(3)} - \rho x_{i-1}^{(3)}, ..., x_i^{(p_{cat})} - \rho x_{i-1}^{p_{cat}}) \text{ where } i = 2, ..., n.$$

3. Base-learners of smooth effects:

   The general formula for base-learners of several smooth effects, such as smoothing spline (S-spline) [47], P-spline, kernel estimation, and local-polynomial regression is given as: $f_j(x_j) = f_{j,smooth}(x_j)$, where x_j is one of the continuous covariates. This type can be used for bivariate smooth effects like spatial terms, cyclic or periodic terms, varying coefficients or interaction terms, etc.

Model fitting by base-learners involves several components of base-learners that effects the model smoothing process. For example, the selection of the appropriate number of knots, *df*, penalty and degree of spline are essential in model smoothing [37].

The *df* is an indicator of the complexity of model fitting, and can be written as

$$df(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} Cov(\hat{y}_i, y_i).$$

It can also be approximated through effective number of parameters [34]. Furthermore, the effective degree of freedom (*edf*) for solution of equation 3.5 of a model fitting is

$$df_{fit} = trace(\mathbf{H}) = \text{number of fitted parameters},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$ is a prediction matrix (or hat matrix) of observation $\mathbf{y}$. The relationship between effective degree of freedom (*edf*) of the fit and smoother matrix in penalized spline is

$$df_{\lambda,fit} = trace(\mathbf{H}_\lambda),$$

with $m + 1 < df_{fit} < m + 1 + k$, where k is a knot and m is the degree of spline. Thus *df* is the degree of the smoother that corresponding to the smoothing parameter λ. The trace of $\mathbf{H}$ tends to the order of penalty if λ increases. As in equation 3.8, let $\mathbf{S} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X} + \lambda\mathbf{H})^{-1}\mathbf{X}^\mathbf{T}$ be smoother matrix of $\mathbf{X}$ [45], then the effective degree of freedom (*edf*) in penalized model can be defined as,

$$df_{\lambda,fit} = tr(2\mathbf{S} - \mathbf{S}^\mathbf{T}\mathbf{S}).$$

There are several approaches in automatic smoothing, such as likelihood to efficiency time, accuracy, balancing the goodness of fit and parsimony aspect for using smoothing parameter λ, the number and position of knots k and the degree of the function basis m as in equation 3.5. The empirical risk (*ER*) of classical linear model related to smoothing of equation 3.9, is defined as

$$SS_E(\lambda) = \sum_{i=1}^{n}(y_i - \hat{f}(x_i, \lambda))^2, \tag{3.10}$$

where λ is the smoothing parameter. The AIC [48] is used for model selection given as

$$AIC(\lambda) \equiv log(SS_E(\lambda)) + \frac{2df_{fit}(\lambda)}{n} = \frac{n}{2}log(SS_E(\lambda)) + df_{fit}(\lambda).$$

In term of the likelihood function, AIC can be formulated as

$$AIC = 2[-log(L) + p], \tag{3.11}$$

where L is the likelihood, p (or df) are the number of parameters and the degrees of freedom respectively, in the model. Cross Validation-risk (CV-risk) is needed to obtain an optimal stopping criterion (m_{stop}) of model fitting, where flexibility of regression P-splines smoothing can explain a large amount of candidate models fitting. We consider the model with a good trade-off between error and goodness of fit of the model. Leave-one-out cross-validation [22, 49] is defined as

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

We compute CV-risk for the leverage of observation by using

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1 - l_i}\right\}^2,$$

where l_i is the leverage statistic

$$l_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \overline{x})^2}.$$

The CV-risk for d-fold cross-validation with $d < n$ is computed as

$$CV_d = \frac{1}{d}\sum_{i=1}^{d} MSE_i.$$

The general model fitting problems would be controlling df, stability (e.g., transformation), and *knots* [50], but the flexibility of the model is attractive [51]. As alluded to in the previous CV-risk and AIC criteria for model selection, the generalized Minimum Description Length

(gMDL) criterion [52, 53] is defined as

$$gMDL = \begin{cases} k\ log(SS_E/(n-p)) + \frac{df}{2}log(F) + log(n), & R^2 \geq \frac{p}{n}; \\ k\ log(\frac{y'y}{n}) + \frac{1}{2}log(n), & otherwise, \end{cases}$$

where n = number of observations, df = vector of degrees of freedom, $k = n/2$, F is the F-ratio for testing the hypothesis. The length of df is the same as the length of SS_E and R^2 is the squared multiple correlation coefficient.

3.4 Boosting for GAM and GAMLSS Models

The relationships between the response and covariates, in the SST data, in complex situations may not be fully determined by classical linear regression models. GAMs are more capable to capture the patterns in the SST data. However, GAM models have the limitation of assuming an exponential family distribution for the response variables. GAMLSS, proposed by [30], are semi-parametric modelling approaches to overcome this limitation of GAM models. In contrast to GAM, GAMLSS relax the family distribution assumption of the response variable.

Moreover, GAMLSS regresses at every distributional parameter, for example, location, scale and shape, to a set of covariates in addition to the expected mean of the model. GAMLSS method is used to estimate the distributions of the parameters corresponding to the covariates (i.e. with additive predictor model that depending upon the covariates additively) or its own predictor [20] by a link function. To fit GAMLSS, based on covariate vector $\mathbf{x}$, the algorithm minimize the risk function to achieve an optimized prediction model

$f^*(\mathbf{x})$ [54, 55]. GAMLSS are designed for the univariate response, possess the flexibility in applying various basis types, covered variety of covariates in functional terms, and a variety of distribution can be handled (over 80 distributions), which are discrete, continuous, and mixed. However, the LSS estimation for a large number of covariates with high dimension can be used in variable selection of model fitting. The traditional fitting procedures for GAMLSSs are inconvenient in case of high dimensional data setup, it requires variable selection based on some information criteria, for example AIC.

Boosting is an ensemble technique used to improve prediction accuracy of an algorithm [41]. It has the capacity to handle various risk functions, simultaneous process between model fitting and variable selection, and addresses multi-collinearity issues. It can be used to improve fitting and prediction of additive models [56–59]. This technique not only estimates the mean, but also the distribution of additive parameters, i.e. location, scale, and shape. To address the issue of variable selection in GAMLSS a boosting technique gamboostLSS is proposed [20]. There are many merits of fitting the model by boosting algorithms, such as efficiency in computational time, capable for GLM, GAM, and complex prediction models with high dimensional data phenomena [19, 44]; gamboostLSS [20], etc.

3.5 Functional Gradient-Based Boosting

Gradient boosting is usually used in the boosting process to minimize risk function with respect to the prediction function $f^*(\mathbf{x})$. To minimize the risk function, P-splines can be used in boosting fit for additive models, one of such approach is proposed by Schmid and Hothorn [27]. [27] used boosting to improve estimates of parameters that are based on

Functional Gradient Descent (FGD).

Gradient estimation through several statistical models is known as component-wise gradient boosting (or boosting) [47, 60, 61]. One of the advantages of boosting by FGD is that the feature selection is done during the process of model fitting by adding the base-learners, there are no separate stages for model fitting and feature selection. This leads to a reduces ER. Boosting models is referred as ”stagewise additive modeling” by Friedman et al. [62], where the term additive means that boosting is an additive combination of estimators (functions). In their work they applied boosting to GAM models.

Later this approach is extended to GAMLSS by Mayr et al. [20], where they observed that fitting GAMLSS by boosting, has a direct effect on the FGD in the form of an additive term. Further they investigated that this approach can be used for each base-learner to fit the covariates component-wise in high dimensional scenarios. In [47] another approach it is proposed that using a boosting algorithm at each step, where in the structural mechanism for each covariate is fitted by the gradient vector, and it is updated only the best of the base-learners performance. We used gradient based boosting in the experiments to increase predictive accuracy of the SST data in high dimensional case. We used squared error as a measure of the risk function to obtain the loss function for model fitting and prediction.

From equation (2), we assume that Y response is univariate and continuous and the loss function ρ is assumed to be differentiable with respect to $f^*(X)$ [27, 61, 63]. To estimate the function $f^*(.)$ minimizing the expected loss function $\rho(.)$, such that

$$\hat{f^*}(.) = \text{argmin}_{f(.)} \mathbb{E}_{Y,X}[\rho(Y_i, f^*(X_i))], \tag{3.12}$$

based on training data $(y_i, x_i), i = 1, ..., n$. Also suppose that $f^*(X, \beta)$ is an approximate function with a set of parameters $\beta \in \mathbb{R}^p$. Due to the expectation in $\hat{f}^*(.)$ being unknown, so we minimize the expectation by the gradient boosting algorithm. Whereas, FGD can be implemented in the boosting algorithm [47], the loss function $\rho(Y, f^*(X, \beta))$, under these assumptions can provide a gradient method. The loss function is used to evaluate the negative gradient for each boosting iteration. Furthermore, the function $f(.)$ can be estimated through a constraint or objective minimization of the empirical risk (ER)

$$ER = \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, f^*(X_i)) \tag{3.13}$$

which is implemented by FGD [47]. In other words, to minimize ρ with respect to f function, instead minimize ER with respect to f using component-wise gradient boosting,

$$\hat{f}^*(.) = \text{argmin}_{f(.)} ER. \tag{3.14}$$

In statistical boosting, the objective is to obtain the function $\hat{f}^*(.)$. The steepest descent algorithm [47], can be used to minimize ER through fitting the negative gradient of the loss function. Gradient boosting is one approach to approximate $\hat{f}^*_M$ of f^* as a sum of M+1 base-learners developed by M boosting iterations,

$$\hat{f}^*_M = \sum_{m=0}^{M} f^*_m.$$

By gradient boosting for $m = 0$ as starting iteration is $\hat{f}^*_0$ and iteratively of steepest descent

implementing gives the negative gradient in order to minimize the loss function ρ is

$$h_{m,i} = \nabla_{f^*_{m-1}} \rho(Y_i, f^*_{m-1}(X_i)),\ 1 \leq i \leq N,$$

where $\hat{f}^*_m = \hat{f}^*_{m-1} - \gamma_m h_m$,

$$\gamma_m = \text{argmin}_\gamma \rho(\hat{f}^*_{m-1} - \gamma h_m).$$

As we know that $\hat{f}^*_0 = f^*_0$ and $f^*_m = -\gamma_m h_m,\ m > 0$, a solution cannot be directly found for $h_{m,i}$. Furthermore, training base-learners (say $\hat{u}_m$), to fit the gradient by a training set $(X_i, h_{m,i})$, where $1 \leq i \leq N$ so that,

$$f^*_m = -\gamma_m \hat{u}_m,\ m > 0 \text{ and } \hat{f}^*_m = \hat{f}^*_{m-1} - \gamma_m \hat{u}_m,$$

$$\gamma_m = \text{argmin}_\gamma \rho(\hat{f}^*_{m-1} - \gamma \hat{u}_m).$$

There are properties to control boosting that can be represented as,

$$\hat{f}^*_m = \hat{f}^*_{m-1} + \nu_{slf}\ \hat{u}_{m-1},$$

where $0 < \nu_{slf} \leq 1$ as step-length of factor (regularization parameter) and the final of control boosting estimate can be represented as,

$$\hat{f}^*_{m_{stop}} = \hat{f}^*_0 + \nu_{slf}\ \hat{u}_0 + ... + \nu_{slf}\ \hat{u}_{m_{stop}-1}\,,\ \hat{f}_j^{m_{stop}} = \sum_{m=1}^{m_{stop}} \nu_{slf} \hat{u}_j^m,$$

where $m = m_{stop}$ stopping iteration parameter. Similarly, the parameter estimates of the jth

base-learners in the m_{stop}th iteration is,

$$\hat{\beta}_j^{m_{stop}} = \sum_{m=1}^{m_{stop}} \nu_{slf} \hat{\beta}_{j,u}^{m},$$

where $\hat{\beta}_{j,u}^{m}$ as in equation 3.8. Considering the flexibility of boosting, the combination of loss functions and base-learners are different and both can form a new model [64,65].

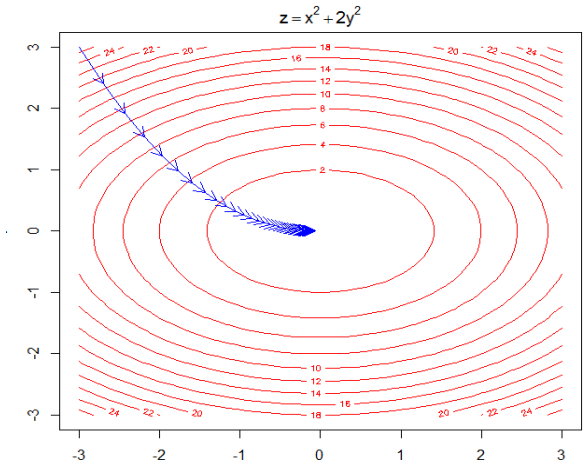


Figure 3.1: *Illustration of Functional Gradient Descent (FGD) for* $z = x^2 + 2y^2$.

A standard of FGD algorithm in [20,27,64,66] for training the dataset is as follows:

Algorithm 1 FGD (Boosting)

a) Initially setup the covariates in the function $\hat{f}^{*[0]} \equiv$ offset values with default value $\hat{f}^{*[0]} \equiv \overline{Y}$ and setup $m = 0$ in the function $\hat{f}_j^{[0]} \equiv 0$. In case SST data set, for time covariates can be represented as numeric.

b) Increase m by 1, account the negative gradients (residuals):

$$u_i = -\nabla f^* \rho(Y_i, f^*(X_i))|_{f_i = \hat{f}^{m-1}(x_i)}.$$

Compute at $\hat{f}^{[m-1]}(x_i), i = 1, ..., n$. The negative gradient vector

$$u^{[m-1]} = u_i^{[m-1]}|_{i=1,...,n} = -\nabla f^* \rho(Y_i, f^*)|_{Y=Y_i, f=\hat{f}^{[m-1]}(x_i)}.$$

(Boosting): evaluate the residuals $u_i = Y_i - \hat{f}^{[m-1]}(x_i), i = 1, ..., n$

c) Fit the negative gradients vector become residuals vector $\mathbf{u} = (u_1, ..., u_n)^T$ as base-learners $\hat{h}^m(.)$. The vector $\mathbf{u}$ is as response and fitted with covariate x; (x_i, u_i), where $i = 1, ..., n$ gives the estimate function $\hat{h}^{[m]}$.

d) Select the best fitting base-learners by using argmin with minimum SS_E:

$$j^* = argmin \sum_{i=1}^{n} (u_i - \hat{u}_{ij})^2.$$

e) Updating the prediction function with using the size length factor (slf) $0 < \nu_{slf} \leq 1$,

$$\hat{f}^{[m]} = \hat{f}^{[m-1]} + \nu_{slf} \hat{h}^{[m]}.$$

f) Repeat iteration step b to e until $m = m_{stop}$ iteration.

3.6 GamboostLSS by considering Time Covariates

To observe the functional and distributional effects in construction of model with time covariates, consider GAMLSS model without random effects

$$g_d(\phi_d) = \beta_{0\phi_d} + \sum_{j=1}^{p_d} f_{j\phi_d}(x_{dj}) = \eta_{\phi_d}, \quad d = 1, 2, 3, 4 \tag{3.15}$$

The above model consists of two terms i.e.

$\beta_{0\phi_d}$, where $d = 1$ represents index for mean μ, 2 represents index for variance σ, 3 represents index for skweness ν, and 4 represents index for shape τ. $\beta_{0\phi_d}$ are the intercept term of the four submodels; and

$\sum_{j=1}^{p_d} f_{j\phi_d}(x_{dj}) = \mathbf{X}_d \beta_d$ as a parametric term;

$f_{j\phi_d}$ are the type of effect the covariate j on the distribution parameter ϕ_d;

ϕ_d and η_{ϕ_d} are vectors,

$\beta_d^T = (\beta_{1d}, ..., \beta_{p_{d'd}})$ is a parameter vector of length $p_{d'}$;

$\mathbf{X}_d$ is a known design matrix of order n x $p_{d'}$;

For instance, $f_j\eta_d(x_{dj})$ is linear effect, smooth effect, categorical effect, and other effects depending on the characteristic of the covariates [20, 33]. Each distribution has a fitting function. Through the link-function like in equation 3.3, precision can be achieved in fitting process [32]. From above GAMLSS equation, we know that $g_d(.)$ is a monotonic link function that is related to distribution parameter ϕ_d with function η_{ϕ_d} given by

$$g_1(\mu) = \eta_\mu = \beta_{0\mu} + \sum_{j=1}^{p} f_{j\mu}(x_j) = \mathbf{X}_1\beta_1 \tag{3.16}$$

$$g_2(\sigma) = \eta_\sigma = \beta_{0\sigma} + \sum_{j=1}^{p} f_{j\sigma}(x_j) = \mathbf{X}_2\beta_2 \tag{3.17}$$

$$g_3(\nu) = \eta_\nu = \beta_{0\nu} + \sum_{j=1}^{p} f_{j\nu}(x_j) = \mathbf{X}_3\beta_3 \tag{3.18}$$

$$g_4(\tau) = \eta_\tau = \beta_{0\tau} + \sum_{j=1}^{p} f_{j\tau}(x_j) = \mathbf{X}_4\beta_4. \tag{3.19}$$

GAMLSS distributional term from equation 3.15, which represented by independent observations $(y_i, \mathbf{x_i}^T)$ for $i = 1, 2, ..., n$ where y_i is response and $\mathbf{x_i} = (x_{i1}, ..., x_{ip})^T$ is a set of the covariates vector. The conditional density function (cdf) $f_Y(y_i|\phi_i)$, depends on

$$\phi_i = (\mu_i, \sigma_i, \nu_i, \tau_i) \tag{3.20}$$

where a vector of four distribution parameters, i.e. μ_i is location parameter, σ_i is scale parameter, ν_i is skewness parameter, and τ_i is kurtosis parameter of observation ith respectively [16, 17, 20, 33]. In general, each distribution parameter is modelled through its own additive covariate η_{ϕ_d} and depend on additively on covariates effects, such as nonlinear, smooth, interaction, etc [32,33]. Location parameter of distribution refers to as the measurement of the center, such as mean, median and mode or modus; scale of distribution refers to the variance or dispersion, such as quantile, percentile, longitude-latitude and levels or layers; shape parameter of distribution refers to skewness and kurtosis. The optimization of the distribution parameters of cdf in equation 3.20 for gamboostLSS models are:

$$(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}) = \text{argmin}_{\eta_\mu, \eta_\sigma, \eta_\nu, \eta_\tau} \mathbb{E}_{Y,X}[\rho(Y_i, \eta_\mu(X), \eta_\sigma(X), \eta_\nu(X), \eta_\tau(X))], \tag{3.21}$$

with loss function $\rho = -L$ the negative log-likelihood of the response distribution and based on training data (Y, X). By equation 3.13, gradient boosting approach to minimize the ER is used,

$$ER = \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, \eta_{\phi_d}). \tag{3.22}$$

Similar equation 3.11, GAIC (or BIC) criteria in variable selection can be written as,

$$AIC(m) = -2[logL(\hat{f}^m) + df(m)],$$

$$GAIC(p) = -2 \sum_{i=1}^{n} log[f(y_i \mid \hat{\phi}_i)] + p\ df,$$

where m is iteration, p is a fixed penalty factor, [20, 45]. Mayr et al. [20] stated several limitations of this criteria, such as a potentially large variance and becoming unstable, a large number of noninformative covariates, and that it tends to be biased. Moreover,

if potentially covariates become candidate GAMLSS fits are very large, then GAIC can cause computational cost or almost impossible for high dimensional data. To overcome this limitation, [20] proposed to use a resampling scheme model fitting. By using this approach, we account for a gamboostLSS model fitting by using gradient boosting.

3.7 Summary

In this chapter, we provided a detailed discussion on additive models. The four models, generalized additive models (GAM), generalized additive models by boosting (gamboost), generalized additive models for location, scale, and shape (GAMLSS), and generalized additive models for location, scale, and shape by boosting (gamboostLSS) are introduced. We focused on the SST data fitting and prediction in case of a large number of missing observations (called gap). GamboostLSS model can deal with continuous, discrete and categorical variables. GamboostLSS also can handle the distribution of conditional location, scale and shape parameters.

Chapter 4

Linear to gamboostLSS Models Fitting for Sea Surface Temperature

4.1 Introduction

In this Chapter we presented linear regression model (LRM) and several additive models, i.e. GAM, gamboost, GAMLSS and gamboostLSS models to fit SST data from one buoy. We applied the same dataset that was used in this Chapter to observe seasonal and annual effects in model fitting. We used methodology of the models as displayed in Chapter 2 and discussed the generalised additive model for location, scale, and shape by boosting (gamboostLSS) for SST data fitting. Then, we compared our propose model with usual GAM, gamboost, and GAMLSS models.

The chapter is further organised as follows. In section 4.2 LRM models to identify the effects of covariates on the SST dataset with and without transformation. Section 4.5 GAM models with P-spline for fitting SST data are discussed. In section 4.6 gamboost models

with P-splines for fitting SST data are presented. Section 4.7 provides a detailed description of the GAMLSS models fitting for SST data. Section 4.8 gamboostLSS models fitting for SST data are given. Finally, in section 4.9 summary of the chapter is reported.

4.2 Fitting of the SST Data by Linear Regression Models

The SST data can be modelled as a linear combination of three climate parameters. These parameters are: air temperature, relative humidity, and rainfall in monthly and yearly effects. The complexity of the relationship between the response and covariates becomes challenging in a dynamic model. In the preliminary SST data observations (days) modelling, we consider the linear assumption of the training data $D = (x_i, y_i), i = 1, 2, ..., n$ where the x_is are covariates and y_i represents the response variable. The relationship between the variables X and Y can be written as a linear model in the matrix form as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{4.1}$$

where $\mathbf{Y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ are the response variables, $\beta = (\beta_0, ..., \beta_p)^T \in \mathbb{R}^{p+1}$ are unknown parameters, $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is a matrix of n rows and $p + 1$ columns of a set of p covariates $X_0, X_1, ..., X_p$ of length n including an intercept and the elements of ε are assumed independent and identically distributed (i.i.d), i.e. normal random variables $\varepsilon \in N(\mathbf{0_n}, \sigma^2\mathbf{I_n})$, where $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$ and $\mathbf{I}_n$ is in the identity matrix. Covariates can be assigned quantitative values, transformations, interaction between covariates, and variable representing nominal factors [41, 48, 67–69]. The conditional expectation illustrates the linear or functional

relationship of the parameters of the model

$$E[Y|X] = \sum_{j=0}^{p} \beta_j X_j, \tag{4.2}$$

or

$$\mu_i = E[Y|X_i] = f(X_i) \quad i = 1, ..., n. \tag{4.3}$$

To demonstrate the importance of the time covariates in the LRM models we use two models. The first model of the function $f(.)$ represented by M0 is a LRM with three continuous covariates and the second model M1 is a LRM with three continuous covariates and two time covariates, i.e. month and year. We assume that the three continuous covariates are linearly related to the SST, i.e. the relationship will be the same for all levels of the time covariates and without interaction between covariates is as follows,

$$SST_i = \beta_0 + \beta_1 Temp_i + \beta_2 Humd_i + \beta_3 Rain_i + \eta_k Month_i + \gamma_l Year_i + \varepsilon_i, \tag{4.4}$$

for $k = 1, .., 12, l = 1, 2, ..., 6$, and $i = 1, ..., n$, where η and γ are parameters vector of time covariates for month and year, respectively.

Table 4.1: *Analysis of variance for M0 regression model*

Predictor	df	SumSq	MeanSq	F-value	Pr(> F)
Temp	1	22.368	22.368	100.768	$< 2.2e-16$
Humd	1	9.477	9.477	42.694	$9.36e-11$
Rain	1	0.243	0.243	1.095	0.296
Residuals	1227	272.368	0.222		

Analysis of variance of M0 model is given in Table 4.1. The table shows that the rainfall covariate is less significant than the other two covariates. The measures of R-squared and adjusted R-squared is 10.54% and 10.32% respectively for the model (not given in the table).

These two measures provide that the model does not fit the data well.

Table 4.2: *Coefficients of M0 regression model*

Coefficients	Estimate	Std.Error	t-value	Pr(>| t |)
Intercept	21.085	0.714	29.511	$< 2e-16$
Effect of *Temp*	0.222	0.019	11.775	$< 2e-16$
Effect of *Humd*	0.026	0.004	6.617	$5.47e-11$
Effect of *Rain*	−0.0005	0.0005	−1.047	0.296

Table 4.2 shows negative effect of rainfall covariate (-0.0005) and positive effects of temperature and humidity covariates. The results reveal that the temperature and humidity have a significant effect as compared to the other covariates in the model.

Table 4.3: *Analysis of variance for the M0 model*

Source	SS	df	MS	F	Pr(>F)
Regression	32.0898	3	7.0224	31.6357	< 2.2e-16
Residuals	272.3672	1227	0.2220		
Total	304.4570	1230			

M0 is a restricted model and M1 is an unrestricted model, where in the M0 model the seasonal effect η_m and the annual effect γ_l are restricted to 0. We evaluated several models with different order of including time covariates in the model. In order to reach an appropriate model we carried out experiments by using possible $2^5 = 32$ models with time covariates as ordered, and possible 32 models with time covariates as factor. We investigated the order of covariates in the model with forward selection, backward elimination, and both methods in stepwise regression. The results of the selected model M1 are displayed in Table 4.4. The results from the table demonstrate that all the covariates are statistically significant except the rainfall covariate. It can be seen from the table that the mean squared for the temperature is higher as compared to the humidity and rainfall

covariates. Whereas, in time covariates mean squared deviations of month is higher than year. By comparing the Table 4.3 and Table 4.4 it can be seen that the residual mean squares are reduced from 0.2220 to 0.1026 by including the time covariates in the model.

Table 4.4: *Analysis of variance for M1 regression model*

Predictor	df	SumSq	MeanSq	F-value	Pr($> F$)
Temp	1	22.368	22.3683	218.1163	$< 2.2e-16$
Month	11	144.063	13.0966	127.7065	$< 2.2e-16$
Year	5	12.612	2.5223	24.5953	$< 2.2e-16$
Humd	1	0.850	0.8501	8.2894	0.004058
Rain	1	0.373	0.3735	3.6416	0.056589
Residuals	1211	124.191	0.1026		

For a more detailed analysis we categorise the covariates in the model in three categories, continuous, seasonal and annual. The results are reported in Table 4.5. We kept January as the base-line for the monthly effects and 2006 as the base-line for the annual effects. In addition, both informations might be useful to determine the SST onset or offset in calendar time. The results from the table show that all the covariates are highly significant except for the rainfall covariate. For seasonal effects the results from the table reveal that the effects for all months are highly significant, except for July. From the month coefficients, February to July show the positive magnitudes (positive estimated effect) and August to January show the nonpositive magnitudes (nonpositive estimated effect). This illustration provides seasonal patterns for dry and wet seasons, i.e. positive estimated effect for the first six months and nonpositive estimated effect for the second six months, as can be seen in the Table 4.5. The months effects are displayed in the Figure 4.1. The graph shows that there is a slight increase in the month effects from January to February and then a rapid increase from February to April. Then the direction is downwards to August.

Table 4.5: *Coefficients of M1 model*

Coefficients	Estimate	Std.Error	t-value	Pr(>| t |)
Intercept	26.3835	0.5715	46.164	$< 2e-16$
Effect of *Temp*	0.0843	0.0144	5.866	$5.74e-09$
Effect of *Humd*	0.0097	0.0030	3.184	0.0015
Effect of *Rain*	−0.0007	0.0003	−1.908	0.0566
Effect of Feb	0.0905	0.0426	2.122	0.0341
Effect of Mar	0.4639	0.0417	11.114	$< 2e-16$
Effect of Apr	0.7945	0.0431	18.440	$< 2e-16$
Effect of May	0.5439	0.0446	12.192	$< 2e-16$
Effect of June	0.2560	0.0462	5.540	$3.70e-08$
Effect of July	0.0095	0.0441	0.216	0.8292
Effect of Aug	−0.2212	0.0479	−4.618	$4.30e-06$
Effect of Sept	−0.2245	0.0481	−4.670	$3.34e-06$
Effect of Oct	−0.3180	0.0484	−6.572	$7.35e-11$
Effect of Nov	−0.2624	0.0479	−5.483	$5.10e-08$
Effect of Dec	−0.3134	0.0485	−6.458	$1.53e-10$
Effect of 2007	−0.3861	0.0564	−6.840	$1.25e-11$
Effect of 2008	−0.5128	0.0622	−8.247	$4.18e-16$
Effect of 2010	−0.4877	0.0574	−8.498	$< 2e-16$
Effect of 2011	−0.3283	0.0568	−5.781	$9.43e-09$
Effect of 2012	−0.3485	0.0662	−5.263	$1.67e-07$

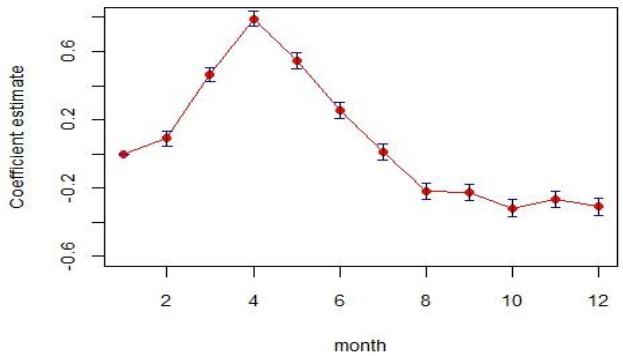


Figure 4.1: *The month pattern and standard error (SE) of seasonal effects*

There is a similar slope change from April to August, but different levels (in magnitudes) from July to August with the base line as January. Interestingly, on July shows statistically insignificant effects at p =0.1. We can see that from February to June shown positive effects significantly and then from August to December shown nonpositive effects significantly.

Thereby, we can stated that July is transition period of seasonal effects.

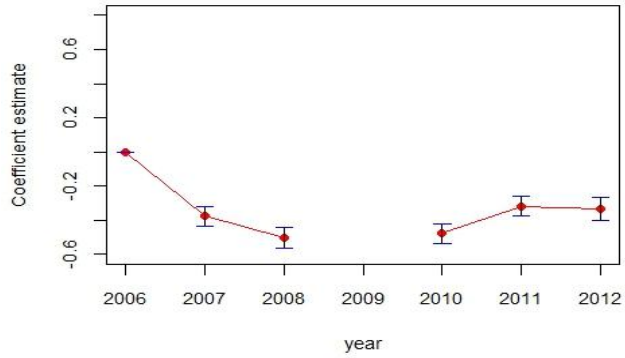


Figure 4.2: *The year pattern and standard error (SE) of annual effects*

The trend of annual effects in the six years is depicted in Figure 4.2. The graph shows a decrease in the trend from 2006 to 2008, there are no observations (gaps) from 2008-2010, then an increase from 2010 to 2011 and almost stable from 2011-2012. In general, overall years have highly significant effects at $p < 0.001$. An effect in the M1 model fitting can produce a change in level or in slope or both for coefficient estimates of parameters in the case of gaps in the data. Level and slope changes can be detected by the sign of the magnitude and direction respectively.

From M0 model the explained variability R^2 is 10.54% and from M1 the R^2 is increased to 59.21% by including the time covariates in the M1 model. The M1 model has adjusted R-squared 58.57%.

Table 4.6: *Analysis of variance for the M1 model*

Source	SS	df	MS	F	Pr(>F)
Adjust time-groups	148.1792	16	9.2612	90.309	< 2.2e-16
Res-within	124.1880	1211	0.1025		
Res-total	272.3672	1227			

The results from the Table 4.6 show that time covariates give significant effects in the M1 model. Addition of the time covariates in the model leads to a reduction in the amount

of unexplained variances in the residuals from 272.37 to 124.19 and also degree of freedom from 1227 to 1211.

4.3 Model Diagnostic

Model M1 is assessed for the nonlinearity of the response - covariates relationship, normality, variance of the error, outliers, and high-leverage points. The plot of the residual for linear fit can be used to detect patterns in the residual against the fitted SST data.

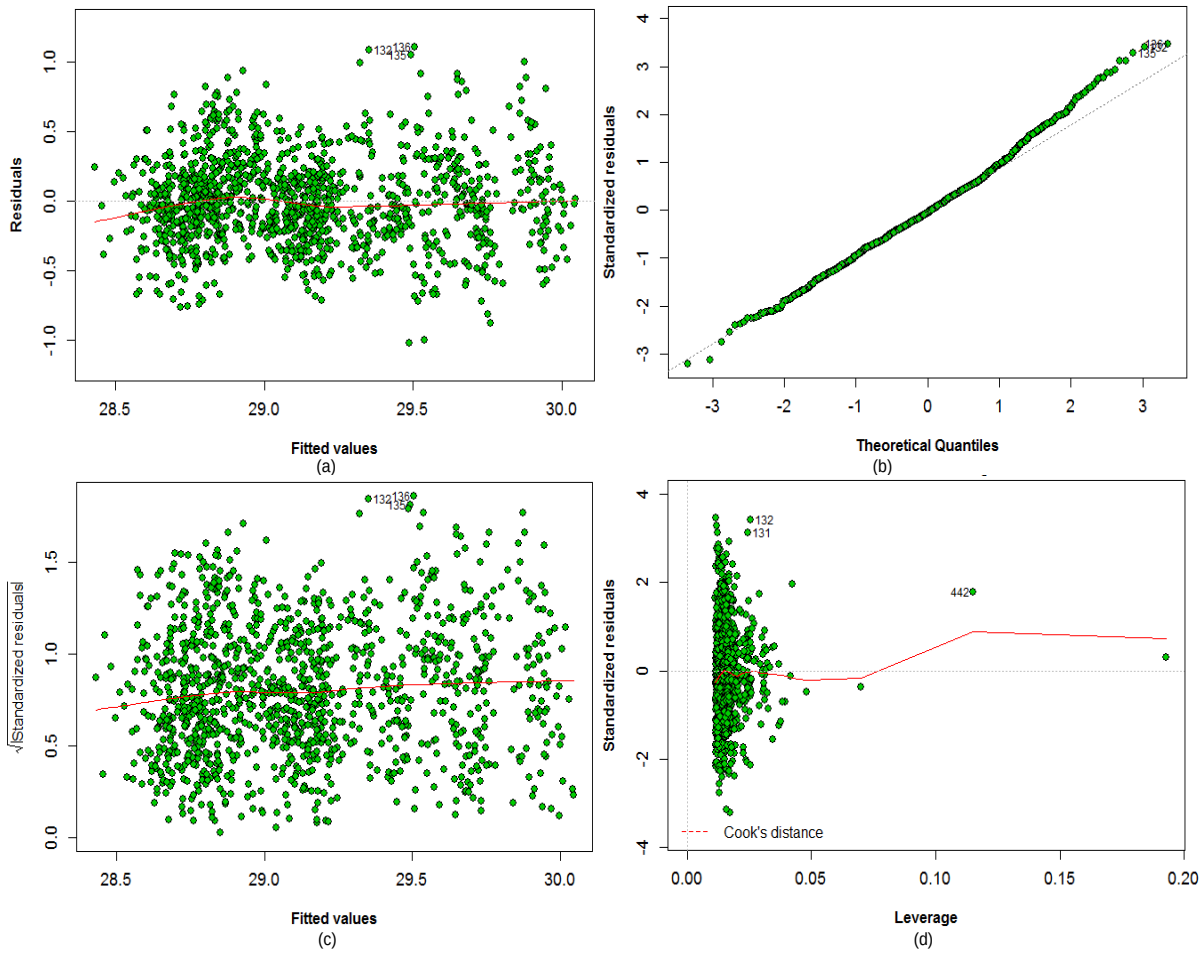


Figure 4.3: *For identifying non-linearity, the model checking for M1: Residual vs Fitted M1 model can be used to identify a trend(a); Normal QQ-plot of M1 model (b); Scale-Location of M1 model (c); Residual vs Leverage of M1 model can be used to identify outliers (d).*

The Figure 4.3 (a) shows a non-linear trend of the residuals of M1 model. From *QQ*-plot in Figure 4.3 (b) the residual can be assumed to be normally distributed. The M1 model has several outliers, in the upper tail, i.e. 131, 132, 135, and 136 in the SST data. The M1 curve indicates residual skewness in the upper tail, so transformation of the SST response is suggested. In addition, changes or extended modification in the M1 model can be done. Figure 4.3 (c) shows almost the similar trend, non-linear, for the residual as in Figure4.3 (a). The M1 model also has more potential influential observations. The scale-location of the M1 model shows that the standardized residual squared root magnitudes with constant variance. The figure shows that there is no clear evidence of heteroscedasticity. From the Figure4.3 (d) the leverage measure shows that the outliers in the continuous covariates and time covariates have influenced the M1 model fitting. The 442th observation has a very high impact on the M1 diagnostic result. Several observations on the leverage effect the M1 model fitting, for example the 131 and 132th observations.

4.4 Linear Model Fitting with Transformed Covariate

In the SST data the rainfall covariate has a large number of 0 values. Therefore, we transformed the rainfall covariate as, $log(Rain + 0.01)$ and investigated the effect of this transformation on the M0 and M1 model.

Table 4.7: *Analysis of variance for M0 regression model with transformed covariate*

Predictor	df	SumSq	MeanSq	F-value	Pr($> F$)
Temp	1	22.368	22.3683	101.0923	$< 2.2e-16$
Humd	1	9.477	9.4771	42.8314	$8.745e-11$
$log(Rain + 0.01)$	1	1.117	1.1175	5.0504	0.0248
Residuals	1227	271.494	0.2213		

The results for the M0 model from Table 4.7 shows that the rainfall covariate is significant at $p < 0.05$ and the other two covariates highly significant at $p < 0.001$. The measures of R-squared and adjusted R-squared is 10.83% and 10.61% respectively for the model (not given in the table). These two measures indicate that the model still does not fit the data appropriately. Transformation effect of rainfall covariate with respect to R-squared and adjusted R-squared in M0 model is increased by 0.31% as compared to pre-transformation.

Table 4.8: *Coefficients of M0 regression model with transformed covariate*

Coefficients	Estimate	Std.Error	t-value	Pr(>| t |)
(Intercept)	21.2417	0.7167	29.638	< 2e-16
Effect of *Temp*	0.2284	0.0188	12.135	< 2e-16
Effect of *Humd*	0.0218	0.0040	5.426	6.95e-08
Effect of $log(Rain + 0.01)$	0.0095	0.0042	2.247	0.0248

Table 4.8 shows positive effect of the rainfall covariate in the M0 model as compared to without transformed covariate in Table 4.2. The effect of the other covariates are almost similar. The results of transformation for M1 model are reported in the Table 4.9.

Table 4.9: *Anova for M1 model with transformed covariate*

Predictor	df	SumSq	MeanSq	F-value	Pr(> F)
Temp	1	22.3680	22.3683	217.9602	< 2.2e-16
Month	11	144.0630	13.0966	127.6151	< 2.2e-16
Year	5	12.612	2.5223	24.5777	< 2.2e-16
Humd	1	0.850	0.8501	8.2835	0.004071
$log(Rain + 0.01)$	1	0.285	0.2845	2.7727	0.096144
Residuals	1211	124.280	0.1026		

The results reveal that the rainfall is significant as compared to pre-transformation in the Table 4.4. The measures of R-squared and adjusted R-squared are 59.18% and 58.54% respectively for the M1 model (not given in the table). The transformation effect of rainfall with respect to R-squared and adjusted R-squared M1 model is decreased by 0.03%.

Table 4.10: *Coefficients of M1 model with transformed covariate*

Coefficients	Estimate	Std.Error	t-value	Pr(>| t |)
(Intercept)	26.5204	0.5752	46.106	$< 2e-16$
Effect of *Temp*	0.0879	0.0144	6.115	$1.30e-09$
Effect of *Humd*	0.0069	0.0032	2.165	0.0306
Effect of $log(Rain + 0.01)$	0.0049	0.0029	1.665	0.0961
Effect of Feb	0.0922	0.0427	2.161	0.0309
Effect of Mar	0.4610	0.0418	11.027	$< 2e-16$
Effect of Apr	0.7869	0.0432	18.231	$< 2e-16$
Effect of May	0.5501	0.0446	12.332	$< 2e-16$
Effect of June	0.2563	0.0462	5.546	$3.59e-08$
Effect of July	0.0150	0.0441	0.339	0.7345
Effect of Aug	−0.2250	0.0479	−4.694	$2.99e-06$
Effect of Sept	−0.2300	0.0481	−4.779	$1.98e-06$
Effect of Oct	−0.3206	0.0484	−6.621	$5.34e-11$
Effect of Nov	−0.2600	0.0479	−5.432	$6.73e-08$
Effect of Dec	−0.3020	0.0486	−6.220	$6.85e-10$
Effect of 2007	−0.3722	0.0566	−6.576	$7.16e-11$
Effect of 2008	−0.5038	0.0623	−8.092	$1.42e-15$
Effect of 2010	−0.4762	0.0575	−8.277	$3.29e-16$
Effect of 2011	−0.3142	0.0569	−5.521	$4.12e-08$
Effect of 2012	−0.3258	0.0664	−4.907	$1.05e-06$

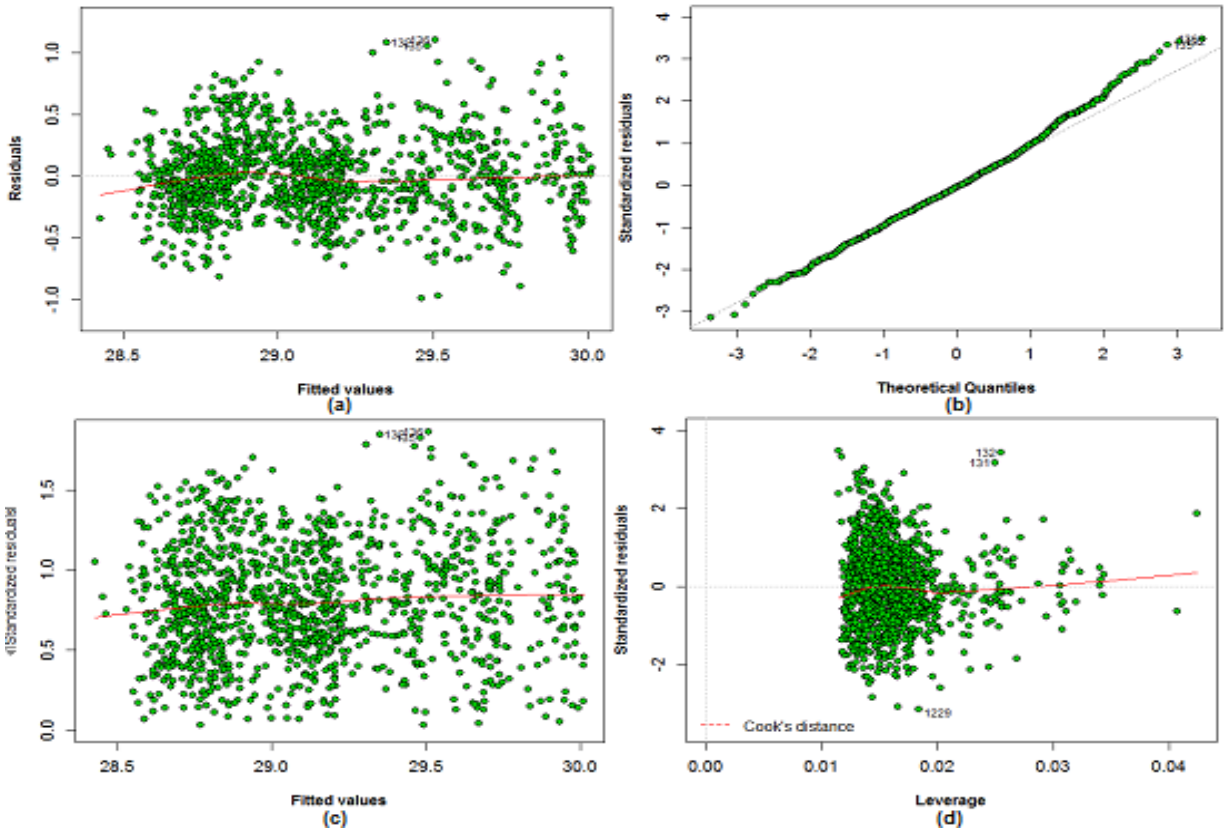


Figure 4.4: *For identifying non-linearity, the model checking for M1 with transformed covariate: Residual vs Fitted M1 model can be used to identify a trend(a); Normal QQ-plot of M1 model (b); Scale-Location of M1 model (c); Residual vs Leverage of M1 model can be used to identify outliers (d).*

Table 4.10 shows temperature highly significance at $p < 0.001$, humidity significance at $p < 0.05$ and rainfall significance at $p < 0.1$ with $log(Rain + 0.01)$. Transformation of rainfall has a little or no change in seasonal and annual effects of the M1 model. The leverage measure from the Figure 4.4(d) shows that the M1 model fitting is not effected by the outliers in the continuous covariates and time covariates.

4.5 GAM Models with P-splines for Fitting SST Data

One of the most significant components in the model fitting of SST data is time effect, which is discussed in detail in Section 4.2. Therefore, in order to get a model we incorporated this into the fitting of additive models. Initially, we used three continuous covariates prior to the model fitting as in Section 4.2. We considered two time covariates, i.e. days of the year (*Doy*) and the number of days (*Nrdays*) before and after the gap.

In this section, we applied GAM models with P-splines basis to fit the SST data. We use this model to obtain smoothness on the SST data fitting.

4.5.1 Results and Discussion

In this section, firstly, we present the AIC values of the SST data by GAM models fitting without and with time covariates, i.e. 1580.344 and 465.8042 respectively. Whereas for the same scenario and with transformation of rainfall in models fitting obtainable AICs are 1582.886 and 458.8409. The transformation shows that drastically decrease of model fitting with time covariates and slightly increase without time covariates.

Secondly, we show the smallest value of the AIC by adjusting degree of freedom, which

AIC does not guarantee an optimal fitting of the SST data using GAM models with P-spline basis. The results of our experiment are depicted in Table 4.11. From Table 4.11 it is presented that the model GM19pre has the smallest AIC, however, it does not fit the data appropriately as seen in Figure A.7 in Appendix A. Models GM6pre to GM9pre and GM17pre have relatively smaller AICs than other models, but higher than for the model GM19. However, these models give appropriate marginal model fitting in time covariates. It means that appropriateness in marginal (local) fitting does not guarantee appropriateness in global model fitting.

Therefore, to avoid choosing arbitrary smoothing parameter then checking the maximum degree or the order of P-spline in the GAM models are one important role. The estimated *df* are the trace of the smoother matrix as *df* = *trace*(**S**)*-1*, where **S** is the smoother matrix. If *df* = 1 implying a linear fit, we suggested to obtain *df* greater than 1. In our experiment we get the maximum *df* of 8 for the *Doy* covariate where the *Doy* covariate has the largest significant effect in the model as in detail see Chapter 2. The second significant effect is the *Nrdays* covariate. Thus, we select the degrees of freedom *df* of the *Nrdays* covariate around the *df* of the *Doy* covariate which can be lower or higher than 8. The results are summarized in Table 4.11.

For the SST data we took the *Doy* covariate by assuming that there are 365 days in each year. There is almost a similar pattern in the *Doy* covariate when changing its *df*, whereas the *Nrdays* shows different patterns for the different values of its *df*. The details are given in Appendix A in Figures A.1 and A.2. In Figure A.1 the *Nrdays* covariate shows variability pattern of long-term trends (annual effects), which are over than 6 *edf* of the pattern have an upward trend (multimodal) and under than 6 *edf* then the pattern of annual effect shows a

Table 4.11: *AIC for GAM models fitting in P-spline without transformation.*

Model	df_{Temp}	df_{Humd}	df_{Rain}	df_{Nrdays}	df_{Doy}	df_{Model}	AIC
GM1pre	6	7	8	4	7	22.7035	620.2200
GM2pre	6	5	4	3	8	21.7301	618.1856
GM3pre	6	5	4	4	8	22.6621	611.5720
GM4pre	6	5	4	5	8	22.9347	612.2723
GM5pre	6	5	4	6	8	23.0704	612.6442
GM6pre	6	5	4	7	8	25.8842	484.8851
GM7pre	6	5	4	8	8	26.8492	486.8411
GM8pre	6	5	4	9	8	27.3706	488.2070
GM9pre	6	5	4	10	8	28.5576	474.1265
GM10pre	8	5	5	4	7	22.8973	618.1874
GM11pre	8	5	5	4	8	23.9244	608.5816
GM12pre	8	5	5	4	9	24.6136	605.5126
GM13pre	8	5	5	3	10	24.3512	606.9454
GM14pre	8	5	5	6	10	25.7254	601.1109
GM15pre	8	5	5	5	10	25.5661	600.8135
GM16pre	8	5	5	4	10	25.2827	600.2266
GM17pre	8	5	5	7	8	26.9152	482.8112
GM18pre	8	5	5	10	10	31.2117	463.8921
GM19pre	8	5	5	10	18	38.1784	413.7819

downward trend like as parabola. The day of year (or seasonal effects) has relative similar trend for GM1pre to GM9pre models. There are mainly four types of time patterns as shown in the mentioned figures. It is indicated from the experimental results where the choice of effective degrees of freedom for *Nrdays* has a significant impact on model fitting.

A good trade-off between *edf* of the time covariates and the AIC is important in order to build an appropriate model. A very small value of *df* results in very high AIC (for example, model GM1pre in Table 4.11), whereas a very large value of *df* leads to a smaller AIC gives an inappropriate model fitting (seen in Figure A.7 for GM19pre). Therefore, a moderate value of *df* is recommended.

We tested the models for different values of *df* for the time covariate, however, for continuous covariates we take fixed values for the *df*. This is due to the continuous covariates having little impact on the response. This is elaborated experimentally in Table 4.12, where zero default values of continuous covariates (air temperature, humidity, rainfall), fixed values of the *Doy*, and changing the *df* for *Nrdays* (in the 2nd column of the table), the AIC decreases and explained deviances increase.

Table 4.12: *AICs for SST data in the GAM models fitting by considering time covariates.*

Model	df composition of covariates					df_{Model}	AIC	$DevianceExplained_{percent}$
	df_{Temp}	df_{Humd}	df_{Rain}	df_{Nrdays}	df_{Doy}			
Mod1	0	0	0	**4**	4	20.4668	686.8576	60.1
Mod2	0	0	0	**5**	4	20.6835	687.5068	60.1
Mod3	0	0	0	**6**	4	20.8644	687.6264	60.1
Mod4	0	0	0	**7**	4	23.5443	603.3747	62.9
Mod5	0	0	0	**8**	4	24.5116	605.0482	62.9
Mod6	0	0	0	**9**	4	25.0177	601.4369	63.0
Mod7	0	0	0	**10**	4	25.8602	599.0413	63.1
Mod8	3	3	3	7	**4**	15.6160	631.8499	61.5
Mod9	3	3	3	7	**5**	16.5987	620.9519	61.9
Mod10	3	3	3	7	**7**	18.7251	520.4324	65.0
Mod11	3	3	3	8	**4**	16.5474	633.6768	61.5
Mod12	3	3	3	8	**5**	17.5334	622.7275	61.9
Mod13	3	3	3	8	**7**	19.6890	522.2987	65.0

For example, the model Mod5 (0,0,0,8,4) in Table 4.12 means that the df for air temperature, humidity, and rainfall are default, and the df for *Nrdays* and *Doy* covariates are eight and four respectively. The model has AIC 605.0482 and df 24.5116, however, the AIC and df are not the smallest and biggest. Table 4.12 shows the smallest AIC and biggest deviance for model Mod7 on (0,0,0,10,4) compared to model Mod1 on (0,0,0,4,4). Similarly, model Mod8 is compared to model Mod10 changing the df for *Doy* shows that increasing the df of *Doy* increases the df of the model and explains the deviances and decrease of the AIC.

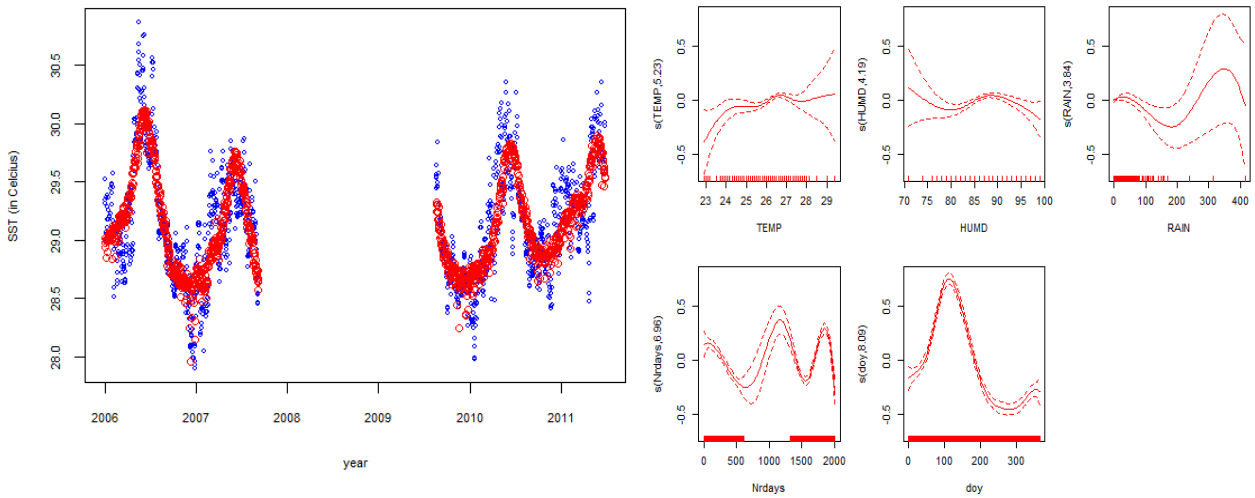


Figure 4.5: *Illustration of Model3 fitting with deviance explained 66.4% (left). The marginal model fitting has optimum composition df of the covariates (right).*

The continuous covariates included in the Model3 fitting, and the results can be seen in Figure 4.5 as an optimum *df* 30.3113 with optimum AIC 492.5577 by optimum composition *df* of covariates (10, 12, 5, 8, 12) for temperature, humidity, rainfall, *Nrdays* and *Doy* respectively. The significance of time covariates in model fitting rather than continuous covariate types as displayed in Figure A.3, Appendix. Figure A.3 shows that the models have used P-spline smoothing by GAM model and without continuous covariates. Although both models fitting show similar pattern, it has different specification. Model (0,0,0,5,7) has *df* 10.04543, AIC 679.4852 and 59.6% deviance explained, whereas model (0,0,0,8,7) has *df* 13.8944, AIC 569.8349 and deviance explained 63.3%.

Statistical references of the continuous covariates and the time covariates for Model3 is reported in Table 4.13. These covariates have significance at *p*-values. The results from the table show that the time covariate has the smallest *p*-value compared to temperature, humidity and rainfall. Moreover, Table 4.13 presents that the time covariate has larger *edf* compared to that of the continuous covariate.

Table 4.13: *The Approximate significance of smooth terms of Model3 fitting.*

	edf	*Ref.df*	*F*	*p*-value
s(*Temp*)	5.231	5.970	3.918	0.000721
s(*Humd*)	4.189	5.094	5.347	6.58e-05
s(*Rain*)	3.842	3.978	2.481	0.042673
s(*Nrdays*)	6.956	6.999	41.587	$< 2e-16$
s(*Doy*)	8.092	8.654	119.378	$< 2e-16$

We applied the models, with the same *df* composition of covariate as in Table 4.11, with transformed rainfall covariate. The results are displayed in Table 4.14. From the table we can see that the AIC for GM6post-GM9post and GM17post is decreased, compared to that shown on Table 4.11. Moreover, there is an overall decrease in the AIC of all the models

observed. A similar explanation in Table 4.14 as in Table 4.11 that a moderate value of *df* is recommended.

Table 4.14: *AIC for the SST data by using GAM models with transformed rainfall covariate.*

Model	df_{Temp}	df_{Humd}	df_{Rain}	df_{Nrdays}	df_{Doy}	df_{Model}	AIC
GM1post	6	7	8	4	7	22.7392	614.3369
GM2post	6	5	4	3	8	21.4319	611.3645
GM3post	6	5	4	4	8	22.3830	603.7943
GM4post	6	5	4	5	8	22.6661	604.5278
GM5post	6	5	4	6	8	22.8226	604.7997
GM6post	6	5	4	7	8	25.2958	477.2821
GM7post	6	5	4	8	8	26.2597	479.2460
GM8post	6	5	4	9	8	26.7899	480.6029
GM9post	6	5	4	10	8	28.0764	465.6658
GM10post	8	5	5	4	7	22.2375	612.9076
GM11post	8	5	5	4	8	23.2057	603.2837
GM12post	8	5	5	4	9	23.7185	599.8643
GM13post	8	5	5	3	10	23.4239	602.7797
GM14post	8	5	5	6	10	24.7946	596.0260
GM15post	8	5	5	5	10	24.6207	595.8440
GM16post	8	5	5	4	10	24.3318	595.1687
GM17post	8	5	8	7	8	25.8784	476.3791
GM18post	8	5	5	10	10	30.0778	457.8599
GM19post	8	5	5	10	18	36.8927	409.6393

Further, we extended the experiments with the transformed rainfall by imposing the initial condition on the *df* of covariates in Algorithm 3. The results given in Table 4.15 show that the *df* for continuous covariates are the same, whereas for the time covariates are different in the model with and without transformation setups. From the results presented in Tables 4.14 and 4.15, it can be concluded that a decrease in AIC can be achieved with transformed rainfall covariate.

Table 4.15: *The smallest AIC of the SST data by using GAM models with and without transformed rainfall.*

Model	df composition	df_{Model}	*AIC*
GMpre	2, 4, 5, 2, 4	32.5858	444.5725
GMpost	2, 4, 5, 5, 3	32.5856	276.2160

In the SST model fitting, the variability of annual effect of the *Nrdays* covariate has various trends on the gap. These trends on the gap can be estimated by the information gathered from the patterns of previous and proceeding years. We observed a high variability of a long gap (high number of missing observations) in the SST model fitting as shown

in Figures A.1 and A.2 in Appendix A. The pattern model fitting of one gap of 1231 the SST dataset can be estimated through the smoothing P-spline basis with specifying several parameters. Furthermore, we use trade-off among AIC, *df*, marginal (local) model, and the global model fitting to get an appropriate model as in Figures 4.5 - A.3 and A.4 - A.7 in Appendix A. Estimating a good trade-off among hyper-parameters in model fitting is an important stage in model fitting.

Therefore, the SST model with a low AIC and the highest degree of freedom is preferred. The results from Tables 4.11 - 4.14 show that the models fitted with seasonal and annual effects have a small AIC and large *df* for GMpost than GMpre models.

The time covariates have a significant impact on model fitting compared to the continuous covariates, especially for the *Doy* covariate. This means that the seasonal and annual effects largely contributed in model fitting for the SST data. A high *df* of *Nrdays* covariate has a tendency to wobble in the gap, whereas with transformation reduces the AIC value compared to without transformation in model fitting, but they have a similar *df* of the model. In the same way, the range values for the covariates of with transformation is smaller than without transformation as in Figure A.4, Appendix A. In this range, the shape of rainfall pattern changes drastically with transformation. The changing in the shape of rainfall covariate are relative with respect to change the range value of its covariate in without and with transformation as shown in Figure A.5 for example. Transformation effect in the GAM model fitting can cause change in the shape of its covariates and also change the range value.

Generally, in investigating the SST data observation by using GAM models without and with transformation, we recommended that ten models can be used to fit the SST data

which are GM6pre to GM9pre and GM17pre and GM6post to GM9post and GM17post models as shown in Tables 4.11 and 4.14. The trade-off in model fitting for the SST data makes no unique solution to the GAM models, and there are nonsmooth model fitting using GAM models for the SST data. Therefore, extending the SST model fitting with a gradient boosting algorithm can be applied by gamboost models.

4.6 Gamboost Model Fitting for SST Data

We applied gamboost to the same SST data as an attempt to achieve improvement in the model fitting of the SST data. Generally boosting can be used to achieve improved prediction accuracy of any learning algorithm [19,41,61,70,71]. We use the AIC and CV-risk as performance measures for evaluation and comparison of the models.

4.6.1 Results and Discussion

In this subsection, we provide SST data fitting by gamboost models without and with time covariates, i.e. AICs -0.5468 and -1.3019 respectively. We use the m_{stop} = 1000 and ν_{slf} = 0.1 in this experiment. In the models fitting for the same scenario previously and with transformed rainfall obtainable AICs are -0.5496 and -1.3100. The gamboost models fitting with transformation also show that AIC decreases with time covariates and slightly decreases without time covariates.

The results of the gamboost model without transformation are reported in Table 4.17. We applied gamboost models without transformation setup in a total of 30 models. The results for the other models as given in Tables 4.16 and 4.17 of Appendix A show a positive

effect of temperature and humidity on the SST data, and surprisingly the impact of rainfall is not shown in this model (i.e. it was not selected in the model fitting by boosting). The smooth effect of time covariates in the model is presented in Figure 4.6. We can see that Table 4.16 consists of appropriate models on global and local fitting for GMboost1pre to GMboost8pre models; and appropriate models on global but not on local fitting for GMboost9pre-30pre models. We recommended to select GMboost1pre-8pre models to fit SST data.

Table 4.16: *AIC of gamboost models using P-spline without transformed rainfall covariate.*

Model	df_{pre}	AIC_{pre}	df_{pre}	$AIC_{Correctedpre}$	df_{pre}	$AIC_{gMDLpre}$
GMboost1pre	9.3089	-1.3017	9.338745	-1.301853	9.30893	-2.205584
GMboost2pre	10.5491	-1.3177	10.54914	-1.317729	10.44939	-2.209858
GMboost3pre	11.6330	-1.3271	11.66759	-1.327309	10.80661	-2.210004
GMboost4pre	9.4497	-1.3025	9.449651	-1.302488	9.44965	-2.205022
GMboost5pre	10.4844	-1.3083	10.55927	-1.308948	10.55927	-2.200746
GMboost6pre	11.7950	-1.3330	12.19130	-1.340700	11.79500	-2.212791
GMboost7pre	10.5275	-1.3087	10.52746	-1.308735	10.52460	-2.200840
GMboost8pre	11.8235	-1.3341	11.82346	-1.334129	11.82346	-2.213628
GMboost9pre	12.7090	-1.3557	12.71496	-1.355875	12.63459	-2.227052
GMboost10pre	12.6693	-1.3569	12.66926	-1.356847	12.66926	-2.228090
GMboost11pre	13.3432	-1.3753	13.37355	-1.376459	13.37355	-2.240832
GMboost12pre	13.9807	-1.3908	13.98068	-1.390821	13.98068	-2.249306
GMboost13pre	14.7529	-1.4152	17.79670	-1.405573	17.79670	-2.228286
GMboost14pre	17.7638	-1.4038	17.33268	-1.395749	17.33268	-2.222891
GMboost15pre	18.5127	-1.4250	18.12795	-1.410517	18.12795	-2.230102
GMboost16pre	18.3934	-1.4209	18.39338	-1.420847	18.39338	-2.237830
GMboost17pre	19.1982	-1.4321	19.15235	-1.431123	19.14144	-2.240964
GMboost18pre	19.9486	-1.4419	19.94858	-1.441847	19.92359	-2.244290
GMboost19pre	21.0751	-1.5109	21.07413	-1.510549	21.04796	-2.301479
GMboost20pre	23.6138	-1.6094	23.61375	-1.609431	23.61375	-2.375225
GMboost21pre	23.7061	-1.6114	23.71177	-1.611529	23.68673	-2.376484
GMboost22pre	23.8549	-1.6150	23.85486	-1.614946	23.84172	-2.378452
GMboost23pre	23.9202	-1.6167	23.96291	-1.618232	23.96291	-2.380657
GMboost24pre	27.5526	-1.6863	27.55260	-1.686320	27.55260	-2.414544
GMboost25pre	27.5565	-1.6863	27.60688	-1.687338	27.57214	-2.415097
GMboost26pre	28.3812	-1.6951	28.38123	-1.695113	28.36581	-2.415674
GMboost27pre	26.0375	-1.7135	26.02261	-1.713153	26.01316	-2.454800
GMboost28pre	26.9929	-1.7225	26.99293	-1.722481	26.82299	-2.455157
GMboost29pre	27.7710	-1.7303	27.75518	-1.730107	27.75518	-2.455501
GMboost30pre	28.5321	-1.7370	28.53208	-1.736964	27.95038	-2.455680

The Figure 4.6 shows that the gamboost model with P-spline smoothing for the days of the year portrays a long-term effect, from 2006 to 2012. The seasonal effect for the SST data over a year is also depicted in Figure 4.6, where we used cyclic and boundary constraints. In Figure 4.6 it can be seen that the cyclic seasonal effect has a peak in around 100 days

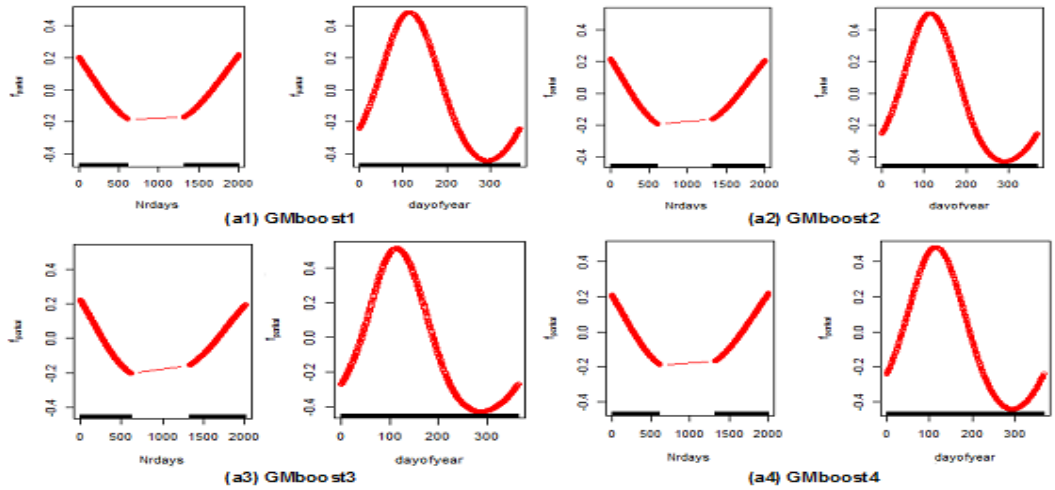


Figure 4.6: *The GMboost1-4 models shows decreasing trends of annual effects before the gap and increasing trends after the gap, whereas seasonal effects show stable patterns.*

or in April, and annual effect (long-term trend over the years) shows decreasing pattern before gap (16 Nov 2006 to 22 July 2008) and increasing pattern after gap (4 July 2010 to 13 May 2012). Both effects indicate a temporal pattern, particularly of the periodical effect for the SST data.

We checked the variability of the gamboost model with P-spline smoothing using different values of the stopping iteration and regularization factor. This variability of the model is also tested for tuning the parameters with different values.

Figure 4.7 shows the pattern of the model with different values of hyper-parameters; stopping iteration (m_{stop}= 1500-2000), the size-length of factor (ν_{slf}= 0.1) and varying values of *df*, knots of *Nrdays* covariates, we considered here are the values of *df* =2.5-3.5 and *knots*= 100-140. As in Figure 4.6, a slightly changed pattern of the *Nrdays* covariate is observed in GMboost7-8 models.

In general, Figures 4.6 and 4.7 show trade-off in hyper-parameters of the gamboost model fitting, such as an increase in the number of knots in model fitting can be a result of variations effects, mainly for *Nrdays* covariate in the gap. Fluctuation of annual effect

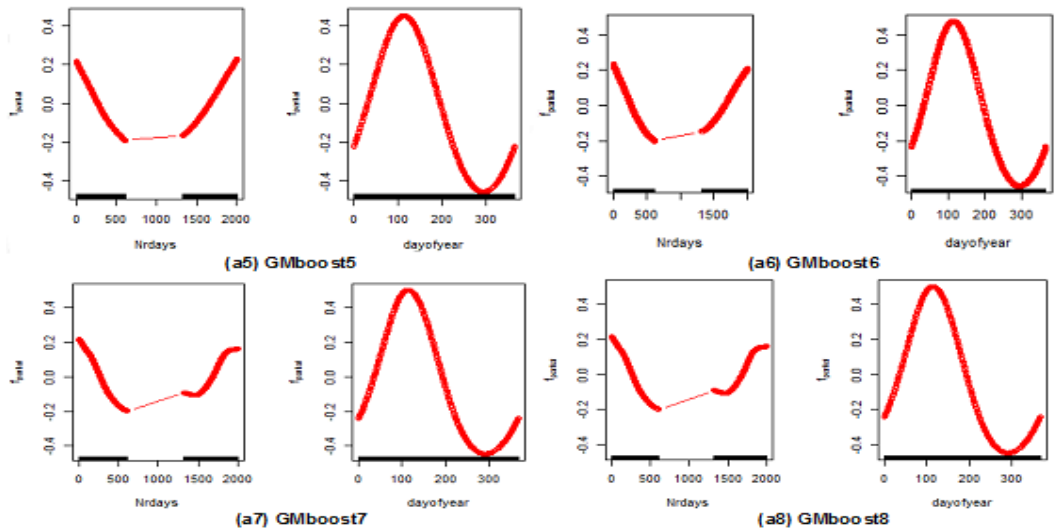


Figure 4.7: *The GMboost5-6 models show decreasing trends of annual effects before the gap and increasing trends after the gap. A slightly changed pattern is observed in GMboost7 and GMboost8 models mainly for annual effect after the gap.*

becomes high if we increase the degrees of freedom of the *Nrdays* covariate as well. These figures describe smoothing P-spline fitting based on the varying number of knots and degrees of freedom mainly at time covariate of the gamboost model. The impact of both parameters being increased leads to wiggling at the *Nrdays* covariate and over smoothing at the *Doy* covariate. The larger number of knots results in smaller equidistant knots at the *Nrdays* covariate, which is presented as a fluctuation or wiggliness curve. Conversely, a cyclic pattern at the *Doy* covariate leads to over smooth approximate or a flat curve. For this case, the relationship between the wiggliness of curve is caused by varying values of smoothing parameters using P-splines, such as the degrees of freedom and the number of knots which can be referred to [72].

Furthermore, we provide some further details performance of appropriate fitting using gamboost model. For example, Figure 4.8 describes the GMboost3 model as one of an appropriate model fitting for the SST dataset with reported intercept at 29.13273 and air temperature, humidity and rainfall in the linear effects. The coefficient values for the

linear effect of air temperature, humidity, and rainfall are 0.06549, 0.00138, and -0.00017 respectively. Air temperature and humidity are also shown having smooth effect as the unimodal curve. For time covariates, the annual effect shows a decrease before the gap and an increase after the gap in fluctuation, whereas the seasonal effect shows a smooth pattern.

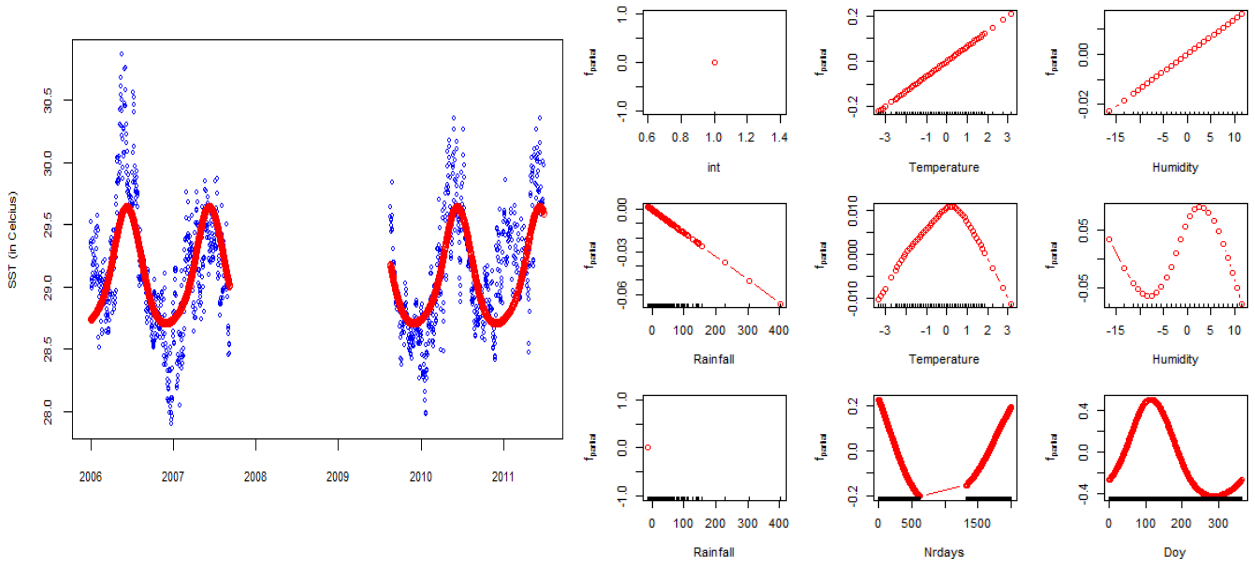


Figure 4.8: *The GMboost3 model fitting in global and local model fitting for the SST data with m_{stop}= 2000. The global fitting shows appropriate model (left) and local fitting with 9 submodels (right).*

Interestingly, Figure 4.8 shows the rainfall covariate with majority zero in smooth term with m_{stop}= 2000 in fitted model. Increasing m_{stop}= 12000 leads to changes in the rainfall covariate from smooth term to polynomial with three outliers as seen in Figure 4.9 clearly. Figures 4.8 and 4.9 show that the solution of the SST data fitting by using gamboost models with 1231 data observation in the gap without transformation is not unique. However, we recommended to select the GMboost model fitting with m_{stop}=12000 better than m_{stop}=2000. Furthermore, as can be seen that the GMboost26-29 models have the appropriate model on global fitting, but the models show inappropriate model on local fitting mainly for the

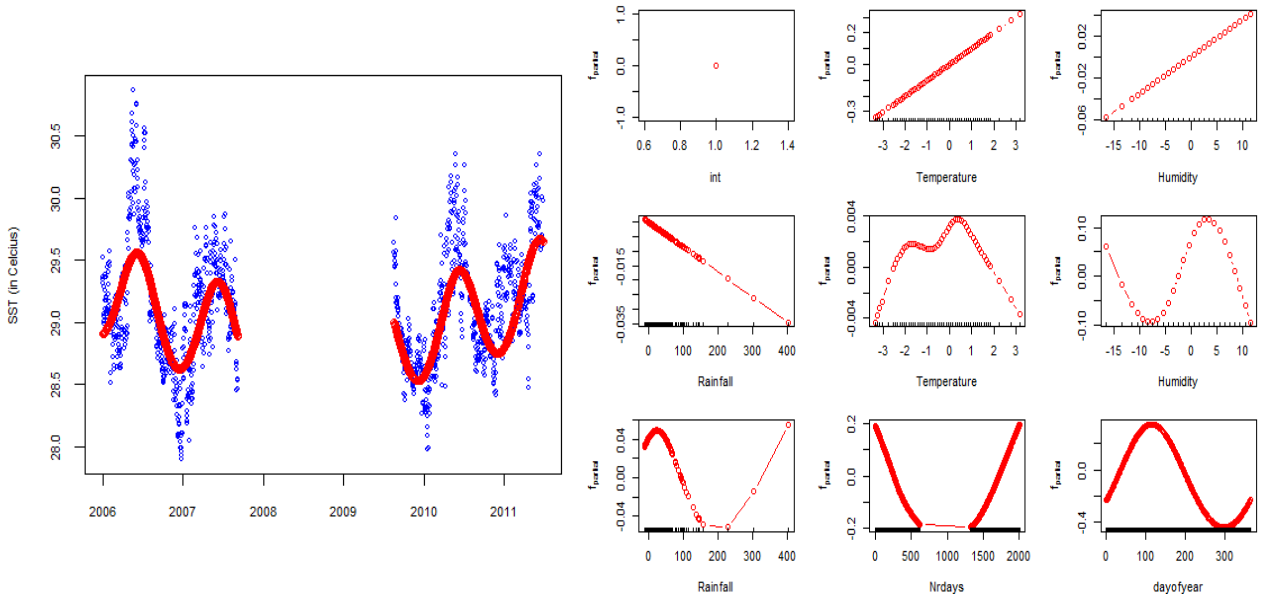


Figure 4.9: *The GMboost model fitting in global and local model fitting for the SST data with m_{stop}= 12000. The global fitting shows appropriate model (left) and local fitting with 9 submodels (right).*

Nrdays covariate as visualization in Figure B.1, Appendix. They have large values of final risk (between 75-92) and CV-risk (between 30-33).

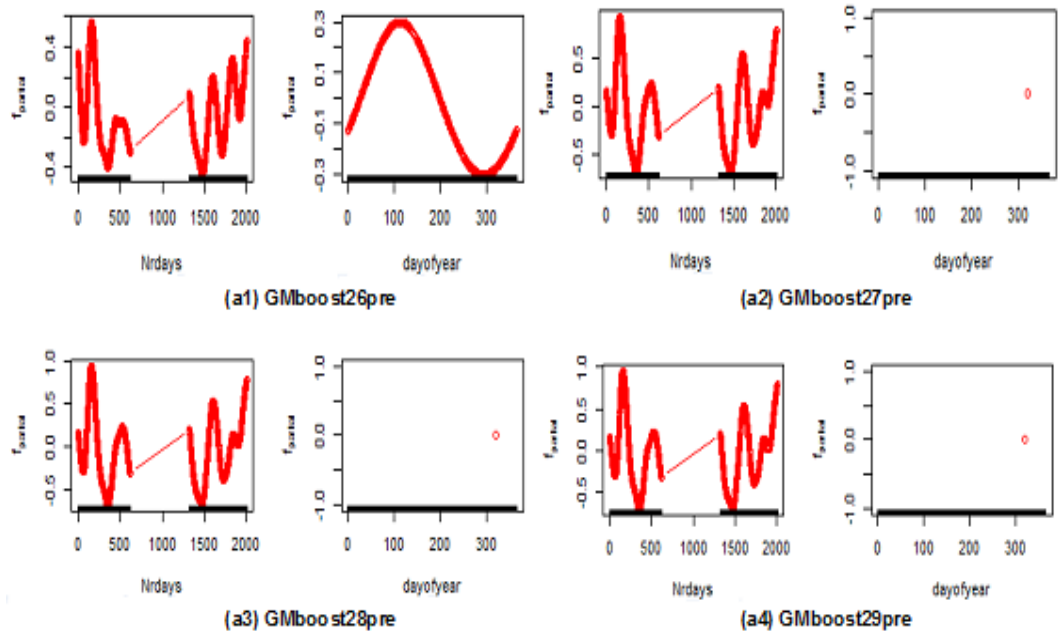


Figure 4.10: *The patterns of time covarites of GMboost26 to GMboost29 models fitting show fluctuation on the Nrdays, cyclic and smooth terms on the Doy covariate.*

The *Nrdays* submodel shows inappropriate of GMboost26-29 models and disappear for

the *Doy* submodel of GMboost27-29 models on local fitting as seen in Figure 4.10. The disappearing of the *Doy* covariate from cyclic term to smooth term because over-fitting model when large value of knots. Therefore, reducing error can be carried out in the gamboost models by transformation of rainfall.

We now investigate model fitting by using the gamboost model with transformation of rainfall to find a better model fitting. The variability of the model is tested for tuning parameters with different values of the covariate specification and control boosting.

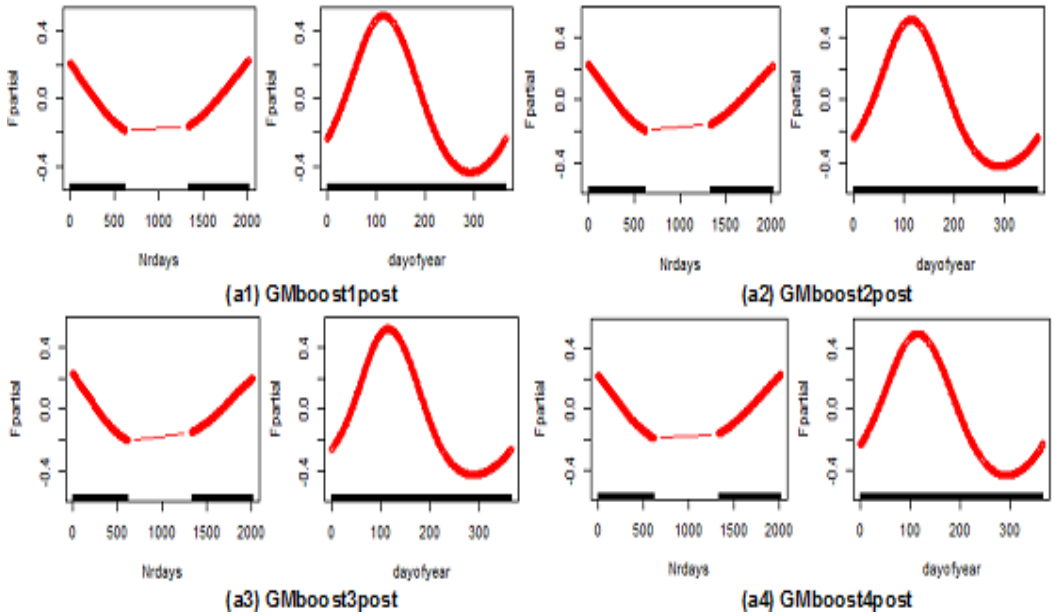


Figure 4.11: *The GMboost1post-GMboost4post show similar decreasing trends of thr Nrdays effect before gap and increasing trends after the gap, whereas the dayofyear effect is stable for all models.*

Figure 4.11 shows the pattern of the model with different values of stopping iteration (m_{stop}= 1000-2000) and the size-length of factor (ν_{slf} = 0.1), with fixed value for difference, and varying values of *df* and *knots* of *Nrdays* covariate. We consider the *df* = 2.5-3.5 and *knots*= 100-140 in the *Nrdays* specification. Similarly, Table 4.17 shows that the AIC can also be decreased with transformed rainfall as compared to without transformation. The transformation of rainfall covariate in the model fitting influences the degree of freedom

and thus provides low AIC. The higher value of the *df* lowers AIC.

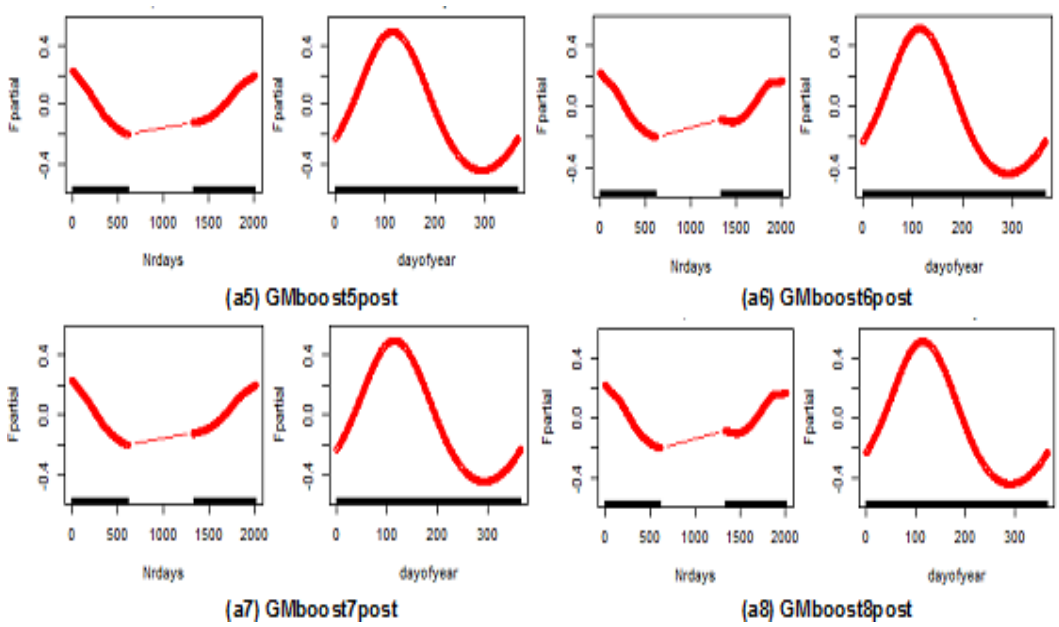


Figure 4.12: *The GMboost5post-GMboost8post show similar decreasing trends of the Nrdays effect before gap and increasing trends after the gap. A slightly changed pattern is observed in GMboost6post-GMboost8post models for the Nrdays covariate, whereas the Doy covariate shows similar patterns for all models.*

As shown in Figure 4.12, a slightly changed pattern is observed in GMboost6post and GMboost8post models with control boosting (m_{stop}= 1500,ν_{slf}= 0.1), *Nrdays* (*df* = 3.5, *diff* = 2, *knots*= 120-140), *Doy* (*df* = 1.5, cyclic, boundary.knots= c(1, 365)), and all base-learners bbs with *df* = 1 for continuous covariates.

Figure 4.13 shows the pattern of the model with different values of stopping iteration (m_{stop} are between 3500-6000) and the size-length factor (ν_{slf}= 0.1), with fixed value *diff* = 2, and varying values of high *knots* and *df* of *Nrdays* covariate, the values of *knots* are between 220-300, and the values of *df* are between 6.5-8.5. As shown in Figure 4.13, GMboost25post and GMboost26post models are similar with slight variation in the pattern. The pattern of *Nrdays* effect is inappropriately fitted in the models with varying control boosting. Further experiments are carried out with different values of m_{stop}= 5500-6500 and the ν_{slf}= 0.1, with fixed value for *diff* = 2, and varying values of high *knots* and *df* of *Nrdays* covariate, the

Table 4.17: *AIC of gamboost models with P-spline in the transformed rainfall covariate.*

Model	df_{post}	AIC_{post}	$df_{Correctedpost}$	$AIC_{Correctedpost}$	$df_{gMDLpost}$	$AIC_{gMDLpost}$
GMboost1post	9.7842	-1.3098	9.80371	-1.309985	9.79830	-2.209054
GMboost2post	10.7005	-1.3260	10.70047	-1.326044	10.70047	-2.216341
GMboost3post	11.7627	-1.3353	11.79642	-1.335506	10.86389	-2.216479
GMboost4post	9.9151	-1.3107	9.91510	-1.310664	9.91510	-2.208640
GMboost5post	11.0509	-1.3162	11.14360	-1.316890	11.14360	-2.203028
GMboost6post	12.1913	-1.3407	12.19133	-1.340694	12.19133	-2.216633
GMboost7post	11.1182	-1.3167	11.12205	-1.316739	11.12205	-2.203085
GMboost8post	12.2405	-1.3418	12.24054	-1.341835	12.24054	-2.217296
GMboost9post	12.9918	-1.3632	12.99758	-1.363347	12.99758	-2.231414
GMboost10post	13.0179	-1.3642	13.01790	-1.364234	13.01790	-2.232099
GMboost11post	13.5949	-1.3824	13.62506	-1.383602	13.61468	-2.245559
GMboost12post	14.2083	-1.3978	14.20833	-1.397762	14.20833	-2.254025
GMboost13post	17.5087	-1.4244	17.17688	-1.410117	17.14203	-2.238638
GMboost14post	17.1416	-1.4084	17.14159	-1.408371	17.10516	-2.237208
GMboost15post	17.8076	-1.4294	17.46496	-1.414999	17.46496	-2.240645
GMboost16post	17.6889	-1.4253	17.68888	-1.425292	17.68888	-2.248719
GMboost17post	18.4357	-1.4364	18.40567	-1.435378	18.40567	-2.252031
GMboost18post	19.0960	-1.4460	19.09595	-1.446026	19.07392	-2.256289
GMboost19post	20.3290	-1.5148	20.31938	-1.514400	20.31938	-2.312161
GMboost20post	23.3514	-1.6118	23.35143	-1.611836	23.21076	-2.380273
GMboost21post	23.4714	-1.6137	23.47726	-1.613903	23.05995	-2.381427
GMboost22post	23.6319	-1.6173	23.6319	-1.617285	23.36884	-2.383001
GMboost23post	23.7123	-1.6190	23.83097	-1.620357	23.43696	-2.384373
GMboost24post	27.9570	-1.6872	27.95703	-1.687221	24.74635	-2.414100
GMboost25post	27.9610	-1.6872	28.01864	-1.688210	25.04738	-2.414629
GMboost26post	28.9616	-1.6957	27.96100	-1.687220	24.75312	-2.414090
GMboost27post	26.2430	-1.7168	26.22806	-1.716641	26.22806	-2.456301
GMboost28post	27.1326	-1.7261	27.13264	-1.726082	27.03282	-2.457406
GMboost29post	27.9021	-1.7339	27.90213	-1.733907	27.90213	-2.457873
GMboost30post	28.6276	-1.7407	28.62755	-1.740647	28.45557	-2.458011

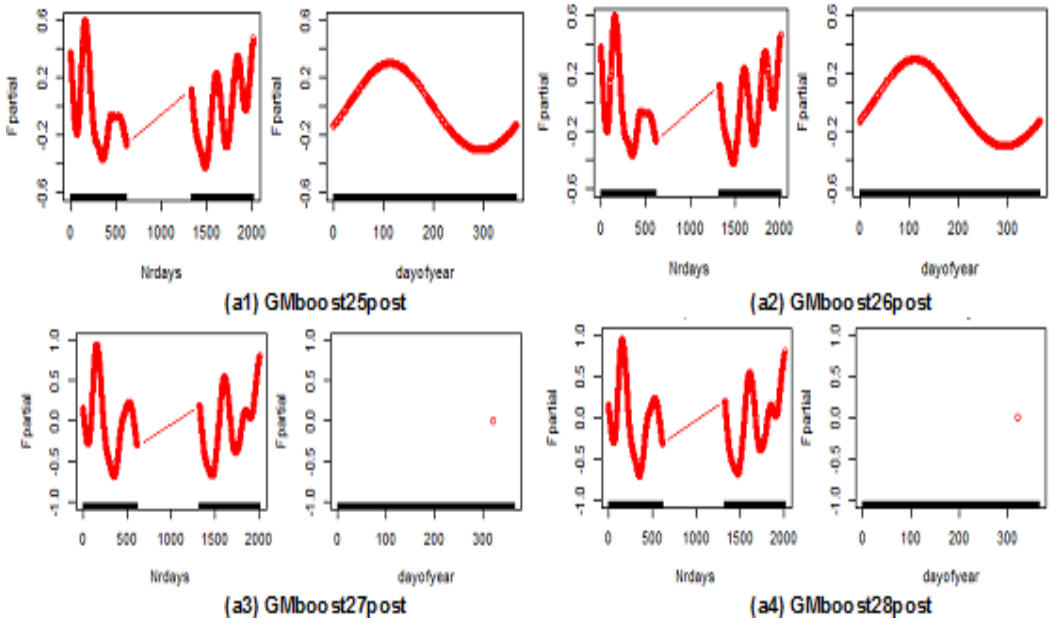


Figure 4.13: *Similar patterns for seasonal effects on models GMboost25post and GMboost26post; The Nrdays effects are not appropriate for models GMboost25post to GMboost28post.*

values of *knots* are between 240-300, and the values of *df* are between 8.5-9.5.

The results of our experiment gamboost model fitting with transformation are displayed

in Figure B.2, Appendix B. The figure shows that the models provide a better global fitting of the SST data. However, the models are not appropriate on local fitting as shown for the *Nrdays* covariate as captured in Figure 4.13. The results of GMboost25post to GMboost28post models with the range FR: 79.85 - 76.21 and CV-risk: 30.78 - 29.83. The SST data fitting is represented as in Figure B.2, Appendix B for GMboost25-28 models. The models show similar smoothing pattern at the beginning smoothing in before the gap and the last smoothing. The appropriate global fitting is not guaranteed appropriate on local fitting.

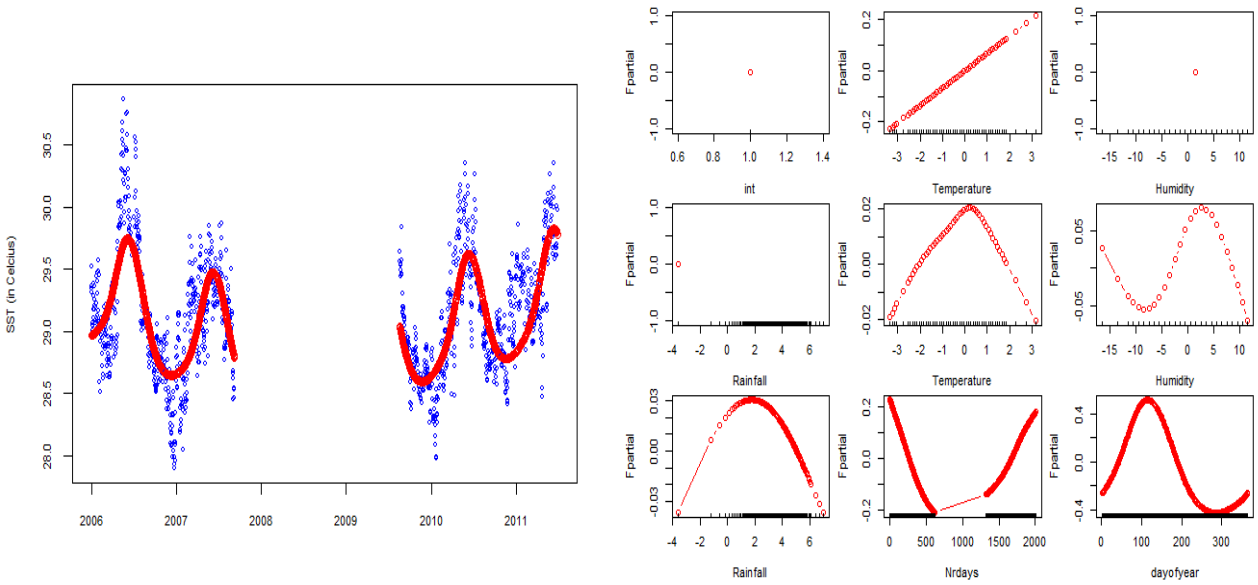


Figure 4.14: *The GMboost3 model fitting in global and local model fitting for the SST data with transformed rainfall covariate and* m_{stop}*= 2500.*

Figures 4.14 and 4.15 show appropriate gamboost model fitting with transformed rainfall, which using a different specification of the hyper-parameters and m_{stop}. The specification of hyper-parameters and high value m_{stop} can reveal submodel on local fitting.

In general, Table 4.17 has similar composition with Table 4.16, which GMboost1post to GMboost8post are appropriate model fitting (global and local fits) and GMboost9post to GMboost30post are inappropriate model fitting (only global fit). Transformation of rainfall

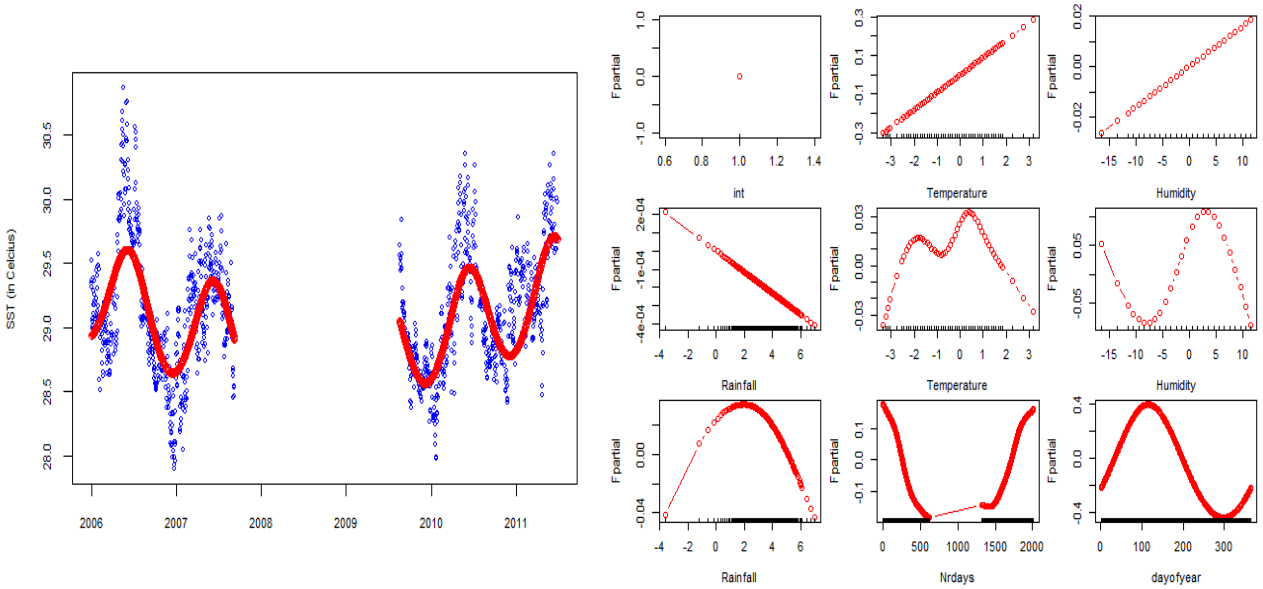


Figure 4.15: *The GMboost model fitting in global and local model fitting for the SST data with transformed rainfall covariate and* m_{stop}*= 19000.*

can increase degrees of freedom (*df*) and decrease AICs on model fitting. Transformation of rainfall can improve model fitting and does not change pattern of rainfall in base-learner for linear models. However, the transformation of rainfall changes pattern in base-learner for smooth model. We can see as comparison this change in Figures 4.8 and 4.9 before transformation and 4.14 and 4.15 after transformation.

Appropriate fitting of the data by the models depends on several factors, for example the base-learner specification for covariates and control boosting. We have two main observations from these experiments. First, the SST model fitting is not unique as different linear combinations of base-learners give several similar curves that precise fittings for the SST data. Second, the basis function with transformation of rainfall allows faster smoothness in model fitting of the SST data as compared to without transformation. The models display simple to complex trends for time covariates effects, mainly for *Nrdays* covariate. Several patterns of covariates effects can be observed through different values of control boosting parameters, i.e. the stopping iteration and regularization factor, and also

by the tuning parameter with various values for base-learner specification.

Variability in the covariates effects can also be affected by their range (in scale unit) and also by specifications of the hyper-parameters, such as the number and position of the knots, the degree of P-spline basis that are used in GAM models, changes in relationship between covariate and response and unit of the scales of the measurements for the covariates [21,73].

This indication can be seen from the previous results for appropriate model fitting of the SST data. The model fitting is influenced by the variability of the time covariate in an additive manner. In model fitting by GAM and gamboost models, we found that in the peak seasons there was a stable pattern. However, the relative seasonal amplitudes (or a range size of seasonal effect) change over time covariates (*Nrdays* and *Doy*) and control boosting specification. We also noticed that the seasonal patterns were more prominent than the annual patterns.

The solution of the SST model fitting by using 1231 observations with one long gap in with transformation is also not a unique model. An illustration for model interpretability is obtainable as shown in Figures B.1 and B.2, Appendix B. Model fitting by gamboost for the SST data, although the model global fitting have similar values for FR and CV-risk, does not guarantee appropriate model local fitting in visualization. Model interpretability becomes essential in the determination of rejecting or accepting appropriate model fitting for SST data.

The presence of a long gap can affect the fluctuation of the annual effect of the *Nrdays* covariate. Gamboost models capture the SST data phenomena better than the simple linear models and GAM in the model fitting. Further there are several aspects to observe seasonal and annual variations by P-spline smoothing in modulation (i.e. cyclic or periodic trend),

such as difference penalties [74]. In this case, the smoothness is tuned using degrees of freedom and number of knots with difference penalties, we fixed this to 2 penalties in *Nrdays* for all the covariates in order to observe seasonal effects. We can conclude, on the basis of our investigations so far, that gamboost models with P-spline for SST data still need to be improved mainly on local fitting. This can potentially be done by using LSS function in GAMLSS and gamboostLSS models.

4.7 GAMLSS Models Fitting for SST Data

In this section, we applied GAMLSS models with P-splines basis to the same SST data. We use this model to reveal location, scale, and shape (LSS) of time covariates of the SST data.

4.7.1 Results and Discussion

Initially we present GAMLSS models fitting without and with time covariates for SST data, i.e. AICs 1580.013 and 697.222 respectively. Whereas for the same scenario and with transformed rainfall in models fitting obtainable AICs are 1582.756 and 691.8568. It means that the models fitting with transformation show drastically decrease in AIC with time covariates and slightly increase without time covariates.

The results of GAMLSS without transformation are displayed in Table 4.19. The experiments that were carried out show that this algorithm can handle large number of covariates. However, it is more time-consuming and takes several days to run the program to evaluate the smallest AIC of GAMLSS models with P-spline as base-learners. The SST data can be fitted by GAMLSS models using P-spline basis function considering several aspects. First,

we consider the model fitting based on degrees of freedom aspect. Second, we observe it based on degrees of freedom and degrees composition of basis for covariates. Third, we investigate the model fitting based on penalized spline (ps) interval, and fourth, we consider the model fitting based on similarity with the second aspect added LSS (location, scale, and shape), as can be seen in Table 4.18. Note that ps.intervals is the number of knots in default 20 [16, 29–32].

In the first stage we used these aspects without transformation, and in the second stage we applied model fitting with transformation covariate. Model fitting of the SST data by GAMLSS models given in Table 4.18 shows that the GM120pre model gives the smallest AIC with the degrees composition 5, 3, 3, 8, 8 for air temperature, relative humidity, rainfall, *Nrdays* and *Doy* covariates, respectively. The GM121post model gives the smallest AIC 423.1292 (or less than 6 points rather than the GM120pre model).

Table 4.18: *AIC of GAMLSS models in P-spline with initial condition.*

Model	*Temp*	*Humd*	*Rain*	*Nrdays*	*doy*	df	AIC	Remarks
GM120pre	5	3	3	8	8	34.00002	429.0574	df
GM121post	5	3	3	8	8	33.99996	423.1292	df
GM122pre	2	2	2	8	8	47.22123	234.6547	df and LSS
GM123post	2	2	2	8	8	47.21247	231.4993	df and LSS

There are several aspects to observe the changes in the AIC using the GAMLSS models with P-spline basis function for fitting model of the SST data. For pre-transformation, GAMLSS0pre to GAMLSS3pre models have very low *df* and the high AIC values as shown in Table 4.19. By P-spline basis function, GAMLSS4pre to GAMLSS7pre models show an increased *df* and decreased AIC and this is significant in the models. Furthermore, GAMLSS8pre to GAMLSS11pre models show that AIC based on degrees of freedom aspect in fitting models increase *df* and lower AIC than previous treatments.

We observed that model fitting by GAMLSS models for SST data provides improvement in terms of the smallest AIC with similar degrees of freedom. Table 4.18 reported the results of GAMLSS models applied for SST data fitting. By using the value of initial condition of 8 as in Algorithm 4 for all covariates with corresponding *df*, we obtained the same composition *df* values of covariates for without and with transformation, as shown in GM120pre and GM121post models, i.e. 5, 3, 3, 8, 8 for *Temp*, *Humd*, *Rain*, *Nrdays* and *Doy*, respectively. The models have the same *df*, i.e. 34. However, both models have different AIC values. Transformation of rainfall covariate can decrease the AIC of the model fitting, i.e. AIC of 429.0574 for without transformation becomes 423.1292 for with transformation.

In the model fitting based on *df* and LSS function, we get the same composition of *df* values of covariates and relatively similar AIC values for without and with transformation. The LSS function results in decreasing *df* for continuous covariates, but it does not change the *df* for time covariates. Likewise, the LSS function can reduce AIC values, as shown in GM122pre and GM123post models, if we compare both previous models in Table 4.20. The trade-off in SST model fitting by this setup is based on the *df* and the LSS function which can decrease the AIC and increase information by *df*, as in Table 4.20.

Figure C.1 in Appendix C using GAMLSS model is similar pattern as in Figure 4.5 by GAM model fitting. The difference by the GAMLSS model fitting is that the AIC values can reach a smaller AIC than by the GAM model. However, the hyper-parameter specification in the GAMLSS15pre and GAMLSS16pre models is insufficient to obtain an appropriate model fitting of the SST data.

Table 4.19: *AIC of GAMLSS models fitting for SST data using P-spline without transformation.*

Model	Links	Terms	df	AIC
GAMLSS0pre	logit(μ)	1+*Temp* + *Humd* + *RAIN* + *Nrdays* + *Doy*	8	1359.959
	log(σ)	1+*Nrdays*		
GAMLSS1pre	logit(μ)	1+*Temp* + *Humd* + *RAIN* + *Nrdays* + *Doy*	8	1353.189
	log(σ)	1+*Doy*		
GAMlss2pre	logit(μ)	1+*Temp* + *Humd* + *RAIN* + *Nrdays* + *Doy*	9	1344.879
	log(σ)	1+*Nrdays* + *Doy*		
GAMLSS3pre	logit(μ)	1+*Temp* + *Humd* + *RAIN* + *Nrdays* + *Doy*	12	1334.689
	log(σ)	1+*Temp* + *Humd* + *RAIN* + *Nrdays* + *Doy*		
GAMLSS with P-spline basis				
GAMLSS4pre	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*RAIN*)+ ps(*Nrdays*)+ ps(*Doy*)	25.99998	783.6255
	log(σ)	1+ps(*Nrdays*)		
GAMLSS5pre	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*RAIN*)+ ps(*Nrdays*)+ ps(*Doy*)	25.99999	717.7558
	log(σ)	1+ps(*Doy*)		
GAMLSS6pre	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*RAIN*)+ ps(*Nrdays*)+ ps(*Doy*)	30.00001	722.6382
	log(σ)	1+ps(*Nrdays*)+ ps(*Doy*)		
GAMLSS7pre	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*RAIN*)+ ps(*Nrdays*)+ ps(*Doy*)	38.00001	764.8103
	log(σ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*RAIN*)+ ps(*Nrdays*)+ ps(*Doy*)		
GAMLSS with P-spline basis and optimal df				
GAMLSS8pre	logit(μ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*RAIN*, df=2)+ ps(*Nrdays*, df=8)+ ps(*Doy*, df=8)	41.00001	301.2090
	log(σ)	1+ps(*Nrdays*, df=8)		
GAMLSS9pre	logit(μ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*RAIN*, df=2)+ ps(*Nrdays*, df=8)+ps(*Doy*, df=8)	41.00002	252.2629
	log(σ)	1+ps(*Doy*, df=8)		
GAMLSS10pre	logit(μ)	1+ps(*Temp*, df=4)+ ps(*Humd*, df=2)+ ps(*RAIN*, df=2)+ ps(*Nrdays*,df=8)+ ps(*Doy*, df=8)	49.00000	230.5008
	log(σ)	1+ps(*Nrdays*, df=8)+ ps(*Doy*, df=8)		
GAMLSS11pre	logit(μ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*RAIN*,df=2)+ ps(*Nrdays*, df=8)+ps(*Doy*,df=8)	62.15796	235.9147
	log(σ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*RAIN*,df=2)+ ps(*Nrdays*, df=8)+ ps(*Doy*, df=8)		
GAMLSS with P-spline basis, optimal df, and knots				
GAMLSS12pre	logit(μ)	1+ps(*Temp*, df=5, ps.interval=20)+ ps(*Humd*, df=2, ps.interval=20)+ ps(*RAIN*, df=2, ps.interval=20)+ ps(*Nrdays*, df=8,ps.interval=100)+ ps(*Doy*, df=8, ps.interval=20)	41.6589	281.3295
	log(σ)	1+ps(*Nrdays*, df=8, ps.interval=100)		
GAMLSS13pre	logit(μ)	1+ps(*Temp*, df=5, ps.interval=20)+ ps(*Humd*, df=2,ps.interval=20)+ ps(*RAIN*, df=2, ps.interval=20)+ ps(*Nrdays*, df=8,ps.interval=100)+ ps(*Doy*, df=8, ps.interval=20)	41.6765	236.6402
	log(σ)	1+ps(*Doy*, df=8, ps.interval=20)		
GAMLSS14pre	logit(μ)	1+ps(*Temp*, df=5, ps.interval=20)+ ps(*Humd*, df=2, ps.interval=20)+ ps(*RAIN*, df=2, ps.interval=20)+ ps(*Nrdays*, df=8, ps.interval=100)+ ps(*Doy*, df=8, ps.interval=20)	50.76528	208.6503
	log(σ)	1 + ps(*Nrdays*,df=8, ps.interval=100)+ ps(*Doy*,df=8,ps.interval=20)		
GAMLSS15pre	logit(μ)	1 + ps(*Temp*,df=5,ps.interval=20)+ ps(*Humd*,df=2,ps.interval=20)+ ps(*RAIN*,df=2,ps.interval=20)+ ps(*Nrdays*, df=8,ps.interval=100)+ ps(*Doy*, df=8,ps.interval=20)	63.43874	198.7397
	log(σ)	1 + ps(*Temp*, df=5, ps.interval=20)+ ps(*Humd*,df=2,ps.interval=20)+ ps(*RAIN*, df=2,ps.interval=20)+ ps(*Nrdays*, df=8, ps.interval=100)+ ps(*Doy*, df=8, ps.interval=20)		

Table 4.20: *AIC of GAMLSS models fitting for SST data using P-spline in without and with transformation.*

Model	Links	Terms	df	AIC
GAMLSS with P-spline basis, optimal df, and knots				
GAMLSS16pre	logit(μ)	1 + ps(*Temp*,df=5,ps.interval=20)+ ps(*Humd*,df=2,ps.interval=20)+ ps(*RAIN*,df=2,ps.interval=20)+ ps(*Nrdays*, df=8,ps.interval=20)+ ps(*Doy*, df=8,ps.interval=20)	77.82624	199.8628
	log(σ)	1 + ps(*Temp*, df=5, ps.interval=20)+ ps(*Humd*,df=2,ps.interval=20)+ ps(*RAIN*, df=2,ps.interval=20)+ ps(*Nrdays*, df=8, ps.interval=20)+ ps(*Doy*, df=8, ps.interval=365)		
GAMLSS17pre	logit(μ)	1 + ps(*Temp*,df=5,ps.interval=20)+ ps(*Humd*,df=2,ps.interval=20)+ ps(*RAIN*,df=2,ps.interval=20)+ ps(*Nrdays*, df=8,ps.interval=100)+ ps(*Doy*, df=8,ps.interval=20)	79.30990	160.1600
	log(σ)	1 + ps(*Temp*, df=5, ps.interval=20)+ ps(*Humd*,df=2,ps.interval=20)+ ps(*RAIN*, df=2,ps.interval=20)+ ps(*Nrdays*, df=8, ps.interval=100)+ ps(*Doy*, df=8, ps.interval=365)		
AIC of GAMLSS models fitting with transformation				
GAMLSS0post	logit(μ)	1+*Temp* + *Humd* + *Rain* + *Nrdays* + *Doy*	8	1356.659
	log(σ)	1+*Nrdays*		
GAMLSS1post	logit(μ)	1+*Temp* + *Humd* + *Rain* + *Nrdays* + *Doy*	8	1348.872
	log(σ)	1+*Doy*		
GAMLSS2post	logit(μ)	1+*Temp* + *Humd* + *Rain* + *Nrdays* + *Doy*	9	1340.794
	log(σ)	1+*Nrdays* + *Doy*		
GAMLSS3post	logit(μ)	1+*Temp* + *Humd* + *Rain* + *Nrdays* + *Doy*	12	1332.725
	log(σ)	1+*Temp* + *Humd* + *Rain* + *Nrdays* + *Doy*		
GAMLSS with P-spline basis				
GAMLSS4post	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*Rain*)+ ps(*Nrdays*)+ ps(*Doy*)	26.00001	645.4679
	log(σ)	1+ps(*Nrdays*)		
GAMLSS5post	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*Rain*)+ ps(*Nrdays*)+ ps(*Doy*)	25.99999	579.3850
	log(σ)	1+ps(*Doy*)		
GAMLSS6post	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*Rain*)+ ps(*Nrdays*)+ ps(*Doy*)	29.99997	564.1076
	log(σ)	1+ps(*Nrdays*)+ ps(*Doy*)		
GAMLSS7post	logit(μ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*Rain*)+ ps(*Nrdays*)+ ps(*Doy*)	42.00002	567.7503
	log(σ)	1+ps(*Temp*)+ ps(*Humd*)+ ps(*Rain*)+ ps(*Nrdays*)+ ps(*Doy*)		
GAMLSS with P-spline basis and optimal df				
GAMLSS8post	logit(μ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*Rain*, df=3)+ ps(*Nrdays*, df=8)+ ps(*Doy*,df=8)	42.00006	297.0516
	log(σ)	1+ps(*Nrdays*, df=)		
GAMLSS9post	logit(μ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*Rain*, df=3)+ ps(*Nrdays*, df=8)+ps(*Doy*,df=8)	41.99998	247.8236
	log(σ)	1+ps(*Doy*, df=8)		
GAMLSS10post	logit(μ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*Rain*, df=3)+ ps(*Nrdays*, df=8)+ps(*Doy*,df=8)	51.00002	228.5520
	log(σ)	1+ps(*Nrdays*, df=8)+ ps(*Doy*, df=8)		
GAMLSS11post	logit(μ)	1+ps(*Temp*, df=5)+ ps(*Humd*, df=2)+ ps(*Rain*, df=3)+ ps(*Nrdays*, df=8)+ps(*Doy*, df=8)	64.14670	233.5817
	log(σ)	1+ps(*Temp*,df=5)+ ps(*Humd*, df=2)+ ps(*Rain*, df=3)+ ps(*Nrdays*, df=8)+ ps(*Doy*, df=8)		

Similarly in GAMLSS model fitting without transformation, we investigated hyper-parameter specification of the covariates to find an appropriate model fitting for the SST data based on LSS function by using GAMLSS model in transformation of rainfall as captured in Table 4.20. For GAMLSS0post to GAMLSS3post models have very low *df* and high AIC values as seen in Table 4.20. By P-spline basis function, GAMLSS4post to GAMLSS7post models show increased *df* and decreased AIC, this change is approximately more than 50%.

Furthermore, GAMLSS8post to GAMLSS11post models show that AIC based on degrees of freedom aspect in fitting models give an increased *df* and lower AIC than previous treatments. In general, several aspects need to be considered in order to observe the changes in the AIC from the GAMLSS models by P-spline basis function, such as AIC based on *df* aspect, AIC based on *df* and degree of P-spline basis function, and AIC is based on similar values to the AIC based on *df* with added LSS parameters, i.e. μ, σ, ν, and τ parameters for LSS aspects. It should also be noted that LSS is exploring the distributional term of GAM models. It can be seen in detail in the Tables 4.19 to 4.20.

Here, we conclude that model fitting using P-spline by setup GAMLSS models for the SST data significantly decreases AIC if the LSS function is incorporated and time covariate is used with varying degrees of freedom. However, in order to get an appropriate model interpretability of the *Nrdays* and the *Doy* it takes time in composition of hyper-parameters. The singularity of GAMLSS models is caused by rainfall covariate with many zero values of the SST dataset. An alternative to reduce the singularity issue is by transformation of rainfall, whereas by using the gradient boosting technique the computational time will be reduced.

4.8 GamboostLSS Models Fitting for SST Data

In this section, we used GaussianLSS families distribution to fit SST data by using GAMLSS models with a boosting technique. The boosting technique is used to optimize the general risk function. Base-learners specifications are used as a framework to select an appropriate model.

4.8.1 Results and Discussion

We observed from our results that the degrees of freedom (*df*), *knots*, and m_{stop} of hyper-parameters specification in base-learners for the time covariates have improved the model fitting. The result of gamboostLSS with various specification for base-learners and for control boosting used without and with transformation are given in tables and figures.

4.8.1.1 Effect of the Degrees of Freedom on GamboostLSS

One of the important parameters in the model is the degrees of freedom *df*. It contributes to the smoothing of model fitting. To get appropriate base-learners we checked for several values of the *df*. We started from the small value for *df* for each base-learner to avoid misfitting as suggested by Hofner [75], mainly for continuous covariates. Hereinafter, we consider values of *df* of the *Nrdays* covariate from 2.01 to 3.05, difference of penalty = 2 and *knots* = 40. The control boosting m_{stop} is 1000 and the size length factor ν_{slf} is 0.1. The values of *df* give more effects on the *Nrdays* than the *Doy* covariate as can be seen from Figure 4.16. The increasing *df* in model fitting causes more fluctuation in the *Nrdays* covariate (as annual effects), however, the pattern of the *Doy* covariate is stable within the μ and σ

paramaters.

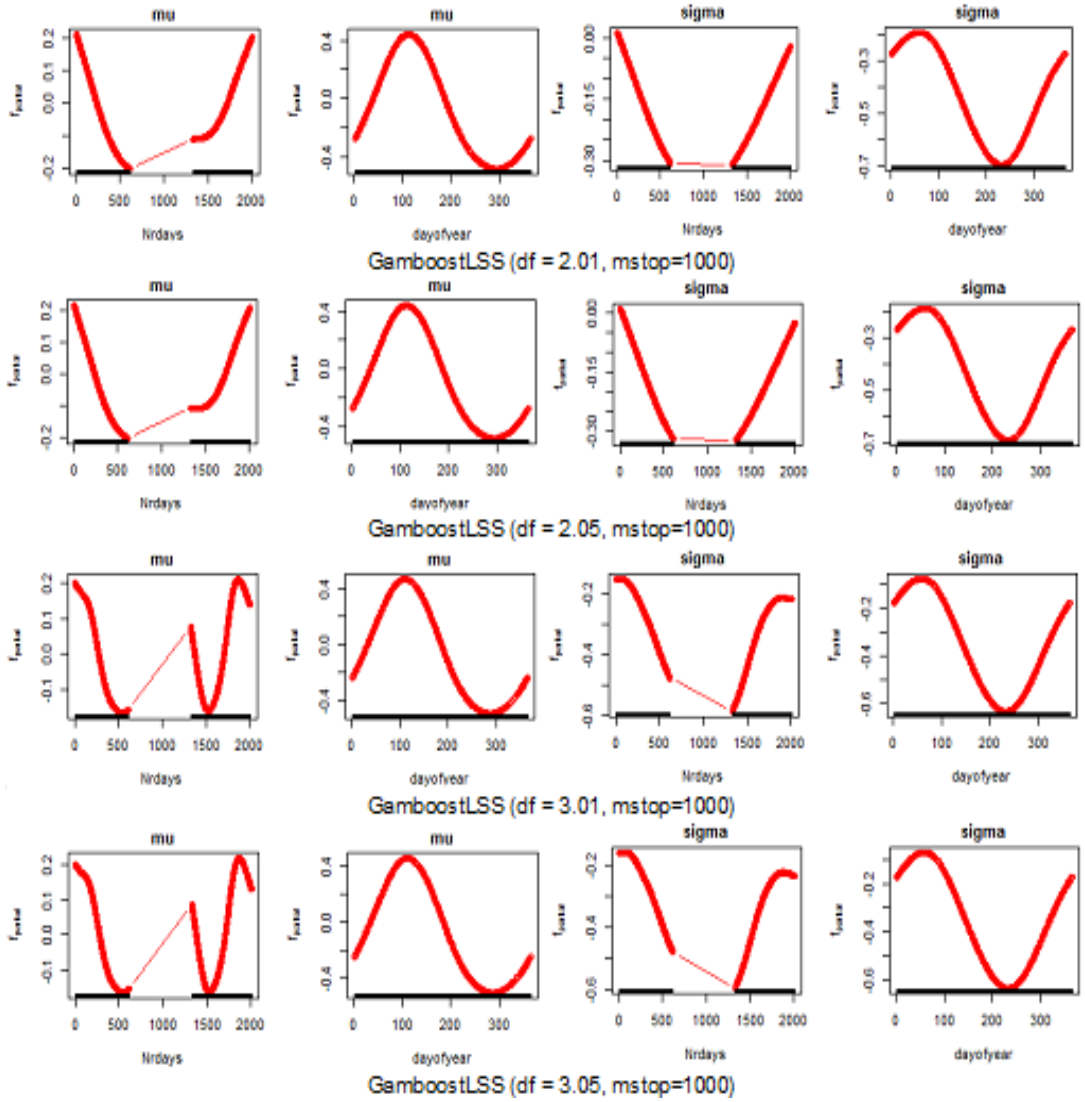


Figure 4.16: *Illustrating different degrees of freedom and fixed m_{stop}= 1000 with respect to time covariates in the SST model fitting using gamboostLSS models.*

We observed that a larger value of degrees of freedom *df* tends to an inappropriate model fitting, as seen in Figure 4.16 for degrees of freedom *df* = 3.01 and 3.05 which shows fluctuation after the gap of the *Nrdays* covariate. The effect of changing the value of *df* shows that the model fitting without transformation is more appropriate for *df* = 2.01-2.05 with 15 submodels compared to *df* = 3.01 and 3.05 with 14 submodels. The results of the models show similar pattern and nonsmooth global fitting, however, we do not present it.

4.8.1.2 Effect of the Stopping Iteration on GamboostLSS models

Selection of an appropriate number of boosting iteration, m_{stop}, is important to avoid misfitting and computational time. We start in model fitting by using the stopping iteration: 500 to 1500 with step 500. Fixed parameters are given for smooth function of the *Nrdays* covariate, where the number of knots is 30, and the degree of freedom *df* is 2.01 and difference of penalty is 2.

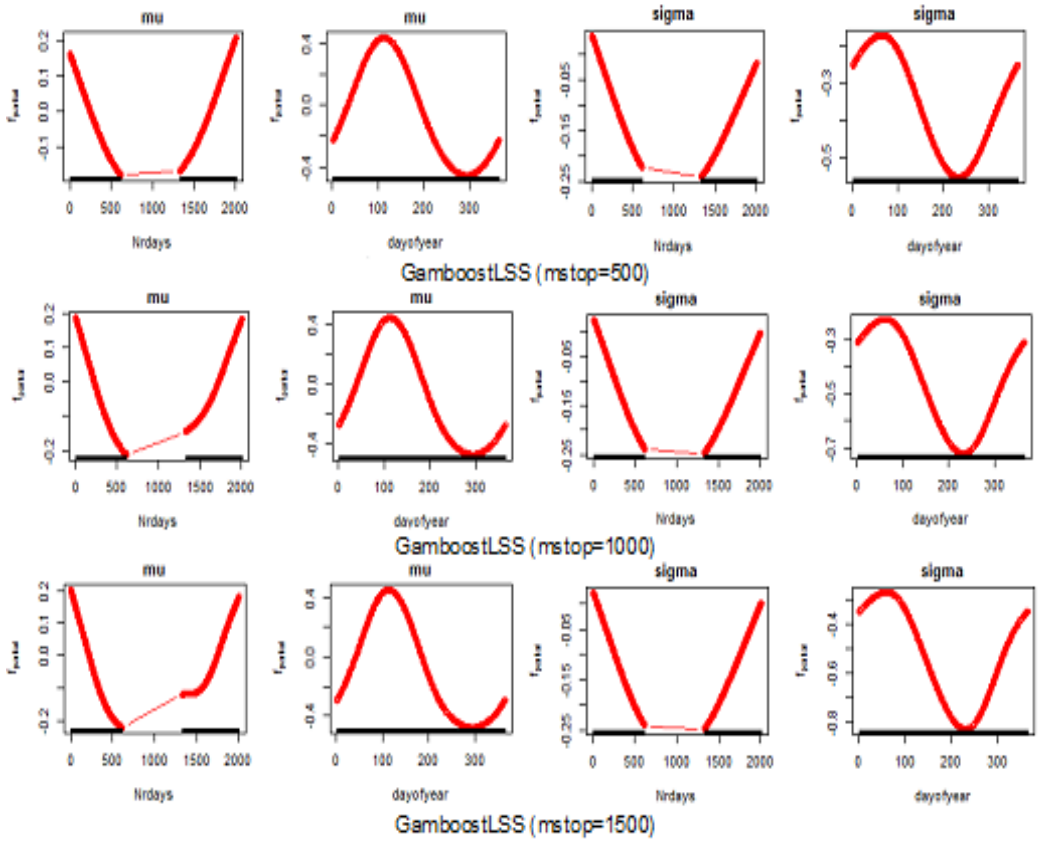


Figure 4.17: *Local model fitting for time covariates by using different m_{stop}= 500-1500 shows similar patterns in μ and σ parameters by using gamboostLSS models.*

The effect of changing the value of m_{stop} (500 to 1500) is more visible on *Nrdays* covariate in local model fitting as captured in Figure 4.17. It can be seen that there is a significant change in the pattern before and after the gap in local model fitting. At the same time, *Doy* covariate is more stable in μ and σ parameters. However, the effect of the m_{stop} for global

model fitting has a relatively similar pattern. Further we consider 3000-5000 values of m_{stop} for gamboostLSS models, which results are depicted in Figure D.1, Appendix D. The graphs show that there is a slight change in the pattern of the *Nrdays* at μ and σ parameters. However, there are similar patterns for *Doy* covariate for all m_{stop}.

Furthermore, we observed range values of m_{stop}= 2000-5000 and ν_{slf} = 0.01 for gamboostLSS models. The results are described in Figure D.2, Appendix D. The graphs show that there is a slight change in the pattern of the *Nrdays* at the μ and σ parameters. This shows there are stable patterns for *Doy* covariate for all the four values of the m_{stop}.

If we compare both groups of the stopping iteration, then model fitting by using m_{stop}= 2000-3000 is better than m_{stop}= 4000-5000 in the local model fitting. We can see that the large m_{stop} values causes a slight change in the annual effects and a similar change in the seasonal effects. For m_{stop}= 4000-5000 it gives 15 submodels larger than m_{stop}= 2000-3000, i.e. 11 and 12 submodels respectively.

4.8.1.3 Effect of the Knots on GamboostLSS Models

We consider 40 to 60 knots with 10 steps in model fitting by using gamboostLSS models. The increasing knots gradually changes the patterns of the time covariates, especially for annual effects, whereas the seasonal effects show a stable pattern within the μ and σ parameters. Time covariates effects of gamboostLSS models show similar patterns for location (μ) of annual effects and for scale (σ) of seasonal effects. In addition, annual effects before and after gap each increase knots show similar trends. Effects of the different number of knots for local model fitting does not change the seasonal pattern and model fitting pattern. The knots effect with respect to the model fit has a similar pattern due to dominance of stability

of the seasonal patterns in the model fitting.

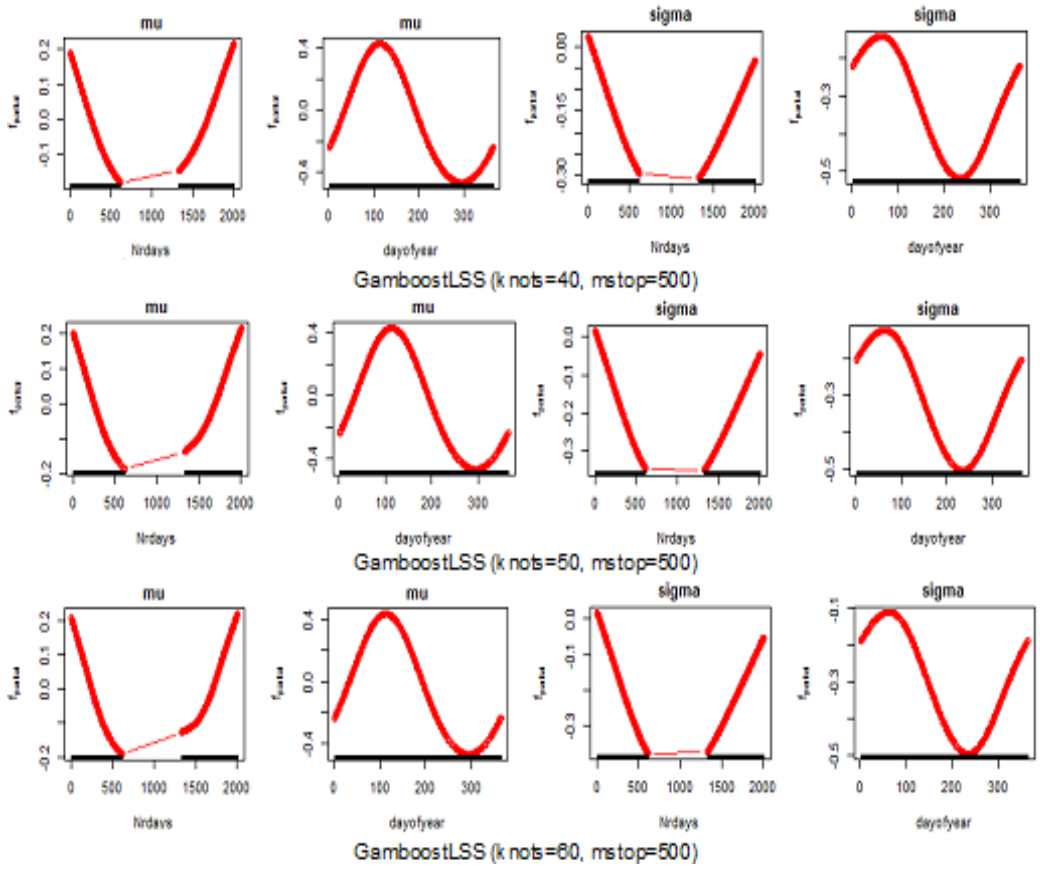


Figure 4.18: *Time covariates effects of gamboostLSS models show similar patterns for location and for scale for annual and seasonal effects. For annual effects before and after the gap, it shows similar trends for each step of knots.*

4.8.1.4 Effect of the *df* GamboostLSS with Transformation

Similarly, without transformation we consider the degrees of freedom *df* of the *Nrdays* covariate from 2.01 to 3.5. The control boosting that used is stopping iteration (m_{stop}= 1000) and the size length factor (ν_{slf}= 0.1). Increasing degrees of freedom *df* in the model fitting causes the increase of fluctuation on the *Nrdays* covariate (as annual effects) and it does not change patterns of the *Doy* covariate (as seasonal effects), as seen in Figure D.3, Appendix D. Transformation of rainfall in model fitting does not change patterns of time covariates as in without transformation as captured in Figure 4.19.

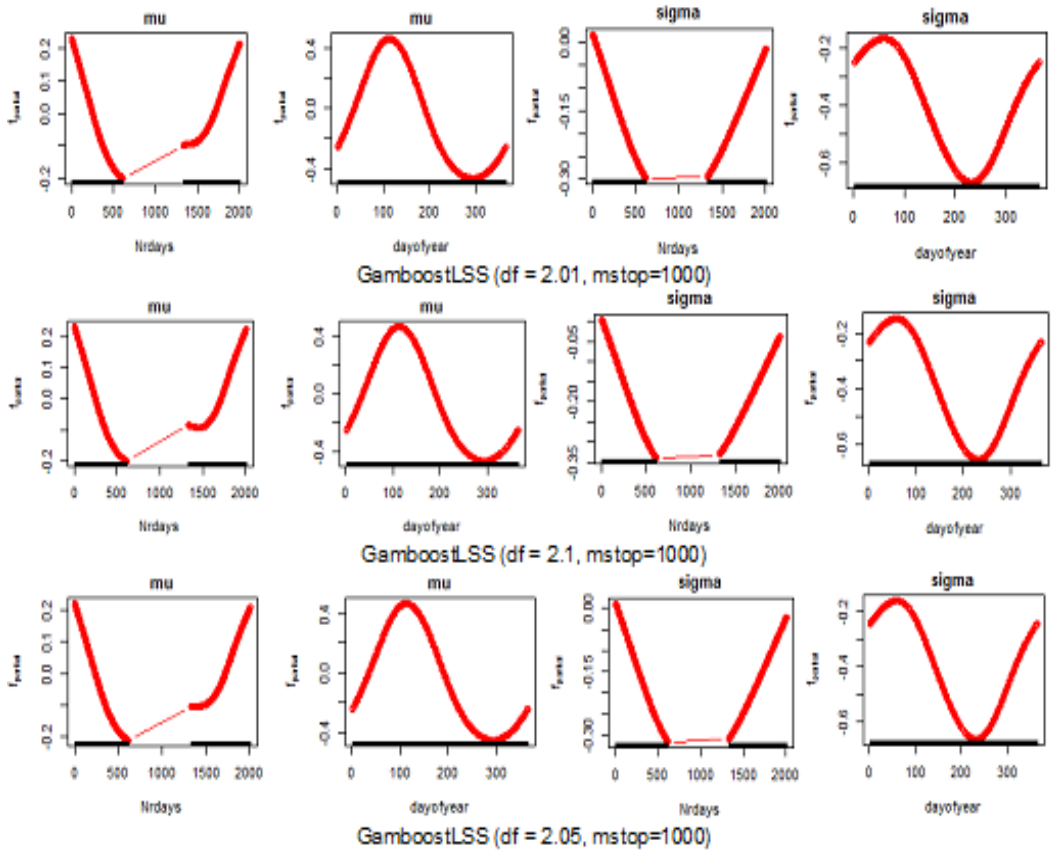


Figure 4.19: *Illustration of local model fitting with different degrees of freedom for time covariates of the SST data fitting with transformation of rainfall covariate.*

This figure shows appropriate local fitting of time covariates by using gamboostLSS models with considering *df*, m_{stop} and transformation. The same pattern of the *Doy* covariate in the μ and σ parameters. The similar pattern of the *Nrdays* in the σ parameter, whereas a slightly changed pattern after the gap in the μ parameter. Different *df* in model fitting shows changed after the gap of the *Nrdays* covariate. The increasing *df* (over specific values) can affect fluctuation of the *Nrdays* mainly after the gap.

Illustration of D.3 in Appendix D shows inappropriate gamboostLSS model on local fitting mainly at the *Nrdays* covariate. The increasing of *df* from 2.5 to 3.5 with fixed m_{stop}= 1000 displays fluctuation at the *Nrdays* after the gap in the μ parameter.

We investigated SST model fitting by using different degrees of freedom with transfor-

mation. The effect of changing the value of *df* shows that the model fitting with transformation is more appropriate on *df* = 2.01-2.1 (Figure 4.19) compared to *df* = 2.5 to 3.5 in Figure D.3, Appendix D). However, the results of the global fitting is nonsmooth, we do not present in graphs. The increasing of *df* in model fitting also causes increasing the number of submodels if we compared to without transformation.

Generally, SST model fitting with transformation gives a similar pattern such model fitting without transformation. In other words, transformation of rainfall does not affect the SST model fitting. The transformation of rainfall has an impact on the same covariate of marginal model fitting (i.e. the submodel rainfall itself), but it does not on global model fitting.

4.8.1.5 Effect of the m_{stop} on GamboostLSS with Transformation

In the transformation, we investigated model fitting by using the stopping iteration similar to without transformation by the same model. The changing parameter is focused on the *Nrdays* covariate with number of the *knots*= 40, *df*= 2.01 and difference of penalty= 2.

The stopping iteration gives impact for marginal model fitting, mainly for the *Nrdays* covariate, which is slightly changed after the gap. Similarities on the annual effects occur each 500 increases of the m_{stop}. For *Doy*, the covariate is more stable in μ and σ parameters. In addition, increasing the m_{stop} values does not influence fluctuation of seasonal effects, but has a slight effect on the fluctuation of annual effects in the model fitting as seen in Figure 4.20.

Figure 4.20 shows the effect of increasing m_{stop} for model fitting which shows a similar pattern in μ and σ parameters. We can see a slight change for *Nrdays* covariate with *knots*=

40 and m_{stop}= 1500 after the gap in the μ parameter.

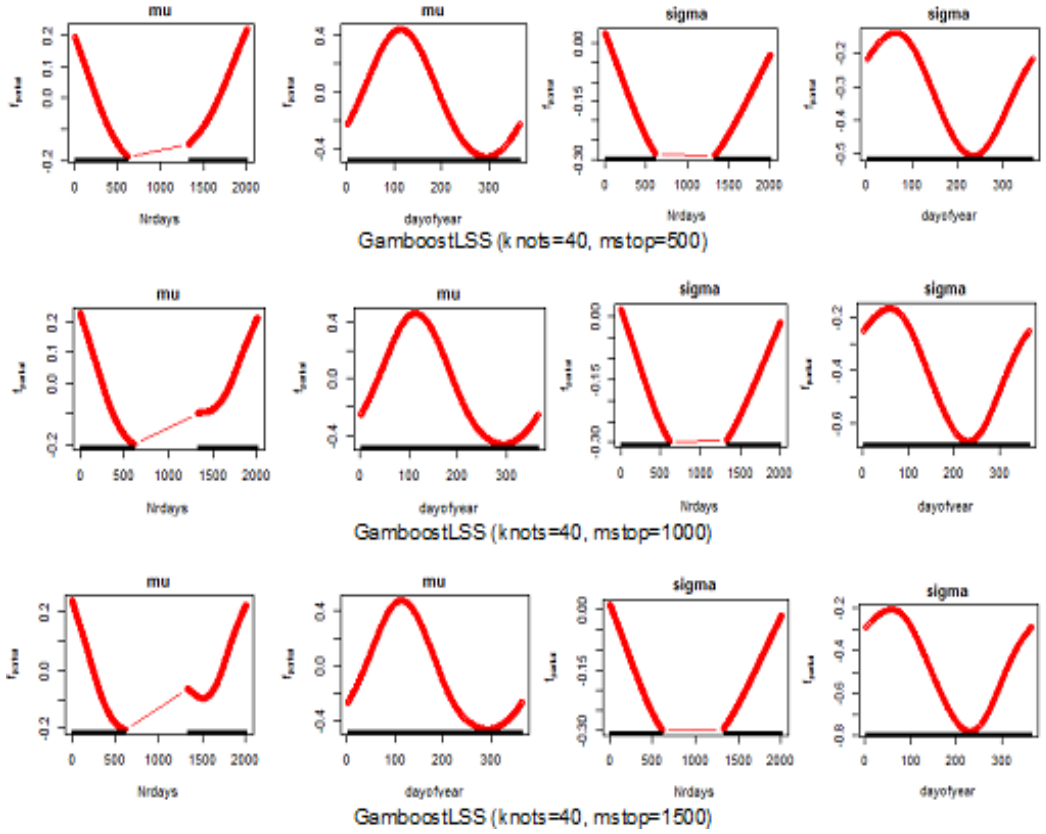


Figure 4.20: *Local model fitting for time covariates with different m_{stop} = 500-1500 and transformation of rainfall gives similar patterns.*

4.8.1.6 Effect of the Knots on GamboostLSS with Transformation

In transformation of rainfall, we consider from 40 to 60 knots with 10 steps in model fitting by using the same GMb5 gamboostLSS model. In general the effect of the knots in the model fitting for without transformation has similar patterns with transformation of rainfall. These similar patterns are in μ parameter from 40 to 50 knots, but at the 60 knots there are differences of patterns in μ and σ parameters as seen in Figure D.4, Appendix D.

Overall, the effect of the number of the knots for local model fitting is unchanged in the patterns of seasonal effects in μ and σ parameters. The knots effect with respect to model fitting has a similar pattern due to dominance of stability of seasonal effects in model fitting.

Hence, the different knots in the *Nrdays* has not impacted significantly on the model fitting for the SST data.

4.8.1.7 Effect of the *df* at the *Doy* covariate on GamboostLSS with Transformation

We investigated the *df* parameter in the *Doy* covariate is *df* = 1.1 to 1.5. We fixed parameters in the *Nrdays* covariate are *df* = 2.01, *difference*= 2, *knots*= 40, and in the control boosting parameters: m_{stop}= 1000 and ν_{slf}= 0.1. The result of this particular investigation is depicted in Figure D.5, Appendix D.

The figure provides various patterns of *df* of the time covariates, regarding fixed m_{stop}= 1000. Similar pattern of annual effects at *Nrdays* covariate in μ and σ, exclude at μ parameter after the gap with *df* is 1.2 to 1.5 shows inappropriate local fitting. In general the same pattern displayed for seasonal effects in μ and σ parameters.

4.8.1.8 Effect of the m_{stop} with respect to the *Doy* covariate on GamboostLSS with Transformation

We observed the m_{stop} parameter on gamboostLSS model with transformation. We fixed parameters in the *Nrdays* covariate: *df* = 2.01, *difference*= 2, *knots*= 40, and in the control boosting parameters: m_{stop}= 500, 1000, and 1500; ν_{slf}= 0.1. The result of this particular observation is displayed in Figure 4.21.

Figure 4.21 describes analog results by using the m_{stop}= 500-1000 with the same specification. Whereas m_{stop} =1500 shows changed after the gap of the *Nrdays* covariate. The increasing m_{stop} (over specific values) can affect fluctuation of the *Nrdays* mainly after the gap. We recommend for the *Doy* covariate using *df* = 1.1 and m_{stop} until it reaches 1000 in

gamboostLSS model fitting. Therefore, we used the $df = 1.1$ at the *Doy* covariate for next investigation of the SST data fitting with considering m_{stop} values. The m_{stop} can be used as threshold point to determine whether model fitting in the gap is appropriate or not.

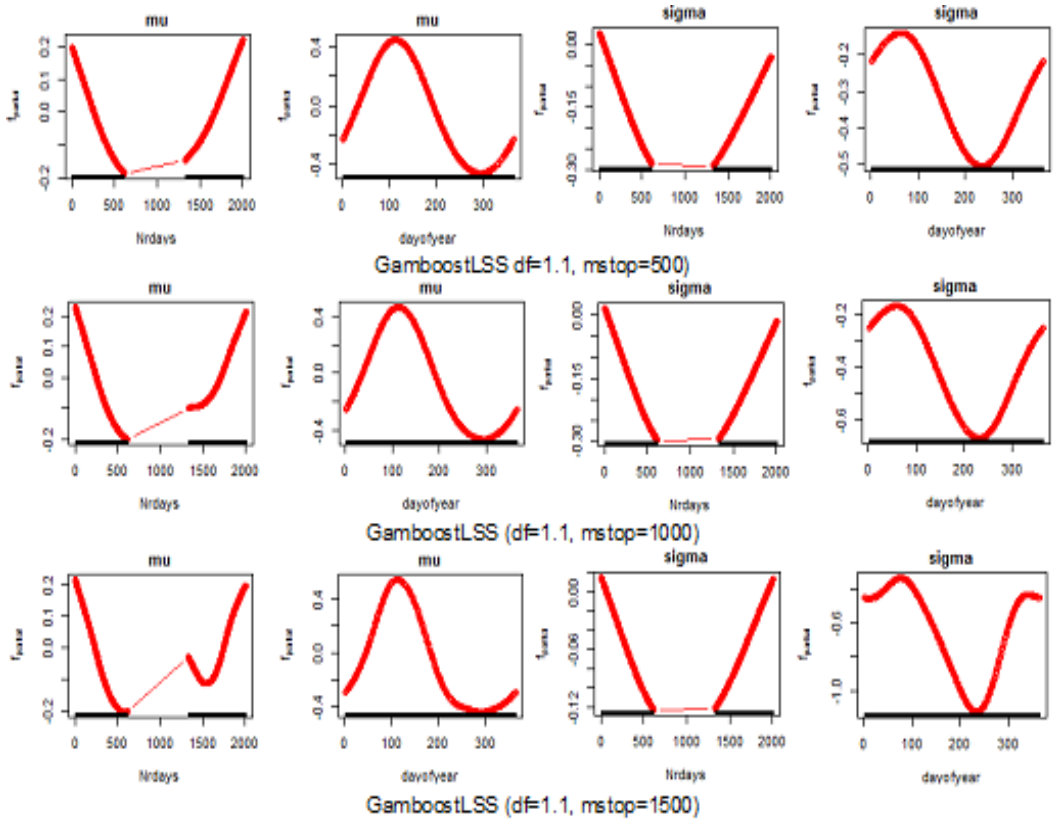


Figure 4.21: *Local fitting of time covariates with different df and m_{stop} = 500-1500 using gamboostLSS model for the SST data.*

Transformation of rainfall covariate can change the type of effect for each covariate on corresponding distribution parameters. We observed from the results that transformation decreases the AIC, but the transformation does not necessarily improve the model fitting. The SST model fitting by gamboostLSS models depends on the optimal choice of the parameters on base-learners and control boosting.

4.9 Summary

In this chapter, for linear model fitting we proposed two models M0 and M1 to investigate the effects of time covariates in the SST data modelling. Two time covariates seasonal and annual along with three continuous covariates are included in the M1 model and the M0 model consists of three continuous covariates. We observed that the time covariates have large influence in the model fitting of SST data. Meanwhile, transformation of rainfall in the M0 and M1 models improved the fit of both models. Our experimental comparisons of M0 and M1 model reveal that an increase in R-squared, F-value and a reduction in residuals can be achieved by inclusion of the time covariates in the model.

The preliminary statistical evidences of the linear models indicate that the proposed M1 model fitting for SST data need further improvement. A more flexible structured model is required to capture the complexity of SST data.

Further we have proposed gamboostLSS models for SST data fitting by incorporating the time covariates in the model. We compared our proposed method with GAM, gamboost and GAMLSS models for SST data fitting. Boosting technique is implemented in gamboost and gamboostLSS models that leads to an increase in the degree of basis composition, minimizes computational time, and improve the fitting process of the SST data. The AIC measure is used for selection of hyper-parameters in the first three models. For the gamboostLSS models, CV-risk is used to choose optimal values for the parameter *mstop*. The wiggliness in the gap is potentially changed based on both of the mentioned factors.

Our experimental results show that a good trade-off between hyper-parameters and model fitting is essential in model selection for SST data. The experimental comparisons

reveal that GAM models provide appropriate model fitting with a comparatively smaller number of hyper-parameters (only three hyper-parameters), however, these models are expensive in terms of computational time. This issue is dealt with by gamboost models where the boosting technique is applied to GAM models. Although, the gamboost models are faster than GAM models, however, the limitation of these models is that sometimes the time covariates are not covered by the models. The results on GAMLSS models reveal that the smaller AIC can be achieved by these models when incorporated with time covariates. However, it takes considerable computational time.

The experimental results on our proposed gamboostLSS models demonstrate that these models are more flexible in terms of smoothing function as compared to other three models. These models are also comparatively faster than the other models. In addition, model fitting by gamboostLSS can handle long gap observations, which is very common in the SST dataset. By implementing gamboostLSS models to the SST data, a good trade-off between computational time and CV-risk can be achieved. Moreover, we carried out experiments on transformed data. The transformation is done on the rainfall. The results of the models reveal that the transformation of rainfall leads to a reduction in final risk.

Chapter 5

GamboostLSS in Autocorrelation Models and Applications for Different Buoys

5.1 Introduction

The SST data derived from the buoys has gaps, sparsity, and irregular patterns. The autocorrelation or serial correlation errors in it can affect the response over time. It is collected from different locations and therefore has different variability. Additionally it possesses two types of autocorrelation, i.e. spatial (or spatio-temporal) and time autocorrelation. The earlier can be from the sources of region in latitude and longitude as a smooth function, interaction terms, and longitude shift. The latter can be derived from periodical time units, such as daily, monthly, seasonal, and annual basis.

In this chapter, we propose the filtering of covariates by generalizing the differencing approach in gamboost and gamboostLSS models to deal with the issue of time autocorrelation in the SST data. The proposed model fitting is dependent on the specification of

base-learners, boosting-control, and family based distribution of the SST data. Nevertheless, to approximate hyper-parameter values of the model, controlled fitting processes are required. These include appropriate model fitting and optimal number of submodels.

The chapter is further organised in two part as follows, Firstly, we presented gamboostLSS models in autocorrelation for one buoy, and secondly, application for gamboostLSS-AR(1) models for different buoys.

In first part, section 5.2 autocorrelation is discussed. In the subsequent section 5.3 gamboostLSS using generalized differencing for AR(1) model is presented. We used experimental setup with gamboost-AR(1) and gamboostLSS-AR(1) models as in sections 2.3 and 2.3 Chapter 2. In section 5.4 results and discussion are given that have been conducted previously.

In second part, we presented application of gamboostLSS and gamboostLSS-AR(1) models 5.5. Section 5.6 we provide results and discussion for different buoys. In section 5.7 we presented marginal prediction interval of gamboostLSS models in autocorrelation. Finally, in section 5.8 the summary of the chapter is reported.

5.2 Autocorrelation

Autocorrelation is a pattern of a sequential relationship between the same types of objects with a lag. For example, autocorrelation between errors ε_t and ε_{t-k}, response y_t and y_{t-k} and covariate x_t and x_{t-k} at k lag, respectively. The term autocorrelation is used when there is a serial correlation of the residuals in a period as well. In this chapter, we focus on equation 3.2 where we assume that ε_t's have a zero-mean and are autocorrelated. Generally,

autocorrelation occurs due to the heteroscedasticity or serially correlated problems when there are possible violations of assumptions,

(a). $E[\varepsilon_t \varepsilon_t^{'}|X] = \sigma^2 I_n$ (b). $E[\varepsilon_t, \varepsilon_{t-1}] = 0$,

Autocorrelation can also be caused by other factors, such as missing important covariates, the misspecification of the model or disturbance terms, systematic errors in measurement of data, spatial ordering, and event inertia. In the SST data the measurements are based on daily observations and therefore autocorrelation arises in the time factor. Several techniques are used to investigate the autocorrelation in the model. Considering autocorrelation of the residual is one way of diagnostic checking and can lead to an effective model fitting.

Recall that in Chapter 2, equation 4.1, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = V\sigma^2$, and V_{nxn} is a matrix of autocorrelations in the errors, then

$$V^{-1/2}\mathbf{Y} = V^{-1/2}\mathbf{X}\beta + V^{-1/2}\varepsilon,$$

and suppose $\mathbf{Y}^* = Z\beta + \delta$, where $\mathbf{Y}^* = V^{-1/2}\mathbf{Y}$, $Z = V^{-1/2}\mathbf{X}$, and $\delta = V^{-1/2}\varepsilon$, then

$$\hat{\beta} = (Z'Z)^{-1}Z'\mathbf{Y}^* = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}\mathbf{Y}, \tag{5.1}$$

where V has the particular correlation errors $v_{ij} = corr(\varepsilon_i, \varepsilon_j) = \rho^{|i-j|}$.

Furthermore, from equation 3.7 in Chapter 3 the P-spline with autocorrelation errors is:

$$PLS(\beta) = (\mathbf{u} - \mathbf{B}\beta)^T V^{-1}(\mathbf{u} - \mathbf{B}\beta) + \lambda W(\beta, m) \tag{5.2}$$

where the correlation matrix $V = [v_{ij}]$ as suggested in [76].

5.3 GamboostLSS using Generalized Differencing for AR(1)

We suggested that a model auto regressive (AR(1)), where generalizing differencing approach is used to investigate autocorrelation in the data by incorporating an autoregressive process. Referring to equation 3.2, we use the AR(1) model in a formulation in our experiments which is,

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad t = 1, 2, ..., n \tag{5.3}$$

If we assume that the u_t's are uncorrelated random errors with zero mean and constant variances, then,

$$E(u_t) = 0, Var(u_t) = \sigma_u^2 \text{ and } Cov(u_t, u_s) = 0, t \neq s, \tag{5.4}$$

and let us assume that $\varepsilon_t \sim N(0, \sigma^2\Lambda)$, where Λ is a correlation matrix defined through an AR(1) with autocorrelation parameter ρ can be implemented in the SST data. Lillard and Willis [77] proposed a model where the error structure is considered, and it generalizes differencing approach which can applied to reduce the error autocorrelation. In equation 5.3, if $\left|\rho\right| < 1$, then ε_t is stable in AR(1) model [78]. When $\rho = 0$, then it is called the white noise error [78–81]. If ρ tends to 1 then AR(1) model becomes a random walk RW(1) model [42].

AR(1) is used to deal with low-order autocorrelation. The ρ in the model is a robust estimator because it represents the sequential (serial correlation) effect of ε_t which depends on ε_{t-1}. Subsequently, ε_{t-2} control the previous effect of ε_{t-1}, [79, 82]. However, in seasonal or quarterly autocorrelation and spatial autocorrelation effects, this approach cannot be applied. In this case each pattern of autocorrelation is not easily adjusted and it is more complex to combine both effects. Furthermore, if the monthly SST data is considered, then the autocorrelation of the residual lag should be checked at lag 12. When autocorrelation

is not significant at lag 12, then it can be assumed that the errors are independent [48].

Specifically, the autocorrelation coefficient of ρ_l in discrete types of covariate x at $l = 0, \pm1, \pm2, \pm3, ...$ is the covariance among pairs of covariates at time lags (or a distance in spatial lags) l that is normalized by discretizing the continuous process,

$$\rho_l = \frac{\pi(|||l|||)}{\sigma_x^2}, \ l = 0, \pm1, \pm2, \pm3, ..., \tag{5.5}$$

whereas the autocorrelation ρ_l in continuous types of covariate x at l is,

$$\rho_l = \frac{\pi(l)}{\sigma_x(n)^2}, \tag{5.6}$$

where $\pi(l) = cov(x_n, x_{n+l}) = E[(x_n - \mu)(x_{n+l} - \mu)]$, and variance $\sigma_x(n)^2 = E[(x(n) - \mu(n)^2]$, [83,84]. In general, to calculate the lth lag sample autocorrelation errors $\hat{\rho}$, it can be formulated as

$$\hat{\rho}_l = \frac{cov(\varepsilon_t, \varepsilon_{t-l})}{\sigma_\varepsilon^2} \tag{5.7}$$

In seasonal data, using the model AR(1) the autocorrelation is $\varepsilon_t = \rho_i \varepsilon_{t-3i} + w_t$, for $i = 1, 2, 3, 4$ per season, where w_t refers 5.4. Thus ρ for all the seasons can be formulated as,

$$\varepsilon_t = (\rho_1 \varepsilon_{t-3} + \rho_2 \varepsilon_{t-6} + \rho_3 \varepsilon_{t-9} + \rho_4 \varepsilon_{t-12}) + w_t.$$

Afterwards, to calculate $\hat{\rho}$ in equation 5.7 and then to transform a covariate x and response y in $x^* = x_t - \hat{\rho} x_{t-1}$ and $y^* = y_t - \hat{\rho} y_{t-1}$, we use relationship formula between covariates and response as follows $y^* = \beta_0 + \beta_1 x^*$.

Consider an additive model $y = f + \varepsilon$ where,

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, i = 1, ..., n, \tag{5.8}$$

then errors ε_t and ε_{t-1} are: $\varepsilon_t = y_t - f_t$ and $\varepsilon_{t-1} = y_{t-1} - f_{t-1}$. We can say that equation 5.8 is a static additive model.

Boyce *et al*. and Anderson [84, 85] suggested that if the response is an observation over time, i.e. periodic process, then the harmonic regression oscillation approach can be utilised in model fitting. Furthermore, we start with the 1231 daily observations in the SST data. To apply the data we use generalized differencing for AR(1) model is translated as:

$$Y_t = \beta_0 + \beta_1 X_{1t} + ... + \beta_{3t} X_{3t} + \gamma_l I_{lt} + \eta_m D_{mt} + \rho\varepsilon_{t-1} + u_t, \tag{5.9}$$

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{1t-1} + ... + \rho\beta_{3t} X_{3t-1} + \rho\gamma_l I_{lt} + \rho\eta_m D_{mt} + \rho^2\varepsilon_{t-2} + \rho u_{t-1}, \tag{5.10}$$

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_{1t} - \rho X_{1t-1}) + ... + \beta_{3t}(X_{3t} - \rho X_{3t-1}), \tag{5.11}$$

$$+ \gamma_l(I_{lt} - \rho I_{lt-1}) + \eta_m(D_{mt} - \rho D_{mt-1}) + u_t, \tag{5.12}$$

where,

Y_t and Y_{t-1} are the SST at t and $t-1$ times (in day unit),

X_{1t} and X_{1t-1} are air temperature; X_{2t} and X_{2t-1} are relative humidity,

X_{3t} and X_{3t-1} are rainfall, I_{lt} and I_{lt-1} are *Nrdays*; D_{mt} and D_{mt-1} are the *Doy* covariates,

and $\rho\varepsilon_{t-1} - \rho^2\varepsilon_{t-2} + u_t - \rho u_{t-1} = \varepsilon_t - \rho\varepsilon_{t-1}$, given $u_t = \varepsilon_t - \rho\varepsilon_{t-1}$ and

ε_t and ε_{t-1} are errors at t and $t-1$ time, respectively. To simplify the model above we get,

$$Y_t^* = \beta_0^* + \beta_1 X_{1t}^* + ... + \beta_3 X_{3t}^* + \gamma_l I_{lt}^* + \eta_m D_{mt}^* + u_t$$

where $Y_t^* = Y_t - \rho Y_{t-1}$; $\beta_0^* = \beta_0(1-\rho)$; $X_{1t}^* = X_{1t} - \rho X_{1t-1}$; ... ; $X_{3t}^* = X_{3t} - \rho X_{3t-1}$; $I_l^* = I_{lt} - \rho I_{lt-1}$; $D_m^* = D_{mt} - \rho D_{mt-1}$; and $u_t = \varepsilon_t - \rho \varepsilon_{t-1}$. By using differencing approach, the additive model (as in Chapter 3, equation 3.3) is:

$$Y_t^* = \beta_0^* + \sum_{j=1}^{p} f_j(X_{jt}^*) + f(\gamma_l^*) + f(\eta_m^*) + u_t. \tag{5.13}$$

Further, we suggest that equation 5.13 is a dynamic additive model with the AR(1) model errors. Another way to express additive models in autocorrelation AR(1) errors of linear models is,

$$Y_t = (\beta^T x_t + \gamma I_t + \eta D_t) + \rho y_{t-1} - \rho(\beta^T x_{t-1} + \gamma I_{t-1} + \eta D_{t-1}) + u_t$$

Additive models in autocorrelation AR(1) errors are translated as,

$$Y_t = (\sum_{j=1}^{p} f_j(x_t) + f(I_t) + f(D_t)) + f(y_{t-1}) - (\sum_{j=1}^{p} f(x_{t-1}) + f(I_{t-1}) + f(D_{t-1})) + u_t$$

There are two ways to approximate ρ in equation 5.9. First, if ρ is known then the estimated Ordinary Least Squares (OLS) regression can be used to obtain a BLUE (Best Linear Unbiased Estimator). This procedure can be implemented to the base-learners approach, such as linear (bols) and smooth (bbs) functions in the gamboost and gamboostLSS models. Second, if ρ is unknown then it can be estimated by samples of n data observations.

5.4 Results and Discussion of Gamboost and GamboostLSS in Autocorrelation

The results of gamboostLSS in autocorrelation models are displayed as follows: in subsection 5.4.1 provides tuning parameters for autocorrelation errors AR(1) model. In subsection 5.4.2 discussed autocorrelation of the gamboost models. Subsection 5.4.3 gamboost-AR(1) models with transformation are given. In subsection 5.4.4 autocorrelation of the gamboostLSS-AR(1) models are reported. In subsection 5.4.5 gamboostLSS-AR(1) models with transformation are discussed in detail. Finally, restriction errors of autocorrelation AR(1) models are reported 5.4.6.

5.4.1 Tuning Parameters for Autocorrelation Errors AR(1)

In this section, we have applied linear models to observe autocorrelation in the SST data as shown in Figure 5.1.

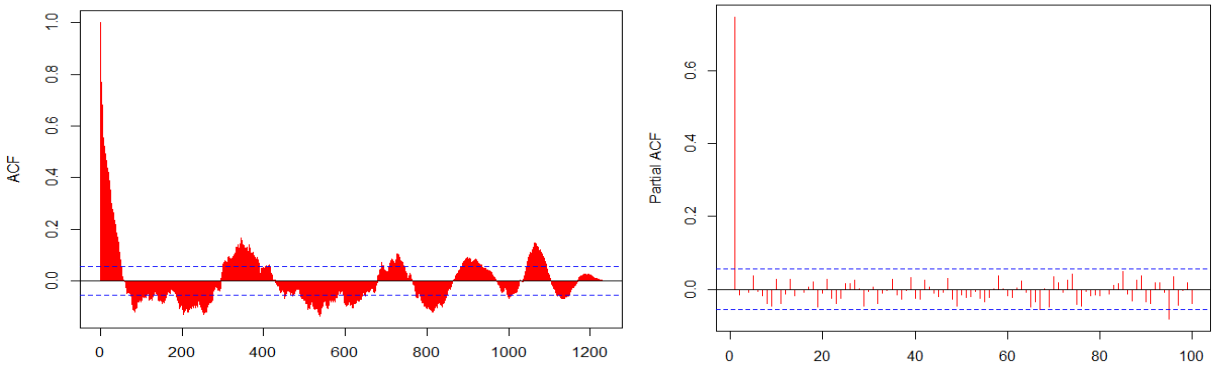


Figure 5.1: *The autocorrelation pattern in the SST data using the M1 linear model.*

Figure 5.1 displays the autocorrelation as a mixture of waves between sinusoidal and exponential functions, where the damped sinusoidal and exponential decaying (left), which shows autocorrelation errors as an autoregressive model. The model is showing the peri-

odic terms in the SST data, where one period shows an upward trend and the other reveals a downward trend. We can see from the Figures that there are two patterns with step 1 a long downward periods at the lag around 100-300 and 500-700. The graph shows the autocorrelation function having degree 1 at the time lag 0 and tends to 0 at the time lag over 1200. We also checked this pattern for local observations of 100 samples as a partial ACF. Figure 5.1 (right) displays the partial autocorrelation with a significant peak at lag 1 and the wiggly trend after lag 1 around 0. This means that all the higher-order autocorrelations are effectively explained by the AR(1).

We applied M1 (detail is given in Chapter 2) model to find the value for the coefficient of autocorrelation, i.e. $\hat{\rho}$ =0.8566652, in equation 5.7. In Chapter 4, we have observed that the time covariates have a large influence in the additive model fitting for the SST data. It is important to know the dynamics of the errors in the model and also to detect specifications of the covariates in the model fitting process. The aim is to investigate the effect of the covariates on the stability of the model for a large data set. Minimizing the autocorrelation using generalizing differencing method can lead to an appropriate model.

5.4.2 Autocorrelation of the Gamboost Models

Assessment of the data for autocorrelation is necessary before model fitting. The mean residuals for daily observation of the SST data is described as first-order serial correlation at lag 1. In this case, the Auto-Correlation Function (ACF) is described in Figure 5.1. We observe a high autocorelation with ρ = 0.85666 in the data. To mitigate the effect of this high autocorrelation we apply the differencing method to the data. Then we implement gamboost models for the data by considering the autocorrelation effect in the model fitting

as captured in Table 5.1. The results are depicted in Table 5.2.

Table 5.1: *Gamboost-AR(1) models specification using P-spline for with and without transformation.*

Model	df_{Nrdays}	$knots_{Nrdays}$	df_{Doy}	m_{stop}
GMboost1-AR(1)	2.5	100	1.5	1000
GMboost2-AR(1)	2.5	100	1.5	1500
GMboost3-AR(1)	2.5	100	1.5	2000
GMboost4-AR(1)	2.5	120	1.5	1000
GMboost5-AR(1)	3.5	120	1.5	1000
GMboost6-AR(1)	3.5	120	1.5	1500
GMboost7-AR(1)	3.5	140	1.5	1000
GMboost8-AR(1)	3.5	140	1.5	1500

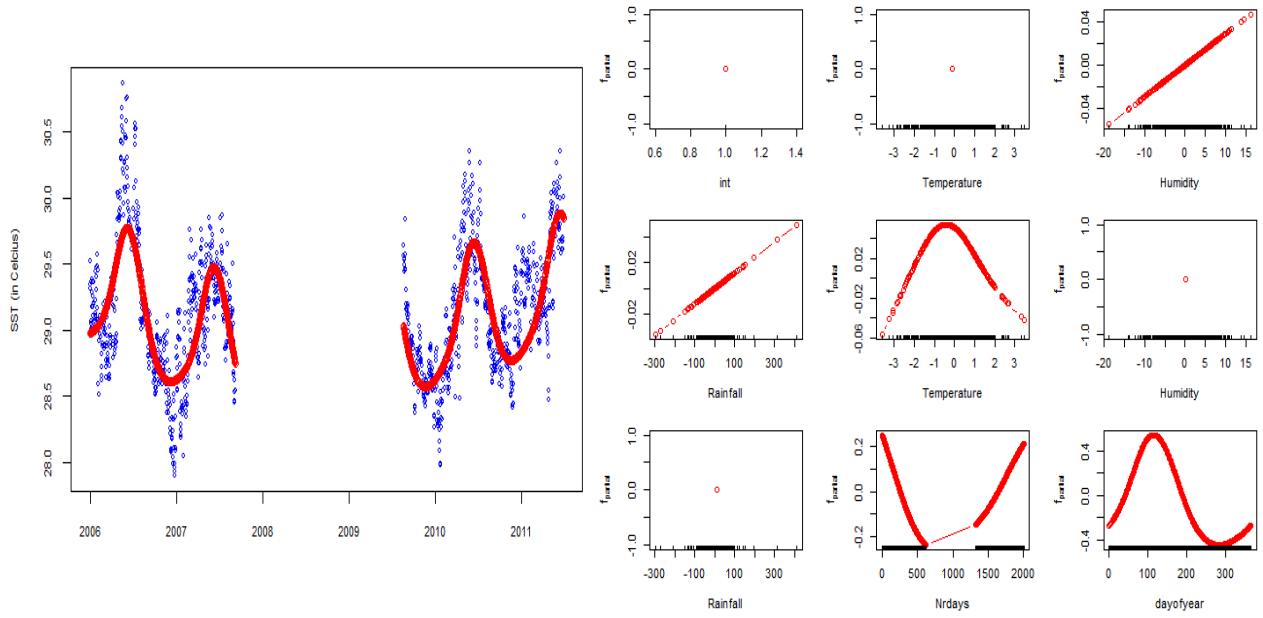
Table 5.2: *AIC of gamboost-AR(1) models using P-spline without transformed rainfall.*

Model	$df_{Corrected}$	$\text{AIC}_{Corrected}$	df_{gMDL}	AIC_{gMDL}	Final Risk
GMboost1pre-AR(1)	8.42453	-1.265255	8.42453	-2.177940	125.7202
GMboost2pre-AR(1)	9.79728	-1.283225	9.71272	-2.183098	123.2010
GMboost3pre-AR(1)	10.75285	-1.294201	10.74375	-2.184230	121.6633
GMboost4pre-AR(1)	8.49151	-1.265940	8.49151	-2.177968	125.6202
GMboost5pre-AR(1)	9.48281	-1.274146	9.48281	-2.176505	124.3587
GMboost6pre-AR(1)	10.8415	-1.299299	10.82986	-2.188418	121.0269
GMboost7pre-AR(1)	9.50973	-1.274359	9.49769	-2.176519	124.3574
GMboost8pre-AR(1)	10.85447	-1.299759	10.74683	-2.189285	120.9687

Table 5.2 shows that GMboost1pre-AR(1) to GMboost8pre-AR(1) models have slightly similar *df*, AIC, gMDL and final risk values. It means that performance of GMboost1pre-AR(1) to GMboost8pre-AR(1) models fitting of the SST data have the similar pattern.

As interpretability we can see the models as displayed in Figures E.1 and E.2, Appendix E. The figures show similar patterns of global fitting using gamboost-AR(1) models of the SST data. These results can be compared with the same specification of gamboost models in Table 4.16. It can be seen from the figure that gamboost-AR(1) model gives more appropriate global fitting than gamboost models. However, in local fitting using the gamboost-AR(1) models produce less number of submodels than in gamboost models. For example, the GMb1 to GMb8 of gamboost model gives more submodels than GMb1-AR(1) to GMb8-AR(2) of gamboost-AR(1) models.

There are effects when removing autocorrelation in the model fitting by using gamboost models as we can see in Figure 5.2. Gamboost-AR(1) model fitting has more flexibility than gamboost models, where we can use the low values of *df* and *knots* specification in the *Nrdays* covariate as representing annual effects. Likewise, boosting iteration of gamboost-AR(1) model fitting can reach a high value of the stopping iteration in reducing empirical risk so that it makes it easy to obtain an appropriate model fitting with many solutions as seen in Figure 5.3.



(a) (b)

Figure 5.2: *The SST data fitting by GMboost3-AR(1) model (left) and submodels (right) with* m_{stop}*=2000.*

Figure 5.2 displays global fitting by GMboost3-AR(1) model with the same specification as in the GMboost3 model (in Chapter 4). Removing autocorrelation in the GMboost3-AR(1) model gives flexibility to improve global model fitting, as can be seen in Figures 5.2 (a) and 4.9 (a) as a comparison on global fitting. In addition, GMboost3 model without autocorrelation produces nine submodels with two smooth submodels and seven linear submodels as seen in Figure 4.9 (b), whereas GMboost3-AR(1) model gives nine submodels with four smooth submodels and five linear submodels as seen in Figure 5.2 (b).

Further we also tested the model by considering flexible specification for stopping iteration (m_{stop}), as displayed in Figures 5.3 and 5.4.

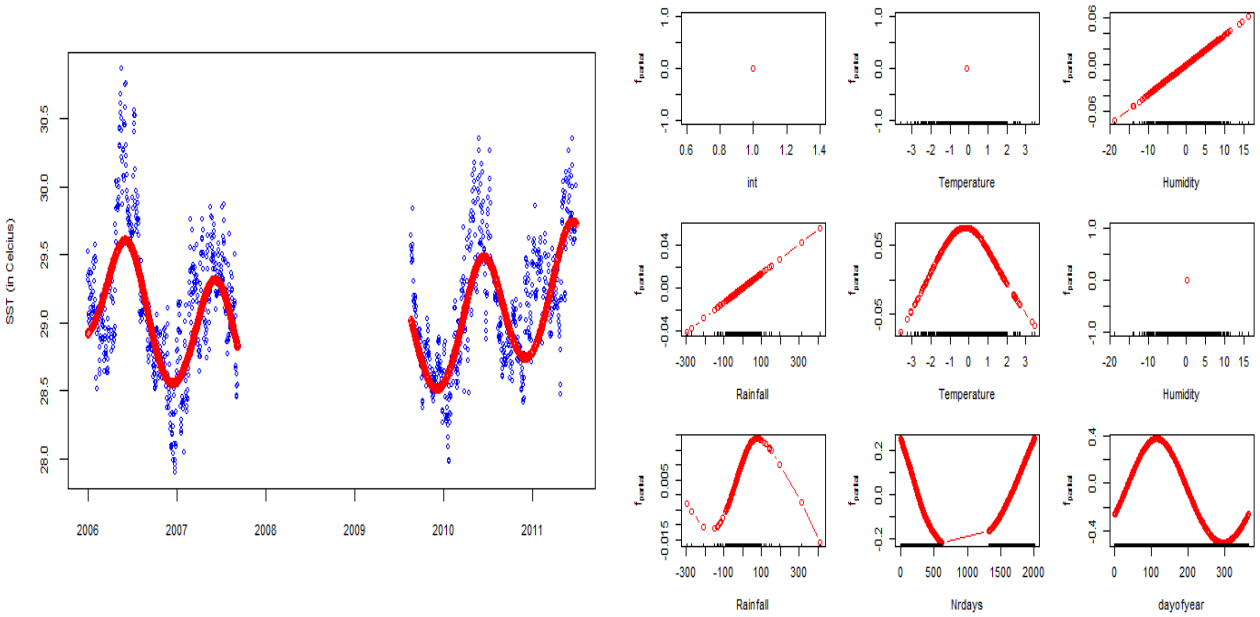


Figure 5.3: *The GMboost-AR(1) model fitting in global and local model of the SST data (m_{stop}= 12000).*

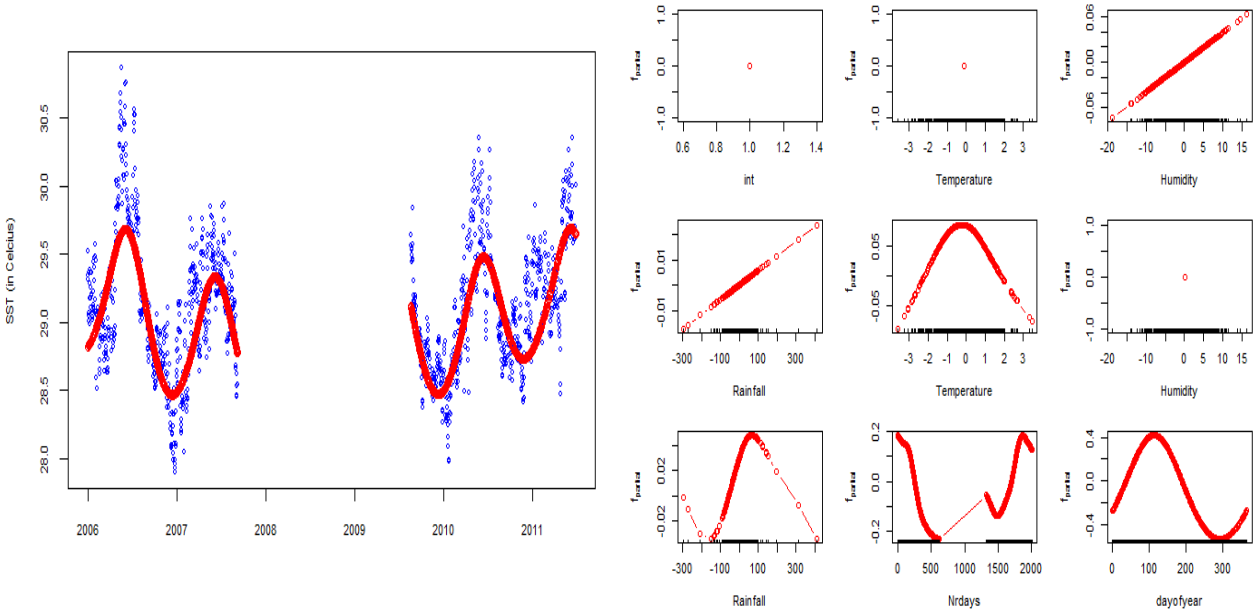


Figure 5.4: *The GMboost-AR(1) model fitting in global and local model of the SST data (m_{stop}= 35000).*

The higher m_{stop}= 12000 produces a precise global model fitting (left), whereas the local model fitting (right) leads to appropriate fitting. The rainfall covariate from smooth curve in the GMboost3-AR(1) model to polynomial curve in the GMboost-AR(1) model. The

higher m_{stop} gives a precise global model fitting (left), while the local model fitting (right) leads to inappropriate fitting as displayed in Figure 5.4. There is a fluctuation in the pattern of annual effects, mainly after the gap. In local fitting, we can see similar patterns between the low m_{stop} and high m_{stop} where the linear effect increases for humidity and rainfall. A polynomial curve in the patterns for temperature and rainfall is shown by the figure.

The finding reveals that pattern of the *Doy* covariate in gamboost-AR(1) is more stable than GAM and gamboost models fitting of the SST data. Furthermore, we observe patterns of time covariate on local model fitting in the μ parameter.

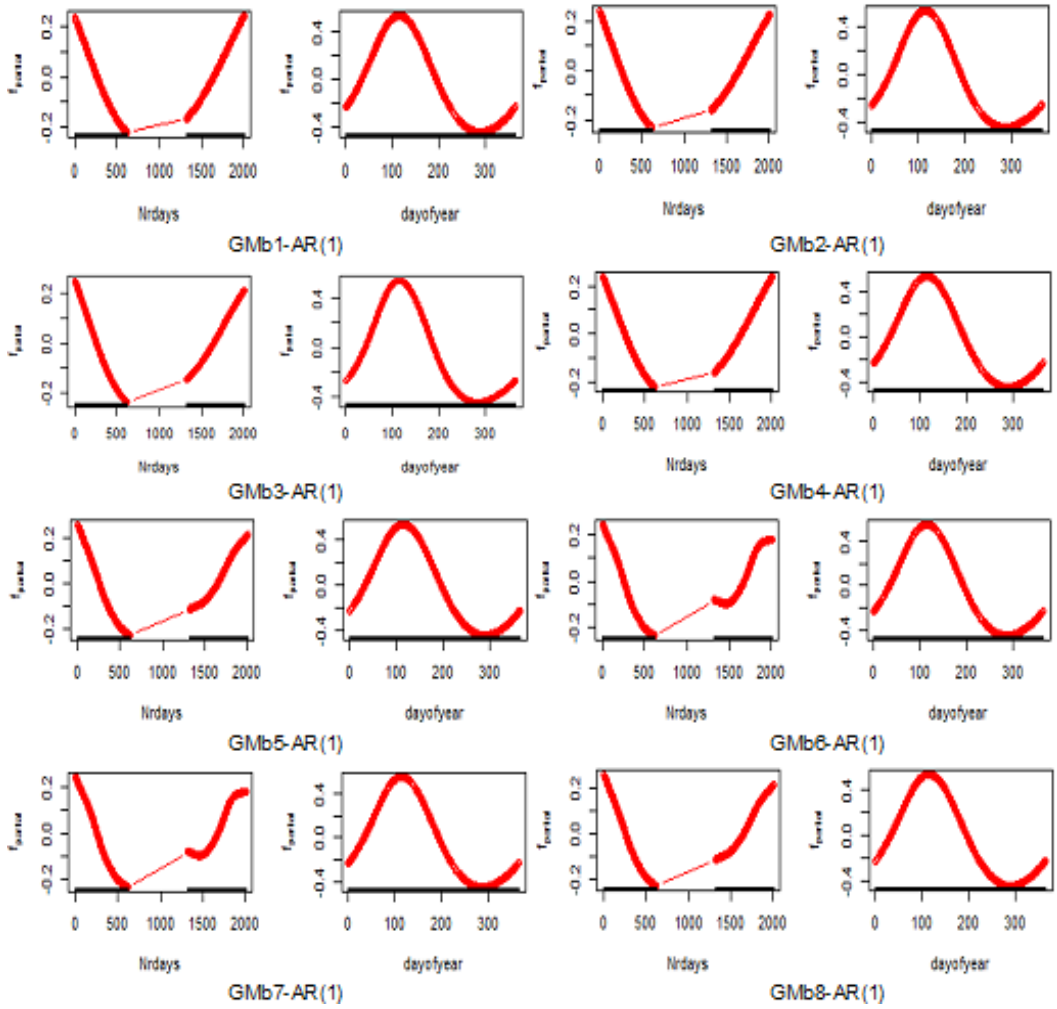


Figure 5.5: *The GMboost1-AR(1) to GMboost8-AR(1) models in local fitting of the SST data, to see in detail refer to Tables 5.1 and 5.2.*

Figure 5.5 shows appropriate local fitting on time covariates of the GMboost1-AR(1) to GMboost8-AR(1) models as in Table 5.2. However, GMb6-AR(1) and GMb7-AR(1) models show a slight change in the pattern of the *Nrdays* covariate after the gap. The slight change due to effect of the m_{stop}= 1500 in the GMb6-AR(1) model and effect of *knots*= 140 in the GMb7-AR(1) model, in detail see Table 5.1. Therefore, trade-off between m_{stop} and *knots* is important role to obtain appropriate local and global models fitting.

We presented gamboost-AR(1) models fitting without transformation, where the models have inappropriate in local fitting but it has appropriate in global fitting.

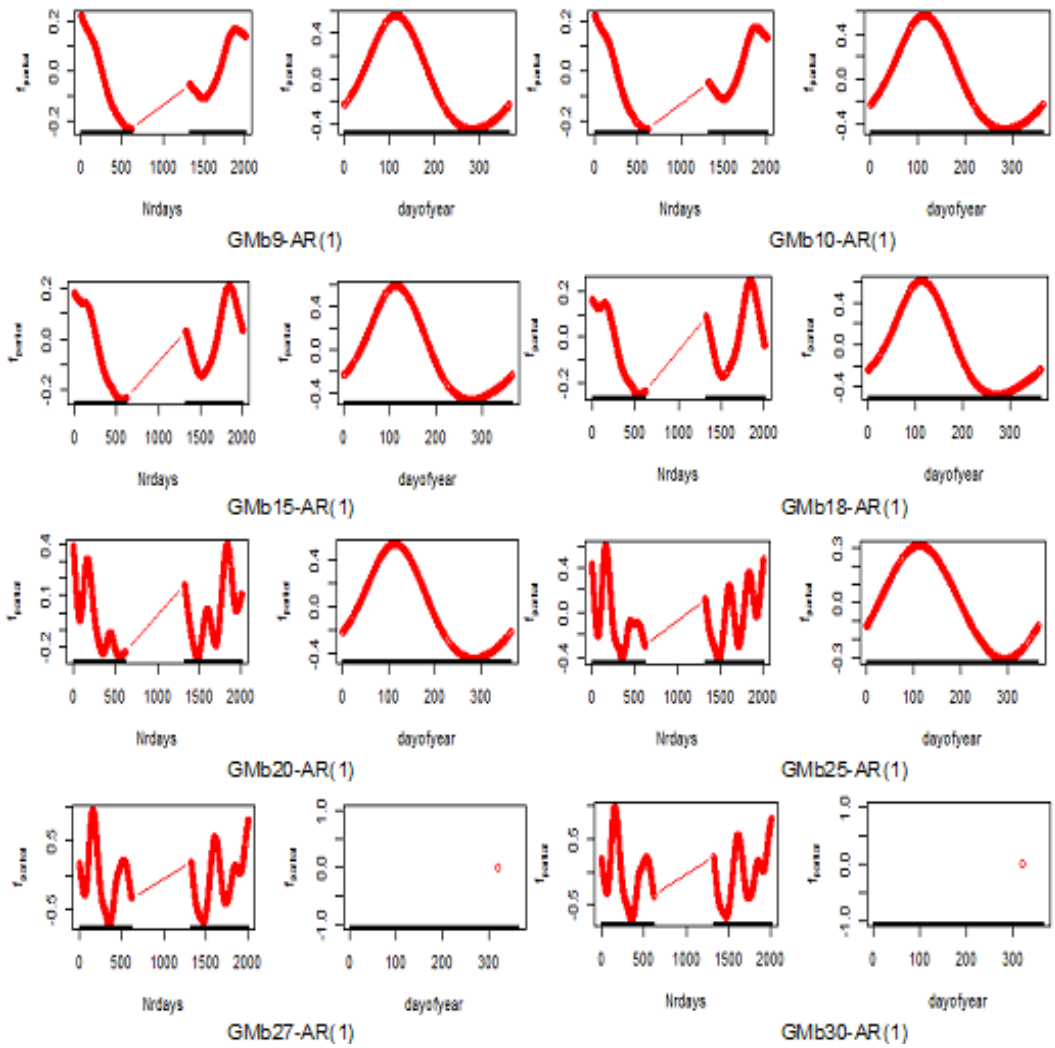


Figure 5.6: *The GMboost9-AR(1) to GMboost30-AR(1) models in local model fitting for the SST data.*

Figure 5.6 shows inappropriate local fitting on time covariates of gamboost-AR(1) models. The inappropriate local fitting on time covariates can be caused by several factors such as; a large value of the *df*, large number of the *knots*, and large values of the stopping iteration. Effects of these factors fluctuate on the *Nrdays* covariate before and after the gap and in smooth term of the *Doy* covariate. However, if the local fitting is not appropriate, then the global fitting is not always automatically inappropriate as well, as captured in Figure E.3, Appendix E.

5.4.3 Gamboost-AR(1) Models with Transformation

The results of the gamboost-AR(1) models with transformation of rainfall are reported in Table 5.3. We applied the models setup in a total of 30 models (GMboost9post-AR(1) to GMboost30post-AR(1) are not given in the table).

Table 5.3: *AIC of gamboost-AR(1) models using P-spline with transformed rainfall.*

Model	$df_{Corrected}$	$AIC_{Corrected}$	df_{gMDL}	AIC_{gMDL}	Final Risk
GMboost1post-AR(1)	10.90136	-1.266204	10.82829	-2.155453	125.0868
GMboost2post-AR(1)	13.02548	-1.283337	11.33569	-2.155538	122.5287
GMboost3post-AR(1)	14.82441	-1.293389	11.33569	-2.155538	120.9398
GMboost4post-AR(1)	10.96797	-1.266898	10.88812	-2.155425	124.9863
GMboost5post-AR(1)	11.68876	-1.275210	11.65217	-2.156604	123.7954
GMboost6post-AR(1)	13.24488	-1.300302	13.24488	-2.166611	120.4235
GMboost7post-AR(1)	11.69417	-1.275476	11.69417	-2.156674	123.7694
GMboost8post-AR(1)	13.22743	-1.300812	13.22743	-2.167279	120.3656

Table 5.3 shows that GMboost1post-AR(1) to GMboost8post-AR(1) models have slightly similar *df*, AIC, gMDL and final risk values. We select these models based on appropriate model fit in time-covariates.

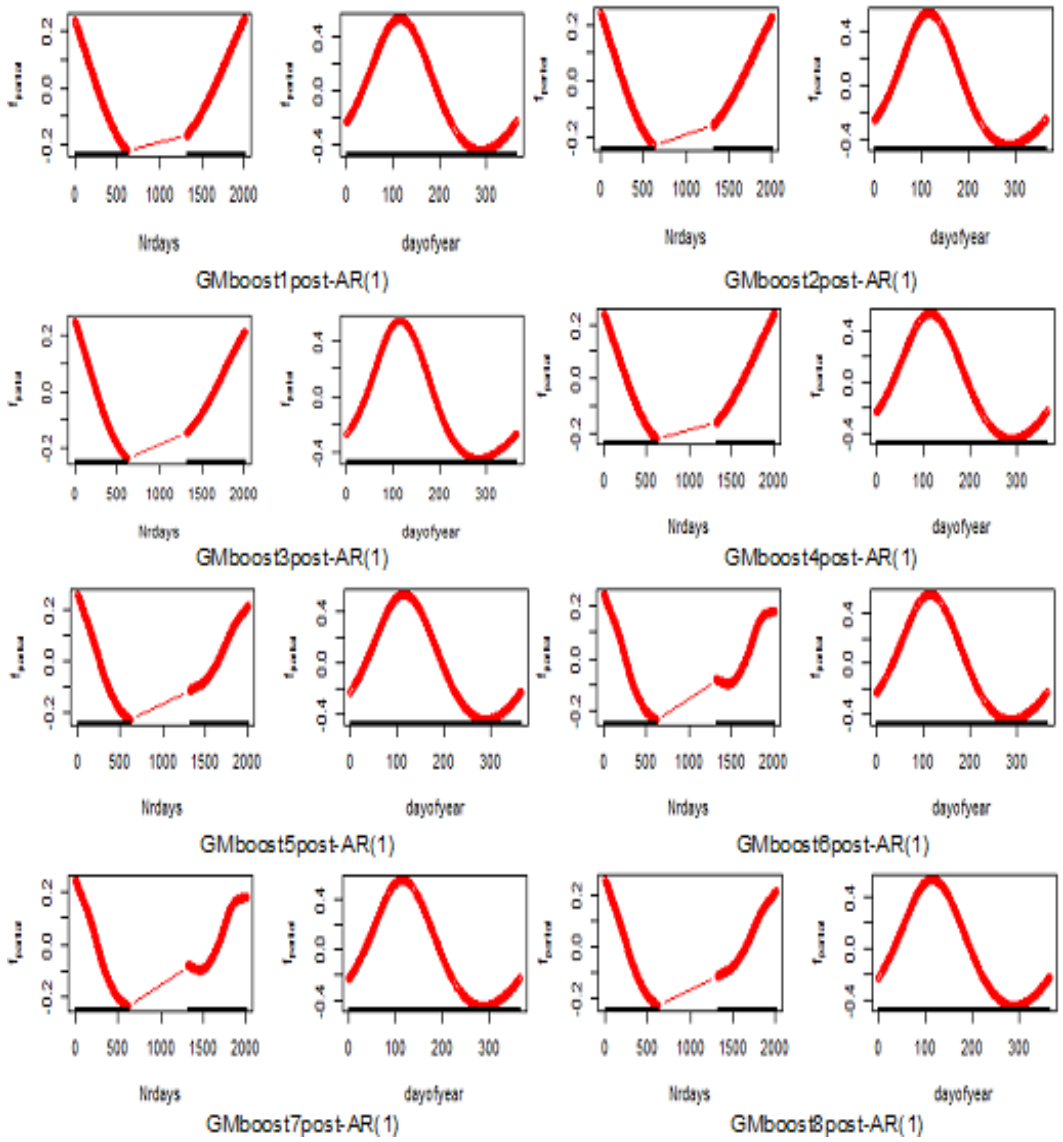


Figure 5.7: *The time-covariates in local fitting for the SST data by gamboost-AR(1) models with transformation of rainfall, to see in detail refer to Tables 5.1 and 5.3.*

Figure 5.7 shows appropriate local fitting on the *Nrdays* and the *Doy* covariates of the GMboost1post-AR(1) to GMboost8post-AR(1) models with transformed rainfall covariate as shown in Table 5.3. The models have a similar pattern with the gamboost-AR(1) models without transformation as displayed in Figure 5.5. It means that time covariates do not change by with or without transformation of rainfall in the gamboost-AR(1) models.

Figure E.8 in Appendix E shows that GMboost1-AR(1) to GMboost4-AR(1) models with transformation of rainfall have similar pattern on global fitting of the SST data. Different

specification of the hyper-parameters on the *Nrdays* and *Doy* covariates does not change pattern of the global fitting, in detail specification see Table 5.1.

We can see patterns of global fitting by using gamboost-AR(1) models without transformation as presented in Figures E.1 and E.2, Appendix E. The patterns are similar with gamboost-AR(1) models with transformation as in Figure E.8, Appendix E. It means that transformation of rainfall does not change appropriate global fitting by using gamboost-AR(1) models.

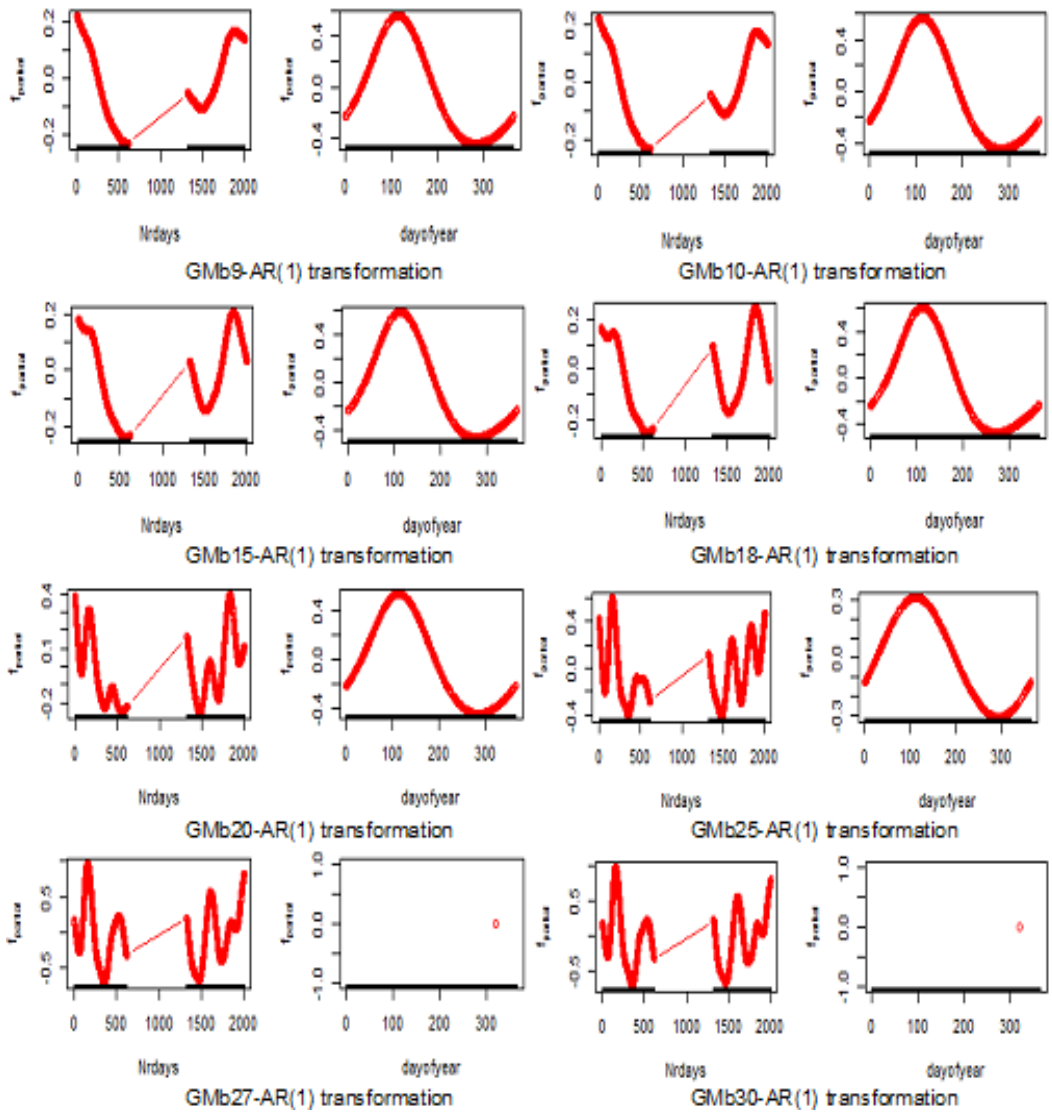


Figure 5.8: *The GMboost9-AR(1) to GMboost30-AR(1) models with transformation in local model fitting for the SST data.*

Figure 5.8 shows inappropriate local fitting on time covariates of gamboost-AR(1) models with transformation. These figures have similar pattern as fitting by using gamboost-AR(1) models without transformation as in Figure 5.6. Nevertheless, inappropriate local fitting is not always followed by gobal model fitting. We presented gamboost-AR(1) models fitting with transformation, where the models have inappropriate in local fitting but it has appropriate in global fitting.

For example, we can see that GMboost3-AR(1) model without transformation has similar pattern on global fitting. However, the model without transformation on local fitting produces 5 linear submodels and 4 smooth submodels as in Figure 5.2. For the same model with transformation on local fitting produces 6 linear submodels and 3 smooth submodels as depicted in Figure 5.2. We can compare the presence of one submodel in GMboost3-AR(1) model fitting as transformation effect, where the rainfall is a polynomial curve in local fitting as captured in Figure 5.9 but it is as smooth term as displayed in Figure 5.2.

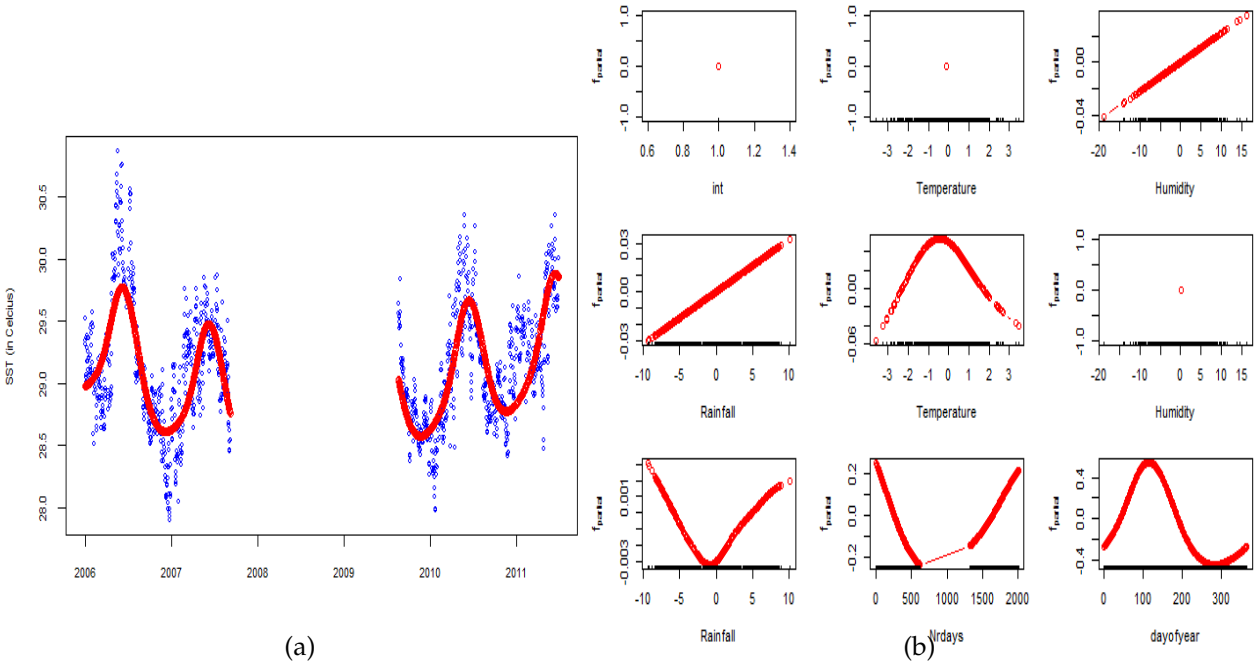
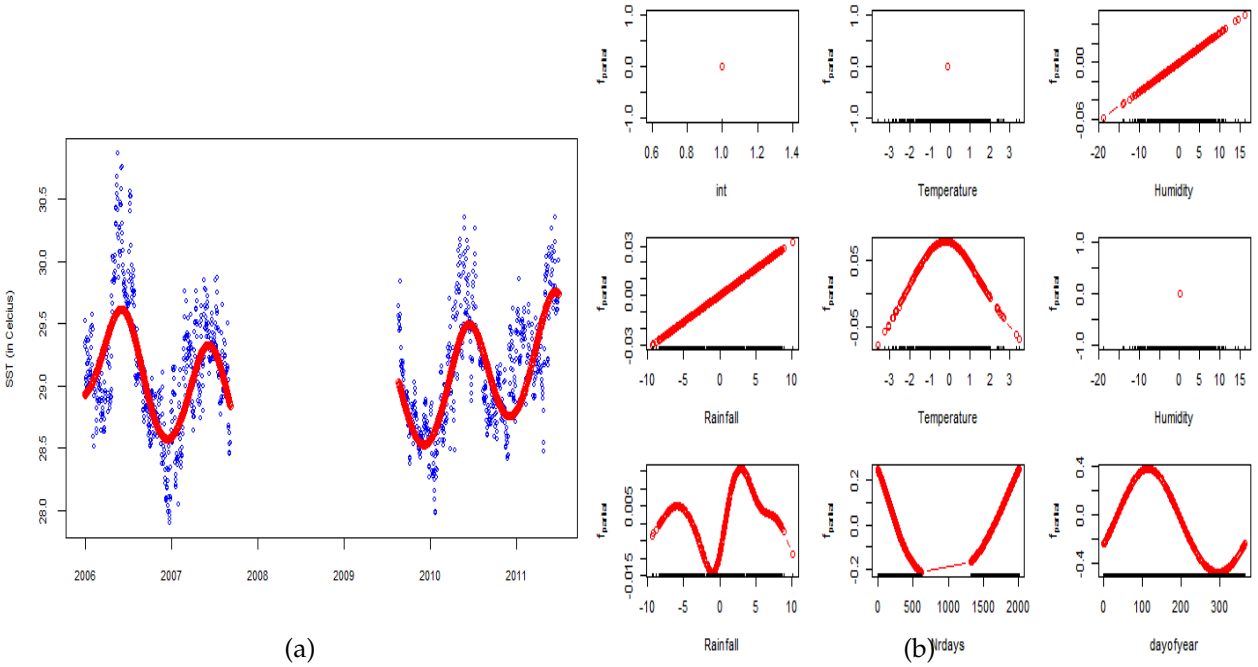
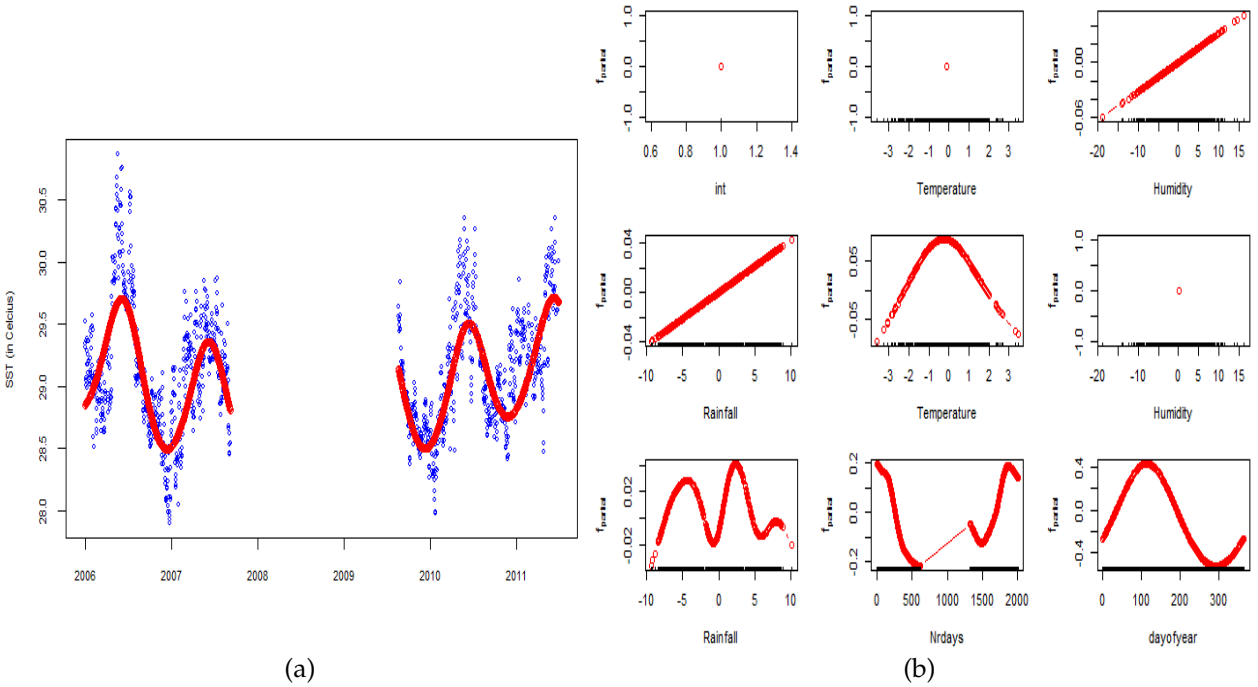


Figure 5.9: *The SST data fitting by GMboost3-AR(1) model with transformation (left) and submodels (right) $m_{stop} = 2000$.*



(a) (b)

Figure 5.10: *The SST data fitting by GMboost-AR(1) model with transformation (left) and submodels (right) $m_{stop} = 12000$.*



(a) (b)

Figure 5.11: *The SST data fitting by GMboost-AR(1) model with transformation (left) and submodels (right) $m_{stop} = 35000$.*

The increase of m_{stop} from 2000 to 12000 and from 1200 to 35000 is not sufficient to reveal linearity pattern of the temperature and humidity covariates as captured in Figures 5.9,

5.10, and 5.11. The different phenomena between gamboost and gamboost-AR(1) models are the stopping iteration tend to extract range of the data values and differencing in AR(1) tend to insert range of the data values. We can see different m_{stop} in gamboost-AR(1) with transformation as displayed in Figures 5.10 and 5.11. Increasing m_{stop} in gamboost-AR(1) with transformation in model fitting is improving the global fitting, whereas on local fitting increasing m_{stop} does change covariate in transformed rainfall and covariate in the gap. The rainfall and *Nrdays* covariates with m_{stop}= 35000 are more fluctuation if we compared to m_{stop}= 12000, the *Nrdays* covariate is more fluctuation after the gap. Trade-off m_{stop} is also required to obtain appropriate model fitting. Therefore, in this case m_{stop}= 12000 is appropriate model fitting (in global and local fitting) better than m_{stop}= 35000 only appropriate on global fitting using gamboost-AR(1) model.

Our experimental results reveal that time covariates, particularly *Doy* covariate, has a significant effect in the model fitting (the details are also given in Chapters 2 and 4). There are several effects when removing autocorrelation in the model fitting by using gamboost-AR(1) models with transformation as follows:

a) Effect transformation of gamboost-AR(1) model fitting is increasing *df* and decreasing AIC, gMDL, and the final risk values as in Table 5.3.

b) Reducing optimal number of boosting iterations, for example, GMb1-AR(1) model with m_{stop} = 1000 to be 993, GMb2-AR(1) and GMb3-AR(1) models with m_{stop} =1500, 2000 to be 1108 in gMDL method.

c) Transformation can accelerate in the fitting process.

d) Transformation can change the number of linear or smooth models in local fitting.

For example, GMb3-AR(1) model with transformation as captured in Figure 5.9 and without transformation as seen in Figure 5.2.

e) The model is also assessed for the transformation effect of rainfall. The results from figures show a similar pattern of the model fitting with and without transformation. However, transformation of rainfall results in a reduced final risk. In addition, the transformation also gives a larger number of submodels in the local model fitting than the non-transformed data for rainfall.

5.4.4 Autocorrelation of the GamboostLSS-AR(1) Models

In our experiment for the SST data fitting by using gamboostLSS-AR(1) models, we consider the degrees of freedom (*df*), *knots*, stopping iteration (m_{stop}) and the step of length factor (ν_{slf}) parameters in the model specification. We use fixed autocorrelation coefficient ρ= 0.8566652 in the gamboostLSS-AR(1) models. The result of the gamboostLSS-AR(1) models have various specification for base-learners and for control boosting used in with and without transformation.

5.4.4.1 Effect of the Degrees of Freedom on GamboostLSS-AR(1) Models

We observed the effect of the degrees of freedom on gamboostLSS-AR(1) models. We chose the degrees of freedom *df* = 2.1 - 2.5 and 2.01 - 2.05, *differences*= 2, and *knots*= 40 of the *Nrdays* and the *df* = 1.1 and 1.5 of the *Doy* covariates specification. The results are displayed as in Figures 5.12, E.4 and E.5, Appendix E.

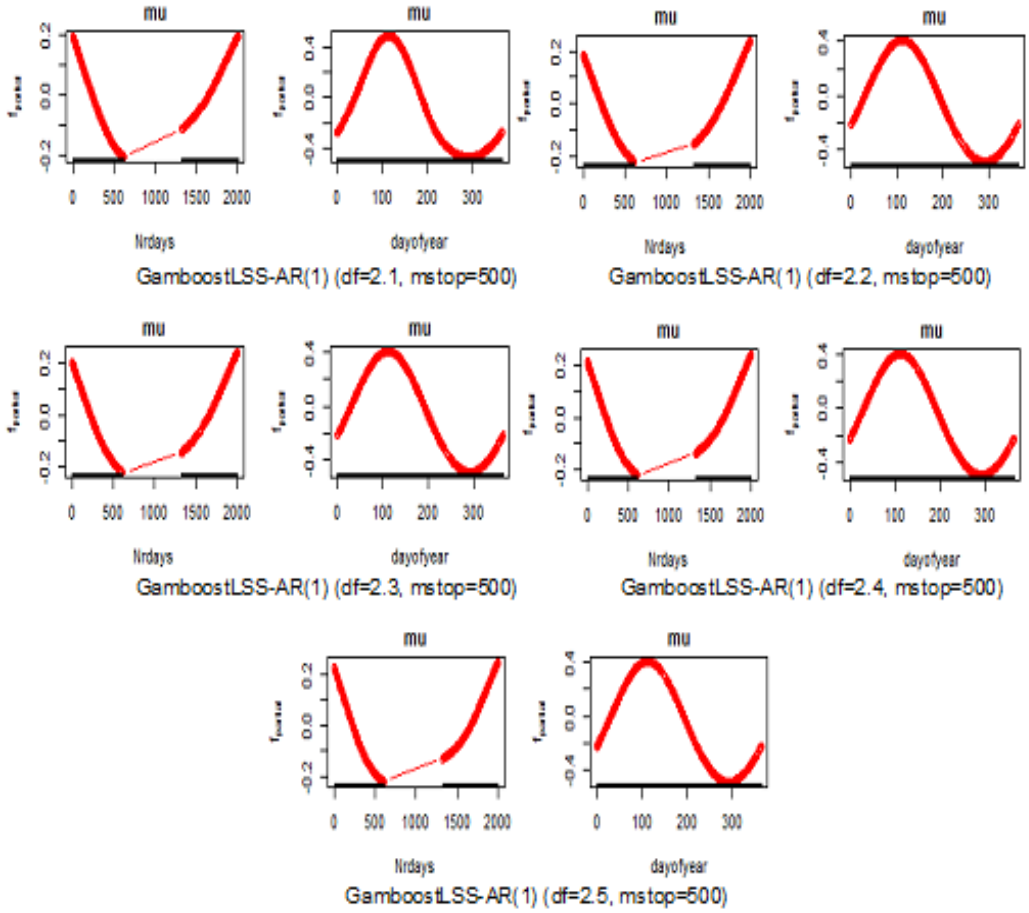


Figure 5.12: *The patterns of time covariates in local fitting using gamboostLSS-AR(1) models. The patterns show a decrease before the gap and an increase after the gap for the Nrdays effect and the same pattern for the Doy effect.*

Figure 5.12 shows local fitting with the same m_{stop}= 500 and different df= 2.1 - 2.5 produces the similar pattern of time covariates. Further we can see local fitting with similar specification m_{stop}= 1000 and different df= 2.1 - 2.5 as depicted in Figure E.4, Appendix E.

We can see from Figure E.4 in Appendix E that gamboostLSS-AR(1) models with stopping iteration m_{stop}= 1000 for time covariates at the degrees of freedom df= 2.4 and 2.5 show inappropriate model fitting. GamboostLSS-AR(1) models with stopping iteration m_{stop}= 1500 for time covariates at the degrees of freedom df= 2.3 to 2.5 show inappropriate model fitting as well as seen in Figure E.5, Appendix E.

Increasing stopping iteration m_{stop} with the different degrees of freedom df tend to change after the gap of the *Nrdays* covariate, whereas the *Doy* covariate shows stable

pattern in gamboostLSS-AR(1) model fitting. Detecting inappropriate time covariate as submodels in local fitting is an important role to avoid misfitting on global fitting of the SST data. Therefore, trade-off degrees of freedom on time covariates are essential role in the SST data fitting by using gamboostLSS-AR(1) models.

5.4.4.2 Effect of the Stopping Iteration on GamboostLSS-AR(1)

Here, we investigated effect of the stopping iteration on gamboostLSS-AR(1) model. We consider the stopping iterations from m_{stop}= 500 to 1500 with step 500 on gamboostLSS-AR(1) models fitting. The results are depicted in Figures E.18, E.19, Appendix E and 5.13.

Figure E.18 in Appendix E shows similar patterns of time covariates with slightly different of *df*'s and the stopping iteration, whereas Figure E.19 in Appendix E shows that gamboostLSS-AR(1) model fitting with fixed m_{stop}= 500 and *df* = 1.5 of the *Doy* and different *df* of the *Nrdays* have similar pattern of time covariates. The *df* = 1.5 of the *Doy* has more impact on smoothing than does the *df* = 1.1, mainly at the *Nrdays* after the gap. In addition to a larger value the *df* tends to nonsmooth the model fitting.

Figure 5.13 shows that gamboostLSS-AR(1) models fitting with fixed m_{stop}= 1000 and *df* = 1.5 of the *Doy* covariate and different *df* of the *Nrdays* covariate have similar pattern of seasonal effects. The annual effects are slightly similar for *df* = 2.1 and 2.2, whereas *df* = 2.3 to 2.5 tends to change after the gap. In general, gamboostLSS-AR(1) model fitting with fixed *df* = 1.1 of the *Doy* and the same class of the stopping iteration m_{stop} of the *Nrdays* have the similar pattern of time covariates. We suggest to use *df* = 2.1 and 2.2 in gamboostLSS-AR(1) model fitting of the SST data.

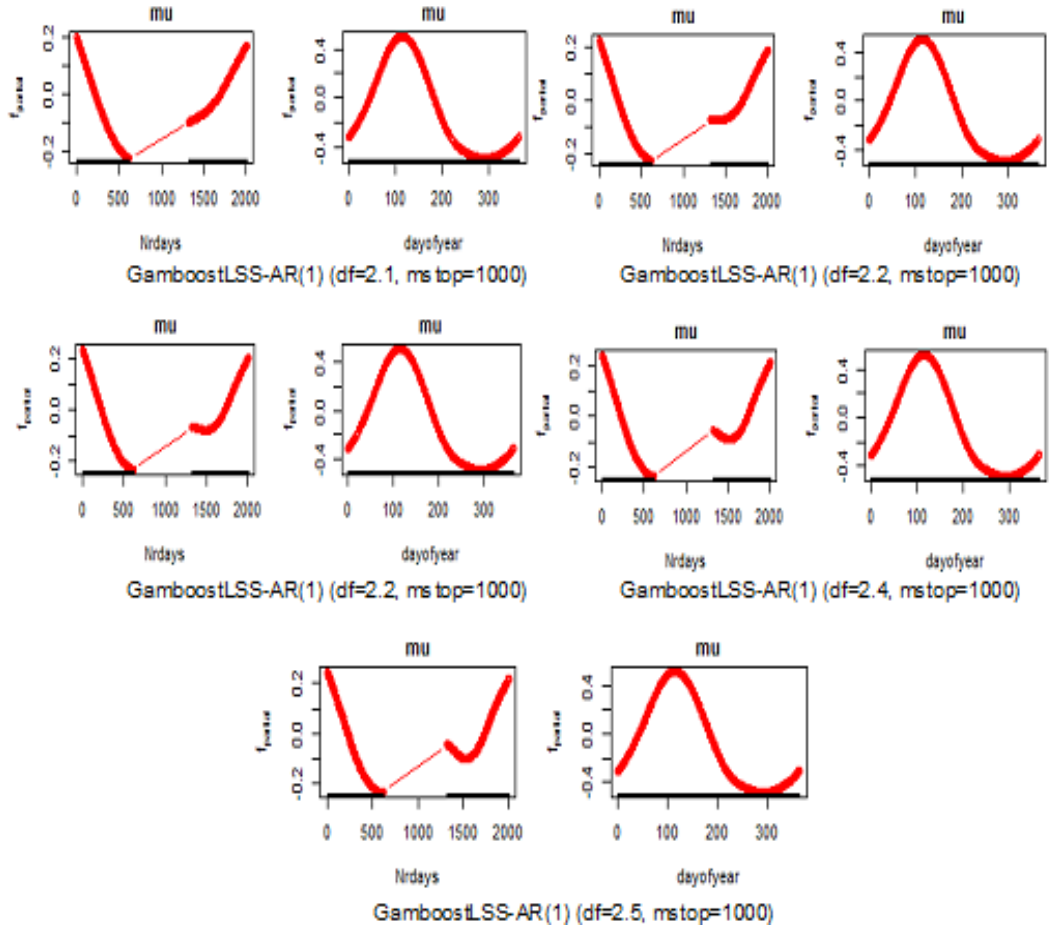


Figure 5.13: *Local fitting of time covariate using gamboostLSS-AR(1) models with different df=2.1-2.5 at the Nrdays covariate and the same mstop=1000.*

5.4.4.3 Effect of the Knots on GamboostLSS-AR(1) Models

We investigated the effect of the *knots* on gamboostLSS-AR(1) models of the SST data fitting. We consider the degrees of freedom df= 1.1 and 1.5 at the *Doy* and different *knots*= 30-60 with each step 10 and df= 2.01, 2.1 at the *Nrdays* covariate. We use the control boosting parameters: m_{stop}= 500 - 1500 and ν_{slf}= 0.1. The results are summarized as in Tables 5.4 and 5.5.

Table 5.4 shows that the effect of *knots* from 30 to 60 with df= 2.1 is more stable than with the df= 2.01, mainly for m_{stop}= 1500, whereas Table 5.5 shows that increasing df= 1.5 at the *Doy* lead to various the number of submodels. Further we observed the local fitting for time covariates with m_{stop}= 1000 at the composition of the df= 2.01 and 1.1; df= 2.01 and

1.5; *df* = 2.1 and 1.1; and the *df* = 2.1 and 1.5 at the *Doy* and *Nrdays* covariates as displayed in Figures 5.14 to E.7, Appendix E.

Table 5.4: *Knots Effects in the GamboostLSS-AR(1) model with df = 1.1 at the Doy for SST data fitting.*

Boosting	*knots*	df=2.1		df=2.01	
m_{stop}		Final Risk	Submodels	Final Risk	Submodels
500	30	307.0386	10	309.4764	10
	40	304.5867	10	306.2531	10
	50	302.2614	10	303.6320	10
	60	300.1213	10	301.3761	10
1000	30	280.0081	13	283.2150	13
	40	276.4000	13	278.8379	13
	50	272.6298	13	274.8928	13
	60	268.8476	13	271.0861	13
1500	30	262.8653	13	266.3752	13
	40	258.2544	13	261.4013	14
	50	253.3687	13	256.1880	13
	60	248.5729	13	251.3355	13

Table 5.5: *Knots Effects in the GamboostLSS-AR(1) model with df= 1.5 at the Doy for SST data fitting.*

Boosting	*knots*	df=2.1		df=2.01	
m_{stop}		Final Risk	Submodels	Final Risk	Submodels
500	30	255.3522	10	257.4524	10
	40	252.7441	10	254.5191	10
	50	250.4077	10	251.7720	10
	60	248.1843	11	249.4940	10
1000	30	230.3242	12	232.8218	12
	40	227.1549	12	229.2931	12
	50	223.6701	12	225.7913	13
	60	219.7801	13	222.1147	12
1500	30	217.8403	14	280.8349	14
	40	213.1461	14	216.3165	14
	50	208.0611	14	211.0960	14
	60	202.8360	14	205.9028	14

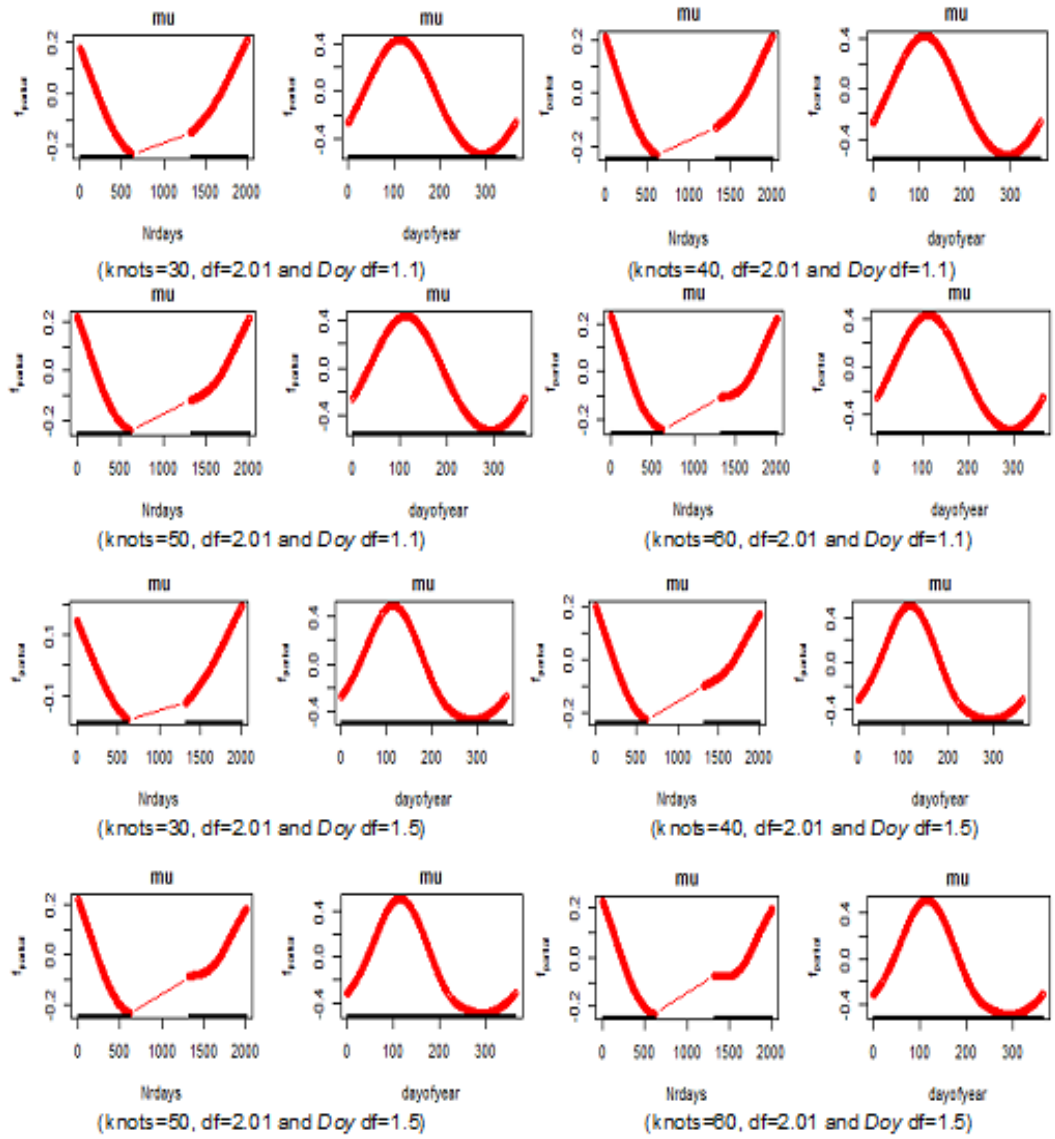


Figure 5.14: *The patterns of time covariate in gamboostLSS-AR(1) models fitting with fixed df= 2.01 and different knots=30-60 of the Nrdays covariate and fixed df= 1.1 and 1.5 at the Doy covariate.*

We can see that fitting at the time covariates with $df = 1.1$ is smoother than using $df = 1.5$ at the *Doy* as captured in Figure 5.14, mainly for the *Nrdays* covariate after the gap. If we compare fitting at the time covariates with $df = 1.1$, as seen in Figure E.6 Appendix E, it is smoother than using $df = 1.5$ at the *Doy* covariate as captured in Figure E.7 Appendix E, mainly for the *Nrdays* covariate after the gap. We can select degrees of freedom $df = 1.1$ instead of $df = 1.5$ at the *Doy* covariate with $df = 2.01$ or 2.1 at the *Nrdays* in the gamboostLSS-AR(1) model fitting for the SST data.

5.4.4.4 Effect of the Size-Length Factor on GamboostLSS-AR(1) Models

We observed the size-length factor ν_{slf}= 0.1-0.4 with step 0.1, and 0.01-0.05 each step 0.01 in gamboostLSS-AR(1) model fitting. The results of observation are displayed in Table 5.6.

Table 5.6: *The Size-Length Factor Effects in the GamboostLSS-AR(1) model with df= 1.1 at the Doy.*

Size-Length Factor	Final Risk	Submodels
0.1	278.8379	13
0.2	247.9946	13
0.3	227.4881	14
0.4	358.2745	10
0.01	521.3064	4
0.02	377.1072	6
0.03	333.2335	9
0.04	316.1602	10
0.05	306.4167	10

The similar patterns of global fitting using gamboostLSS-AR(1) model can be achieved with df= 1.1 at the *Doy* covariate and (ν_{slf})=0.1-0.4 as seen in Table 5.6. However, the size of length factor (ν_{slf}) in the gamboostLSS-AR(1) model gives impact fitting of smoothness, which the larger values of the ν_{slf} tends to nonsmooth global fitting. Nevertheless, the size of length factor ν_{slf}= 0.1 and 0.2 produces the same number of submodels, i.e. 13 on the local fitting. The largest number of submodels (i.e. 14) with ν_{slf}= 0.3 gives lowest final risk and then smallest number of submodels (i.e. 10) with ν_{slf} = 0.4 produces highest final risk for class ν_{slf} = 0.1-0.4.

Further we observed the size of length factor effect (ν_{slf})= 0.01-0.05 with df= 1.1 at the *Doy* covariate in global fitting as displayed in Figure 5.15.

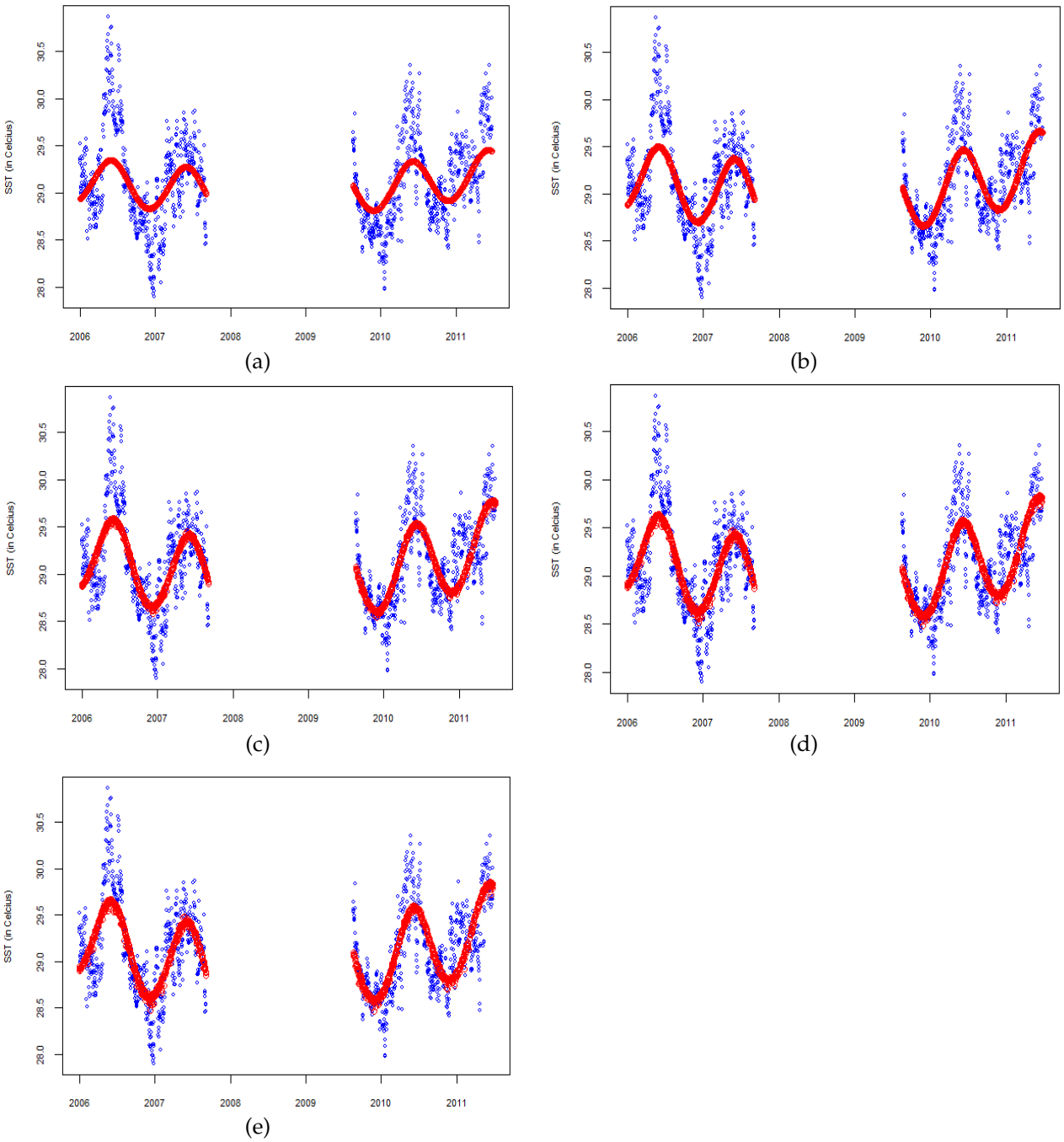


Figure 5.15: *The SST data fitting by gamboostLSS-AR(1) models with the same m_{stop} = 1000 and different ν_{slf} = 0.01 to 0.05 for (a)-(e) respectively.*

The ν_{slf} values increase from ν_{slf}= 0.01 to 0.05 lead to the growth of fitting process, as depicted in Figure 5.15. Furthermore, gamboostLSS-AR(1) model fitting with ν_{slf}= 0.01 produces smallest submodels and it is not optimal in local fitting. The ν_{slf}= 0.02 value in gamboostLSS-AR(1) model fitting gives six submodels and the ν_{slf}= 0.03 value gives

a larger number of submodels. However, this model shows slightly nonsmooth global fitting, whereas the v_{slf}= 0.04 and 0.05 values produce the same number of submodels, i.e. 10 in gamboostLSS-AR(1) model fitting.

As the class v_{slf}= 0.01 to 0.05 give impacts in global and local fitting, then we recommend to select v_{slf}= 0.01 and 0.02 with using higher m_{stop} in the gamboostLSS-AR(1) models for SST data fitting. For example, we applied m_{stop}= 2000-3000 with v_{slf}= 0.01 and 0.02 to fit SST data by using gamboostLSS-AR(1) model as depicted in Figure 5.16.

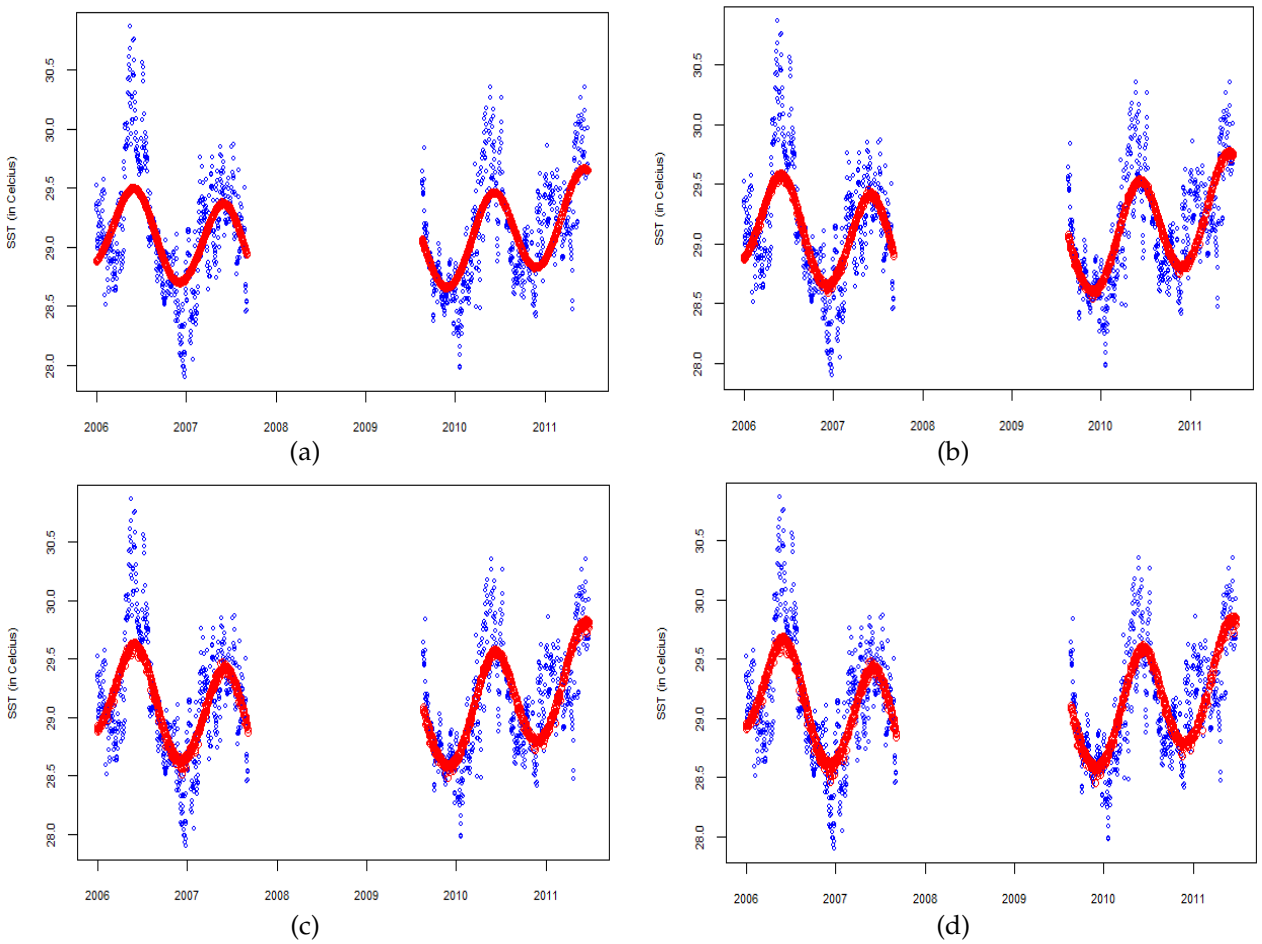


Figure 5.16: *The SST data fitting by gamboostLSS-AR(1) models with different m_{stop} and v_{slf} with the models as follows:(a) 2000, v_{slf}=0.01, (b) 3000, v_{slf}=0.01, (c) 2000, v_{slf}=0.02, and (d) 3000, v_{slf}=0.02.*

The increase of m_{stop} values in model fitting gives a larger number of submodels on local fitting and improves on global fitting as depicted in Figure 5.16. Figure 5.16 (a) shows

global fitting with v_{slf} = 0.01 produces 6 submodels and (b) global fitting with 9 submodels. In the same figure, graph (c) with v_{slf} = 0.02 produces 10 submodels and graph (d) with v_{slf} = 0.03 produces 12 submodels. Increasing m_{stop} values in the gamboostLSS model also gives impact to global and local model fitting, by comparing the results in Figures 5.15 and 5.16. However, trade-off the m_{stop} values are needed to avoid over-fitting or under-fitting on global and local fitting.

5.4.5 GamboostLSS-AR(1) Models with Transformation

We observed the gamboostLSS-AR(1) models with transformation of rainfall covariate. We use specification of the *Nrdays* covariate: *knots*= 20-70, *diff*= 2, *df*= 2.01-2.2 and m_{stop}= 1000-3000 with ρ= 0.8566652. The results of our experiment are reported in Table 5.8. We applied the models setup in a total of 30 models in which the GMbLSS13tr-AR(1) to GMbLSS30tr-AR(1) are not given in the table.

Table 5.7: *Specification of GamboostLSS-AR(1) models with transformation.*

Model	df_{Nrdays}	$knots_{Nrdays}$	df_{Doy}	m_{stop}
GMbLSS1post-AR(1)	2.1	40	1.47	1000
GMbLSS2post-AR(1)	2.1	70	1.20	1500
GMbLSS3post-AR(1)	2.1	60	1.15	2000
GMbLSS4post-AR(1)	2.1	60	1.40	1000
GMbLSS5post-AR(1)	2.2	50	1.40	1000
GMbLSS6post-AR(1)	2.2	60	1.18	1500
GMbLSS7post-AR(1)	2.2	65	1.35	1000
GMbLSS8post-AR(1)	2.2	30	1.20	1500
GMbLSS9post-AR(1)	2.2	40	1.12	2000
GMbLSS10post-AR(1)	2.2	70	1.11	2000
GMbLSS11post-AR(1)	2.1	20	1.11	2500
GMbLSS12post-AR(1)	2.01	30	1.08	3000

Table 5.8 shows that GMbLSS1post-AR(1) to GMbLSS12post-AR(1) models have a similar pattern on global fitting and the same number of submodels. The following pattern of

Table 5.8: *GamboostLSS-AR(1) models fitting using P-spline with transformed rainfall.*

Model	Submodel	Final Risk	Model	Submodel	Final Risk
GMbLSS1post-AR(1)	8	323.5800	GMbLSS7post-AR(1)	8	326.4281
GMbLSS2post-AR(1)	8	331.5986	GMbLSS8post-AR(1)	8	338.6751
GMbLSS3post-AR(1)	8	330.2864	GMbLSS9post-AR(1)	8	349.0427
GMbLSS4post-AR(1)	8	322.8192	GMbLSS10post-AR(1)	8	351.8073
GMbLSS5post-AR(1)	8	323.3653	GMbLSS11post-AR(1)	8	359.8472
GMbLSS6post-AR(1)	8	338.7529	GMbLSS12post-AR(1)	8	356.8358

time covariates of these models can be seen in Figure E.9, Appendix E. The pattern of time covariates in gamboostLSS-AR(1) models with transformation is more stable for the *Doy* covariate and is slightly changed for the *Nrdays* covariate after the gap.

Transformation effects on global fitting by gamboostLSS-AR(1) models is shown as depicted in Figures E.16 and E.17, Appendix E. We suggest using the ν_{slf}= 0.01 or 0.02 for gamboostLSS-AR(1) models with transformation. Improving the model fitting by transformation of rainfall gives significant effect in the gamboostLSS-AR(1) models. We can see the similar patterns of the model in local fitting and global fitting as displayed in Figures E.9, E.16, and E.17, in Appendix E.

5.4.5.1 Effect of the Degrees of Freedom on GamboostLSS-AR(1) Models with Transformation

We observed the effect of the degrees of freedom on gamboostLSS-AR(1) models with transformation. We choose the degrees of freedom *df*= 2.1 - 2.5 and 2.01 - 2.05, *differences*= 2, and *knots*= 40 of the *Nrdays* and the *df*= 1.1 and 1.5 of the *Doy* covariates specification. The result of this observation is displayed as in Figures E.10, E.11, and E.12, Appendix E.

We can see increase stopping iteration from m_{stop}= 500 to 1500 and transformation do not change patterns of the *Nrdays* and *Doy* covariates drastically in the μ and σ parameters.

A slight changes pattern of the *Nrdays* covariate with *df*= 2.1 and the *Doy* covariate with increase *df* from 1.2 to 1.5 and m_{stop}= 1500 after the gap in the μ parameter, and also, in the beginning fitting for the *Doy* covariate with *df* from 1.2 to 1.5 in the σ parameter. Generally the effect of degrees of freedom on gamboostLSS-AR(1) models with transformation occurs at the *Doy* covariate with increase *df* from 1.2 to 1.5 in σ parameter for m_{stop}= 500-1500.

5.4.5.2 Effect of the Stopping Iteration on GamboostLSS-AR(1) Models with Transformation

We observed the effect of the stopping iteration on gamboostLSS-AR(1) models with transformation. We choose the degrees of freedom *df*= 2.1 - 2.5 and 2.01 - 2.05, *differences*= 2, and *knots*= 40 of the *Nrdays* and the *df*= 1.1 and 1.5 of the *Doy* covariates specification. The result of this observation is displayed as in Figures E.13 to E.15, Appendix E.

Figure E.13 in Appendix reveals that the pattern of time effects in the μ and σ parameters is almost the same, except after the gap of the *Nrdays* effect with m_{stop}= 500. Similarly, Figure E.14 and E.15 in Appendix E show almost the same patterns of time effects, except after the gap of the *Nrdays* effect with m_{stop}= 1000 and 1500. We can see that the effect of increase stopping iteration from m_{stop}= 500 to 1500 and transformation do not change patterns of the *Nrdays* and *Doy* covariates drastically in the μ and σ parameters. However, a slight changes pattern of the *Nrdays* covariate with *df*= 2.2-2.5 and m_{stop}= 1500 after the gap in the μ parameter.

Generally the effect of stopping iteration on gamboostLSS-AR(1) models fitting with transformation of 1231 SST data occurs at the *Nrdays* covariate with increase *df* from 2.2 to 2.5 in μ parameter for m_{stop}= 1500.

5.4.5.3 Effect of the Knots on GamboostLSS-AR(1) Models with Transformation

We investigated the effect of the *knots* on gamboostLSS-AR(1) models with transformation. The degree of freedom df= 1.1 at the *Doy* and different *knots*= 30-60 with each step 10 and df= 2.01, 2.1 at the *Nrdays* covariate is considered. We use the boosting parameters: m_{stop}= 1500-3000 and ν_{slf}= 0.01. The results of our experiment are summarized as in Table 5.9.

Table 5.9: *Knots Effects in the GamboostLSS-AR(1) model with transformation and df=1.1 at the Doy.*

Boosting	m_{stop}	df=2.1		df=2.01	
knots		Final Risk	Submodels	Final Risk	Submodels
30	1500	422.0076	4	444.5972	4
	2000	372.0361	7	389.0417	**6**
	2500	345.3396	8	357.3865	8
	3000	329.5268	9	338.1158	9
40	1500	431.8036	4	428.5862	4
	2000	378.9423	7	376.6308	7
	2500	349.9913	8	348.3310	8
	3000	332.7192	9	331.5514	9
50	1500	415.1589	4	419.1801	4
	2000	367.6151	7	370.1872	7
	2500	342.5344	8	344.1315	8
	3000	327.6831	9	328.7482	9
60	1500	410.3917	**5**	413.1963	4
	2000	364.7553	7	366.4300	7
	2500	340.8095	8	341.7845	8
	3000	326.4176	9	327.1607	9

Table 5.9 shows that increasing *knots* with slight different degrees of freedom df does not change the number of submodels in the gamboostLSS-AR(1) models fitting with transformation. Interestingly, the number of submodels 4 at the m_{stop}= 1500 show the patterns of time covariates in the μ and σ parameters. For *knots*= 60 with stopping iteration m_{stop}= 1500 reveals that the patterns for a linear submodel of the rainfall covariate and four submodels of time covariates, whereas *knots*= 30 with stopping iteration m_{stop}= 2000 shows that the patterns for the linear submodels of the rainfall and humidity covariates and four submodels of time covariates.

5.4.5.4 Effect of the Size-Length Factor on GamboostLSS-AR(1) with Transformation

We investigated the size-length factor v_{slf}= 0.01-0.05 each step 0.01, m_{stop}= 1500 in the gamboostLSS-AR(1) model fitting with transformation. The visualization gamboostLSS-AR(1) models with different v_{slf} is displayed in Figures E.20, Appendix E and 5.17.

Figure E.20 in Appendix E shows almost the same pattern of time covariates in the μ and σ parameters. Each pattern has specification as follows: the *Nrdays* and *Doy* covariates with df= 2.1, df= 1.3, v= 0.01 gives 8 submodels; df= 2.01, df= 1.3, v= 0.01 produces 8 submodels; df= 2.1, df= 1.1, v= 0.02 gives 9 submodels; and df= 2.01, df= 1.1, v= 0.02 produces 9 submodels. The effect of increasing size-length factor from 0.01 to 0.02 with the same class specification gives the same number of submodels.

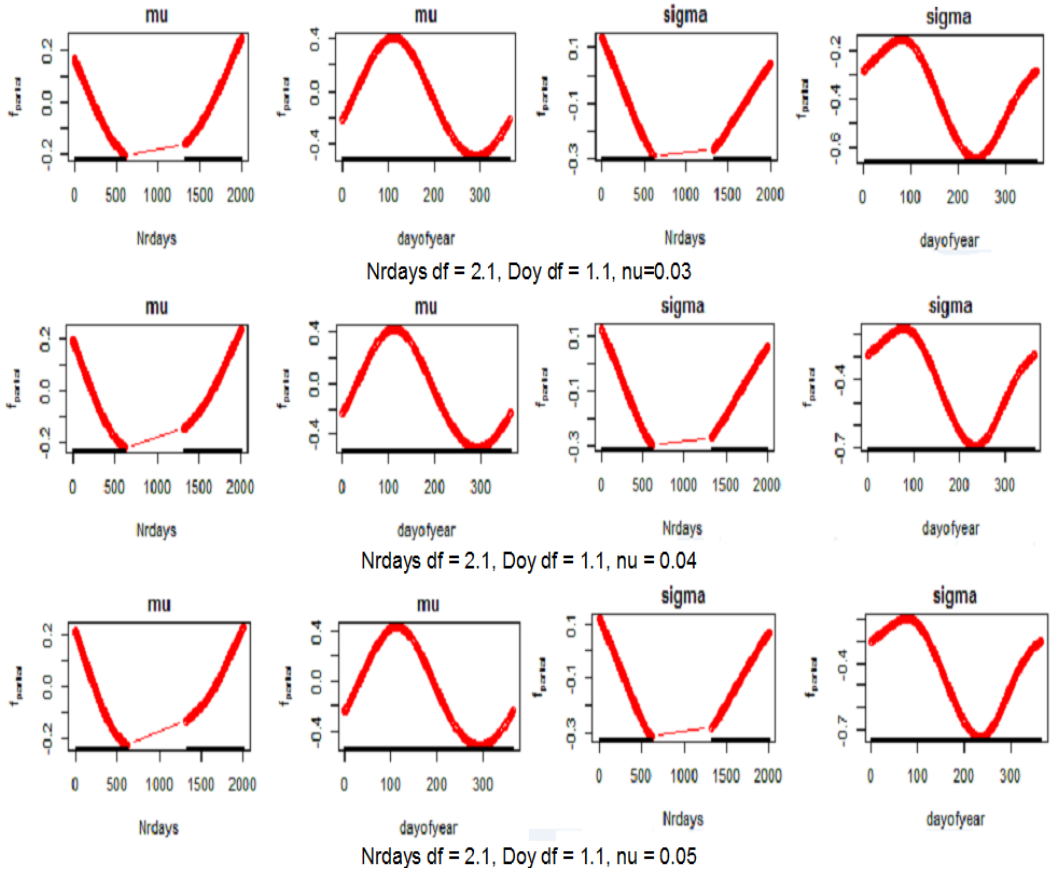


Figure 5.17: *Local fitting using gamboostLSS-AR(1) models with transformation of time covariates, where the time shows almost the same pattern for the Nrdays and Doy effects.*

Figure 5.17 shows that almost the same pattern of time covariates in the μ and σ parameters where each pattern has specification as follows: the *Nrdays* and *Doy* covariates with $df = 2.1$, $df = 1.1$, $\nu = 0.03$ gives 10 submodels; $df = 2.1$, $df = 1.1$, $\nu = 0.04$ produces 11 submodels; and $df = 2.1$, $df = 1.1$, $\nu = 0.05$ gives 12 submodels. The effect of increasing size-length factor from 0.03 to 0.05 with the same specification tends to increase the number of submodels.

Furthermore, we present the best model fitting of appropriate gamboostLSS, gamboostLSS -AR(1), and gamboostLSS-AR(1) models with transformation of the SST data. There are appropriate gamboostLSS model fitting class with considering autocorrelation as follows:

1) Appropriate model in global fitting but inappropriate model in local fitting.

2) Appropriate model in global fitting but low the number of submodels in appropriate local fitting.

3) Appropriate model in global fitting but not optimal number of submodels in appropriate local fitting.

4) Appropriate model in global fitting and appropriate in local fitting.

5) Appropriate model in global fitting and optimal number of submodels in appropriate local fitting.

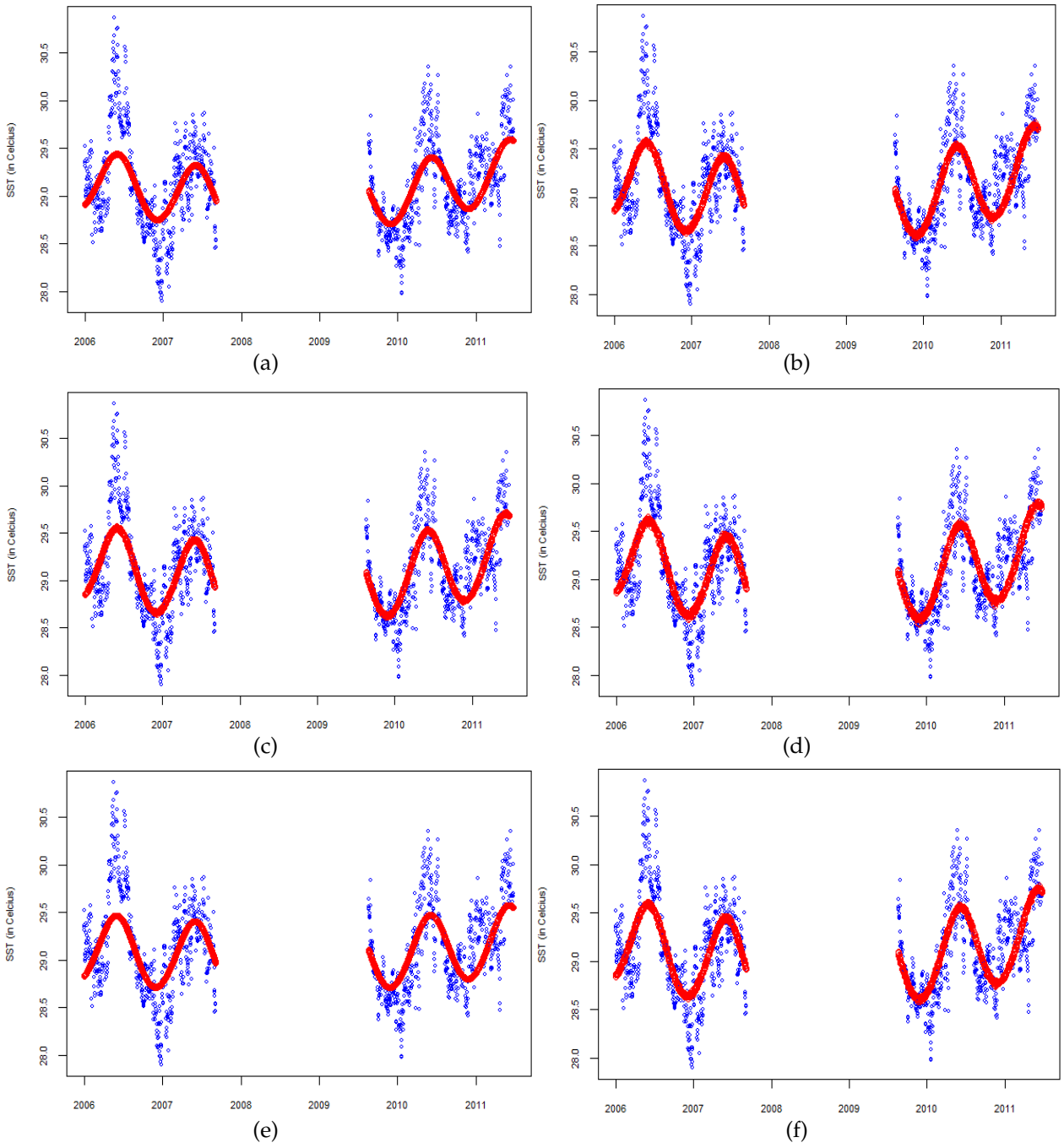


Figure 5.18: *The similar patterns in global fitting by appropriate gamboostLSS-AR(1) models for the Nrdays covariate with df=2.1 fixed, and for thr Doy covariate with (a). df=1.1, m_{stop}=1500; (b). df=1.2, m_{stop}=1500;(c). df=1.3, m_{stop}=1000;(d). df=1.3, m_{stop}=1500;(e). df=1.4, m_{stop}=500; and (f). df=1.4, m_{stop}=1000.*

Figure 5.18 shows that appropriate gamboostLSS-AR(1) models in global fitting have low the number of submodels in appropriate local fitting as captured in graphics (a) and (e) with 4 submodels, whereas no optimal number of submodels in appropriate local fitting

as depicted in graphics (b), (c), (d) and (f) with 7, 6, 8, and 6 submodels respectively.

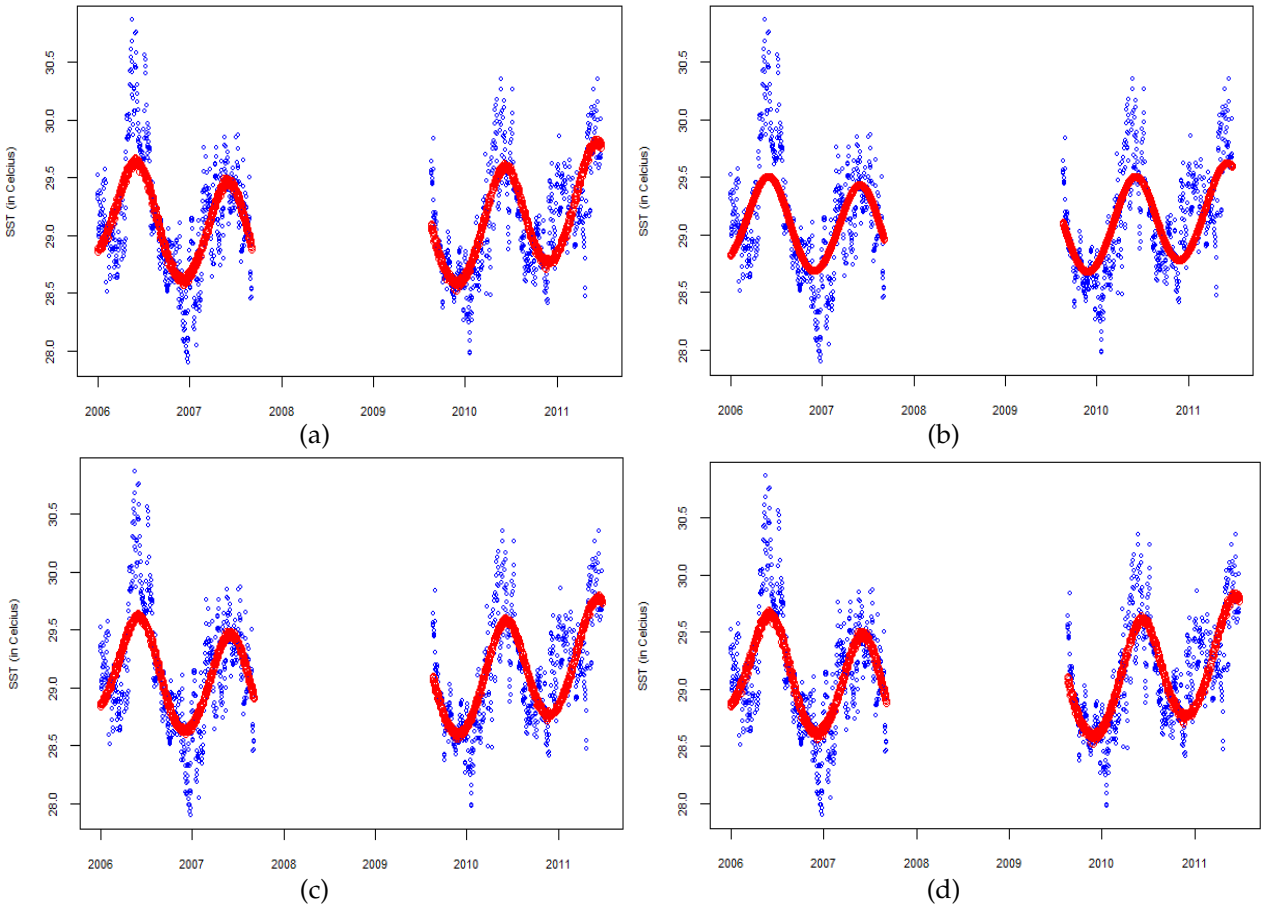


Figure 5.19: *The similar patterns in global fitting by appropriate gamboostLSS-AR(1) models for the Nrdays covariate with df=2.1 fixed, and for the Doy covariate with (a). df=1.4, m_{stop}=1500; (b). df=1.5, m_{stop}=500;(c). df=1.5, m_{stop}=1000; and (d). df=1.5, m_{stop}=1500.*

Figure 5.19 shows that appropriate gamboostLSS-AR(1) models in global fitting have low the number of submodels in appropriate local fitting as displayed in graphics (a) with 4 submodels, whereas no optimal number of submodels in appropriate local fitting as seen in graphics (b) and (c), both with 8 submodels. Further we use appropriate model in global fitting and appropriate in local fitting class to visualize the best gamboostLSS model fitting of the SST data as captured in Figure 5.20 with submodels as displayed in Figure 5.21.

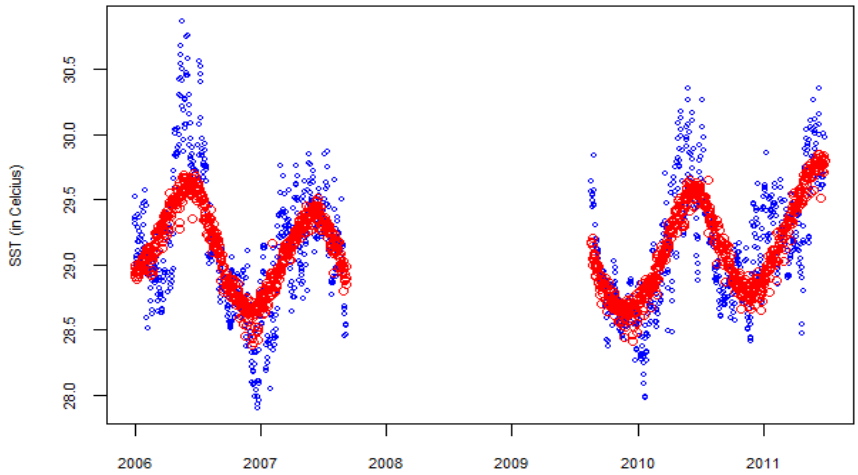


Figure 5.20: *An illustration of 1231 SST data fitting by using gamboostLSS model without transformation, with boosting parameters m_{stop}= 300, ν= 0.1.*

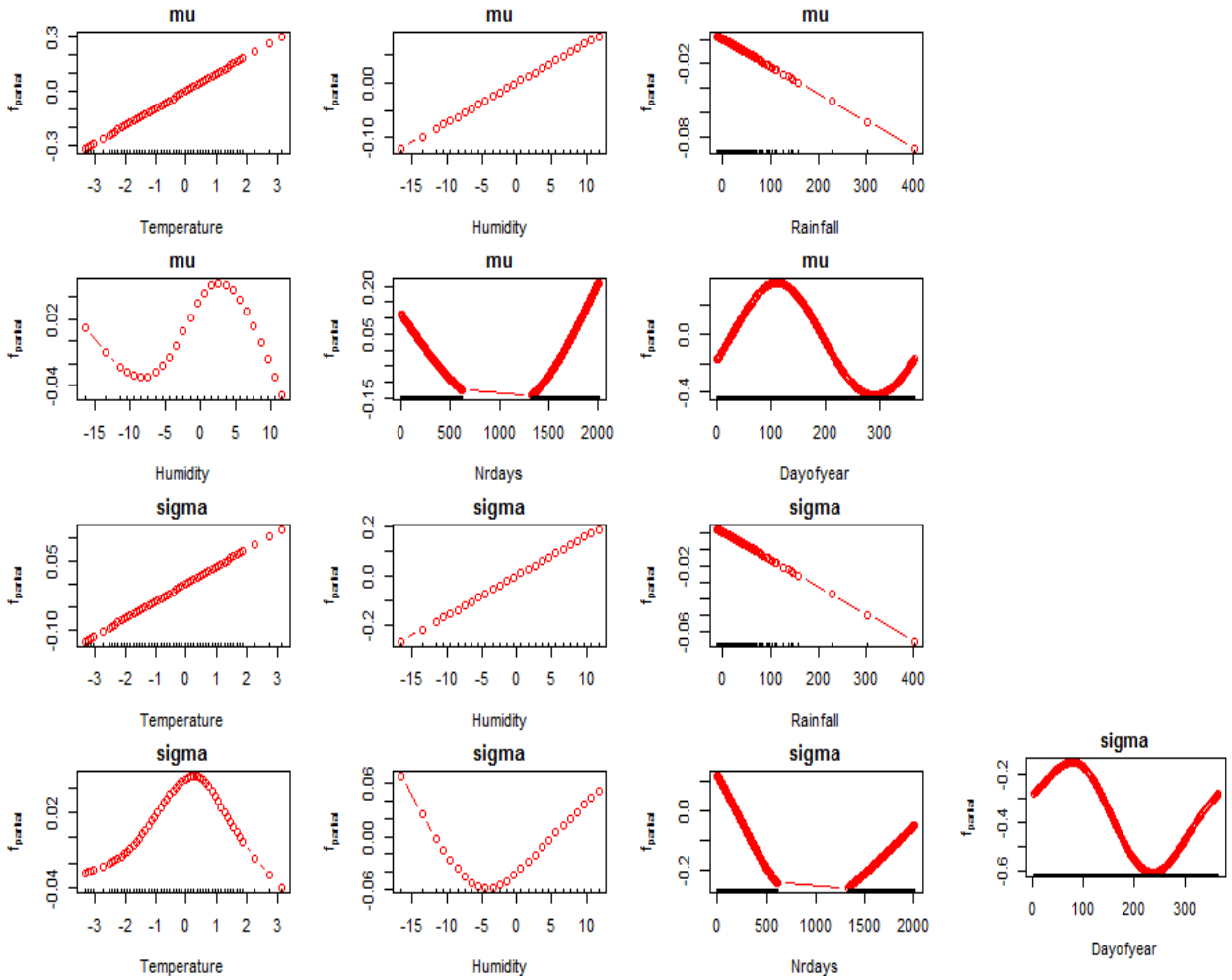


Figure 5.21: *Local fitting of the gamboostLSS model fitting for 1231 SST data produces 13 submodels. It is shown that temperature and humidity have similar trends in μ and σ parameters, but show opposite trends with rainfall in both parameters. Humidity has a polynomial curve in the μ and σ parameters, whereas temperature has a downward curve in the σ parameter. The Nrdays covariate have similar trends before and after the gap in both parameters, whereas the Doy covariate has a sinusoidal curve in both parameters as well.*

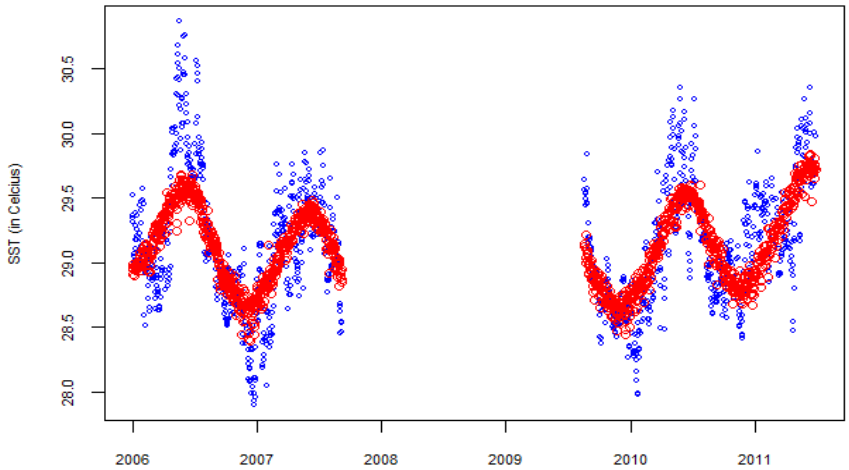


Figure 5.22: *GamboostLSS model fitting with transformation for 1231 SST data, m_{stop}= 250, ν= 0.1.*

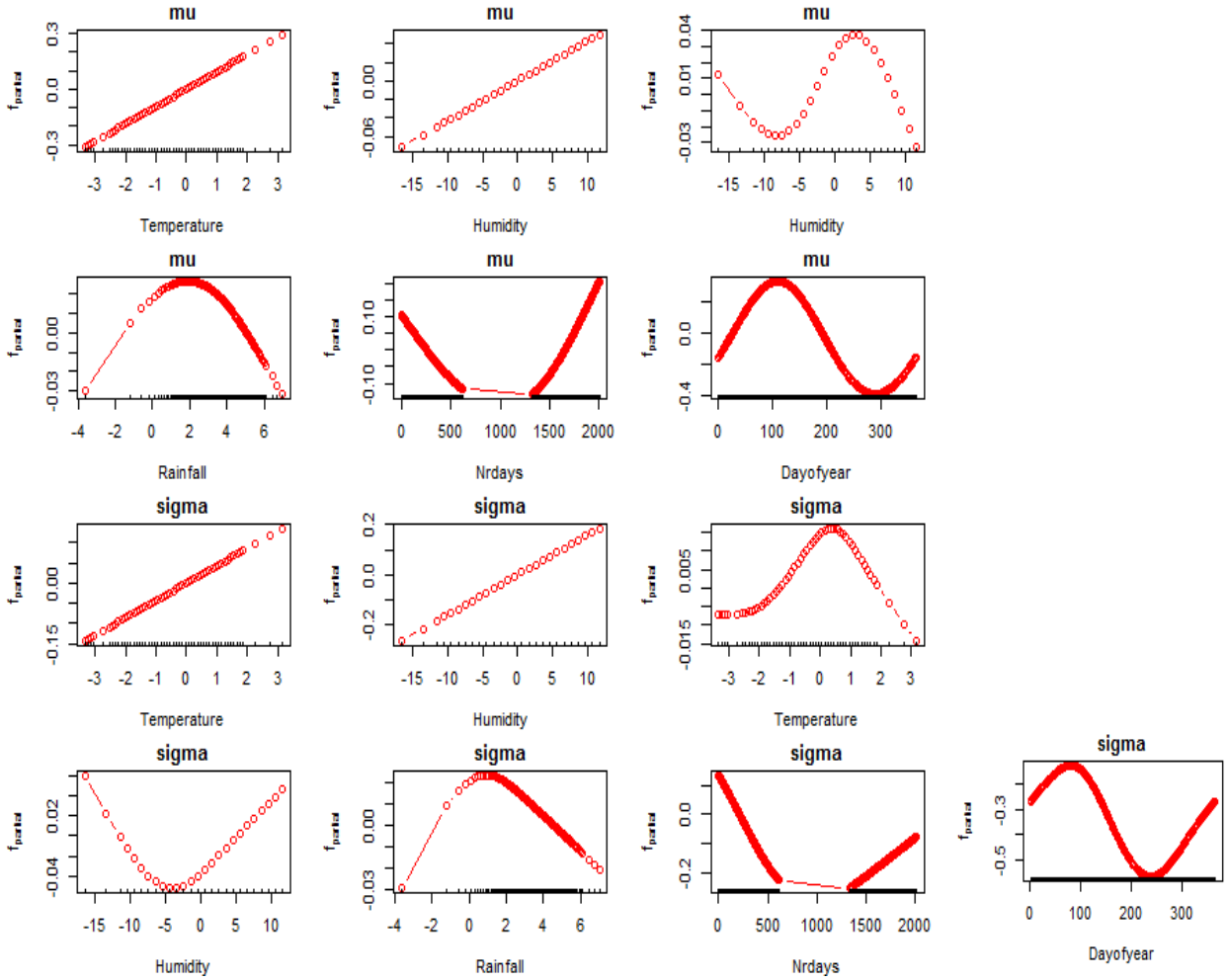


Figure 5.23: *Local fitting of the gamboostLSS model fitting with transformation for 1231 SST data gives 13 submodels. The submodels show the similar patterns and trends for all covariates, excluding rainfall if we compared with local fitting without transformation.*

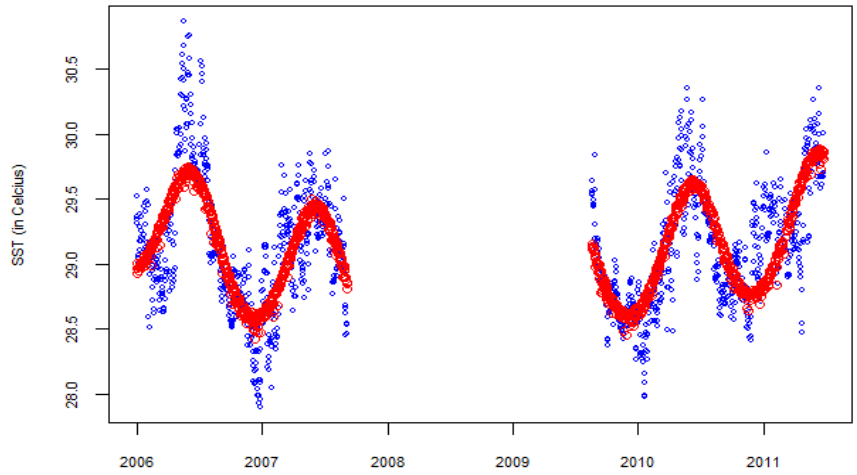


Figure 5.24: *The SST data fitting using gamboostLSS-AR(1) model without transformation, with* $m_{stop} = 1000$.

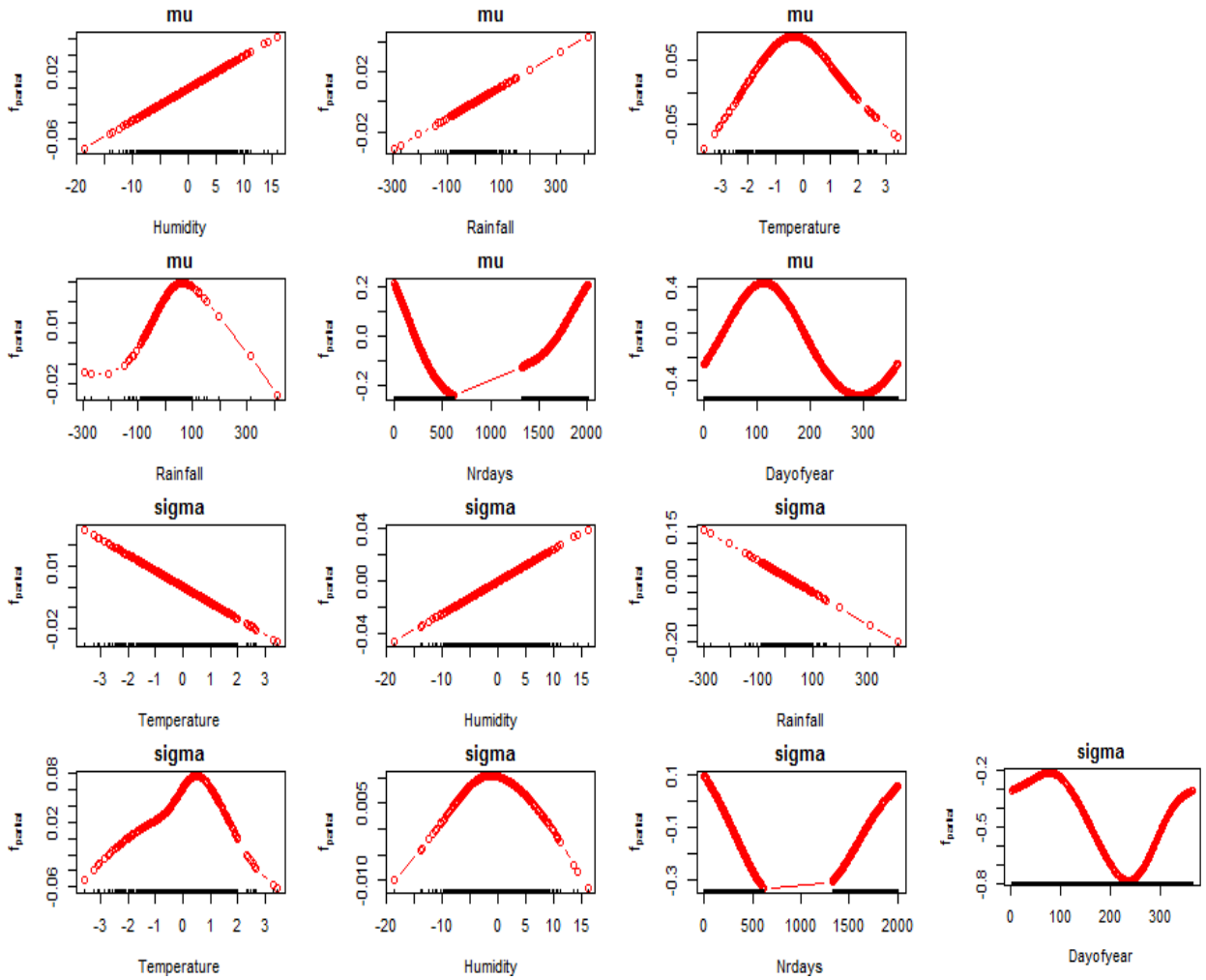


Figure 5.25: *Local fitting in the gamboostLSS-AR(1) model without transformation for 1231 SST data gives 13 submodels. Autocorrelation effect does not change patterns and trends of time covariates, Nrdays and Doy in* μ *and* σ *parameters.*

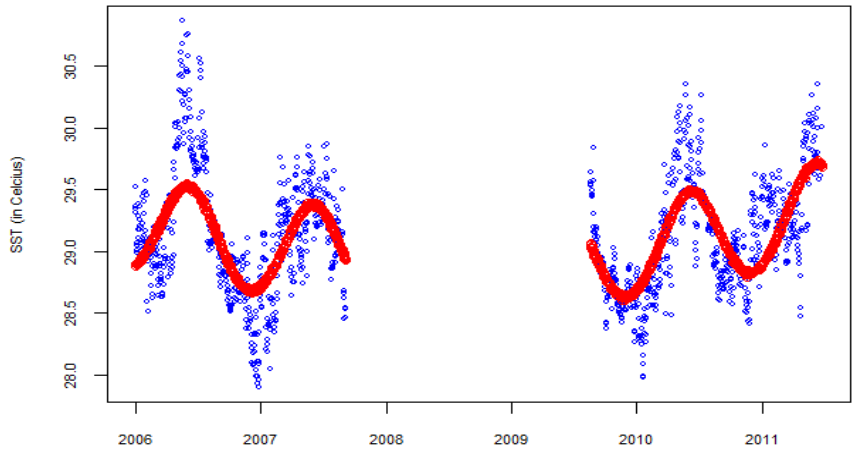


Figure 5.26: *The gamboostLSS-AR(1) model fitting with transformation of 1231 SST data, m_{stop}=2250.*

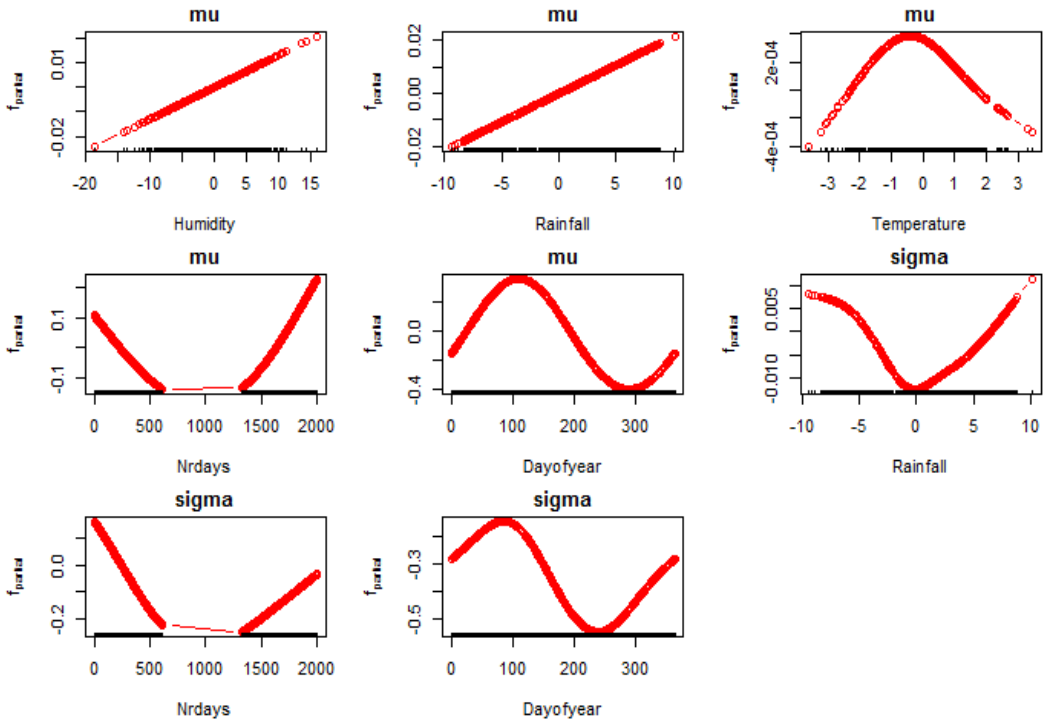


Figure 5.27: *Local fitting of 1231 SST data using the gamboostLSS-AR(1) model fitting with transformation produces 8 submodels. Autocorrelation and transformation effects do not change patterns and trends of time covariates in both parameters, but it has large effects in global and local fitting, such as the best smoothing on global fitting can be achieved.*

There are advantages of the gamboostLSS-AR(1) model fitting with transformation as follows:

a) It provides a more stable in fitting process in global and local fitting.

b) It is more robust to fit SST data.

c) It gives improved fitting in the SST data.

The gamboostLSS-AR(1) models have smoothing terms by P-spline and filtering terms by the AR(1) process. The smoothing terms of gamboostLSS-AR(1) can be expressed as global and local model fitting. Extensions to filters of continuous covariates and time covariates expressed in generalized differencing can handle autocorrelation issues.

5.4.6 Restriction Errors of Autocorrelation AR(1) Models

In this section, we provide the analysis of the restriction errors of the AR(1) models. Pseudo code for removing autocorrelation error is reported in Algorithm 2. The results of implementation of AR(1) model for removing autocorrelation, as seen in Figure 5.28, show that the autocorrelation tends to zero. The lag here can be referred to as the observations of the SST data, in this case it is in daily units. The common autocorrelation function is usually a lag with the length n-1, where n is the number of observations.

Plots (a), (c), (d), and (e) in Figure 5.28 have a similar pattern. The error has a cyclic trend at the starting lag. This cyclic pattern at the time lag 200 is shown in plot (b). Plot (f) and (h) have similar pattern as well. There is no cyclic pattern of error at any lag and has a fluctuation at the end lag. In general, the autocorrelation of AR(1) models has characteristic patterns. For example, a cyclical pattern at beginning time lag, several fluctuations autocorrelation ρ at the middle time lag, i.e. 600, 800-1000 and the last time lag is dependent on restriction subset of errors. It means that there is a relation between behaviour time covariates (i.e., *Nrdays* and *Doy*) in the model fitting and pattern of errors in the modelling errors with AR(1) model. The following is algorithm and visualization of autocorrelation AR(1) model as captured in Algorithm 2 and Figure 5.28.

Algorithm 2 Autocorrelation AR(1) model

Main Program:
1) Construct linear model with a subset data n as model 1 (LM1).
2) Construct linear model with a subset data n-1 as model 2 (LM2).
3) Construct linear model of residual LM1 and LM2 as a coefficient autocorrelation Rho 1.
4) To obtain Rho 2 by using iteration as follows:
   for (i in 2:niter)
   rhoproc = out1Rho
   xproc = out1x
   yproc = out1y
   out1 = FRho(rhoproc, xproc, yproc)
   Rho[i]=out1Rho

Pseudo Program:
1) To develop function
   FRho = function(Rho, x, y)
   n = length(y)
   u1= abs(y[2:n] - (Rho * y[1:n-1]))
   v1= abs(x[2:n,1] -(Rho* x[1:n-1,1]))
   v2= abs(x[2:n,2] -(Rho* x[1:n-1,2]))
   v3= abs(x[2:n,3] -(Rho* x[1:n-1,3]))
2) Setting autocorrelation model of time covariates
   Nr = NULL
   doy = NULL
   for (j in 2:n)
   Nr[j-1]= x[j,4]- (Rho*x[j-1,4])
Submodel doy
Classic Pattern doy[j-1]= (x[j,5]- (Rho*x[j-1,5]))

3) Transformation covariates:
   newx = as.matrix(cbind(v1, v2, v3, Nr, doy))
   $B = lm(u1 \sim newx)$
   beta0 = B$coef$[1]/(1 − Rho)
   beta1 = B$coef$[2]/(1 − Rho)
   beta2 = B$coef$[3]/(1 − Rho)
   beta3 = B$coef$[4]/(1 − Rho)
   beta4 = B$coef$[5]/(1 − Rho)
   beta5 = B$coef$[6]/(1 − Rho)
4) Construct a new dataset by LRM in a subset n.
   Y2hat = beta0 + beta1*x[2:n,1] + beta2*x[2:n,2] + beta3*x[2:n,3] + beta4*x[2:n,4] + beta5*x[2:n,5]
   e2 = y[2:n]- Y2hat
5) Construct a new dataset by LRM in a subset n-1.
   Y3hat = beta0 + beta1*x[1:n-1,1] + beta2*x[1:n-1,2] + beta3*x[1:n-1,3] + beta4*x[1:n-1,4] + beta5*x[1:n-1,5]
   e3 = y[1:n-1] - Y3hat
   $R2 = lm(e2[2 : n] \sim e3[1 : n - 1])$
   NewRho = R2$coef$[2]
   out1=list(Rho = NewRho, x = newx, y = u1)
   return(out1)

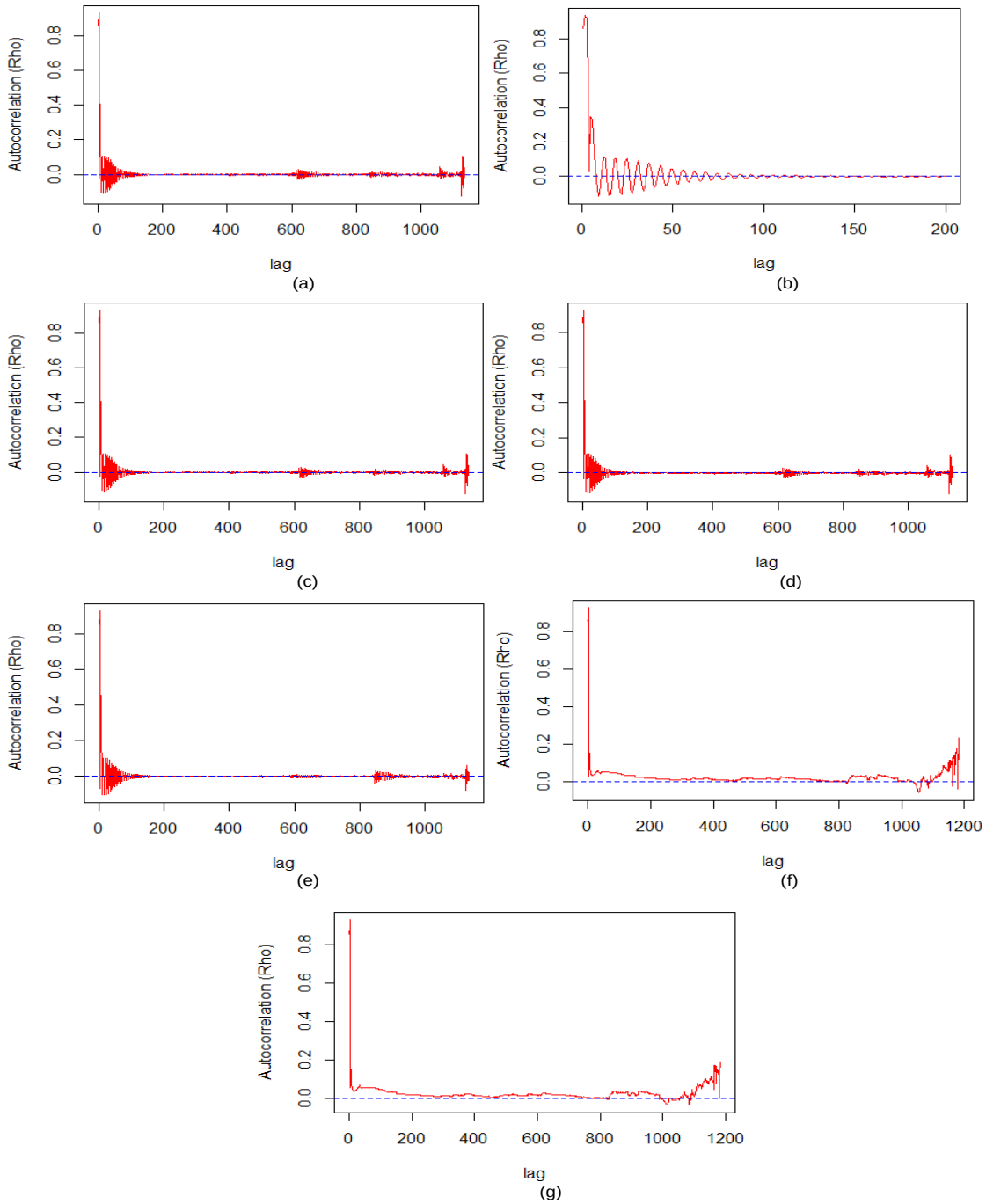


Figure 5.28: *An illustration of restriction errors of the autocorrelation AR(1) model.*

Figure 5.28 shows a cyclical behaviour of the *Doy* covariate that represents seasonal effects indicates an appropriate model or not in the model fitting by gamboost-AR(1) and gamboostLSS-AR(1) models. This indication can also be seen as a cyclical pattern of autocorrelation errors within the AR(1) model.

5.5 Applications for Different Buoys

In this Section, the SST data are extended by including ocean data from two other buoys. The data were collected at two locations, in the Indian Ocean and Meulaboh land station, during the period of 2006 to 2012. In addition, the covariates of the day of the year (*Doy*) and the number of days (*Nrdays*) are also included in the model as time covariates.

The emphasis of this Section is to fit the SST data, we used statistical inferences which were based on gamboostLSS-AR(1) models beforehand in Section 5.1 to 5.4. The results show that there are several procedures in the fitting data by using gamboostLSS-AR(1) models for different buoys. Our approach begins with statistical description, scatterplot in time period, and ACF plot of the SST data. We emphasize on using this approach to understand performance of SST data in magnitudes, measurements, positions, gaps, and patterns that potentially have the plausible climate features for appropriate data fitting.

Previously, we have investigated that the proposed gamboostLSS-AR(1) model shows better results in five different applications (see Chapter 4). Firstly, we used the proposed model for linear regression model. Secondly, we also used the model for additive model class, such as GAM, gamboost, GAMLSS, and gamboostLSS models. The results are assessed by other models using AIC, gDML, Global Deviance, and CVrisk as the performance of the fitting. Graphically, the improvement of the fitting of the proposed model has been demonstrated using SST data from one buoy. In this investigation, we experimented the gamboostLSS-AR(1) model fitting with and without transformation. The optimal number of submodels, fitting time covariate in both local and global model fitting performances, were conducted in the experiment. The result showed that it gives better fitting perfor-

mance than other models.

In the previous work, we have investigated by considering time covariates, our experiment is useful in improving the local and global fitting. This improvement is implicitly remove autocorrelation and thus decrease or keep it the same number of submodels. This condition also is depend on the complexity of data structure. The results showed that the hyper-parameters (the degrees of freedom (df), $knots$, stopping iteration (m_{stop}) and the step of length factor (ν_{slf}) parameters), play important roles in obtaining the appropriate model specification of the SST data fitting. In addition, controlling boosting parameters are also important for achieving appropriate model fitting with and without transformation of rainfall.

The transformation can increase the number of submodels in gamboostLSS-AR(1) model fitting. There are many potential benefits by combining the two approaches, which are removal autocorrelation and transformation. The benefits include more stable in fitting process, more robust to fit SST data, and more improved than other models mentioned above. The first differencing AR(1) model reduces autocorrelation of the SST data, whereas transformation of rainfall can reduce scale of outlier consequent.

In general, we applied ρ to implement the gamboostLSS-AR(1) models. First of all, we tune the hyper-parameters in model specification and autocorrelation coefficients ρ of the SST data for each buoy. We then fit SST data by using ρ. Next, we apply the value of ρ for SST data fitting from three buoys. Finally, we determine the time covariate fitting of submodels to get the appropriate model fitting. In this section, we will look at the use of ρ's in gamboostLSS-AR(1) model fitting and marginal prediction interval (MPI), and we will consider time-autocorrelation at lag 1 (so called MPI-AR(1)) for the SST data.

Moreover, we investigate considerable achievement in the fitting performance for SST data in the gamboostLSS and gamboostLSS-AR(1) models with and without transformation. We compare time effects in the μ and σ parameters of the SST data. Also, MPI-AR(1) can be applied to the SST data from three buoys.

The section is further organized as follows. As in section 2.2 Chapter 2, we provide the data and experimental setup. Section 5.6 discusses the results and the application of the gamboostLSS-AR(1) models for three buoys. In subsections 5.6.1, 5.6.2, and 5.6.3 we describe gamboostLSS models fitting at buoys 1, 2, and 3 without transformation, respectively. Subsection 5.6.4 we describe similarities time effects by gamboostLSS model fitting at three buoys. In subsection 5.6.5 we apply gamboostLSS-AR(1) model fitting with autocorrelation coefficient ρ. Subsections 5.6.6, 5.6.7, and 5.6.8 we present gamboostLSS-AR(1) models fitting at buoys 1, 2, and 3 without transformation, respectively. Subsection 5.6.9 we describe similarities time effects by gamboostLSS-AR(1) model fitting at three buoys. Then we present marginal prediction interval of gamboostLSS models in autocorrelation as in section 5.7. Finally, we summarize this chapter in section 5.8.

5.6 Results and Discussion for Different Buoys

In this section, we present the numerical results of the gamboostLSS-AR(1) model fitting with and without transformation of the SST data set. These results are compared with the gamboostLSS model to fit the same data which was previously discussed in Chapter 5. The results are recorded in tables and capture the global and local model fitting graphically.

5.6.1 The Results of GamboostLSS Fitting Model at Buoy 1

We present the gamboostLSS fitting at buoy 1 where the number of observation are the smallest, i.e. 1460 data. We also capture the local and global models as seen in Figures 5.29, 5.30, and 5.31, respectively.

5.6.1.1 The GamboostLSS Fitting Model at Buoy 1 without Transformation

We experimented the different control boosting parameters. We consider the size-length factor v_{slf} from 0.01 to 0.05 with steps 0.01 and 0.1. We also consider the different stopping iterations m_{stop} for each gamboostLSS model fitting of the SST data at buoy 1. We implemented m_{stop} to obtain optimal submodels and appropriate local and global fitting. The result of this particular experiment is displayed in Table 5.10.

Table 5.10: *The control boosting effects on the gamboostLSS model fitting of the SST data at buoy 1.*

Parameter	m_{stop}	Submodel	Final Risk	Parameter	m_{stop}	Submodel	Final Risk
$v_{slf} = 0.01$	1000	8	806.474	$v_{slf} = 0.04$	1000	11	627.540
	2000	9	739.374		2000	13	493.515
	3000	10	679.754		3000	15	437.867
	4000	11	627.565				
	5000	12	583.155				
	10000	14	459.199				
	15000	15	416.604				
$v_{slf} = 0.02$	1000	9	1230.757	$v_{slf} = 0.05$	1000	12	583.501
	2000	11	627.552		2000	15	460.083
	3000	12	546.656				
	4000	12	493.286				
	5000	15	459.457				
$v_{slf} = 0.03$	1000	9	679.513	$v_{slf} = 0.1$	1000	14	460.504
	2000	12	547.051		2000	15	419.963
	3000	13	474.766				
	4000	15	437.935				

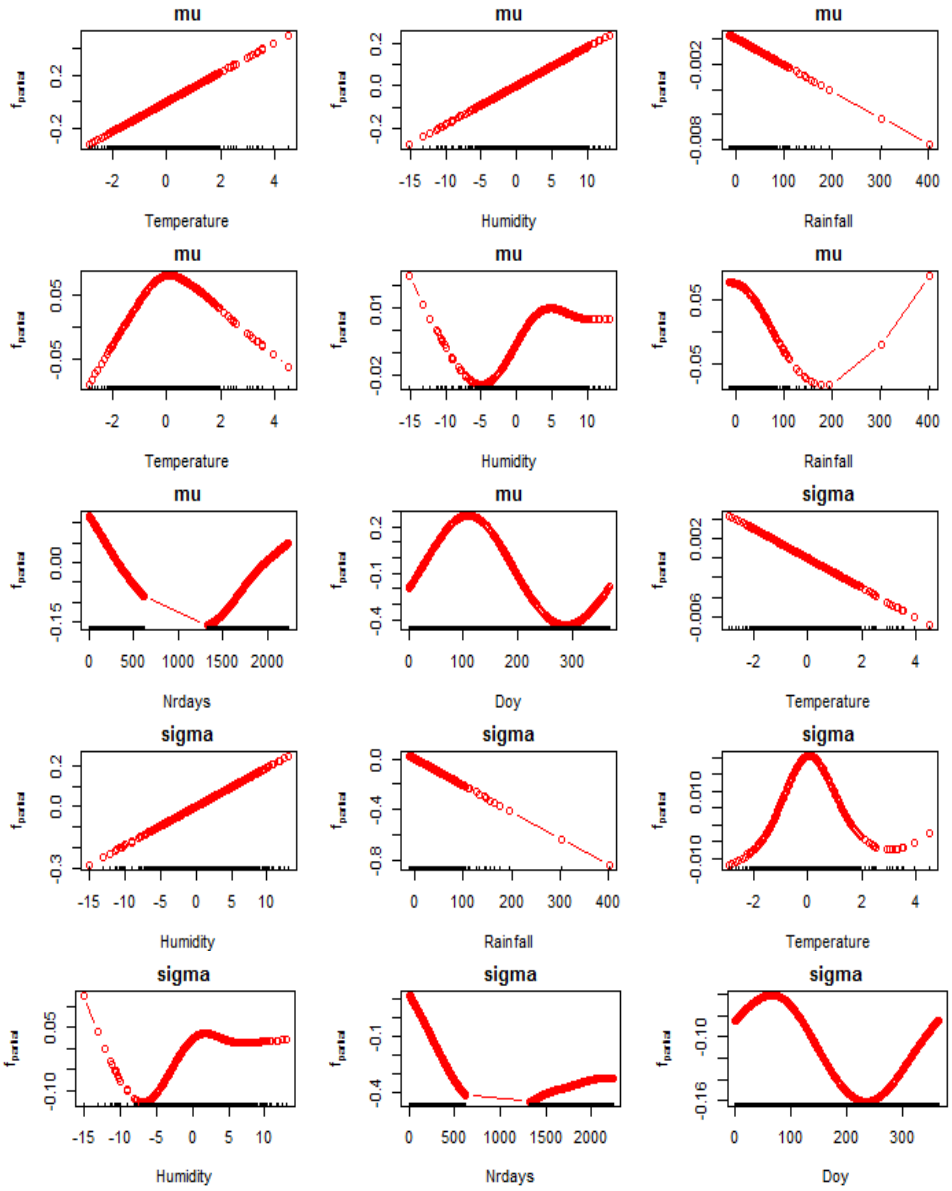


Figure 5.29: *Local fitting by gamboostLSS models without transformation of the SST data produces 15 submodels.*

Figure 5.29 illustrates the local fitting of gamboostLSS models at the buoy 1, as represented from global fitting as seen in Figure 5.31 (left). The figure consists of 15 submodels, each of which presents the climate features and time covariates. It can be seen that for the climate features in the μ parameters of temperature and humidity have a similar trend, however, the σ parameters of temperature and humidity have an opposite trend, regarding to the linear base-learner. The rainfall has an opposite trend of the temperature and

humidity in the μ parameter but it has similar trend to temperature in the parameter σ . When we used the smooth base-learner, the μ parameters of temperature and rainfall have a quadratic curve and an opposite trend. The humidity has the similar patterns in the μ and σ parameters, whereas the temperature has a downward curve in μ and σ parameters.

Furthermore, for the time covariate of the annual effects, the parameters μ and σ decreased before the gap, and increased after the gap. On the other hand, for the seasonal effects, parameters μ and σ show a sinusoidal pattern with one peak season.

5.6.1.2 The GamboostLSS Fitting Model at Buoy 1 with Transformation

Here, similar approach is applied to fit the SST data with transformation of rainfall. The results are recorded in Tables 5.11 and 5.12. The later table is obtained by the first table by considering the first 15 submodels. It illustrates the comparison of submodels with transformations and without transformations. Transformed rainfall in the gamboostLSS models does change the pattern of rainfall covariate from an upward curve to downward curve in the μ and σ parameters as captured in Figures 5.29 and 5.30 but it does not change other covariates in both parameters.

Furthermore, our experiment shows that time covariates between with and without transformation in the μ and σ parameters shows similar effects of different values of ν_{slf} which are from 0.01 to 0.03, and different values of m_{stop} which are 15000, 5000, and 4000, respectively. Transformation of rainfall also does not change time covariates pattern in this control boosting.

Selection of boosting parameter is essential to reveal information from submodel in local model fitting. Also, to get appropriate global fitting of the SST data, we propose an

Table 5.11: *The control boosting effects on the gamboostLSS model fitting with transformation of the SST data at buoy 1.*

Parameter	m_{stop}	Submodel	Final Risk	Parameter	m_{stop}	Submodel	Final Risk
$\nu_{slf} = 0.01$	1000	9	800.4935	$\nu_{slf} = 0.04$	1000	11	618.2469
	2000	10	732.1456		2000	13	484.2205
	3000	10	671.3646		3000	**15**	430.1754
	4000	11	618.2059		4000	15	405.8644
	6000	13	536.4907		6000	16	379.4355
	10000	14	450.3402				
	11000	**15**	438.8946				
	20000	15	391.2930				
	30000	16	363.6534				
$\nu_{slf} = 0.02$	1000	10	731.8544	$\nu_{slf} = 0.05$	1000	11	573.8829
	2000	11	618.2797		2000	13	451.1741
	3000	13	536.9263		3000	**15**	410.5813
	4000	14	483.8639		4000	15	391.3832
	5000	14	450.6098		5000	16	376.6471
	6000	**15**	429.9423				
	10000	15	391.3462				
	11000	16	385.2203				
$\nu_{slf} = 0.03$	1000	10	670.9519	$\nu_{slf} = 0.1$	1000	13	451.2929
	2000	13	536.3975		1500	**15**	410.3765
	3000	14	464.8827		2000	16	391.1545
	4000	**15**	429.7095				
	8000	16	379.3055				

Table 5.12: *The change of the boosting effects on the gamboostLSS model fitting with and without transformation of the SST data at buoy 1.*

ν_{slf}	m_{stop}	Submodel without transformation	Submodel with transformation	Effect
0.01	20000	15	15	equal
0.02	5000	15	14	decrease
0.03	4000	15	15	equal
0.04	3000	15	15	equal
0.05	2000	15	13	decrease
0.1	2000	15	16	increase

approach to model based on assessment of time covariates (annual and seasonal effects). Considering time covariates in the model can be found in Chapter 3 and technically approach in the Chapter 4, whereas regarding time-autocorrelation is provided in detail in Chapter 5. A cyclic trend represents to seasonal effect while a non periodic long term trend represents annual effect. We can see clearly both trends at buoys 1, 2 and 3 with or without

transformation of rainfall in the gamboostLSS and gamboostLSS-AR(1) models fitting as seen in local model fitting.

Further transformation of rainfall in the gamboostLSS model does not affect on smoothing of global fitting of the SST data as seen in Figure 5.30. However, the transformation gives effect to increase the number of submodel and decrease m_{stop} as depicted in Figures 5.29 and 5.31.

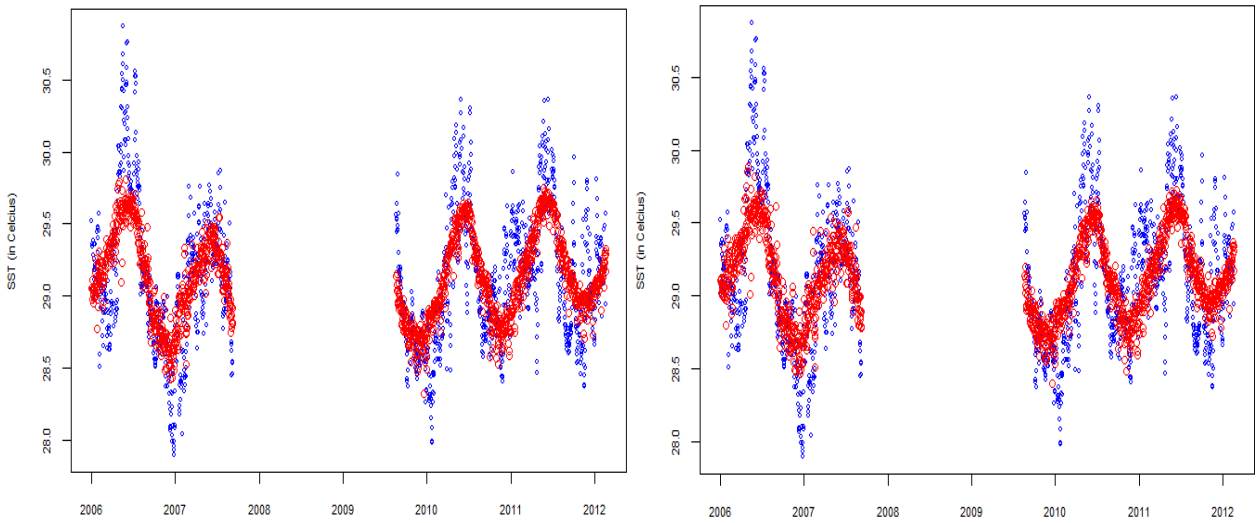


Figure 5.30: *GamboostLSS model fitting without transformation in boosting parameters, $\nu_{slf} = 0.01$ and $m_{stop} = 15000$ (left), and with transformed rainfall in the $\nu = 0.01$ and $m_{stop} = 11000$ of the SST data at buoy 1 (right).*

Figure 5.30 illustrates the global fitting of gamboostLSS model with different control boosting at the buoy 1 in detail as seen in Table 5.10, it mainly produces 15 submodels. It is clearly visible that the curve has a long gap from 2008 to 2010. This illustrates SST data with regular pattern on 2006 to 2007 period and irregular pattern on 2010 to 2012 period. The other results which can be seen in Figure 5.30, show similar pattern and non smooth of the global fitting. It means that the transformation of rainfall covariate does not affect in gamboosLSS model fitting.

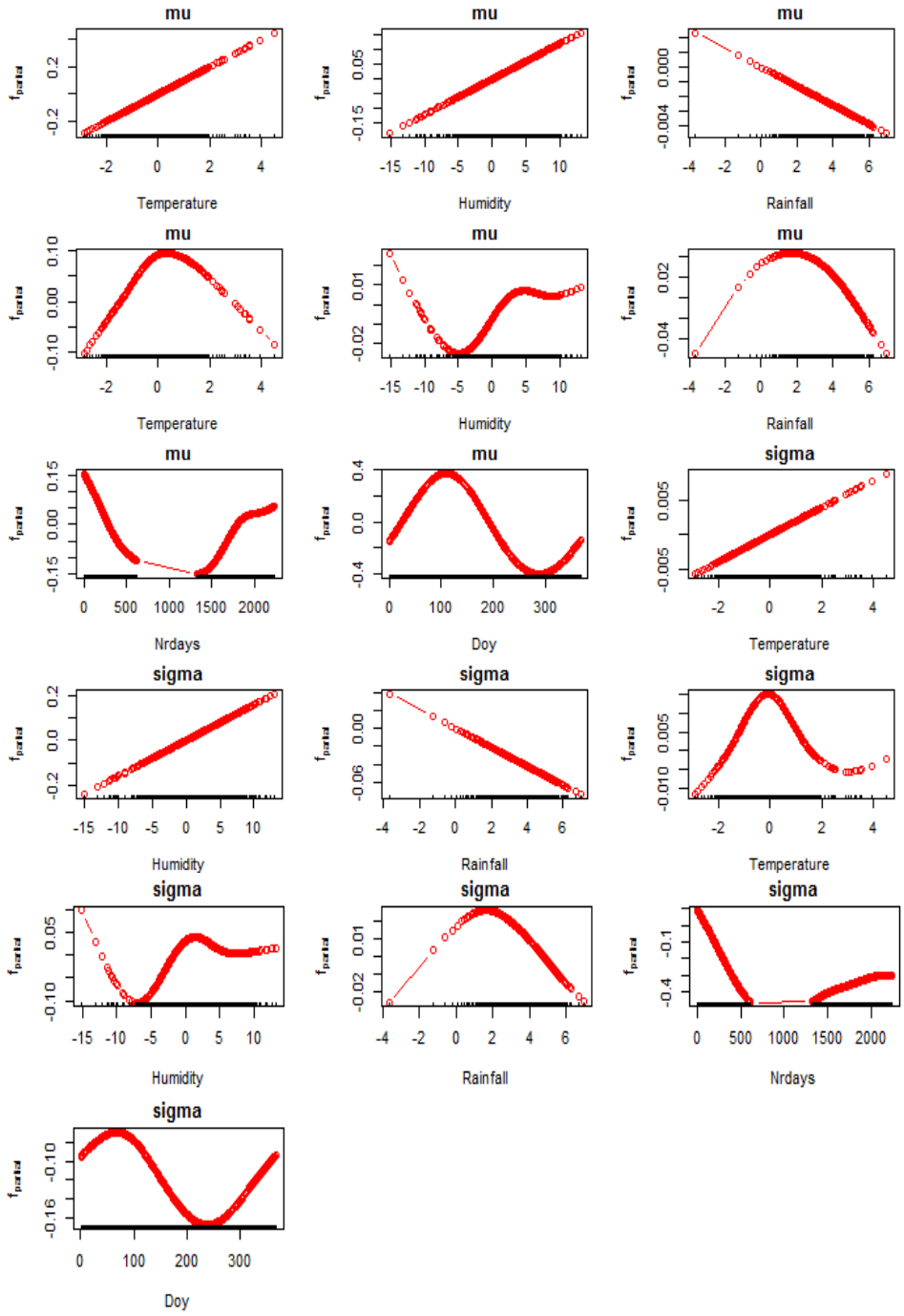


Figure 5.31: *Local fitting by gamboostLSS models with transformation of the SST data at buoy 1 gives 16 submodels.*

In Figure 5.31, we can refer in detail of explanation as captured in Figure 5.29. Furthermore, there are some advantages of transformed rainfall covariate in the gamboostLSS model fitting. They are as follows:

a) it can reduce the final risk value;

b) it sometime reduces number of submodels, however, it sometimes increases the submodels, depending the complexity of the data structure;

c) it can change the patterns of rainfall covariate itself (see Figures 5.29 and 5.30) and it does not change the pattern of another covariates, includes time covariates.

d) it can accelerate the fitting process;

e) it gives many combinations between ν_{slf} and m_{stop} on the control boosting parameters which provides many solutions of the model fitting, (see Tables 5.10 and 5.11).

f) it can accelerate to reach optimal submodel in model fitting (see Figures 5.29 and 5.30).

5.6.2 The Results of GamboostLSS Fitting Model at Buoy 2

Here, we specifically present the gamboostLSS fitting at buoy 2 where the numbers of data observations are the largest, i.e 2066, this position is explained in the previous section. The global and local model are captured differently. Figures 5.32, 5.33, and 5.34 illustrate the local and global models fitting of SST data at buoy 2.

5.6.2.1 The GamboostLSS Fitting Model at Buoy 2 without Transformation

We observed gamboostLSS model fitting with different size of length factor ν_{slf} = 0.01-0.05, 0.1 and different values of stopping iteration m_{stop} for SST data at buoy 2. The result of this

experiment is recorded and we selected the optimal number of submodels based on the appropriate local and global fitting.

The large value of m_{stop} to reach the optimal number of submodels 16, mainly for the size of length factor ν_{slf}= 0.01; 0.02; 0.03; 0.04; 0.05; 0.1, is the stopping iteration m_{stop}= 90000; 50000; 30000; 25000; 20000; and 8000 respectively.

It is clearly visible that the pattern appears regularly from 2006 to 2007 and from 2009 to 2010, whereas the pattern appears otherwise from 2008 to 2009 and from 2011 and 2012. In addition, the results display a short gap between the end of 2010 and the beginning of 2011. Although the values of control boosting ν_{slf} and m_{stop} are different, the similar pattern of global fitting by using gamboostLSS models of the SST data at buoy 2 can be achieved. This is clearly visualized in Figure 5.34.

Figure 5.32 illustrates the patterns and the trends of 15 submodels of local fitting. It is shown that temperature and humidity have the similar trends in parameter μ but opposite trends with rainfall in the same parameters. In addition, the rainfall shows three outliers in both parameters, linear and smooth base-learners. Humidity has opposite curve in μ and σ parameters, an upward and downward curves respectively, whereas temperature has the similar curve between both parameters. Furthermore, the *Nrdays* covariate in parameter μ shows the increasing trend that occurs before the gap, while the opposite trends occur after the gap. Differently, for parameter σ, the increasing trend happens after the gap. In the *Doy* covariate, furthermore, the μ parameter has a cyclic curve with a seasonal peak, whereas σ parameter has the different trend.

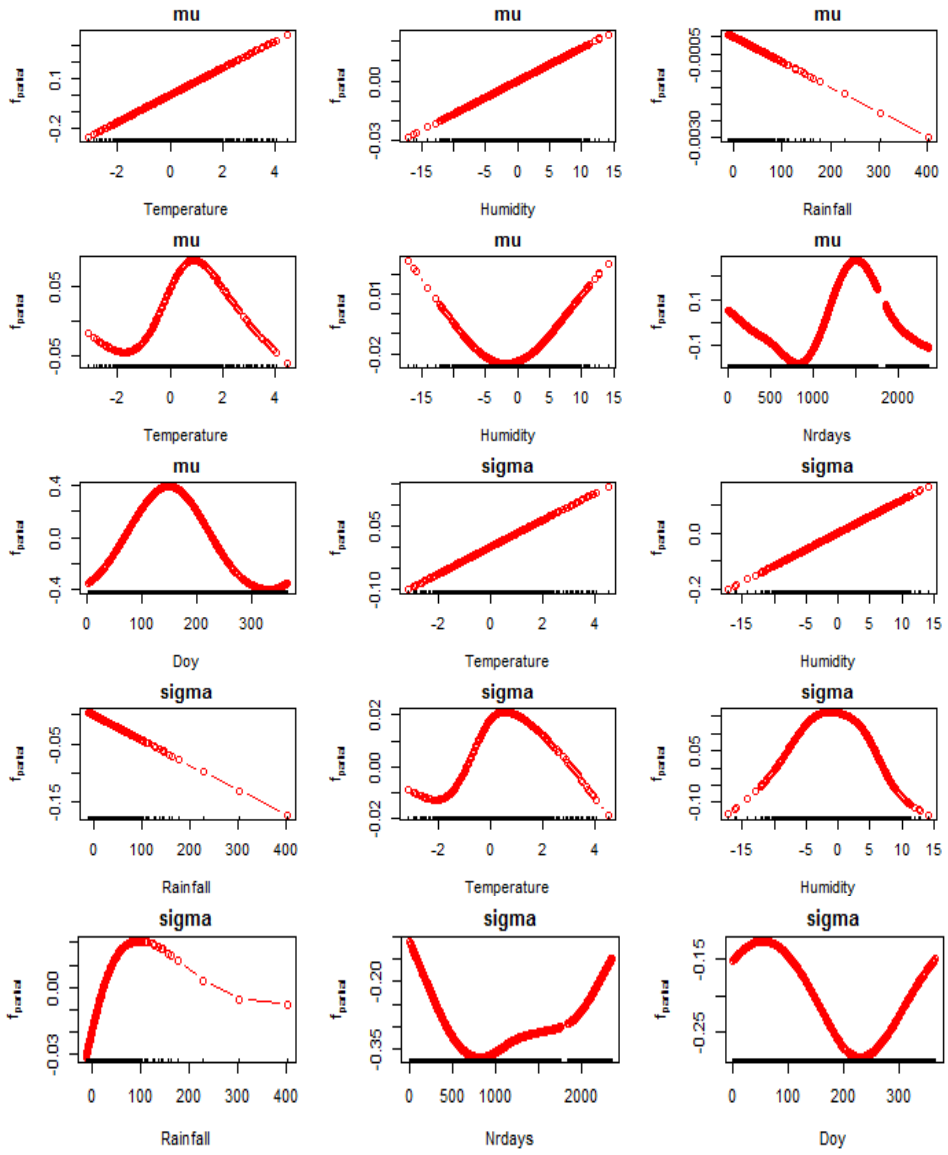


Figure 5.32: *Local fitting of the gamboostLSS model without transformation of the SST data at buoy 2 produces 15 submodels.*

5.6.2.2 The GamboostLSS Fitting Model at Buoy 2 with Transformation

As in the previous section, we used transformation to obtain the optimal submodels of the gamboostLSS model and the appropriate local fitting at buoy 2. The optimum numbers of submodels with transformed rainfall can be reached with lower values of m_{stop}.

The following result describes the control boosting effects on gamboostLSS model with transformation and without autocorrelation of the SST data at buoy 2. The model with

transformation shows that the appropriate model fitting gives 16 optimal submodels. These can be obtained by m_{stop} values of 50000, 25000, 15000, 15000, 10000 and 5000 with ν_{slf} = 0.01 to 0.05 and 0.1 respectively.

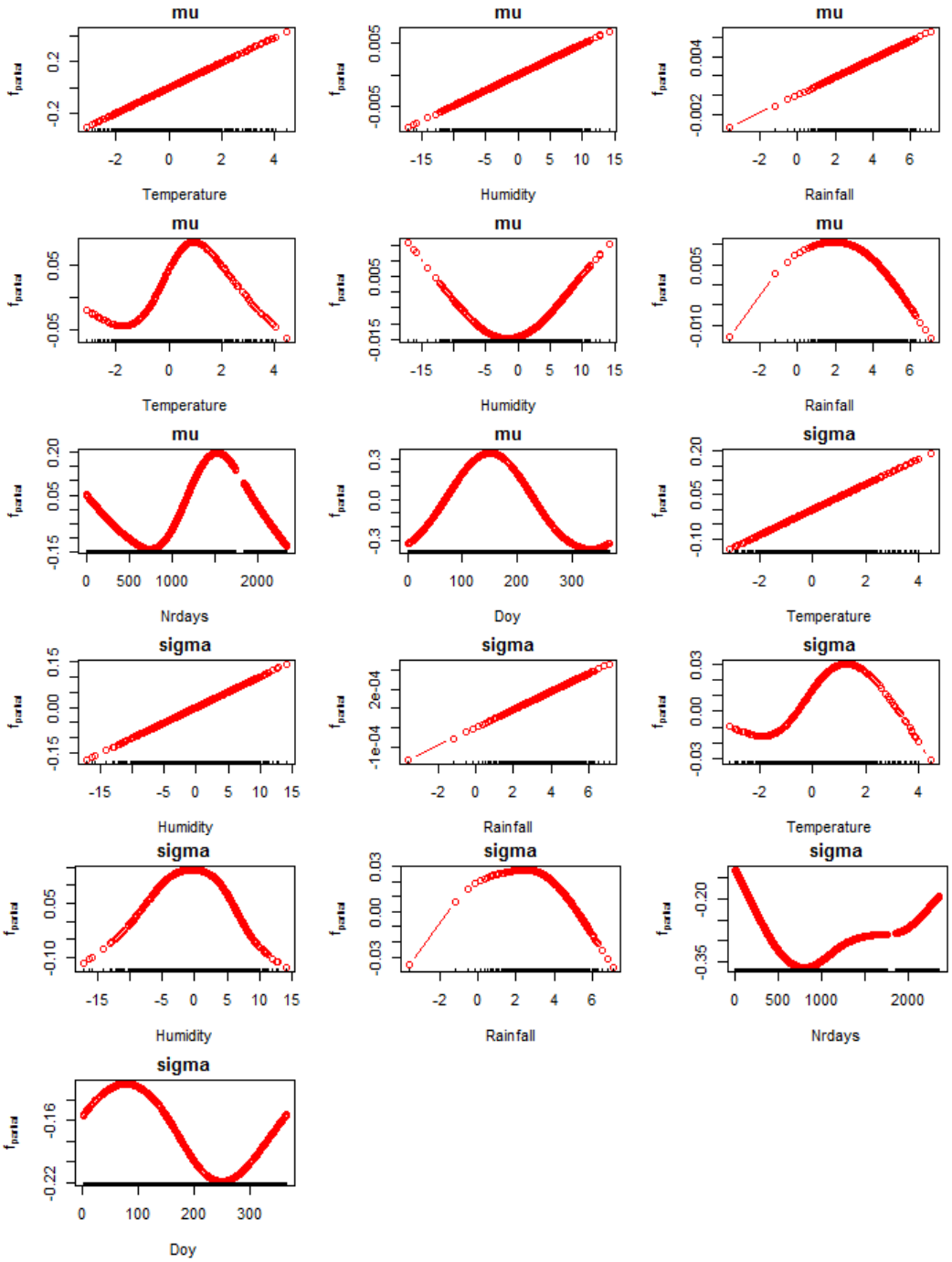


Figure 5.33: *Local fitting of the gamboostLSS models with transformation of rainfall of the SST data at buoy 2 produces 16 submodels.*

Figure 5.33 presents the local model fitting of the SST data set at buoy 2. In general, it consists of some figures which present the climate features such as temperature, humidity, and rainfall, each of which represents submodel of gamboostLSS model. It can be seen that the μ and σ parameters of humidity, temperature, and rainfall show similar curves which are linear, regarding linear base-learner.

Interestingly, the μ and σ parameters of temperature have similar curves which are increase trend in linear base-learner and unique with unimodal curve in smooth base-learner respectively, whereas the μ and σ parameters of humidity and rainfall have different curves. The rainfall has the similar trend on the μ and σ curves. However, the transformation of rainfall changes direction in both parameters of linear base-learner. On the other hand, it has the downward curve in both parameters of smooth base-learner.

In this figure, we also captured the *Nrdays* and *Doy* covariates to determine the annual and seasonal effects, respectively, of fitting model of the SST data. For *Nrdays* curve, the μ parameter shows decrease and increase trends before the gap, then decrease after the gap. It also has a peak before the gap. For the *Doy* covariate, the μ parameter shows a unimodal curve, whereas the σ parameter has a sinusoidal curve. This curve is as an implication of alteration difference penalty with entering a sinus term in P-splines as cyclic penalties [36, 37, 45, 74]. It can be seen here that the transformation of rainfall reduced the final risk as well as of m_{stop} values (see Tables 5.13 and 5.14). This also affects the number of submodels that are achieved.

Figure 5.34 shows the similar pattern of global fitting by gamboostLSS with and without transformation. However, transformation of rainfall in the model does not have effects in the global fitting of the SST data at buoy 2. The fitting also shows nonsmooth pattern.

Table 5.13: *The change of the boosting effects on final risk of the gamboostLSS model fitting with and without transformation of the SST data at buoy 2.*

ν_{slf}	m_{stop}	Submodel without transformation	Final Risk	Submodel with transformation	Final Risk
0.01	10000	13	1243.098	13	1240.930
0.02	10000	14	1096.310	15	1093.304
0.03	5000	13	1153.353	13	1150.646
0.04	5000	14	1096.191	15	1093.203
0.05	5000	14	1057.276	16	1054.233
0.1	3000	14	1027.071	16	1023.841

Table 5.14: *The change of the boosting effects on m_{stop} of the gamboostLSS model fitting with and without transformation of the SST data at buoy 2.*

ν_{slf}	m_{stop}	Submodel without transformation	m_{stop}	Submodel with transformation
0.01	40000	15	17000	15
0.02	20000	15	9000	15
0.03	20000	15	6000	15
0.04	10000	15	5000	15
0.05	10000	15	4000	15
0.1	4000	15	2000	15

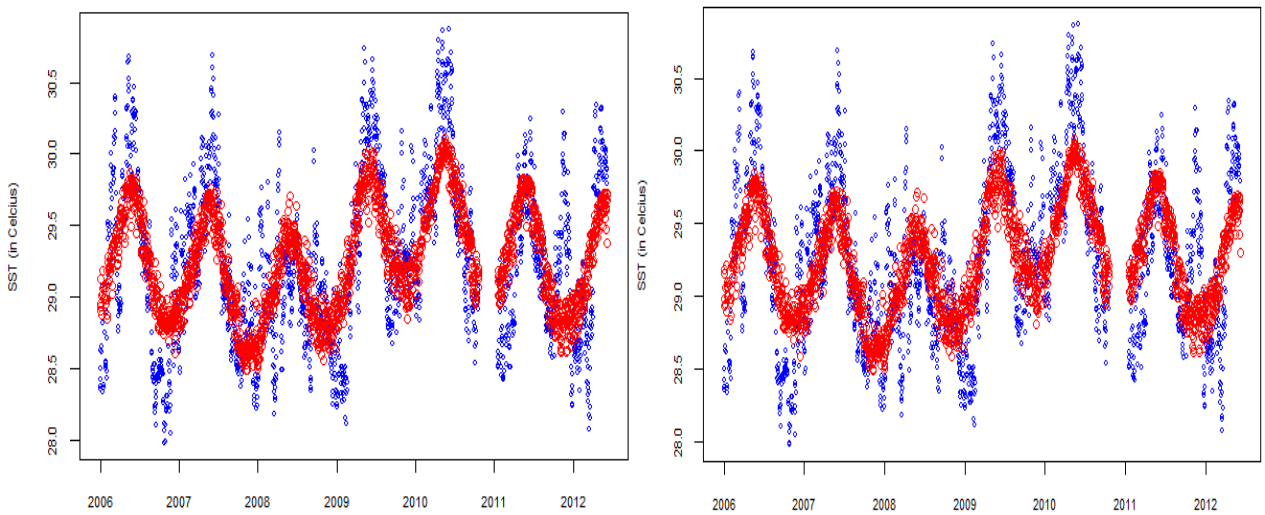


Figure 5.34: *GamboostLSS models without transformation in boosting parameters (ν_{slf} = 0.01 and m_{stop} = 90000) (left), and with transformed rainfall (ν_{slf} = 0.01 and m_{stop} = 50000) (right) for the SST data from buoy 2.*

5.6.3 The Results of GamboostLSS Fitting Model at Buoy 3

Lastly, we present the gamboostLSS fitting at buoy 3 where the number of data observations are 1606.

5.6.3.1 The GamboostLSS Fitting Model at Buoy 3 without Transformation

We observe different ν_{slf} and m_{stop} of control boosting parameters on the gamboostLSS model fitting without ρ of the SST data at buoy 3. We summarize the results in the following: the ν_{slf}= 0.01 to 0.05, 0.1 with m_{stop}= 30000, 20000, 10000, 10000, 10000, and 3000 respectively gives 15 submodels. The different size of length factor ν_{slf} is used in gamboostLSS models without transformation and the model achieved optimal submodels with different m_{stop} for the SST data fitting. However, the model includes inappropriate model fitting, therefore we need to select the gamboostLSS models fitting.

Figure 5.35 describes the local fitting of gamboostLSS model at buoy 3. As can be seen here, the μ parameters of temperature and humidity have similar trends. Similarly, the σ parameters of temperature, humidity, and rainfall have the same curves which are linear, regarding the linear base-learner. In contrast, the σ parameters of temperature and humidity have opposite curves. The rainfall covariate in the μ and σ parameters for linear and smooth base-learners show increase trend and downward curve with three outliers.

The μ and σ parameters of gamboostLSS model show that temperature, humidity, and rainfall have similar trends. Temperature and rainfall have similar curves on μ and σ curves, whereas humidity has opposite curves on both curves. The annual effects increase before and after the gap, and the peak occurs at seasonal term on μ parameter. Temperature and humidity have similar smooth curves on μ, but different patterns on σ curve for humidity.

In *Nrdays* covariate, increasing trend is clearly visible for the μ parameter, whereas for σ parameter, the curve forms a sinusoidal wave. For *Doy* covariate as well as the μ parameter, the curve forms a bimodal.

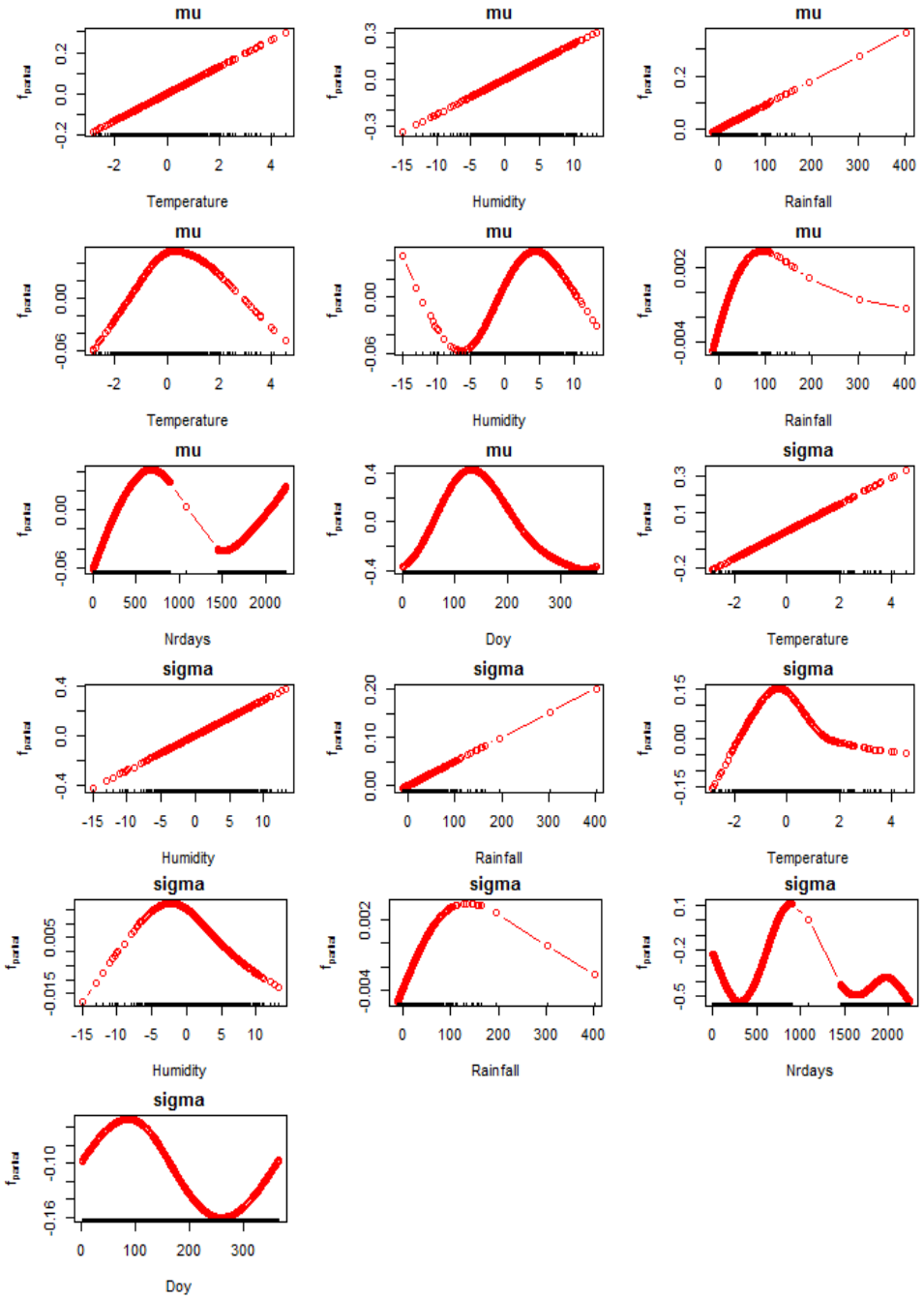


Figure 5.35: *Local gamboostLSS model fitting of the SST data from the buoy 3 displays 16 submodels in boosting parameters ($v_{slf} = 0.1$ and $m_{stop} = 3000$).*

In our investigation, time effects in gamboostLSS model fitting with transformation shows the similar pattern. The annual and seasonal effects have the similar pattern in gamboostLSS models fitting without transformation. It means that the pattern of time effects and global fitting by different control boosting parameters does not change in gamboostLSS models fitting without transformation. The result shows that time effects in local model

fitting can be fitted by using fixed size of length factor and different stopping iteration.

5.6.3.2 The GamboostLSS Fitting Model at Buoy 3 with Transformation

Similarly, we consider the size of length factor and different stopping iteration parameters to observe control boosting effects with respect to gamboostLSS model fitting for the SST data at buoy 3. The result of this experiment is as follows: the ν_{slf}= 0.01 to 0.05, 0.1 with m_{stop}= 30000, 20000, 10000, 10000, 10000, and 3000 respectively for 15 submodels.

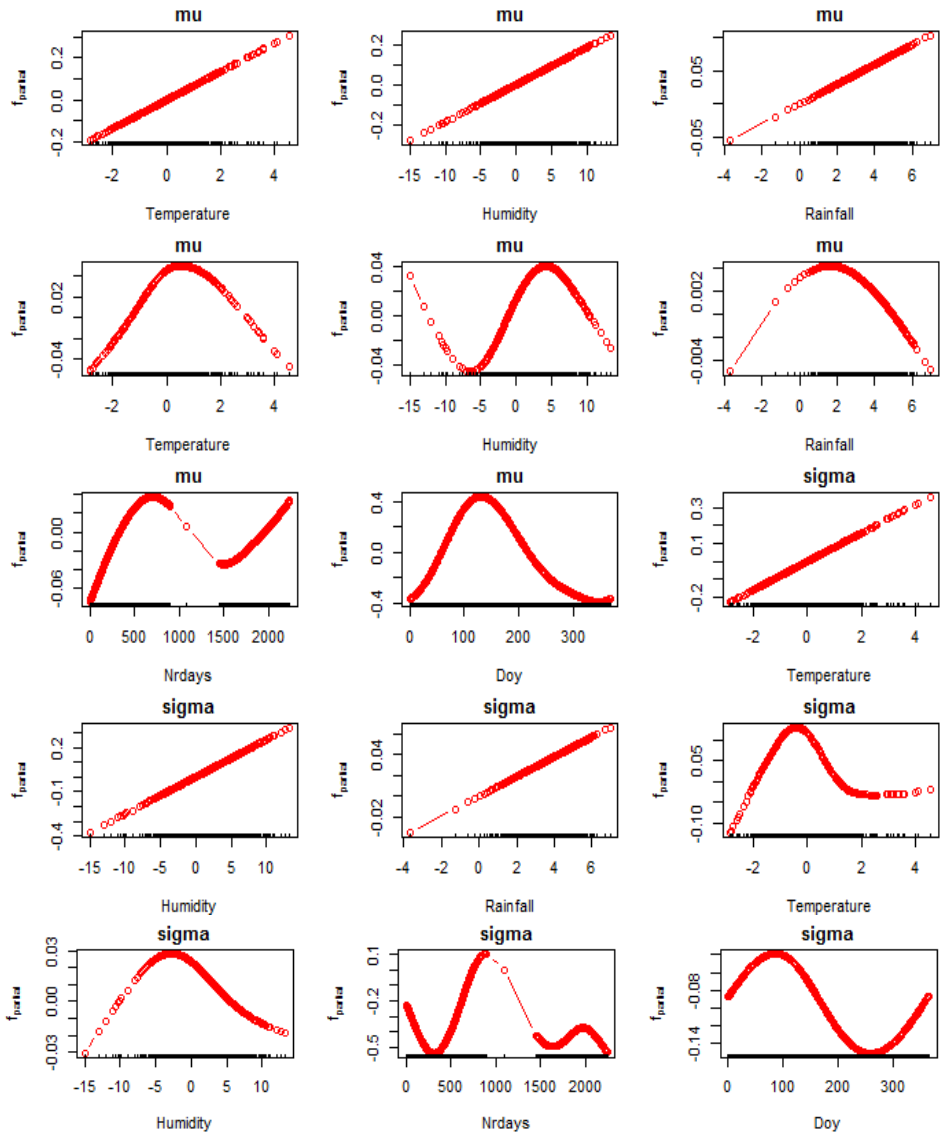


Figure 5.36: *Local fitting by gamboostLSS model of the SST data from buoy 3 displays 15 submodels.*

Transformation of rainfall covariate in the μ and σ parameters for linear and smooth base-learners show increase trend and downward parabolic curve as seen in Figure 5.36. Transformation of rainfall in the gamboostLSS model by using the same parameters ν_{slf} and m_{stop}, gives smaller number of submodels and final risk values compared to the one without transformation.

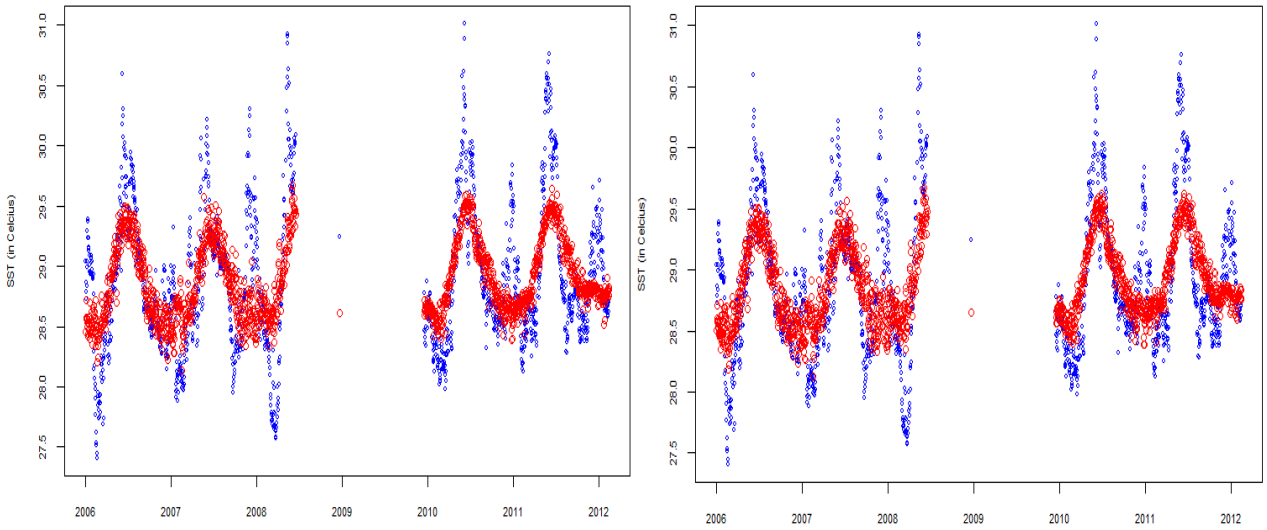


Figure 5.37: *GamboostLSS model fitting without transformation in boosting parameters: $\nu_{slf} = 0.1$ and $m_{stop} = 3000$ (left) and with transformation of the SST data from buoy 3 ($\nu_{slf} = 0.1$ and $m_{stop} = 3000$) (right).*

The global model fitting as captured in Figure 5.37 shows similar patterns. Both models have the same boosting parameters and different approaches (with and without transformation). The effect of control boosting in gamboostLSS model fitting with transformation shows that the optimal number of submodels can be achieved with lower m_{stop} compared to before transformation.

5.6.4 Similarities Time Effects by GamboostLSS Model Fitting at Buoys

Here, we investigated the similarity of the time of μ and σ parameters as the effect of techniques with and without using transformation of rainfall at buoys 1, 2, and 3.

5.6.4.1 Similarities Time Effects by GamboostLSS Model Fitting at Buoys 1, 2, and 3 Without Transformation

We present the results of time similarities without using transformation of rainfall in gamboostLSS model fitting.

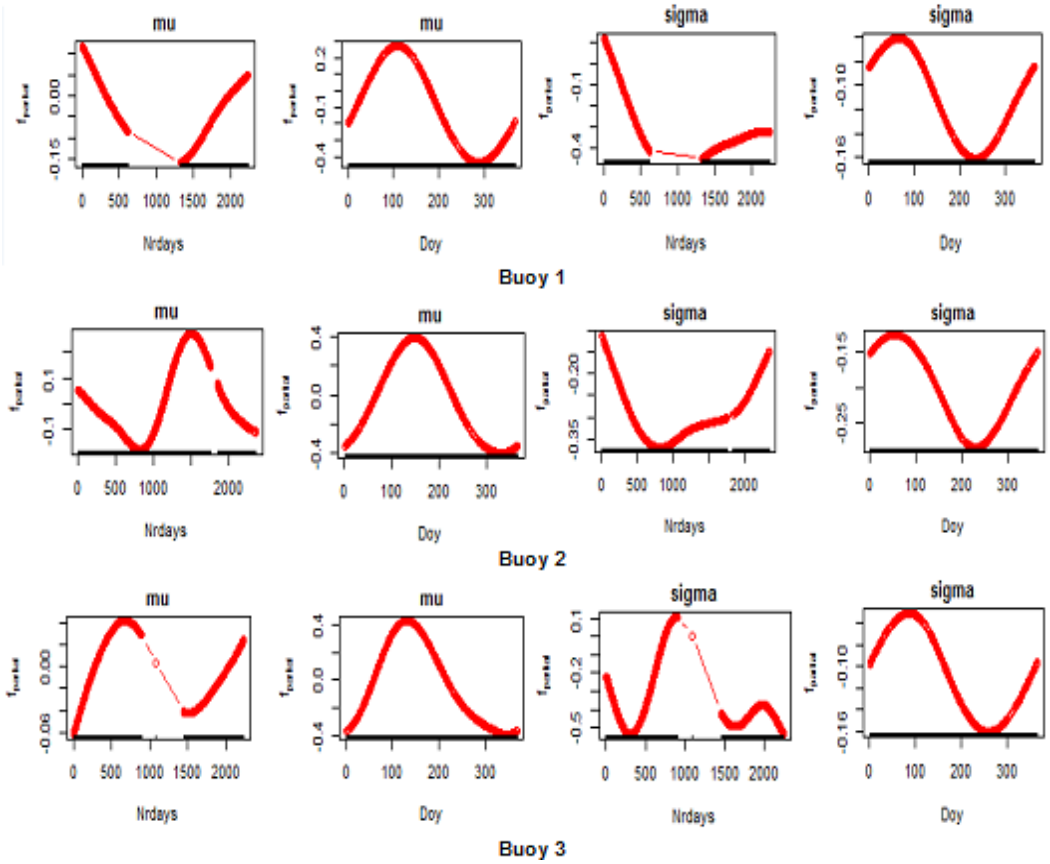


Figure 5.38: *The annual and seasonal patterns of the μ and σ parameters at buoys 1, 2, and 3 without transformation using the same specification gamboostLSS model.*

Figure 5.38 illustrates the cyclic curves and unimodal of the seasonal patterns. The seasonal effects of the σ parameter for all buoys show a similar curve. The annual effects of the μ and σ parameters of the smooth base-learner are vary. However, it shows the increasing trend at buoy 3 before and after the gap on the μ parameter.

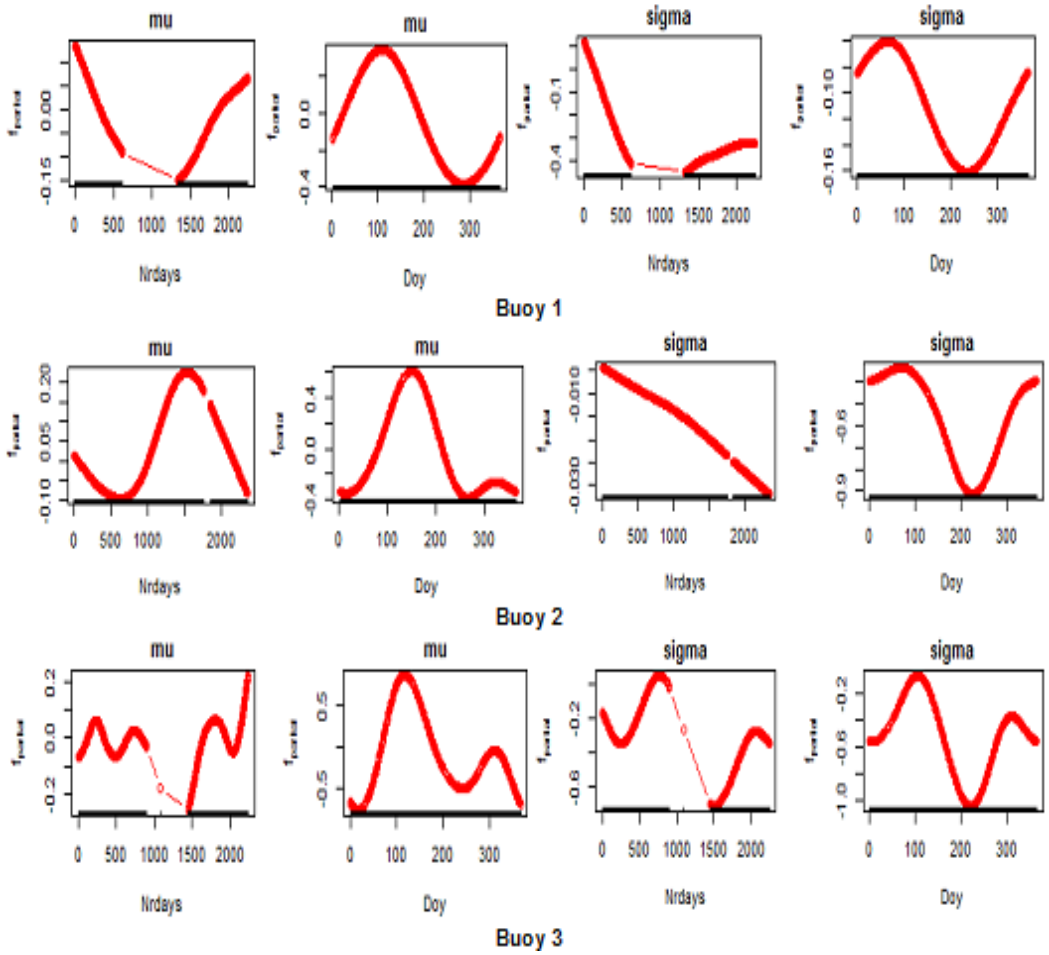


Figure 5.39: *The annual and seasonal patterns of the μ and σ parameters at buoys 1, 2, and 3 without transformation using the different specification gamboostLSS model.*

Figure 5.39 describes annual and seasonal patterns. On the μ and σ curves, it shows the different pattern for annual effects and similar pattern for seasonal effects, such as bimodal curve in μ parameter for buoys 2 and 3 and as letter "V" in σ parameter for all buoys.

5.6.4.2 Similarities Time Effects by GamboostLSS Model Fitting at Buoys with Transformation

In this section, we discuss the results of time effects similarity by using transformation in gamboostLSS model fitting. We recorded them graphically as seen in Figure 5.40. Here, we displayed the results as two parts, the similarities of the seasonal patterns at buoys 2 and 3 on the μ parameter, and the ones at buoys 1 and 2 on the σ parameter.

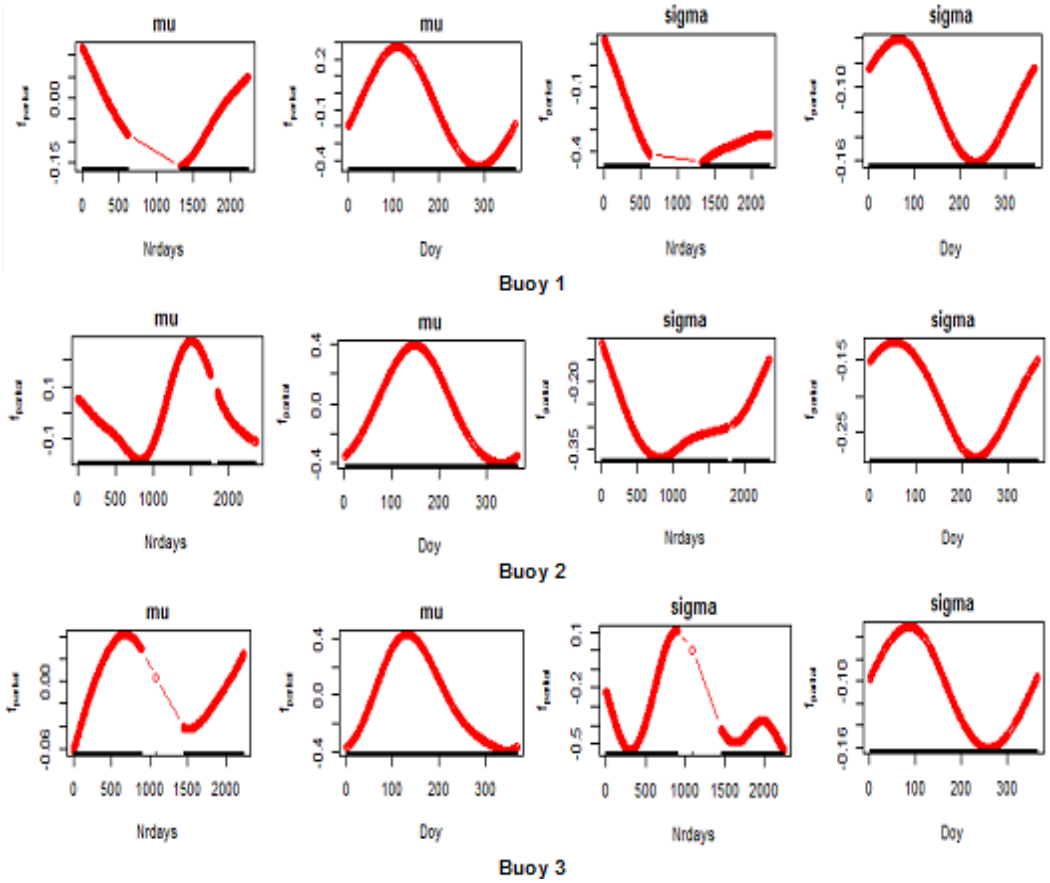


Figure 5.40: *The annual and seasonal patterns of the μ and σ parameters at buoys 1, 2, and 3 with transformation using the same specification gamboostLSS model.*

For μ parameter, the seasonal pattern forms the unimodal curve and it has one peak. On the other hand, for σ parameter, it has a slightly different trend at beginning of the seasons.

We conclude that the seasonal pattern of the same specification of the gamboostLSS models, has a unimodal form and unique pattern for annual effects on the μ and σ parameters. This is different from the seasonal pattern of different specification of the gamboostLSS models as can be seen in Figure 5.41. In this figure, the seasonal patterns at buoys 2 and 3 show a bimodal curve on the μ parameter. On σ parameter, it does not form such a curve, however, it only has similar patterns. In addition, the annual seasons of buoys 2 and 3 are vary smoothing, The similarity of the seasonal effects, however, occurs only on param-

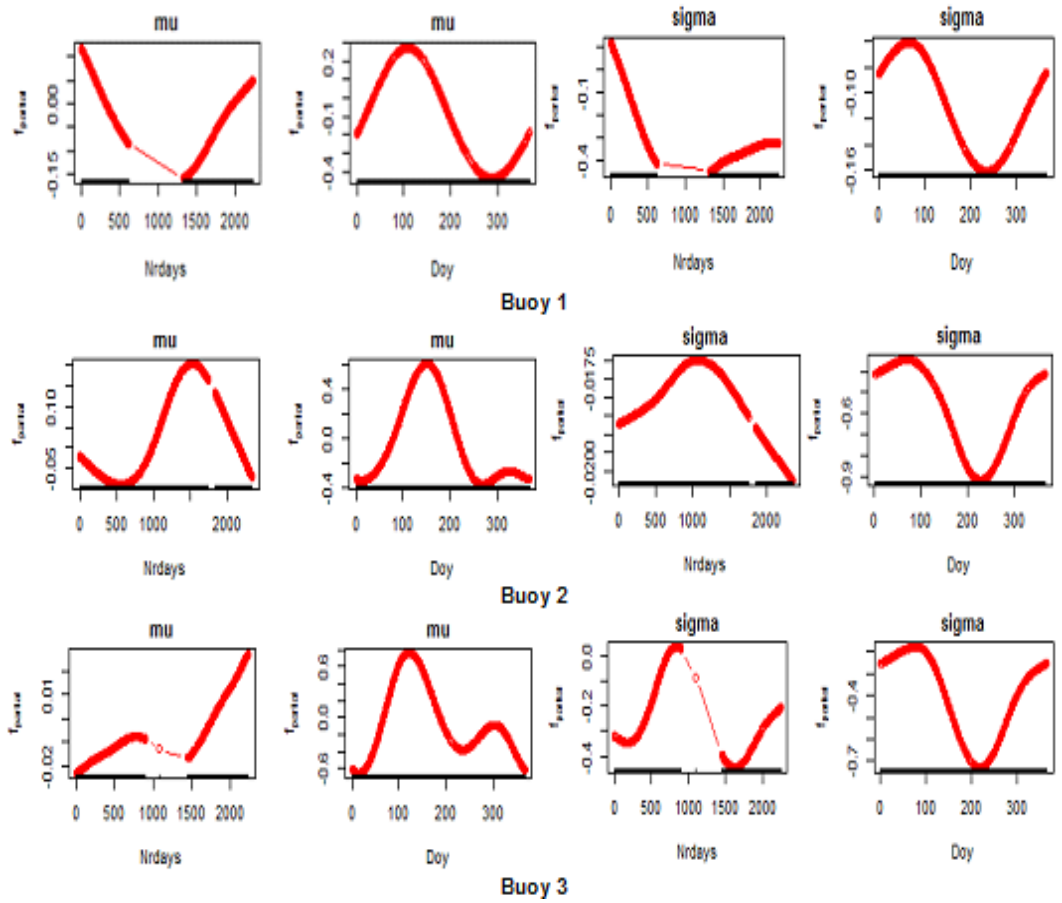


Figure 5.41: *The annual and seasonal patterns of μ and σ parameters at buoys 1, 2, and 3 with transformation using different specifications gamboostLSS models.*

eter σ for buoys 1, 2, and 3. For further works, we suggest to use different specification gamboostLSS model fitting with transformation when the SST data have different patterns.

5.6.5 Application of GamboostLSS-AR(1) Model Fitting with Autocorrelation Coefficient ρ

In this subsection, we investigate the application of gamboostLSS-AR(1) to the global model fitting of the three data sets which are given in Model 5.6.5. In our experiments, we used the different values of autocorrelation coefficient, $\rho_{buoy-1} = 0.8477007$, $\rho_{buoy-2} = 0.8835944$, and $\rho_{buoy-3} = 0.9466932$, for each data set of the buoys. The results are depicted in Tables

and Figures in this subsection. The figures show that gamboostLSS-AR(1) model fitting is achieved for these values of ρ's.

$$
\begin{aligned}
Model = gamboostLSS(SST \sim\ & bols(int, intercept = FALSE)+ \\
& bols(Temperature, intercept = FALSE)+ \\
& bols(Humidity, intercept = FALSE)+ \\
& bols(Rainfall, intercept = FALSE)+ \\
& bbs(Temperature, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
& bbs(Humidity, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
& bbs(Rainfall, center = TRUE, knots = 20, df = 1, degree = 3, differences = 2)+ \\
& bbs(Dayofyear, df = 1.01, cyclic = TRUE, boundary.knots = c(1, 365))+ \\
& bbs(Nrdays, df = 2.1, degree = 2, knots = 40), \\
& families = GaussianLSS(), \\
& control = boost_{control}(mstop = 1000, nu = 0.1, trace = TRUE), data = databr)
\end{aligned}
$$

5.6.6 The Results of GamboostLSS-AR(1) Fitting Model at Buoy 1

We present the gamboostLSS-AR(1) fitting model with and without transformation of rainfall of the SST data at buoy 1.

5.6.6.1 The GamboostLSS-AR(1) Fitting Model without Transformation at Buoy 1

We consider different values of the size of length factor $\nu_{slf} = 0.1$, and from $\nu_{slf} = 0.01$ to 0.05 with step 0.01. We also consider different values of the stopping iteration m_{stop} of the control boosting parameters in the gamboostLSS-AR(1) models fitting for the SST data at buoy 1. The result of this experiment is summarized in the following: fixed $\nu_{slf} = 0.01$ to 0.05, 0.1 with different $m_{stop} = 30000$, 15000, 10000, 6000, 5000, and 3000 respectively. In this result, we choose 13 submodels to obtain the appropriate model in global and local fitting. This selection is for all of ν parameters with different m_{stop}. The result of particular size of length factor $\nu_{slf} = 0.01$ in global fitting can be seen clearly in Figure 5.44.

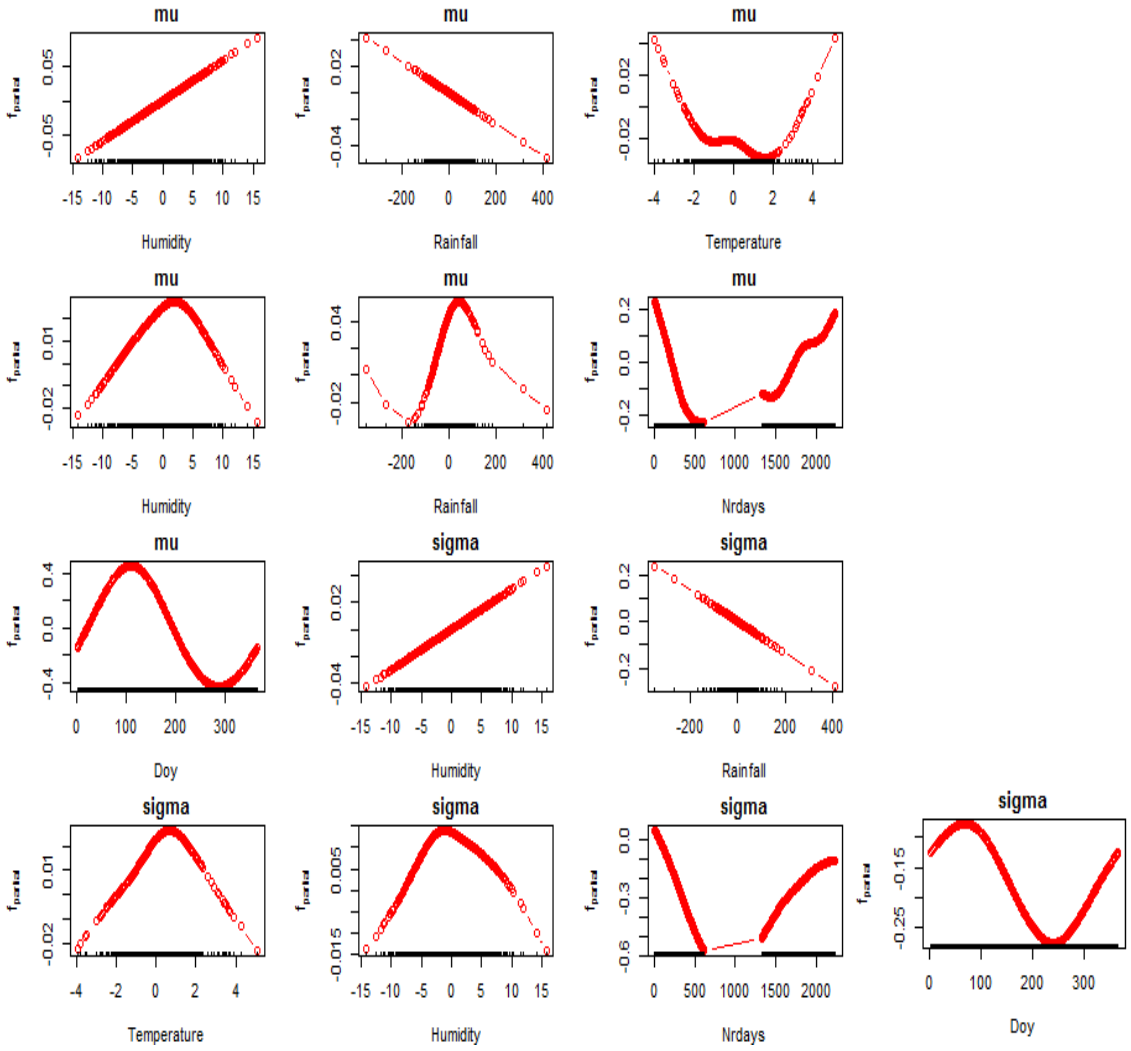


Figure 5.42: *Illustration of 13 submodels of the gamboostLSS-AR(1) model fitting without transformation for the SST data at buoy 1.*

In Figure 5.42, the rainfall forms unimodal curve smoothly and linearly, while humidity and temperatures have a variety of curves. More specifically, the μ and σ parameters of gamboostLSS-AR(1) model show that the humidity and rainfall have opposite trends. The humidity and rainfall shows a downward curve and the temperature shows an upward curve on the σ parameter. The temperature and humidity have the similar downward curve on the σ parameter. The interpretation of this figure is that the annual effects decrease before the gap and increase after the gap on the parameters μ and σ. The seasonal effects show a sinusoidal wave on μ and σ parameters.

5.6.6.2 The GamboostLSS-AR(1) Fitting Model with Transformation at Buoy 1

We present gamboostLSS-AR(1) model fitting with transformation of the SST data at buoy 1, firstly, we used the ν_{slf}= 0.01 to 0.05, 0.1 with different m_{stop}= 25000, 15000, 10000, 7000, 5000, and 3000 respectively. Secondly, we apply this result to get the global and local fitting which are visualized in Figures 5.43 and 5.44. Thirdly, we summarized as in Table 5.15. The interesting one of this table is that we can clearly see the appropriate model based on the 13 submodels.

Table 5.15: *The change of the boosting effects on m_{stop} of the gamboostLSS-AR(1) model fitting with and without transformation of the SST data at buoy 1.*

ν_{slf}	m_{stop}	Submodel without transformation	Final Risk	m_{stop}	Submodel with transformation	Final Risk
0.01	30000	13	411.1054	25000	13	418.8557
0.02	15000	13	411.0993	15000	13	405.3620
0.03	10000	13	411.0888	9000	13	413.2136
0.04	6000	13	427.2254	7000	13	410.5385
0.05	5000	13	424.3410	5000	13	418.8191
0.1	3000	13	410.9750	3000	13	405.3114

Figure 5.43 shows the local fitting of gamboostLSS-AR(1) model fitting with transformation. The humidity and rainfall have similar trends in the μ parameter but it has different trends in the σ parameter. The rainfall has a bimodal curve in the μ parameter. It reaches a peak and off seasons in the σ parameter as the seasonal effects in the μ parameter for smooth base-learner. In the σ parameter, the humidity and rainfall have opposite trend for linear base-learner. For annual effects show decrease before the gap and increase after the gap in both parameters, whereas the seasonal effects show a sinusoidal wave in the σ parameter. Differently, in Figure 5.43, on μ parameter, the rainfall forms bimodal curve, while on σ parameter, the rainfall forms a sinusoidal curve. It is obvious that the time covariate does not change for both with and without transformation.

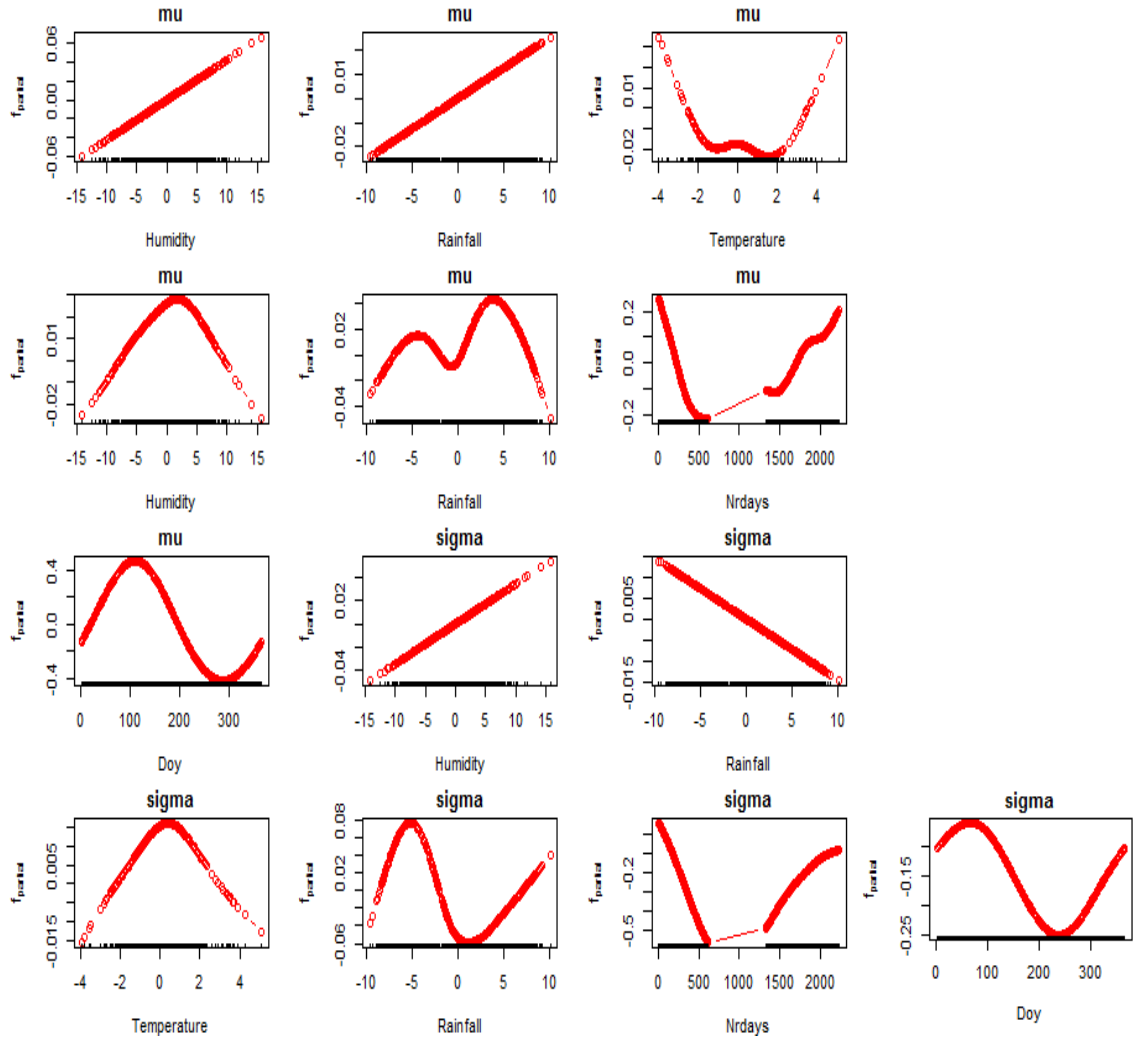


Figure 5.43: *Illustration of 13 submodels of gamboostLSS-AR(1) model fitting with transformation of rainfall for the SST data at buoy 1.*

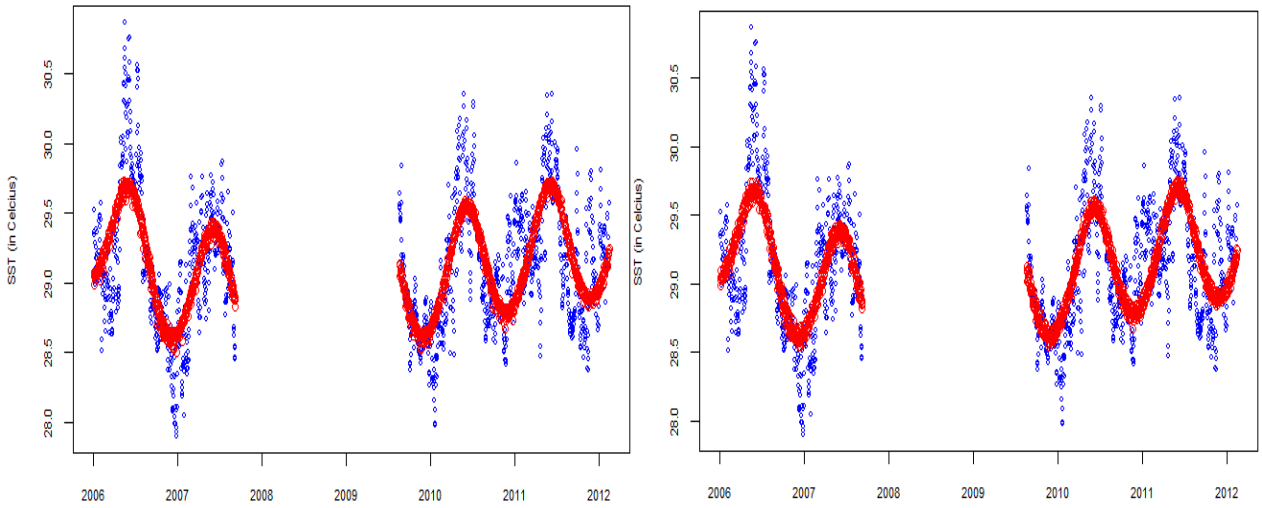


Figure 5.44: *Global fitting for the SST data from buoy 1 shows similar patterns of the gamboostLSS-AR(1) models without transformation (ν=0.01 and m_{stop}=30000) (left) and with transformation (ν=0.01 and m_{stop}=25000)(right).*

Different m_{stop} and fixed ν_{slf} parameters in the gamboostLSS-AR(1) model do not change pattern of the global fitting as captured in Figure 5.44.

5.6.7 The Results of the GamboostLSS-AR(1) Fitting Model at Buoy 2

In this subsection, we present the gamboostLSS-AR(1) model fitting with and without transformation of the SST data at buoy 2.

5.6.7.1 The GamboostLSS-AR(1) Fitting Model without Transformation at Buoy 2

We observed different values of the ν_{slf}= 0.1, 0.01 to 0.05 and m_{stop} parameters of the control boosting in the gamboostLSS-AR(1) models fitting for the SST data at buoy 2. In the results of observation, we recorded that the optimal number of submodels reach 16 where the ν_{slf}= 0.01 to 0.05, 0.1 with m_{stop}= 110000, 60000, 40000, 30000, 25000, and 15000 respectively.

Figure 5.45 displays the parameters μ and σ of the gamboostLSS-AR(1) model showing that temperature and rainfall have a similar trend. The temperature and humidity have opposite linear effects in the μ and σ parameters. The peak of annual effects occur around the gap and decrease after the gap in the μ parameter. The temperature has a similar curve such seasonal effects of the (*Doy*) covariate in the μ parameter. Decreasing of annual effects occur before the gap and steeply increase after the gap in the σ parameter. The peak effects occur at the seasonal term in the μ parameter and decrease in the σ parameter.

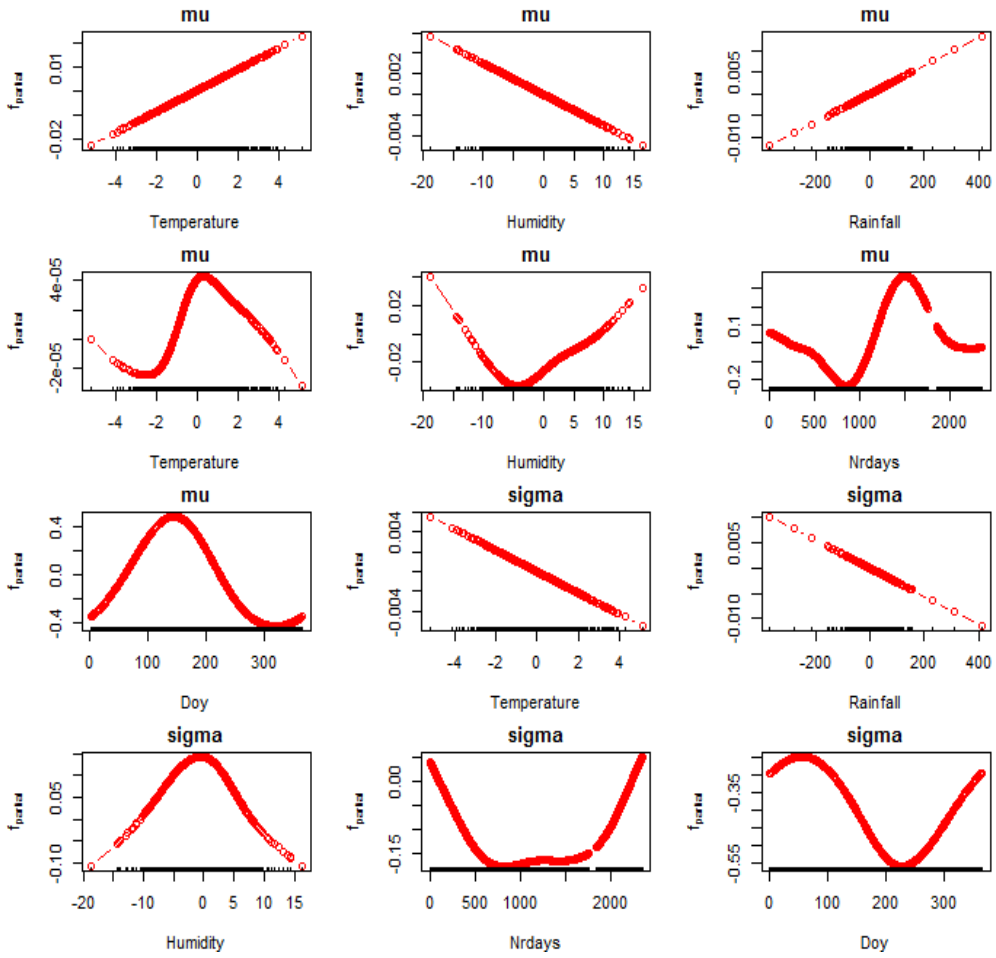


Figure 5.45: *Local fitting using gamboostLSS-AR(1) model for the SST data at buoy 2 displays 12 submodels.*

5.6.7.2 The GamboostLSS-AR(1) Fitting Model with Transformation at Buoy 2

We reported gamboostLSS-AR(1) model fitting with transformation of the SST data at buoy 2 reach optimal number of submodels as follows: the fixed ν_{slf} = 0.01 to 0.05, 0.1 and different m_{stop} = 80000, 40000, 25000, 20000, 15000, and 7000 respectively. If we compare to without transformation in the same ν_{slf}, then gamboostLSS-AR(1) model fitting with transformation can significantly reduce the m_{stop} values to obtain optimal number of submodels and appropriate model fitting. We used this report to get the global and local fitting which are displayed in Figures 5.46 to 5.47.

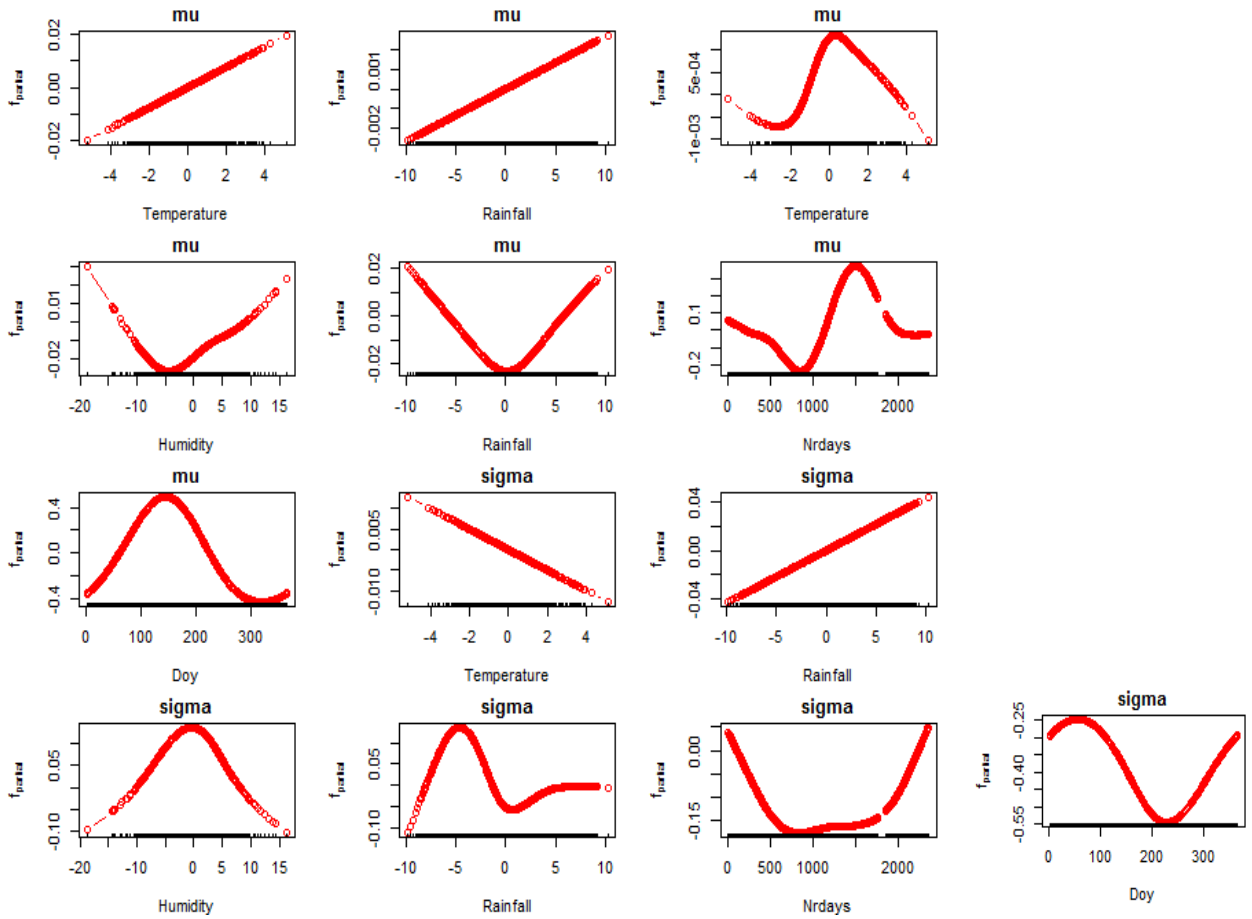


Figure 5.46: *Local model fitting of the SST data at buoy 2 using gamboostLSS-AR(1) model and transformed rainfall describes the optimal number of submodels.*

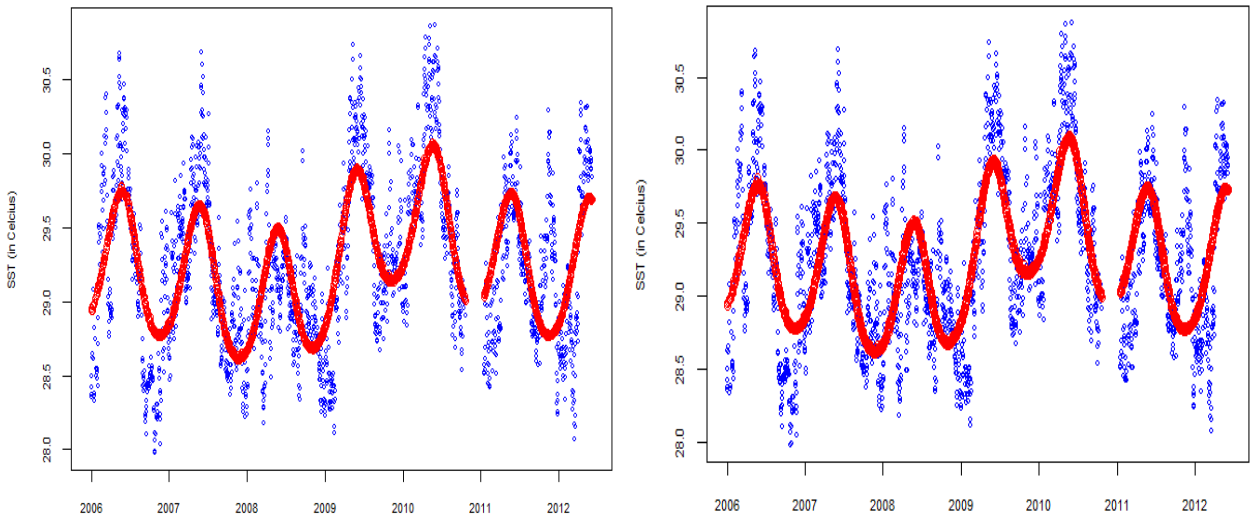


Figure 5.47: *An illustration of the gamboostLSS-AR(1) model fitting without transformation (left) and the model with transformed rainfall of the SST data at buoy 2 (both models in the $\nu = 0.01$ and $m_{stop} = 60000$ parameters) (right).*

Figure 5.47 shows appropriate global fitting of the SST data at buoy 2 using gamboostLSS-

AR(1) model. The global model fitting describes the pattern with two year intermittent and a seasonal peak.

Removing time-autocorrelation and transformation in the gamboostLSS-AR(1) models fitting gives smooth annual and seasonal effects. These effects have the similar patterns given fixed v_{slf} and different m_{stop} values in the control boosting parameters.

5.6.8 The Results of the GamboostLSS-AR(1) Fitting Model at Buoy 3

In this subsection, we specifically present the gamboostLSS-AR(1) model fitting with and without transformation of rainfall at buoy 3. The global and local model are captured differently.

5.6.8.1 The GamboostLSS-AR(1) Fitting Model without Transformation at Buoy 3

To assure fitting process comparability, we set the v_{slf}= 0.1, 0.01 to 0.05 and selected m_{stop} parameters of the control boosting in gamboostLSS-AR(1) models fitting for SST data at buoy 3. The results are depicted as follows: the v_{slf}= 0.01 to 0.05, 0.1 with m_{stop}= 90000, 50000, 30000, 22000, 18000, and 9000 respectively.

Figure 5.48 shows annual and seasonal patterns in the μ and σ parameters of the gamboostLSS-AR(1) model with transformation of rainfall. Different stopping iteration m_{stop} values and fixed the size of length factor v_{slf} in the gamboostLSS-AR(1) model fitting produces the similar patterns of time covariates. We suggest to use this specification to obtain appropriate model fitting for the SST data from buoy 3.

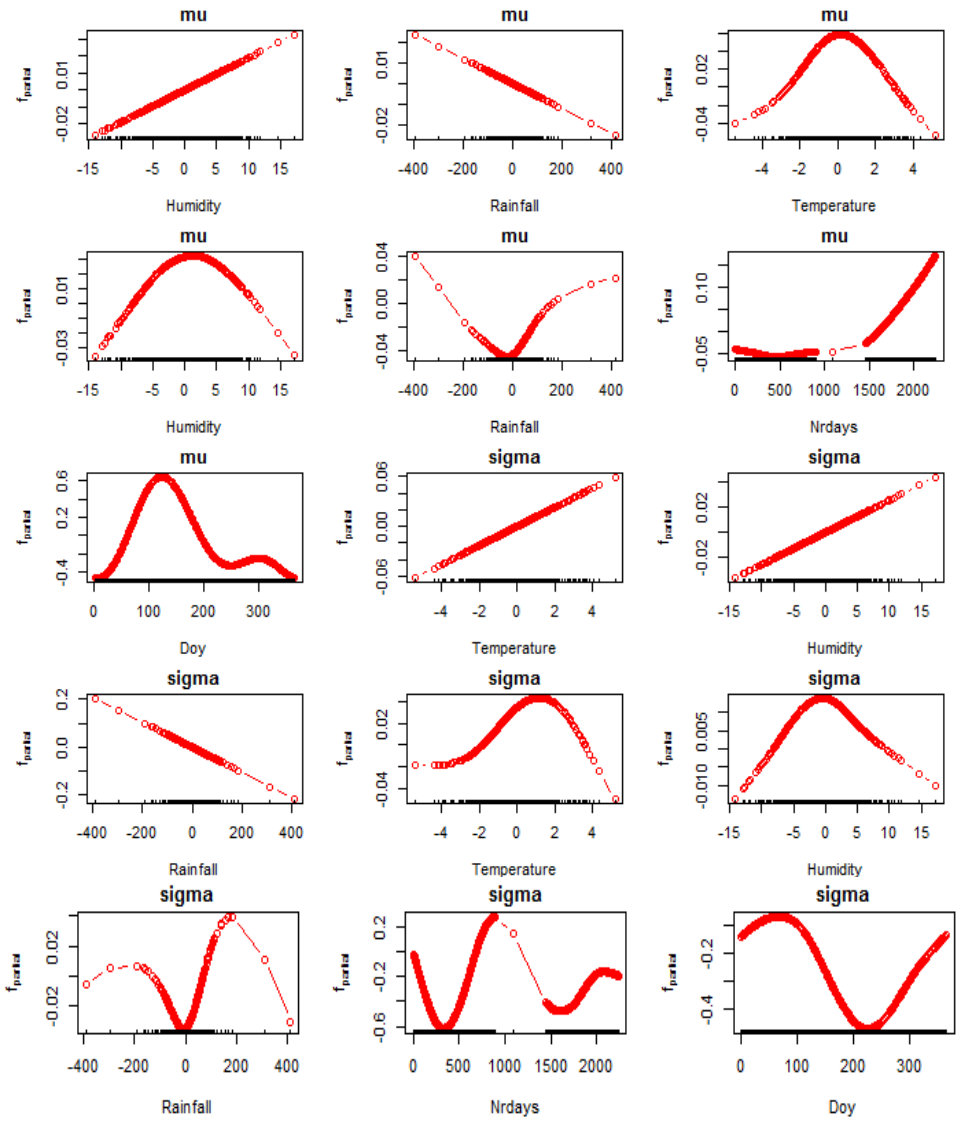


Figure 5.48: *Local model fitting with transformation for the SST data at buoy 3 gives 15 submodels in boosting parameters, $\nu_{slf} = 0.01$ and $m_{stop} = 90000$.*

5.6.8.2 The GamboostLSS-AR(1) Fitting Model with Transformation at Buoy 3

Similar approach is implemented to fit the SST data with transformation. We also modelled each size of length factor ν_{slf}= 0.01 to 0.05, 0.1 with different m_{stop}= 50000, 25000, 20000, 15000, 10000, and 5000 respectively. We consider the first 15 submodels to obtain appropriate local and global fitting.

In Figure 5.49, the gamboostLSS-AR(1) model fitting shows that the humidity and rain-

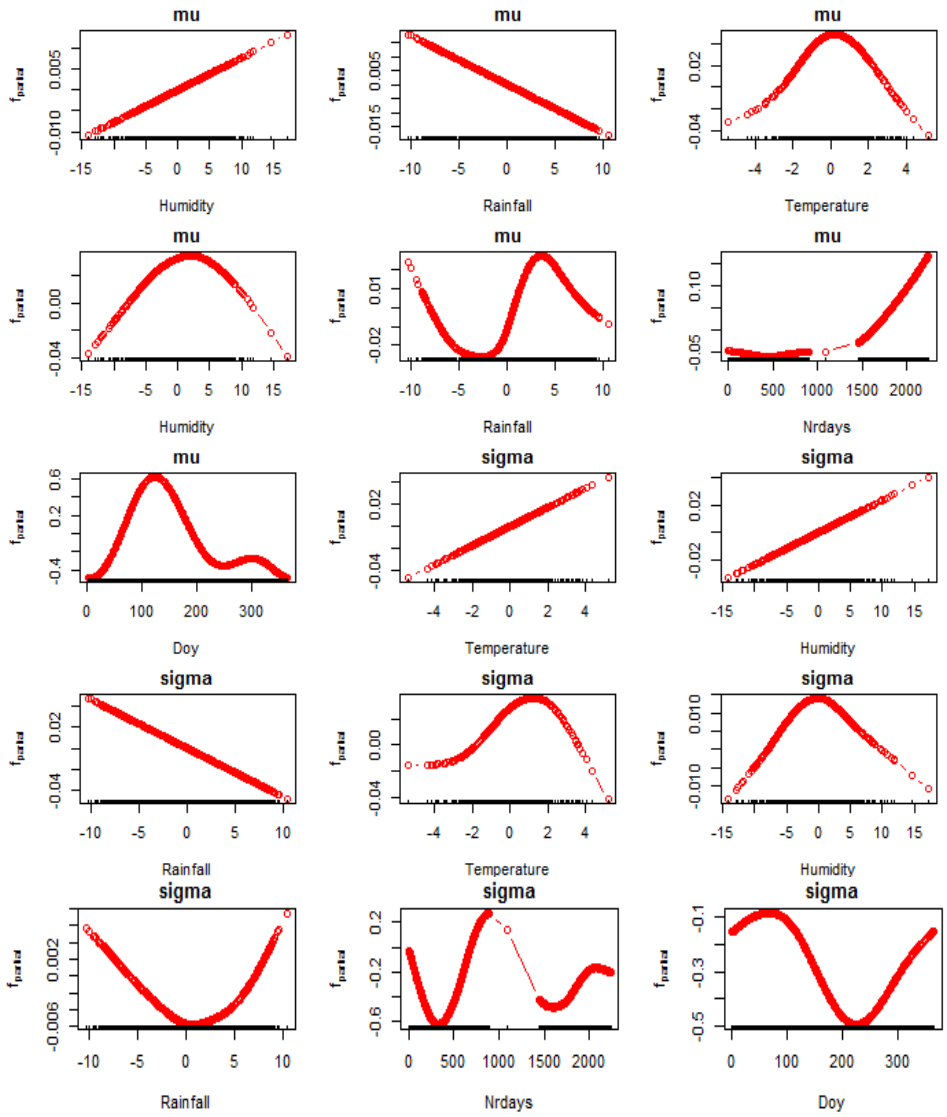


Figure 5.49: *Local model fitting with transformation for the SST data at buoy 3 displays 15 submodels (v_{slf} = 0.01 and m_{stop} = 90000).*

fall have opposite trends in the μ and σ parameters for linear base-learner. The temperature and humidity have similar smooth curve in the parameters μ and σ. The different pattern of the rainfall is an upward curve in the σ and a wave curve in the μ parameters. A trend of annual effects was stable before the gap and increases after the gap in the parameter μ, whereas in the parameter σ it shows decrease and increase before the gap and slightly decrease and increase after the gap. The seasonal effects shows a bimodal curve for μ parameter and seen as letter ”V” for σ parameter.

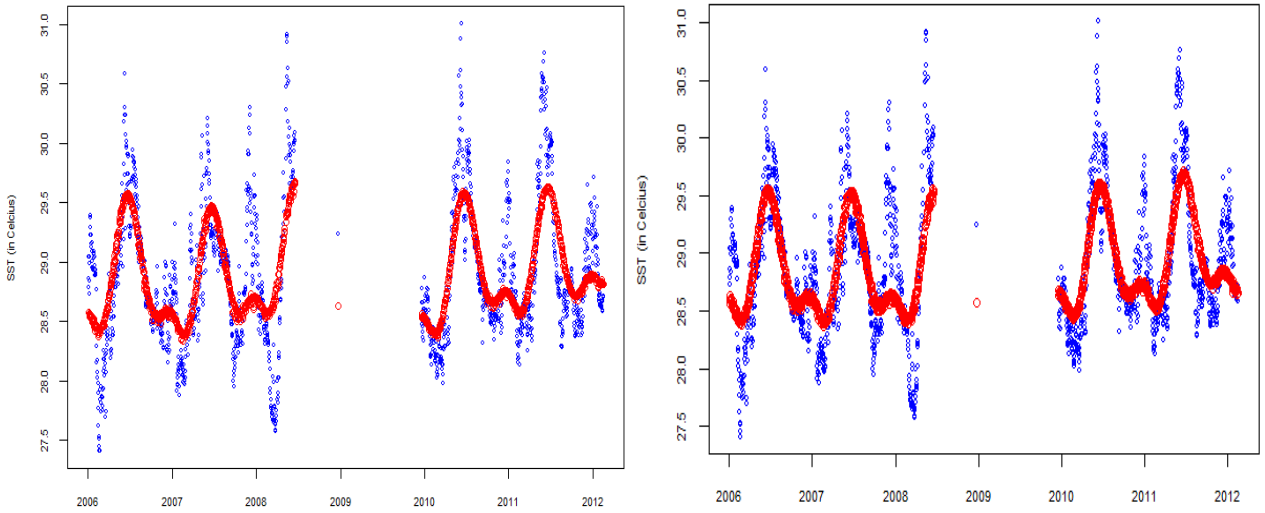


Figure 5.50: *GamboostLSS-AR(1) model without transformation (left) and with transformation (right), both models show similar patterns of global fitting for the SST data at buoy 3 (v_{slf} = 0.01 and m_{stop} = 90000).*

Transformation effect of rainfall in the gamboostLSS-AR(1) model fitting of the SST data increased the number of submodels. Figure 5.50 describes the appropriate global fitting using gamboostLSS-AR(1) for both models.

5.6.9 Similarities Time Effects of GamboostLSS-AR(1) Model Fitting

In this subsection, we present the similarity of the time effects in the μ and σ parameters as the effect of with and without transformation of the SST data at buoys 1, 2, and 3.

5.6.9.1 Similarities Time Effects by GamboostLSS-AR(1) Model Fitting at Three Buoys without Transformations

We investigated time covariates based on different specification of the gamboostLSS-AR(1) models fitting without transformation for buoys 1, 2, and 3. The results are recorded in Figure 5.51. In this Figure 5.51, we applied different values of the size of length factor v_{slf} and the stopping iteration m_{stop} parameters to obtain appropriate local fitting. We can see that the similar pattern of seasonal effect in the σ parameter is for all buoys, whereas the

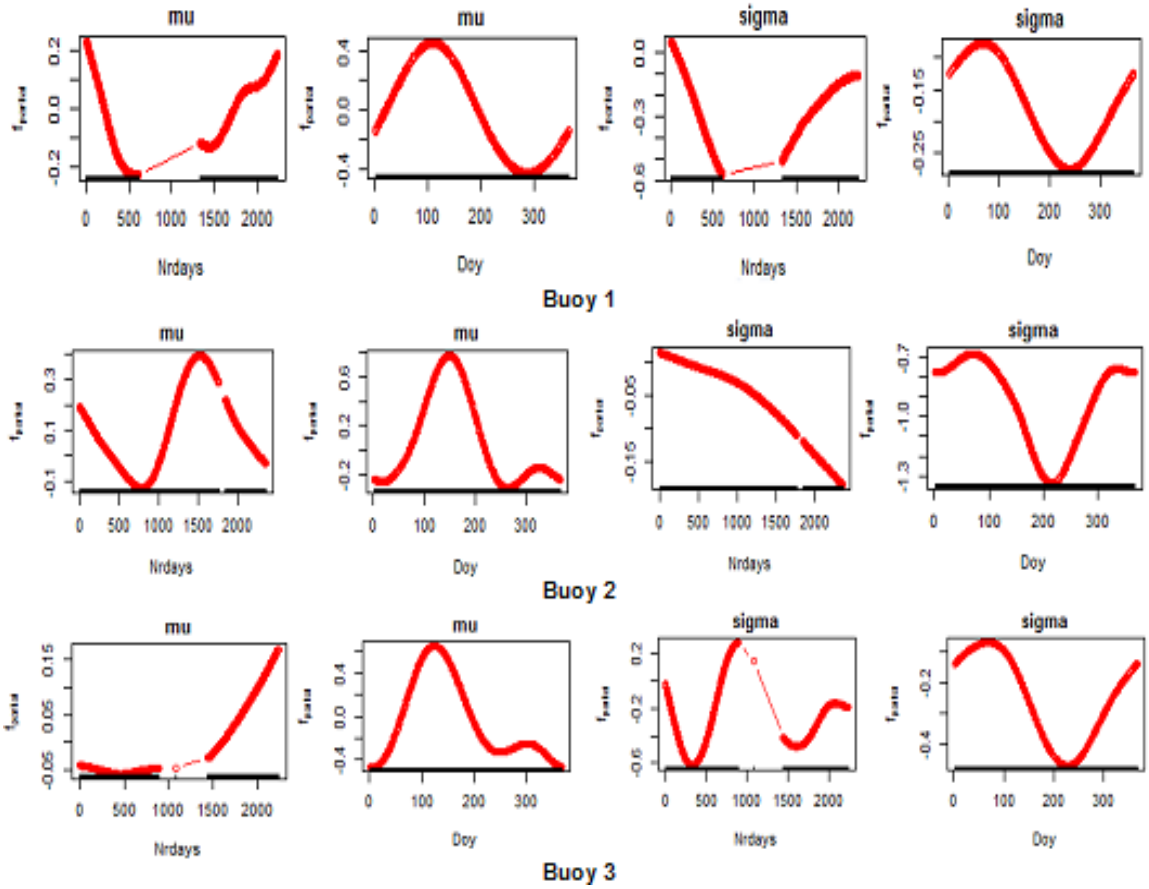


Figure 5.51: *The annual and seasonal patterns using gamboostLSS-AR(1) models in the μ and σ parameters at buoys 1, 2, and 3 without transformation of rainfall.*

similar pattern of seasonal effect in the μ parameter is for buoys 2 and 3. In addition, to remove autocorrelation and transformation of rainfall do not change the patterns of time covariates.

5.6.9.2 Similarities Time Effects of GamboostLSS-AR(1) Model Fitting at Three Buoys with Transformation

Here, we experimented time covariates over different specification of the gamboostLSS-AR(1) models with transformation for buoys 1, 2, and 3. The results are recorded in Figures 5.51 and 5.52. In both figures, the annual and seasonal patterns do not change for both with and without transformations in the μ and σ parameters.

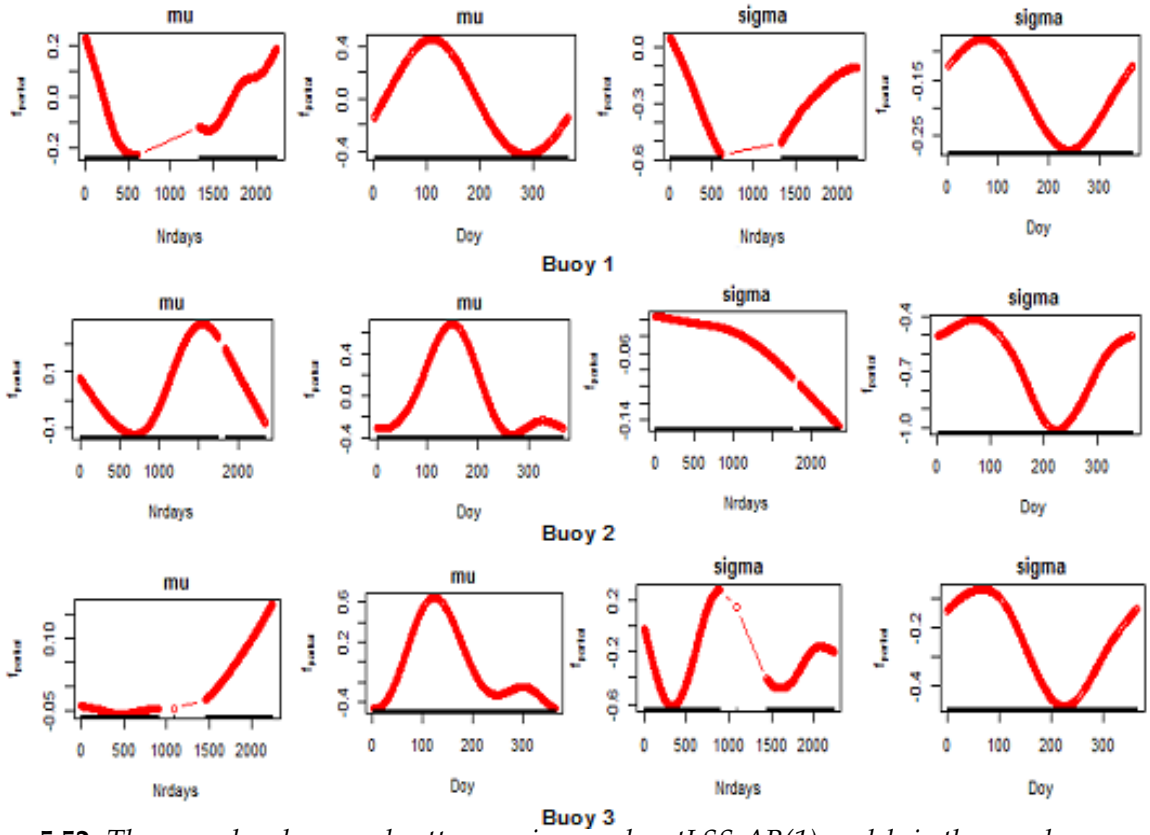


Figure 5.52: *The annual and seasonal patterns using gamboostLSS-AR(1) models in the μ and σ parameters at buoys 1, 2, and 3 with transformation of rainfall.*

5.7 Marginal Prediction Interval of GamboostLSS-AR(1)

In the previous section, we have discussed the annual and seasonal patterns in local fitting. The results explained on Section 5.6 need be to investigated further, particularly to predict the interval of time respect to SST. Therefore, a tool such as marginal prediction interval (MPI) is needed for this aim. In subsections 5.7.1 and 5.7.2, we presented MPI of the gamboostLSS models and subsections 5.7.3 and 5.7.4 for MPI of the gamboostLSS-AR(1), both models without and with transformation.

5.7.1 MPI of the GamboostLSS Models without Transformation

MPI was investigated in [20,86], for GAMLSS without considering autocorrelation in model fitting.

Here, we investigated MPI with autocorrelation at lag 1 of gamboostLSS-AR(1) model fitting. We particularly consider the median prediction to determine MPI. We applied gamboostLSS-AR(1) model in SST data using similar approach as explained in Algorithm 5.6.5. In this section, we investigate MPI of gamboostLSS models for both with and without transformations at all buoys. We used the fixed values of ν_{slf} and m_{stop} parameters. All results are presented graphically, whereas they numerically are not shown. It can be seen in the figures that the resulted models, which have different values of the step of length factor ν_{slf} and the stopping iteration m_{stop}, have the similar MPI patterns. Further we can see these similarities with step of length factor $\nu_{slf} = 0.01$ to 0.05, 0.1 and different stopping iteration m_{stop}. This is interesting because the different values of control boosting parameters do not change MPI patterns. However, we do not present plots of the MPI patterns because they are structurally similar to those obtained from the gamboostLSS model fitting. Further in our investigation, we consider 80% and 95% of confidence intervals for the MPI of the SST data for all buoys. The results show that the prediction of interval for the seasonal effects are appropriate fitting. It means that most of the data are covered in the range of the interval. This can be seen in Figure 5.53. The curves of MPI show the pattern of change over time (annual and seasonal effects) of the SST data. The MPI with 80% and 95% of confidence intervals gives insight that makes it possible to predict the level (in Celcius) of the SST data of annual and seasonal times ahead and to put realistic confidence bounds interval around those predicted. In order for the pattern of change around median of annual effects using gamboostLSS models from highest are SST data at buoys 2, 1, and 3, whereas the pattern of change around median of seasonal effects from highest are SST data at buoys 3, 2, and 1.

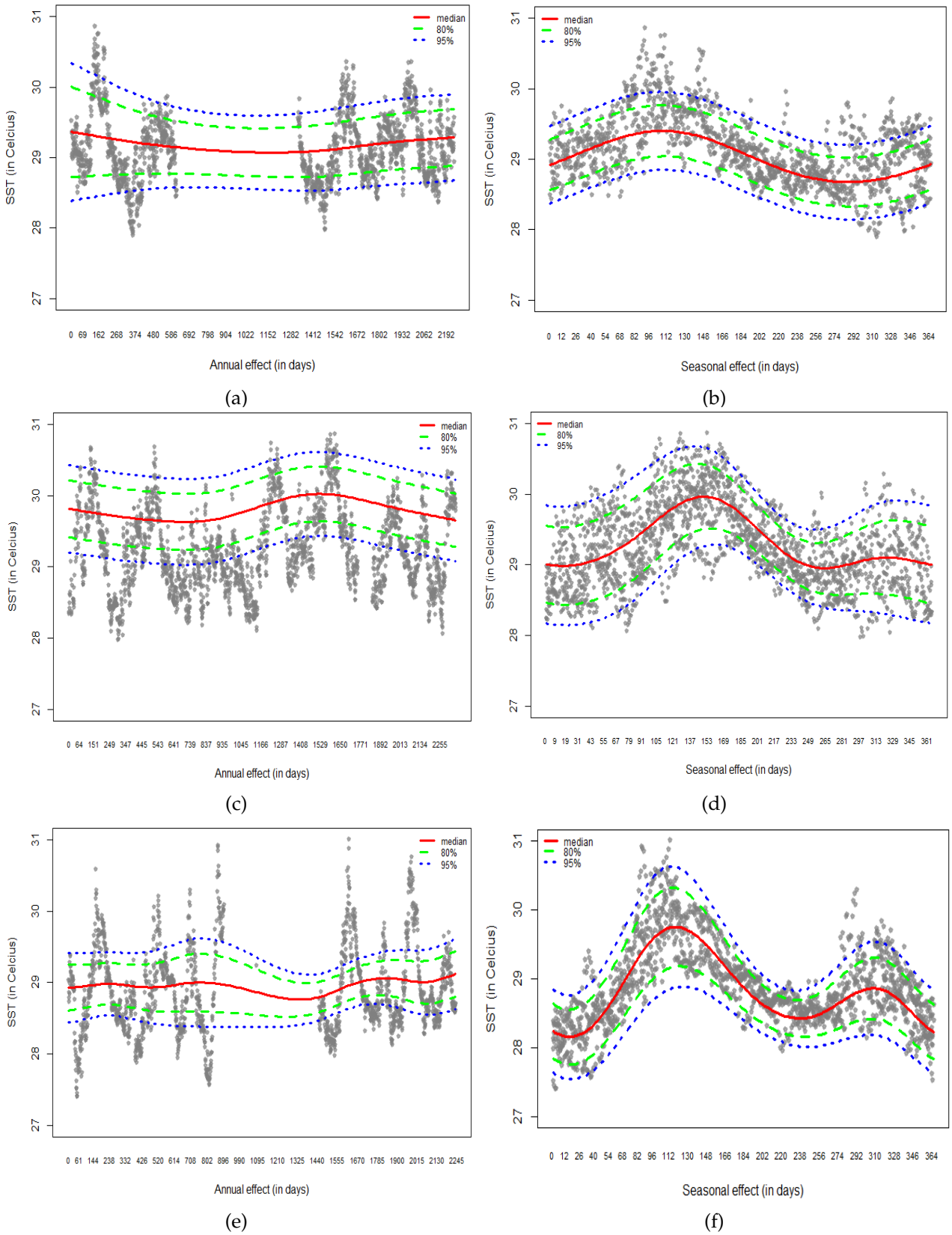


Figure 5.53: *MPI of the SST data fitting at buoys 1, 2, and 3 shows a similar pattern for seasonal effects using gamboostLSS models without transformation in the size of length factor* $\nu_{slf} = 0.01$.

Figures 5.53 show the highest of the seasonal peak on buoy 3 (figure f) as seasonal effect, however, the lowest has annual effect is shown at buoy 3 (closer to median line), (figure e). It means that buoy 3 has low effect in annual scale (long term) but not for the seasonal scale (short term). Moreover, the lowest seasonal effect is shown at buoy 1, whereas the highest annual effect is shown at buoy 2. In general, three buoys show the similar seasonal patterns with peak season around April and the second peak around October. All of these are interpreted as prediction interval over time in the gamboostLSS without transformation and without removing autocorrelation.

We discuss MPI of gamboostLSS models for both with and without transformations, particularly at all buoys. Similarly, 80% and 95% of confidence intervals for the MPI of the SST data at all buoys are as captured in Figures 5.53. The behavior of these figures are similar to the ones which are in the previous section. The annual effect and seasonal effect, for instance, have similar pattern although v_{slf} values are different. Interestingly, the seasonal effect curves forms bimodal and the fit of the annual effect seems to be shifted upwards. The MPI of annual effect at buoy 2 higher than MPI of annual effect at buoys 1 and 3.

Here, besides the figures have the same behavior as mentioned in the previous section, the seasonal effect curves in this particular buoy, have bimodal curve which are higher than the ones which are in buoy 2. Visibly, they covered more data than the previous one. It means that MPI in buoy 3 is more precisely fitting than in both buoys 1 and 2.

We conclude here that MPI using gamboostLSS models for the annual and seasonal effects of the SST data at buoy 2 is wider than MPI at both buoys 1 and 3. This is seen graphically, that the annual curves at buoys 1 and 3 have longer gap than the ones at buoy

2. In addition, the number of observations at buoy 2 is larger than those in buoys 1 and 3.

5.7.2 MPI of the GamboostLSS with Transformation

In this section, we observed MPI of gamboostLSS models with transformation for buoys 1, 2, and 3. The results of all buoys are captured in Figures 5.54. The results of MPI of the gamboostLSS models with transformation of the SST data at buoy 1, as seen in Figures 5.54 (a) and (b), show similar pattern with MPI of the same models without transformation as captured in Figures 5.53. In this section, we also observed MPI of the gamboostLSS models fitting with transformation of the SST data at buoy 2. The results of the MPI are captured in Figures 5.54 (c) and (d).

The results of MPI of the gamboostLSS models fitting with transformation of the SST data at buoy 3 are displayed in Figures 5.54 (e) and (f). The annual effect seems smooth fitting and to be shifted upwards, whereas the seasonal effect seems a bimodal form.

In general, the curves of annual and seasonal effects, as a result of applying MPI in gamboostLSS models with transformation, show slightly wider than the one without transformation. Transformation effect in MPI can enhance wider prediction interval of the annual and seasonal effects. This is more clearly seen in Figures 5.54 (e) and (f) which are the results at buoy 3. We do not explain in detail for each figure here. We suggest to refer to the previous section for the explanation. We conclude here that MPI of gamboostLSS with transformation is better than MPI of gamboostLSS without transformation in terms of the prediction of interval.

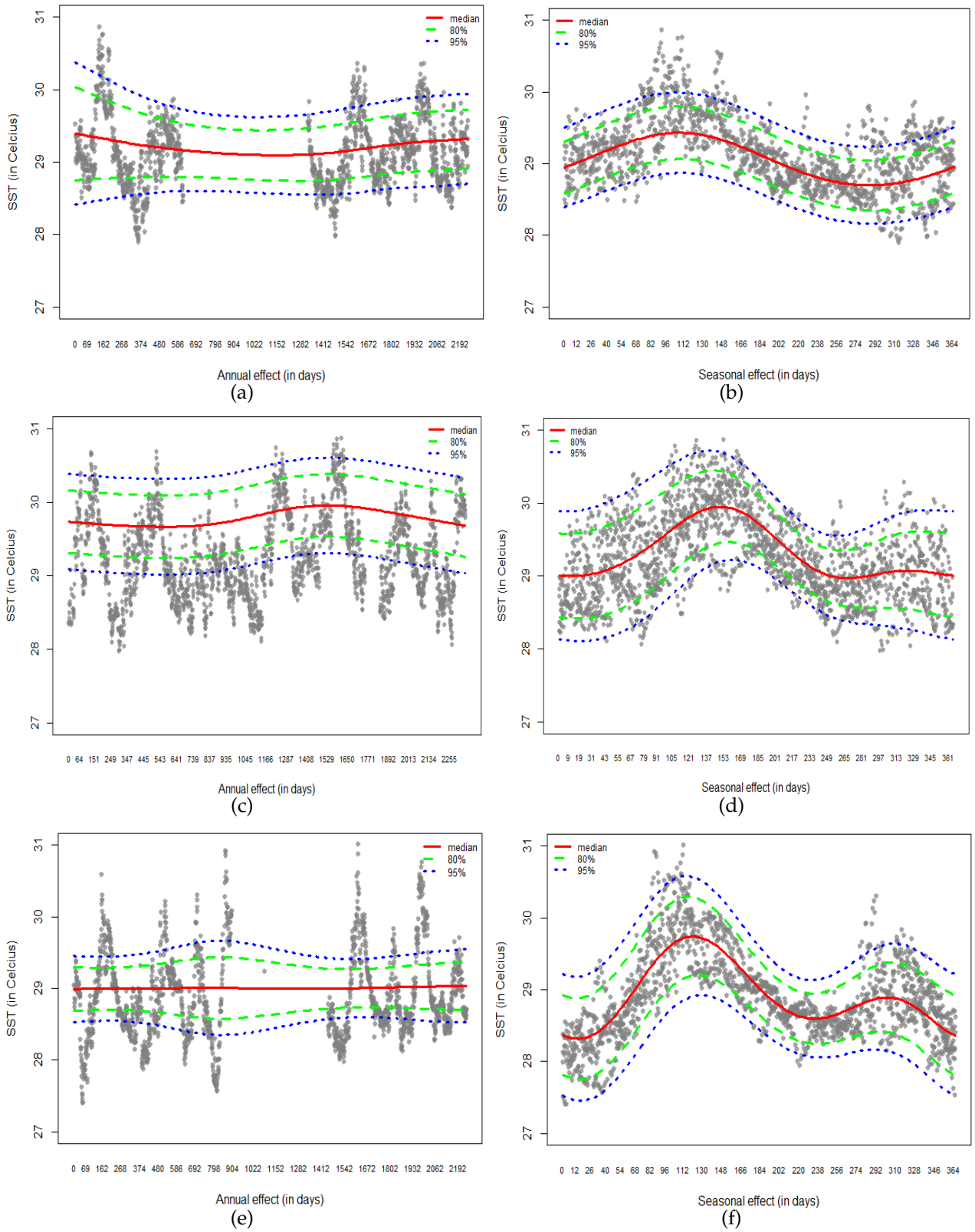


Figure 5.54: *MPI of the SST data fitting at buoys 1, 2, 3 shows a similar pattern of seasonal effects using gamboostLSS models with transformation in the size of length factor* $\nu_{slf} = 0.01$.

5.7.3 MPI-AR(1) of the GamboostLSS-AR(1) Models without Transformation

We investigated MPI of gamboostLSS-AR(1) which aim to remove autocorrelation in SST data. This approach is called MPI-AR(1). Similar to the previous section, we discuss MPI-AR(1) with and without transformation to compare the prediction of the interval. We also discuss MPI-AR(1) at buoys 1, 2, and 3. We compute the results of MPI-AR(1) for the SST data at three buoys, using autocorrelation coefficient ρ's which are found in Section 5.6.5. We use step of length factor ν_{slf} = 0.01 to 0.05, 0.1 and different stopping iteration m_{stop} to obtain MPI-AR(1) for each buoy. It can be seen in the figures that the resulted models have different values of the ν_{slf} and m_{stop}, have the similar MPI-AR(1) patterns. This is also interesting because the different values of control boosting parameters do not change MPI-AR(1) patterns. However, we do not present plots of the MPI-AR(1) patterns because they are structurally similar to those obtained from the gamboostLSS-AR(1) model fitting.

Furthermore, the results are presented using the size of length factor ν_{slf} = 0.01 as depicted in Figures 5.55. As can be seen from the Figures that the annual effects curves (see Figures 5.55, (a), (c), (e)) seem wider when the data are available. In other words, the curves of the missing data (gap) are closer to each other. The seasonal effects, on the other hand, remains the same as in MPI without AR(1). MPI-AR(1) of the seasonal effects at buoys 2 and 3 show a bimodal curve. In addition, Figures 5.55 (e) showing the unique patterns are seen in between the gap. The results of MPI-AR(1) without transformation of the SST data at buoy 1, 2, 3 are depicted in Figures 5.55 (a)-(b); (c)-(d); and (e)-(f) respectively. Removing autocorrelation effect on MPI-AR(1) without transformation shows significant effects.

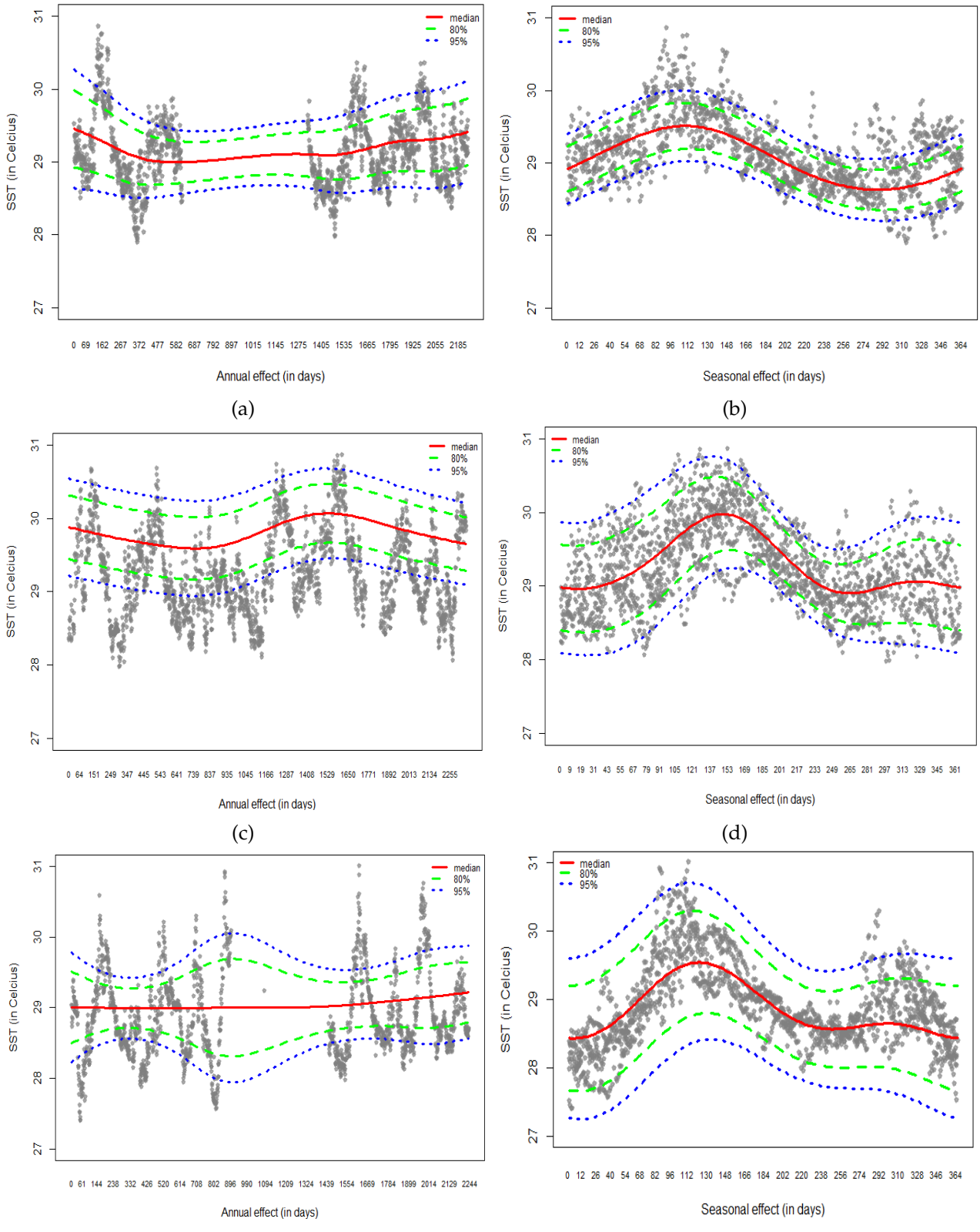


Figure 5.55: *MPI-AR(1) of the SST data fitting at buoys 1, 2, and 3 using gamboostLSS-AR(1) models without transformation, in the size of length factor* $\nu_{slf} = 0.01$.

5.7.4 MPI-AR(1) of the GamboostLSS-AR(1) with Transformation

To determine smoother curves of MPI-AR(1), we need to transform the rainfall data. In this experiment, we used the same autocorrelation coefficient ρ's at buoys 1, 2, and 3. We used the fixed size of length factor v_{slf} and different stopping iteration m_{stop} values. We compute the results of MPI-AR(1) with transformation for the SST data at three buoys, 1, 2, and 3, using autocorrelation coefficient ρ's which are found in Section 5.6.5.

As can be seen in the Figures, the annual effects curves (see Figures 5.56, (a), (c), (e)) seem wider when the data are available. In other words, the curves of the missing data (gap) are closer each other. In addition, Figures 5.56 (a), (c), (e), show that the unique patterns are seen in between the gap for annual effects, whereas for the same figures (b), (d) and (f) show MPI-AR(1) with transformation cover of the available SST data.

The results of MPI-AR(1) at buoy 1 are displayed in Figures 5.56 (a) and (b). The results show similar curves as those of MPI-AR(1) without transformation. The results of MPI-AR(1) at buoy 2 are presented in Figures 5.56 (c) and (d), and the ones at buoy 3 are presented in Figures 5.56 (e) and (f). They also have the same patterns as MPI-AR(1) without transformation at the same buoy.

We conclude that the transformation of rainfall can reduce the stopping iteration (m_{stop}) values more significantly, compared to the one without transformation. Reduce stopping iteration can affect in decrease empirical risk. We also conclude that the investigation of MPI-AR(1) at buoy 3 is more visible in terms of the model fitting results, compared with buoys 1 and 2. However, the seasonal effects do not seem to be better since the curve did not follow the pattern of data. Particularly, the curve did not reach the peaks of the data.

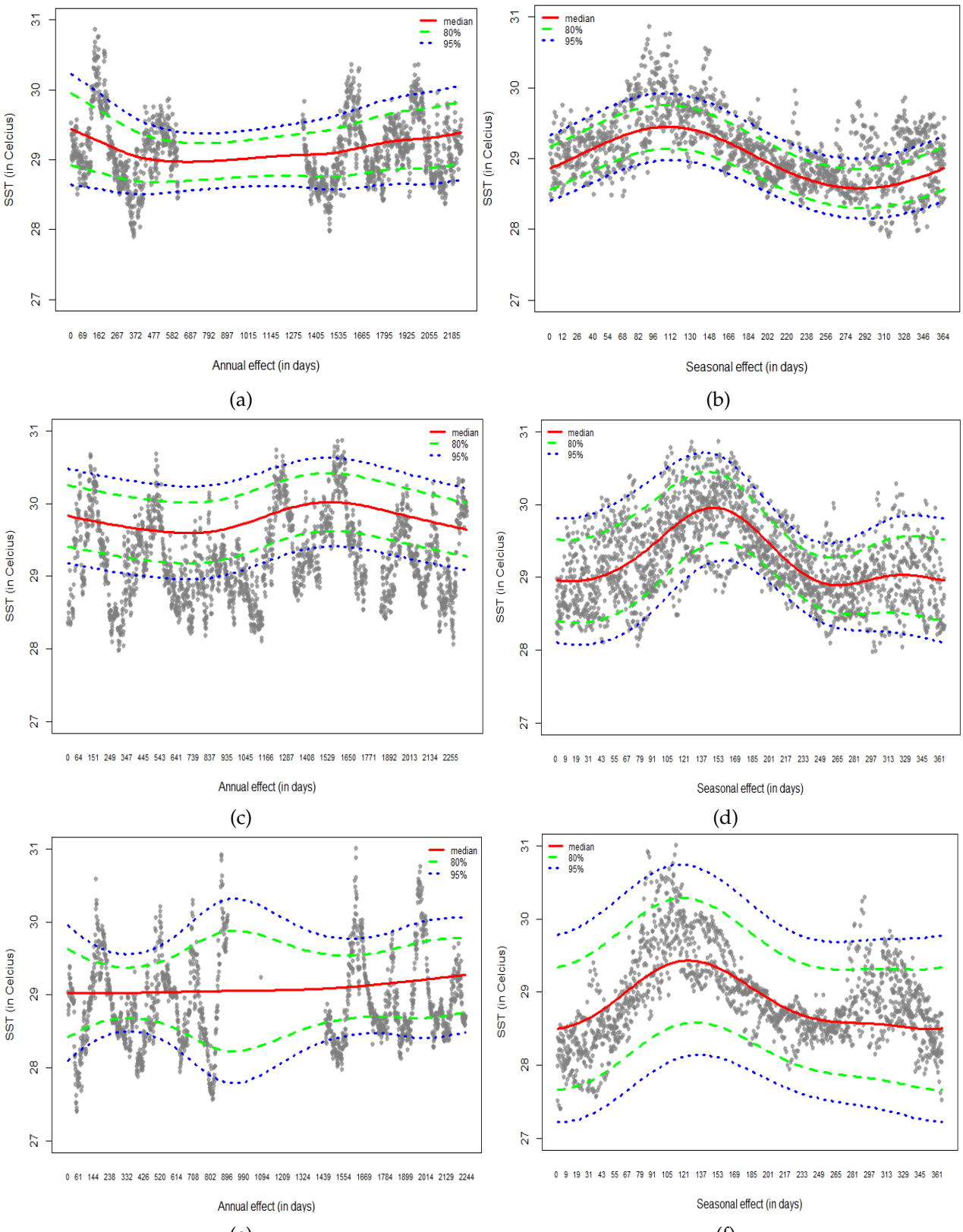


Figure 5.56: *MPI-AR(1) of the SST data fitting at buoys 1, 2, and 3 using gamboostLSS-AR(1) models with transformation, in the size of length factor* $\nu_{slf} = 0.01$.

5.8 Summary

The purpose of this chapter is to present gamboostLSS-AR(1) models for the SST data fitting. The proposed models take into consideration the autocorrelation in the SST data. We applied generalized differencing technique to reduce time autocorrelation in the SST data. The gamboostLSS-AR(1) and gamboost-AR(1) models are presented for the SST data fitting. Our experimental results demonstrate that gamboostLSS-AR(1) models provide more appropriate data fitting with a larger number of submodels than gamboost-AR(1) models with or without transformation.

From our experimental results, it can be concluded that by removing autocorrelation from the data, an appropriate model fitting can be achieved with reduced CV-risk. This can be done by using generalized differencing and/or transformation covariate of the SST dataset. Application of the gamboostLSS-AR(1) models for the SST data fitting has a similar pattern between with and without transformation. The results of transformation of rainfall in the gamboostLSS-AR(1) model show a reduced final risk. In addition, the transformation also gives a larger number of submodels in the local model fitting than the non-transformed data for rainfall.

In general, the autocorrelation leads to misfitting of the data, particularly in sparse data. The model fitting process becomes more complex in the presence of autocorrelation in the data. Moreover, the fitting process becomes bias due to time sequence in correlated data. Optimized cut-offs can be one way to reduce bias in the fitting process. Local fitting can also be considered to gain information for global model fitting. Removing autocorrelation contributes to reducing bias in the model. This evidence shows that gamboostLSS model

fitting by AR(1) approach can help to obtain an appropriate SST model.

Furthermore, we observed that removing autocorrelation errors with an AR(1) model has a large impact on global and local model fitting and also on modelling the error, especially derived from time covariates. The gamboostLSS-AR(1) models have powerful procedure to distinguish between plausible alternative solution in model fitting. Therefore, our suggested gamboostLSS-AR(1) models are an advanced technique for removal autocorrelation. The proposed model can also be applied to the other complex data sets. It is especially useful in situation where the data have various gaps, sparsity, irregular peaks and autocorrelation.

Furthermore, we proposed gamboostLSS-AR(1) models. We applied the model into the SST data from three buoys of various positions. We compared the gamboostLSS-AR(1) model with the existing gamboostLSS model at the buoys. The results show that the gamboostLSS-AR(1) gives smoother model fitting than gamboostLSS models fitting of the SST data.

We also investigated the application of MPI in both gamboostLSS and gamboostLSS-AR(1), with and without transformations. In general, the 80% and 95% of confidence interval for the MPI and MPI-AR(1) of the SST data for each buoy reveals similar patterns of the annual and seasonal effects. The results show that transformation affects the range of the MPI and MPI-AR(1) by using gamboostLSS and gamboostLSS-AR(1) models fitting.

The results of MPI-AR(1) are better than MPI in gamboostLSS model fitting for the SST data. The fitting of MPI-AR(1) model can follow the pattern of the SST data smoother than the fitting of MPI. From our investigation of MPI-AR(1) at buoy 3 is more visible in annual compared with buoys 1 and 2. The seasonal effects, however, do not seem better since the

curve did not follow the pattern of data, particularly, the curve did not reach the peaks of the data. Therefore, we suggest to adjust the hyper-parameters to make the fitting better. MPI-AR(1) can be reached by transformation and removing autocorrelation to estimate the optimal prediction interval. For further work, we suggest to use MPI-AR(1) to determine marginal prediction interval of time covariate in gamboostLSS model fitting by considering autocorrelation.

Chapter 6

General Discussion

6.1 Introduction

In this chapter, we started our general discussion of some findings of specific methods for linear to gamboostLSS-AR(1) models fitting of the sea surface temperature data. Although assumption of these methods is different in structural term for the unknown regression function, but they have the same assumption for Gaussian distribution in our experiment. We presented seven related methods: linear regression models (LRM), generalized additive models (GAM), GAMLSS, gamboost, gamboost-AR(1), gamboostLSS, and gamboostLSS-AR(1) models as in Chapters 4 and 5.

From the experiment results, we are trying to delve into gamboostLSS-AR(1) models fitting for SST data by considering autocorrelation through fitting performance. Typically, the model fitting of the SST data will be appropriate when there is high appropriate time effects in submodels (called local fitting). Although the model is not yet overall an optimal model fitting, we have several interesting results associated with our experiment.

6.2 Linear and Additive Models Fitting

In our experiment, we started by experimenting a simple model and small set of SST data. M1 model is useful for preliminary identification of patterns and trends in the SST data such as seasonal and annual effects as in Chapter 4 (see Figure 4.1 and 4.2) by using linear regression model fitting. Further by fitting GAM (see Figure 4.5), gamboost (see Figures 4.6, 4.7, 4.11, and 4.12), gamboostLSS (see Figure 4.17), gamboost-AR(1) (see Figures 5.5, 5.7) and gamboostLSS-AR(1) models (see Figures 5.12 and 5.17) of the SST data at buoy 1 shows that seasonal effects demonstrates strong similarity pattern and trend as depicted in (Figure 3 in [87]). This statistical evidence shows that different periods and positions, between 1961-1990 and 2006-2012, and at 5^oS-5^oN, 150^o-90^oW in the Nino3 region in the Pacific Ocean and at 4N90E in the Indian Ocean respectively can produce the same pattern and trend of the SST data. Although our model is focused to obtain pattern and trend of seasonal effects with respect to SST variability, whereas the model that represented in Figure 3 in [87] is the same concern for seasonal effects but it related to magnitude of SST data.

Therefore, we can use our model fitting albeit in small datasets, limited area and short time period but the model can reveal annual cycle or seasonal effects, strong condition dry and wet episodes with large data sets, long time period, and the larger area. Different scope data from small to large can be used to gain models fitting with sophisticated method as discussed (see Breiman [88]). The pattern and trend also shows that there is relationship between condition of sea surface temperature in the Pacific and the Indian oceans. Although it is a different time period and region (or called time-spatial), it has similarities and cyclic

phenomena properties. In addition our model is more aesthetically smooth than the model fitting as in (Figure 3 in [87]).

6.2.1 P-splines basis in various gaps

P-splines has 14 pros compared to similar class smoother (e.g. polynomial, cubic spline) and handle fitting to semiparametric models [36]. This basis is adaptable to fit SST data with our models used in the experiment, mainly in gamboost-AR(1) and gamboostLSS-AR(1) models fitting. We assumed that hyper-parameters specification in continuous covariates fixed, i.e., degrees of freedom is 1, the number of knots is 20 (default), degree is 3, and differences is 2 (see Algorithm 5.6.5). We used this assumption to investigate practical selection of hyper-parameters and also to easily observe patterns and trends of seasonal and annual effects in the model fitting.

For base-learners with smooth functions we used the 2nd order differences, where it is used as a penalty for continuous covariates that related to 2nd derivative of the spline. Whereas degree of the regression P-splines is 3. The gamboostLSS-AR(1) model can dampen nonsmoothness of the fit and interpolate the data in various gaps, where the model used P-splines basis although we need efforts to reach the peaks of the SST data. The results of this specification can capture time effects in local model fitting. However, to obtain optimal performance in global model fitting then another setting for continuous covariates is required.

6.2.2 The Degrees of Freedom

Degrees of freedom is one of the hyper-parameters rapidly changing smoothness in gamboostLSS-AR(1) model fitting. Each term of the additive model uses the degree of freedom (*df*) that correspondence to the number of parameters, where it associated with trace of hat matrices. In [48] related to residuals stated that RSS is not adequate to become a selector of the model, so that we used cross-validation risk to compute residuals, where we also used in the hat matrix (see Section 3.3 Chapter 3, as refer in [45]).

We can see the number of *df* that used in GAM models fitting with and without transformation (see Tables 4.11 and 4.14 in Chapter 4) is larger than gamboost models fitting (see algorithm 2.3). Then the *df* values used in gamboost models is similar gamboost-AR(1) models fitting (see Table 5.1 in Chapter 5) in the same scenario (with and without) transformation. Whereas the *df* values used in gamboost-AR(1) models fitting is larger than in gamboostLSS models (see sections 4.8.1.1 and 4.8.1.4 in Chapter 4). Similarly, the *df* values used in gamboostLSS-AR(1) models fitting (see sections 5.4.4.1 and 5.4.5 in Chapter 5) is smaller than in gamboostLSS models. Low *df* in the gamboostLSS-AR(1) models gives less wigglines in the model, especially the *df* of time covariates before and after the gaps reduces variance in the model.

6.2.3 The Knots

As our experiments cannot directly obtain the number of knots in gamboostLSS-AR(1) model fitting with the use of P-splines basis, when the position of the knots are selected, we use adjustment method. Previously we are not using knots in GAM model fitting due

to longer computational time needed to get the *df* composition in the model. In addition, the composition of *df*'s already represents, the patterns and trends of time effects in the GAM model fitting (see Section 4.5). Hence, in GAM model fitting we do not use knots immediately.

Initially, we use 100 knots in gamboost models for *Nrdays* covariate. Then we increase the number of knots about 20 till 140 in with and without transformation (see Section 4.6). Similarly for GAMLSS and GAM models fitting, they have computational time problem. We start by using knots (or called ps.interval) 20 for each covariate in the GAMLSS model without transformation. Increasing knots 100 in the *Nrdays* covariate is to accommodate the gaps of the global model fitting. This number of knots is also used in the location, scale, and shape (LSS) functions. By increasing knots we can obtain changing positions of the knots effects with respect to SST data fitting. Unfortunately, the model is still not revealed to visualize submodels of time covariates. However, the experiment shows very significant results to raise *df* and reduce *AIC* values (see Tables 4.19 and 4.20 as in Chapter 4).

Interestingly, the number of knots that were used in gamboostLSS model fitting is lower than in gamboost and GAMLSS models fitting. In addition to low knots (40-60) in the model fitting, effects of the knots in the model with and without transformation does not change patterns and trends of time covariates in the μ and σ parameters (see Figures 4.18 in Chapter 4 and Appendix D.4). In addition, the number of knots (40-60) guaranties flexibility of gamboostLSS-AR(1) models. The structure of the knot locations for SST data consists of 40-60 equidistant grids points. Therefore, this specific range of knot values can be used to obtain basis construction for appropriate model fitting of the SST data.

6.2.4 Transformation and Stability

Transformation of the response can be used to stabilize the variance of the regression model fitting [69,89]. Our study shows that stability of variance comes from transformation of the rainfall covariate. Although our model shows that model diagnostic in Chapter 4 section 4.3 transformation of the response is suggested, in this experiment we are more focused to explore covariates effects (i.e., time effects and rainfall covariate) with respect to sea surface temperature variability. There are reasons to choose rainfall covariate to be transformed, such as large of range value, leverages of the rainfall covariate, and large number of the data are zero.

In other words, transformation effect gives stability in gamboostLSS-AR(1) models fitting. Although, the experiment results show that global model fitting of SST data does not reach the peaks fully. Transformation of rainfall can prevent outliers so that it is not affecting SST model fitting. Transformation does change pattern and trend of rainfall itself and it does not change another submodels in model fitting, for example, transformation effects see Figures 5.29 and 5.31 in Chapter 5 from buoy 1. Therefore, temporary removal of outliers is needed in addition to transformation before model fitting. Now we have three scenarios related to outliers in model fitting of the SST data, i.e. transformation of rainfall, temporary removal of outliers, and combination of both.

Further we need to localize or impute outliers whether they gives affects to global and/or local models fitting. So that we can obtain structure of residual autocorrelation AR(1) before and after localized outliers.

For P-splines base-learners in the continuous covariates we use degree of the regression spline 3, where the cubic model has bias from estimate an additional parameter, its

prediction has slightly random variation so that less precise compared to fitting from the quadratic model with degree of the regression spline 2.

6.2.5 Different Measurements

In observational experiment like SST data we found many different measurements on the response and/or covariates (see Sections 1.1, 2.1 and 2.2 in Chapters 1 and 2 respectively). Different measurements can be related to times, positions, tools, magnitudes, instruments, scenarios, types and sizes of data, scales, and so on. These measurements are giving variability effects to the data structure. Therefore, there is distribution of the response in covariates effects.

Regarding variability effects we can reveal the data in the form of location (μ), scale (σ), and shape (τ) parameters with and without transformation as seen in figures in Sections 5.4.4 and 5.6.5, Chapter 5. Therefore, there are variability in pre-fitting from the data and in fitting process from hyper-parameters specification in the model. Consequently, a procedure for pre-fitting of the data and current fitting is essential steps of the model fitting process, see Section 2.3 in Chapter 2. GamboostLSS-AR(1) model can accommodate various effects due to these differences.

6.3 Robustness of GamboostLSS-AR(1) Models

Autocorrelation AR(1) process in the gamboostLSS-AR(1) model fitting makes the model more robust over time. Here, robust is in the context of structure best fitting in the majority of the SST data. Outliers can affect the gamboostLSS-AR(1) model when anomalous data of

rainfall covariate has large difference between minimum and maximum values for a large number of observations. The evidence shows that this residual outlier is very small or has insignificant effect on the LRM model (see Chapter 2, section 2.5 and 2.6), where by [90] these outliers are include X- and Y-outliers, i.e. leverages and residuals respectively.

One of the advantages of AR(1) type in the gamboostLSS-AR(1) model fitting is contributed in the development of robust model. Tables 5.8 and 5.15 (see Sections 5.4.5) and 5.6.6 in Chapter 5) show the same number of submodels with similar pattern on global fitting, thus indicating that there is a constant conditional variance supported by the data fitting via autocorrelation AR(1) model. Tables 5.8 and 5.15 in Chapter 5 show that there is 8 submodels of 1231 complete dataset and 13 submodels of 1460 complete dataset. Robustness of gamboostLSS-AR(1) models depends on the number of submodels, autocorrelation coefficients, hyper-parameters in the models and its complex data structure.

6.3.1 Boosting and Autocorrelation Effects

In addition to boosting and autocorrelation approaches, we use function estimation by time effects approach in gamboost-AR(1) and gamboostLSS-AR(1) models fitting. The time effects is important variables in both models, but also the functional form of the dependence on these variables. The implementation of this methodology provides a very broad investigation into the properties of this approach in the context of extended gamboostLSS models with autocorrelation.

Our research shows that presence of autocorrelation errors like AR(1) process through gamboostLSS-AR(1) models fitting can reveal the annual and seasonal effects of the SST variability. In this model, additive does mean that the model fitting is formed by an additive

combination of base-learners with considering location, scale, and shape functions and also autocorrelation AR(1). In [89] stated that standard deviation decreases when we consider AR(1) to be computed in model fitting. In contrast, although the gamboostLSS-AR(1) model is using several approaches, we can see that the SST data fitting physically is not optimal fitting. Model fitting does not reach the irregular peaks, which implies indicator of the model is not an optimal fitting performance.

We have coined these terms (gamboost-AR(1) and gamboostLSS-AR(1) models) because our results apply to a group of estimation problems for autocorrelated data. While our initial motivation for introducing the gaps were to achieve an optimal fitting, we found that this framework also allows us to improve in reaching peaks in irregular data of the associated methodology like with interaction among covariates and transformation of the response. Note that to develop interaction function we suggested to avoid the effect modifier for time covariates as highlihted in our experiment related to autocorrelation errors AR(1) model.

6.4 Balance in GamboostLSS-AR(1) Model Fitting

The results obtained here suggest that the interaction among covariates are needed to increase R-square in linear models fitting or deviance explained and decrease AIC in additive models. There are several reasons to interact components, firstly, each additional coefficient parameter estimated adds to the variance of the LRM model. Secondly, in additive models, additional linear combination of estimators adds to the variance of the model. By adding one or more functions via the interaction, it can increase deviance

explained or decrease AIC values of the model. While the proportion of the variation in the response variable can also be explained by the covariates. Interaction components can be constructed by relationship between continuous covariates and/or between continuous and time covariates. However, it does not guarantee that an additivity model provides the best fit of the data [89].

Initially, we assume that climate data have relationship with each other and mutually dependent, including sea surface temperature (see Chapter 1). In addition, there are similar properties among climate features, for example, humidity and rainfall, where humid and rain are related to dew and liquid of precipitation, respectively. Therefore, interaction among climate features of SST data can be added in gamboostLSS-AR(1) models fitting, so that they are given joint and individual effects in fitting process.

A balance between goodness of fit and parsimony of the gamboostLSS-AR(1) model is needed. Although better fits data can be achieved by adding more functions (or parameters), but simplicity and interpretability in detail also lead to more precise model fitting and prediction. Moreover, in the model fitting, submodels selection are also needed to variance reduction and simplicity without less accuracy fitting. However, it leads to increased bias. Boosting and AR(1) techniques can be used to fit, predict, and select of variables simultaneously.

There are evidence to understand SST variability over time in the Indian Ocean by using gamboostLSS-AR(1) models fitting. Although our experiments using small dataset and limited period time, we obtain a benchmark model to deal with large dataset and longer time period with complex data structure and high dimensional data. SST observation shows that irregularity with various gaps of missing data can be dealt with fitting complex

models such as neural networks as suggested in [91], yet this model cannot reveal LSS of the data.

With the extension to identify pattern of time covariates in model fitting as in Section 4.2 Chapter 4, we can construct a regression tree instead of the stepwise regression. The reason to construct the model is to handle large number of observation and high dimensional data in finding the best composition of regression models, for instance, by 2^p submodels where p is the number of parameters in the stepwise regression and $2p - 1$ sweeps in the regression trees [92].

6.5 Seasonal and Annual Effects in GamboostLSS-AR(1) Model Fitting from Different Buoys

For seasonality by SVM and NN have reported in [93], also both methods related to sea surface temperature anomaly (SSTA) [94]. However, both methods only show smooth curve fitting and not do consider location, scale and shape parameters.

As depicted in Figures 5.51 and 5.52 in Chapter 5 with and without transformation, seasonal effects as shown at buoy 1 for μ and σ parameters where the sinusoidal curve of sine function. Whereas at buoys 2 and 3 in μ parameter show first and second peak seasons. This approach has pros not only to follow the pure sinusoidal curve but also it can be constructed to follow data structure and complex nature. Flexibility of GAMLSS using gradient boosting and P-splines basis with cyclic function can capture cyclic phenomena like seasonal effects [20, 33, 36, 37, 45] as depicted in the mentioned parameters and prove that gamboostLSS-AR(1) by considering autocorrelation is flexible model as well.

From three different buoys position where the parameter μ of the seasonal effects show that the highest peak season in the Indian Ocean for 2006 to 2012 period is around 100 days (on April) and the second lower peak season is at 300 days (on September). For annual effects, the figures show similar increase after the gap at buoys 1 and 3 in μ and σ parameters. In general, three buoys give information that the patterns and trends of seasonal and annual effects of sea surface temperature, with and without transformation of rainfall scenario by gamboostLSS-AR(1) model fitting do not change the patterns and trends in LSS parameters. The variability of sea surface temperature of different buoys represent that various positions and local weather between land and sea have similar condition in seasonal and annual effects.

We found that by association ENSO phenomena via oceanic nino index (ONI) in the Pacific Ocean (PO) as depicted in the Table from *ggweather.com/enso/oni.htm* is not interpretable. However, when we association ENSO with variability of sea surface temperature using annual effects via gamboostLSS-AR(1) models fitting is more interpretable. For example, SST data from using buoy 1 in the Indian Ocean (IO) from 2006 to 2007 is weak El Nino in the PO, 2007 to 2008 is moderate La Nina in the PO, previously similar of 2010 to 2011, and weak La Nina in the PO (see Figures 4.2 for linear model and 5.51 in μ and σ parameters for gamboostLSS-AR(1) model). The 2007 to 2008 period has similar situation with 2010 to 2011 period but it has different trends and directions. By visualizing the variability of phenomena as represented with annual effects, they are more interpretable than with the Table. Nevertheless, we still need more buoys to interpret this phenomena via LSS parameters. Therefore, MPI-AR(1) models fitting can be used to predict future levels of the SST data over time, albeit with various gaps, sparsity, irregular peaks, and

autocorrelation.

6.6 The GamboostLSS-AR(1) as a Benchmark Model Fitting

The results of the previous section indicate that the gamboostLSS-AR(1) models can be a benchmark model fitting by location, scale and shape (LSS) considering autocorrelation. The LSS functions do not only reveal information behind the available SST data but also deal with the gaps, so that the model requires more complex fitting and prediction methods. However, in its application this approach encounters some challenges to produce optimal fitting performance. There are several reasons for our experiment to use these approaches as complimentary to our model, such as:

(a) Support Vector Machine (SVM) for regression (SVM-R):

SVM (or called support vector network, SVN) is one of the supervised machines learning, where the large number of experiment is applied to climate data. SVM-R has a structural form as additive model with some constraints. It can be constructed to semiparametric models by using hyperplane base. The model can deal with sparse data, error models, gaps between clusters [95], incorporate penalization technique, large datasets and class [41, 96–98], and also large optimization task, handling large scale of linear and nonlinearity. There are similar properties between gamboostLSS-AR(1) model and SVM, such as SVM is constructed to minimize empirical risk (ER) in loss function and overall risk (OR). It also is robust, does not take computational cost, handling high nonlinearity relationship, flexible, and capable in predictive accuracy. It also is robust noise due to SVM able to compress outliers [99].

As we consider autocorrelation in gamboostLSS-AR(1) models, SVM with autocorrelation is represented to least squares-SVM (LS-SVM) [100], where it accommodated by linear AR-AR(X) and nonlinear autocorrelation NAR-AR(X). For instance, neural network (NN) considers autocorrelation in applying the global climate model (GCM) via artificial neural network (ANN) and (ARIMA) models [101].

(b) Random Forests for regression (RF-R):

Random forest for regression (also called RF or regression forest) is a machine learning technique that contain a set with elements of individual regression trees (see subsection 6.6 (c)), [102]. The model has several excess, such as conditional variable importance as in gradient-boosted, variable selection for parsimonious prediction model, complex interaction, deal with outliers, handle $n < p$ where n is the number of observation and p is the number of parameter, less misfitting, accuracy to detect bias and noise, capable to prediction point, and so on [41, 103, 104].

A method for regression based on the combination of tree covariates as a classifier with independent identical distribution (i.i.d) random vector. Two strengths of random forest are that each individual tree is a classifier, and there is relationship between them (correlation) in the forest. Pros for random forest, such as it relatively robust for outliers and noise detection, computationally efficient than bagging or boosting on large data, can be parallelized, combine trees, data prediction, partial and multidimensional scaling plots, and has several measurement of error, strength, correlation and variable importance.

RF-R can deal with large decision trees, estimating missing data, and comparable to

boosting in error rate. It is also related to nonlinear classifiers, such as SVM-R and artificial neural nets (ANNs) and has trees level so that it can be avoid misfitting to training dataset. RF-R is potential to interpret data by exploring them graphically [95].

(c) Regression trees:

Regression trees has similar structure form as additive models, where the model can easily deal with missing data via partition trees and estimate the best linear combination by splitting technique with different variables and produce tree classifier. The model can be incorporated into gradient boosting technique [41, 69, 95].

GamboostLSS-AR(1) model can accommodate trees effects of covariates, as inherited from gamboostLSS model as in [105]. The model can also potentially cover various types of covariates, gaps (missing data), nonlinear and interaction relationships. Whereas boosting is the cleverest mean of trees compared to random forest, there are weaknesses with regression trees, such as large trees are difficult to interpret and predict feasible performance results [67, 106].

(d) Neural networks for regression (NN-R):

Neural networks have similar structure with regression trees but it has multi-layer perceptrons (MLP) as base. They are models with pros as follows: able to handle nonlinear relationships in highly interconnected setting (weighted connections), interactions among variables, simulated via neurons for inputs, outputs, and feedbacks. This model is better to deal with complex structure of model, multitask, and multisystem networks. Neural networks have also similarly to our model as additive models. The model can deal with large class and hidden layers, where it can also be

incorporated with boosting technique [41,69]. Among the cons of neural networks is that it can be adjusted with weights for each input to optimize output at each neuron, complex computationally, and sometimes need parallel computing.

(e) Decision trees for regression (DT-R):

This model is very common to be applied in climate data with large attributes. Decision trees and RF-R have structure relationship (see subsection 6.6 (b)), where each of the decision develops the forest. Decision trees in regression context are the process of predictive outcome which allow both qualitative and quantitative simultaneously, aimed to obtain optimal decision trees of a dataset. This approach has similarities with our model where it can handle SST data properties, but it tends to weak for interactions (mixed models), sensitive for noise and irrelevant attributes, and handle nonlinearity in various types of data [41,107,108].

There are steps to make decision trees, such as growing, splitting, pruning, and tree selection. These steps growth with many decision inducers, e.g. ID3, C4.5, M5, RETIS, CART, CHAID, QUEST, CAL5, FACT, LMDT, T1, PUBLIC, MARS, FDT, SDT, etc, where several inducers can capture issues of SST data. For example, decision tree by using C4.5 technique has been used to sea surface temperature regarding tropical cyclone construction with several attributes of climate data [109]. We expected that decision trees can handle large attributes and datasets that commonly used in climate data (e.g. sea surface temperature) as our experiment context.

(f) Fourier transform:

To represent cyclic phenomena and errors pattern, we can use fourier transforms with

splines expression as seen in [110]. In [74,111] modulation model with P-splines is used to represent seasonal patterns via exponential, sine and cosine functions. These functions can be incorporated to additive models. Advantages by this approach can be controlled by convergence, error bands, efficient, flexible smoothing, and computing AIC. It can also detect trends, frequency, phase, and amplitudes of seasonal model via trigonometric functions.

In model fitting, one of the requirements to model SST data smoothing is specification of the basis function that is used to construct smoothness curve. The curve can come from non periodic and periodic data structures. In [112] state that basis functions with non periodic curves can use splines basis and periodic curves with Fourier series. While in our experiment, to construct periodic model such seasonal factor as P-splines basis is used as cyclic penalties [36,37,45,74]. The model applicable to reveal variability of seasonal effects is depicted in Chapters 4 and 5.

6.7 Summary

The analysis of gamboostLSS-AR(1) model fitting with respect to sea surface temperature data from different buoys in the Indian Ocean and climate data from stations in the Sumatra island shows that it is still not an optimal fitting performance. However, the model is capable to describe the phenomena of sea surface temperature variability in seasonal and annual effects via location, scale and shape parameters. The model shows using low hyper-parameters for degrees of freedom and knots. There are advanced of the gamboostLSS-AR(1) model fitting, such as transformation and stability, robustness,

boosting and autocorrelation effects, and balance in the model.

Two points of the model have not yet reached peak values of the SST data, first, structure of the model is using configuration without interaction among covariates. This situation automatically given a construction of the model is also without interaction in functional term. Second, the model is sparsely complicated by missing data (the gaps) due to informative dropout in time period. Whereas measurement times are regular grid points of sea surface temperature observation. This condition can be affected by span of P-spline smoothing curve such as influences with respect to intercepts and slopes of model fitting.

In the general finding we obtain a benchmark model fitting (called gamboostLSS-AR(1)) that reveals phenomena which caused by sea surface temperature variability over time (i.e. seasonal and annual effects) with focused on estimating location, scale and shape parameters. The finding of marginal prediction interval with respect to autocorrelation is also given contribution to predict interval data for each covariate in local model fitting.

The popular methods for regression, such as SVM, random forest, regression trees, neural networks, decision trees and fourier transform are not specific to LSS functions simultaneous as in gamboostLSS-AR(1) models. Throughout this discussion, our emphasis has been on the practical aspects of LSS function in gamboostLSS-AR(1) with autocorrelation context. Finally, we mention some modifications and extensions that have been applied to the gamboostLSS-AR(1) algorithm and discuss the aspect of optimal fitting performance and estimate the gaps from a LSS point of view.

Chapter 7

Conclusion

7.1 Conclusion

In this thesis we have given a detailed study of the various factors that are influencing sea surface temperature (SST) thus indirectly affecting earth's climate variability over time. Several statistical models have been used to correctly identify the patterns of features involved in the sea surface temperature data. These factors are represented by five features (air temperature, relative humidity, rainfall and two time covariates) that are considered in this work. The main issues faced while analysing the data were the incomplete nature of the data, sparsity, irregular peaks, autocorrelation and periodicity. A step by step review of the work done throughout the thesis is summarised in the following lines.

We started with identification and analysing the effect of time covariates on the SST dataset by simple linear models. To this end we investigated seasonal and annual effects of time covariates on the response (SST). Initially we used linear regression models (LRM) to identify the basic effects of the covariates on the response. From this we observed that

both the seasonal and annual effects are significant. However, statistically the seasonal effects are much higher than the annual effect. The LRM models could only give us information about a single pattern corresponding to each of the covariates. In the case of seasonal effects, the fits revealed that SST is high in some months (March-June) while low in the others. Sea surface temperature was seen in increasing pattern from December to April (with peak value) and decreasing pattern onwards until August. After August the decrease in the pattern is very low. Therefore, there is a large variability of seasonal effects in December to April (increase), in April to August (decrease), and little variability in August to December. The measures of R-squared and adjusted R-squared is 59.18% and 58.54% respectively for the LRM model fitting. Similarly in the case of annual effects, both increasing and decreasing patterns are seen, however, due to the incomplete nature of the data, strong statistical arguments can not be established.

To fix the issue of incomplete data, we moved on to other sophisticated statistical models to model the data that could cater for incomplete structure of the data. Along with using penalised splines (P-splines), we used generalised additive models (GAM) to fit the data. Our experiments reveal that better fits were obtained as compared to LRM by covering linear and smooth effects of the covariates by smoothing splines. However, it is hard to correctly identify the patterns of annual effects using GAM models. Other problems related to fitting the GAM models with P-spline functions we faced were pertaining to AIC (indirectly affecting the fitting) and computational cost in case of having higher degrees of freedom in determining degree composition of covariates. In addition, fits obtained were wiggly and had a large amount of fluctuations in the patterns for long gap observation, especially in those of time covariates. Our experimental results using GAM model fitting

show an increase in explained deviance to 65.5% from 59.18% and *df* 18.9732 and a decrease in AIC 505.3375.

For time efficiency we used generalised additive models by boosting (gamboost) to fit SST data. The fits obtained by this model somehow cured the fluctuations in the seasonal effects (not fully cured, in some cases seasonal effects were not shown clearly by the model). However, the annual effects were still hard to be identified correctly due to wiggle on the long gap, mainly after the gap of the SST data. We also considered transformation of the rainfall covariate in gamboost models. The result reveal a reduction in final risk and CV-risk by the transformation. The final risk is reduced to 67-79 and CV-risk 28-30 from 75-78 and 29-31. A reduction in AIC from -1.68 to -1.82 and increase in *df* from 28 to 44 is also achieved by the transformation. Our next step was using generalised additive models for location scale and shape (GAMLSS) in the quest of getting a better fit. This model was useful in getting reduced values of AIC implying a better fit and getting higher degrees of freedom. However, the involvement of computational cost with this approach makes it inefficient, especially if the degrees of freedom for smoothing is enlarged and several specification of parameters for each covariate and LSS function, such as the degrees of freedom, the number of knots, and the degree of penalty are used. We used transformation of rainfall covariate, which by applying transformation, a reduction in the AIC is achieved, and we increase in the *df* is also observed. Moreover, further reduction in the AIC and an increase in *df* is achieved by using optimal values for the parameters *df* and *knots*.

After these attempts, we ultimately proposed generalised additive models by boosting for location scale and shape (gamboostLSS) considering the assumption of having a Gaussian distribution for location scale and shape (GaussianLSS). We incorporated penalised

splines for smoothing. The use of LSS function was aimed at getting a clear visualisation of the covariates effects. This led us to having much better fits than all the models considered before. The main focus in the SST data fitting process was to obtain crystalised patterns for the time covariates. Investigation of the marginal gamboostLSS model fitting reveals that the seasonal effects and the continuous covariates have similar patterns. Therefore, the annual and seasonal effects can be used as an indicator to obtain an appropriate model fitting by using gamboostLSS models.

We utilized P-splines smoothing and gradient boosting to investigate the underlying variability structure of time covariates in modelling the SST data. We carried out experiments by considering different specifications for the gamboostLSS model. By SST data experiment with and without transformation of rainfall, we obtain how P-splines smoothing property and gradient boosting can help to discover an underlying variability structure of time covariates.

One of the issues in SST data is the presence of autocorrelation in the data. Therefore, we proposed gamboostLSS-AR(1) model to deal with this issue. We applied generalized differencing technique to reduce the time autocorrelation of the SST data in fitting process. Removing autocorrelation with AR(1) model has a large impact on global and local model fitting. By tuning hyper-parameters, which are flexible and interpretable estimation of a long-scale (annual) and a medium-scale (seasonal) trends in climate features, we can achieve the appropriate gamboostLSS-AR(1) models. The proposed model can be used in further investigation of the effects of the time covariates in location, scale and shape parameters. We carried out experiments by considering various values of the size length of factor ν_{slf} and different stopping iteration m_{stop} in the model fitting. We observed one

small value of ν_{slf}'s from 0.01 to 0.05 and $\nu_{slf} = 0.1$ and with different value of m_{stop} to obtain appropriate global and local model fitting of the SST data.

We also computed marginal prediction interval with autocorrelation (MPI-AR(1)) of the model. MPI-AR(1) of the gamboostLSS-AR(1) model can be used to predict the missing data on the various gaps and to obtain performance prediction interval of submodels. The results of MPI-AR(1) at buoy 3, particularly, is more visible annually compared to buoys 1 and 2. However, for seasonal effect in this case, the model does not seem to follow the pattern of data.

7.2 Future Research

A number of issues are found in this study:

(1) GamboostLSS-AR(1) models with P-splines basis for the SST data fitting show that it has not been optimal properties on fitting performance. Therefore, the model fitting to reach irregular peaks can be applied, most probably by the simple interaction among covariates to complex interactions via additive mixed models with variety of settings, and/or estimated the gaps by the mimic functions to get the best performance of model fitting and prediction.

(2) Due to derivation of gamboostLSS models then the gamboostLSS-AR(1) models have inherent properties, so that it can also extend to spatial (longitude, latitude, layer) study in the context of autocorrelation AR(1) model. Time and spatial-autocorrelation are related to missing data in 1-dimensional and 2-dimensional grid points as refer in [7].

(3) Furthermore, we construct the gamboostLSS-AR(1) models to distinguish noise in the data, where it comes from time and spatial-autocorrelation effects in additive noise model. Perhaps both autocorrelation effects can be computed by measurement errors. Therefore, we can avoid loss information and bias from misfitting model so that we obtain accuracy fitting and statistical inferences. Further we need to diagnostic autocorrelation of SST data to avoid underestimated standard errors.

(4) So far, we have seen that gamboostLSS-AR(1) models can reveal a univarite distribution of SST data with one of depth level. Further we extend to multivariate distribution in the model fitting for one and multi-levels of depth SST data in multiple scenario of low to high dimensions and with and without transformation. This extension refers to a situation in the Indian Ocean is not a single phenomenon.

(5) Extending points (1) to (3) are major concern for the gamboostLSS-AR(1) models to be dynamic modelling, where this model has flexibility in visualizing results with updated time automatically. Dynamic model can be used not only in global fitting but also in local fitting, as dynamic model fitting in computational system.

(6) Underlying extension of the gamboostLSS-AR(1) models by using strengths of each mentioned method in the general discussion (see Chapter 6) we can build an advanced model with multi tasks, multi interpretable, multi predictable, effective, efficient via automatic smoothers and high performance computing.

Bibliography

[1] F. A. Schott, S.-P. Xie, and J. P. McCreary, "Indian ocean circulation and climate variability," *Reviews of Geophysics*, vol. 47, no. RG1002, pp. 1–46, 2009.

[2] D. Dommenget and M. Jansen, "Notes and correspondence: Prediction of Indian ocean SST indices with a simple statistical model: A null hypothesis," *Journal of Climate*, vol. 22, no. 18, pp. 4930–4938, 2009.

[3] G. R. North and M. J. Stevens, "Detecting climate signals in the surface temperature record," *Journal of climate*, vol. 11, no. 4, pp. 563–577, 1998.

[4] C. Deser, M. A. Alexander, S.-P. Xie, and A. S. Phillips, "Sea surface temperature variability: Patterns and mechanisms," *The Annual Review of Marine Science*, vol. 2, no. 10.1146, pp. 115–143, 2010.

[5] B. P. Kumar, J. Vialard, M. Lengaigne, V. S. N. Murty, M. J. McPhaden, M. F. Cronin, and K. G. Reddy, "Evaluation of air-sea heat and momentum fluxes for the tropical oceans and introduction of tropflux," *CLIVAR exchanges*, vol. 58, pp. 1–9, 2012.

[6] S.-P. Xie, C. Deser, G. A. Vecchi, J. Ma, H. Teng, and A. T. Wittenberg, "Global warming pattern formation: Sea surface temperature and rainfall," *Computer Sciences Technical*

Report, vol. 23, pp. 966–986, 2010.

[7] J. R. Magnus, B. Melenberg, and C. Muris, "Global warming and local dimming: The statistical evidence," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 452–464, 2011.

[8] E. Aldrian and R. D. Susanto, "Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature," *International Journal of Climatology*, vol. 23, no. 12, pp. 1435–1452, 2003.

[9] L. Fan, S.-I. Shin, Z. Liu, and Q. Liu, "Sensitivity of Asian summer monsoon precipitation to tropical sea surface temperature anomalies.," *Journal of the Climate Dynamics*, vol. 46, no. 1-2, pp. 1–14, 2016.

[10] H. H. Hendon, "Indonesian rainfall variability: Impacts of ENSO and local air-sea interaction," *Journal of Climate*, vol. 16, no. 11, pp. 1775–1790, 2003.

[11] N. Saji, B. N. Goswami, P. Vinayachandran, and T. Yamagata, "A dipole mode in the tropical Indian Ocean," *Nature*, vol. 401, no. 6751, pp. 360–363, 1999.

[12] K. Ashok, Z. Guan, and T. Yamagata, "A look at the relationship between the ENSO and the Indian Ocean Dipole," *Journal of the Meteorological Society of Japan. Ser. II*, vol. 81, no. 1, pp. 41–56, 2003.

[13] C. Ihara, Y. Kushnir, M. A. Cane, and V. H. De La Peña, "Indian summer monsoon rainfall and its link with ENSO and Indian Ocean climate indices," *International Journal of Climatology*, vol. 27, no. 2, pp. 179–187, 2007.

[14] C. Reason, R. Allan, J. Lindesay, and T. Ansell, "ENSO and climatic signals across the Indian Ocean basin in the global context: Part I, interannual composite patterns," *International Journal of Climatology*, vol. 20, no. 11, pp. 1285–1327, 2000.

[15] R. A. Rigby and D. M. Stasinopoulos, "The gamlss project: a flexible approach to statistical modelling," *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, pp. 249–256, 2001.

[16] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 3, pp. 507–554, 2005.

[17] D. M. Stasinopoulos and R. A. Rigby, "Generalized additive models for location scale and shape (GAMLSS) in R," *Journal of Statistical Software*, vol. 23, no. 7, pp. 1–46, 2007.

[18] P. Bühlmann and T. Hothorn, "Twin boosting: improved feature selection and prediction," *Statistics and Computing*, vol. 20, no. 2, pp. 119–138, 2010.

[19] P. Bühlmann and T. Hothorn, "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, vol. 22, no. 4, pp. 477–505, 2007.

[20] A. Mayr, N. Fenske, B. Hofner, T. Kneib, and M. Schmid, "Generalized additive models for location, scale and shape for high dimensional data — a flexible approach based on boosting," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 61, no. 3, pp. 403–427, 2012.

[21] B. Hofner, T. Hothorn, T. Kneib, and M. Schmid, "A framework for unbiased model selection based on boosting," *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 956–971, 2012.

[22] T. Hastie and R. Tibshirani, "Generalized additive models," *The Journal of the Royal Statistical Scince*, vol. 1, no. 3, pp. 297–318, 1986.

[23] T. Hastie and R. Tibshirani, *Generalized Additive Models*. London: Chapman and Hall, 1990.

[24] S. N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 73, pp. 3–36, 2011.

[25] S. N. Wood, *Package mgcv: Mixed GAM Computation Vechicle with GCV/AIC/REML smoothness estimation.* University of Bath, 2012.

[26] T. Hothorn, P. Buhlmann, and M. Hothorn, "The mboost package," 2007.

[27] M. Schmid and T. Hothorn, "Boosting additive models using component-wise P-splines," *Computational Statistics & Data Analysis*, vol. 53, no. 2, pp. 298–311, 2008.

[28] A. Mayr, B. Hofner, and M. Schmid, "The importance of knowing when to stop: A sequential stopping rule for component-wise gradient boosting," *Methods Inf Med*, pp. 178–186, 2012.

[29] B. Rigby and M. Stasinopoulos, "Statistical modelling using GAMLSS in R.," 2006.

[30] D. Stasinopoulos and R. Rigby, "Generalized additive models for location, scale, and shape (GAMLSS) in R," *Journal of Statistical Software*, vol. 23, no. 7, pp. 1–46, 2007.

[31] D. Rigby, RA ; Stasinopoulos and C. Akantziliotou, "A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse gaussian distribution," *Computational Statistics & Data Analysis*, vol. 53, no. 2, pp. 381–393, 2008.

[32] B. Rigby and M. Stasinopoulos, *A flexible regression approach using GAMLSS in R*. University of Athens, 2010.

[33] A. Mayr, N. Fenske, B. Hofner, T. Kneib, and M. Schmid, "GAMLSS for high-dimensional data-a flexible approach based on boosting," *Ludwig-Maximilians-Universitat Munchen*, no. 098, pp. 1–29, 2010.

[34] C. Belits and S. Lang, "Simultaneous selection of variables and smoothing parameters in structured additive regression models," *The Journal Computational Statistics and Data Analysis*, vol. 53, pp. 61–81, 2008.

[35] C. de Boor, *A Practical Guide to Splines. Revised Edition. Applied Mathematical Sciences*, vol. 27. Springer-Verlag, New York, 2001.

[36] P. H. C. Eilers and B. D. Marx, "Flexible smoothing with B-splines and penalties," *Statistical Science*, vol. 11, no. 2, pp. 89–121, 1996.

[37] P. H. C. Eilers and B. D. Marx, "Spline, knots, and penalties," *Wiley Interdisciplinary Reviews*, pp. 1–26, 2010.

[38] C. de Boor, "On calculating with B-splines," *Journal of Approximation Theory*, vol. 6, no. 1, pp. 50–62, 1972.

[39] C. de Boor and K. Hollig, "B-splines without divided differences," *Computer Sciences Technical Report*, pp. 1–8, 1985.

[40] S. N. Wood, *Generalized Additive Models:An Introduction with R*. CRC, 2006.

[41] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Stanford, California: Springer, 2008.

[42] L. Fahrmeir and T. Kneib, *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*, vol. 521. Oxford University Press, 2011.

[43] L. Fahrmeir, T. Kneib, and S. Lang, "Penalized structured additive regression for space-time data: A bayesian perspective," *Statistica Sinica*, vol. 14, pp. 731–761, 2004.

[44] T. Hothorn, P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner, "Model-based boosting 2.0," *The Journal of Machine Learning Research*, vol. 11, pp. 2109–2113, 2010.

[45] B. Hofner, *Boosting in Structured Additive Models*. PhD thesis, Ludwig-Maximilians-Universitat Munchen, 2011.

[46] J. Gertheiss and G. Tutz, "Penalized regression with ordinal predictors," *Journal of Statistical Review*, vol. 77, no. 3, pp. 345–365, 2009.

[47] P. Buhlmann and B. Yu, "Boosting with the l_2 loss: regression and classification," *Journal of the American Statistical Association*, vol. 98, pp. 324–339, 2003.

[48] D. Ruppert, M. P. Wand, and R. J. Carroll, "Semiparametric regression during 2003–2007," *Electronic Journal of Statistics*, vol. 3, no. 935-7524, pp. 1193–1256, 2009.

[49] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Application in R*, vol. 112. Springer, 2013.

[50] L. Breiman, "Fitting additive models to regression data," *Journal of the Computational Statistics and Data Analysis*, vol. 15, pp. 13–46, 1993.

[51] T. Kneib, "Beyond mean regression," *Journal of the Statistical Modelling*, vol. 13, no. 4, pp. 275–303, 2013.

[52] M. Hansen and B. Yu, "Model selection and minimum descripion length principle," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[53] N. Kramer and M. Sugiyama, "The degrees of freedom of partial least squares regression," *Journal of the American Statistical Association*, vol. 106, pp. 697–705, 2011.

[54] M. Schmid, "Model-based boosting in R: Introduction to gradient boosting," *Statistical Computing*, pp. 1–35, 2011.

[55] G. Tutz, *Regression for Categorical Data*. Cambridge University Press: Cambridge Series in Statistical and Probabilistic Mathematics, 2012.

[56] H. Druker, "Improving regressors using boosting techniques," *Proceedings of the fourth International Conference on Machine Learning, Monmouth University*, pp. 107–115, 1997.

[57] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression:a statistical view of boosting (with discussion)," *The Annals of Statistics*, vol. 2, no. 28, pp. 337–374, 2000.

[58] J. H. Friedman, "Stochastic gradient boosting," *Journal of Computational Statistics and Data Analysis*, vol. 38, pp. 367–378, 2002.

[59] T. Hothorn and P. Bühlmann, "Boosting with componentwise least-squares: Software and examples," *Ensemble Workshop, Friedrich-Alexander-Universitat Erlangen-Nurnberg*, 2006.

[60] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear estimation and classification*, pp. 149–171, Springer, 2003.

[61] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, 2013.

[62] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[63] C. Henning and M. Kutlukaya, "Some thoughts about the design of loss functions," *REVSTAT-Statistical Journal*, vol. 5, no. 1, pp. 19–39, 2007.

[64] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[65] N. Robinzonov, *Advances in boosting of temporal and spatial models*. PhD thesis, Ludwig-Maximilians-Universitat Munchen, 2012.

[66] P. Bühlmann and T. Hothorn, "Boosting: A statistical perspective," pp. 1–57, Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2006.

[67] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[68] S. Weiberg, *Applied Linear Regression: Third Edition*. New Jersey: John Wiley & Sons, Inc, 2005.

[69] J. J. Faraway, *Extending the Linear Model with R*. CRC Press, 2011.

[70] R. E. Schapire, "A brief introduction to boosting," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, vol. 99, pp. 1401–1406, 1999.

[71] M. Schmid, S. Potapov, A. Pfahlberg, and T. Hothorn, "Estimation and regularization techniques for regression models with multidimensional prediction functions," *Statistics and Computing*, vol. 20, no. 2, pp. 139–150, 2010.

[72] L. Fahrmeir and T. Kneib, *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. New York: Oxford University Press, 2011.

[73] B. D. Marx and P. H. C. Eilers, "Direct generalized additive modelling with penalized likelihood," *Journal of Computational Statistics and Data Analysis*, vol. 28, pp. 193–209, 1998.

[74] P. H. C. Eilers, J. Gampe, B. D. Marx, and R. Rau, "Modulation models for seasonal time series and incidence tables," *Journal of the Wiley InterScience*, vol. 14, no. 27, pp. 3430–3441, 2008.

[75] B. Hofner, A. Mayr, N. Robinzonov, , and M. Schmid, "Model-based boosting in R: a hands-on tutorial using the R package mboost," *Journal of Statistics and Computing*, vol. 12, pp. 1–33, 2012.

[76] P. J. Diggle and M. F. Hutchinson, "On spline smoothing with autocorrelated errors," *Australian and new Zealand Journal Statistics*, vol. 31, no. 1, pp. 166–182, 1989.

[77] L. A. Lillard and R. J. Willis, "Dynamic aspects of earnings mobility," *Econometrica: Journal of the Econometric Society*, vol. 46, no. 5, pp. 985—-1012, 1978.

[78] J. M. Wooldridge, *Introductory Econometric*, vol. 910. South-Western College Pub, 2012.

[79] M. Panik, *Regression Modeling*, vol. 814. Chapman and Hall/CRC, 2009.

[80] W. H. Greene, *Econometric Analysis*, vol. 1054. Prentice Hall, 2011.

[81] S. Sriananthakumar, "Testing linear regression model with AR(1) errors against a first-order dynamic linear regression model with white noise errors: A point optimal testing approach," *Economic Modelling*, vol. 33, pp. 126–136, 2013.

[82] P. M. T. Broersen, *Automatic Autocorrelation and Spectral Analysis*. Springer, 2006.

[83] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.

[84] R. L. Anderson, "The problem of autocorrelation in regression analysis," *Journal of the American Statistical Analysis*, pp. 113–129, 2012.

[85] M. S. Boyce, J. Pitt, J. M. Northrup, A. T. Morehouse, K. H. Knopff, B. Cristescu, and G. B. Stenhouse, "Temporal autocorrelation functions for movement rates from global positioning system radiotelemetry data," *Journal of the Royal Society*, vol. 10, no. 365, pp. 2213–2219, 2010.

[86] N. Fenske, L. Fahrmeir, T. Hothorn, P. Rzehak, and M. Hohle, "Boosting structured additive quantile regression for longitudinal childhood obesity data," *Journal of Biostatistics*, vol. 9, pp. 1–18, 2013.

[87] T. M. Smith and R. W. Reynolds, "A high-resolution global sea surface temperature climatology for the 1961-90 base period," *Journal of Climate*, vol. 11, pp. 3320–3323, 1998.

[88] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[89] D. Ruppert, M. P. Wand, and R. J. Carroll, "Semiparametric regression," vol. 386, 2009.

[90] T. P. Ryan, *Modern Regression Methods*, vol. 642. Wiley Series in Probability and Statistics, 2009.

[91] B. Kedem and K. Fokianos, *Regression Models for Time Series Analysis*, vol. 344. Wiley Series in Probability and Statistics, 2002.

[92] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*, vol. 557. Wiley Series in Probability and Statistics, 2003.

[93] P. Cortez, "Sensitivity analysis for time lag selection to forecast seasonal time series using neural networks and support vector machines," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–8, IEEE, 2010.

[94] T. Asefa, M. Kemblowski, M. McKee, and A. Khalil, "Multi-time scale stream flow predictions: the support vector machines approach," *Journal of Hydrology*, vol. 318, no. 1, pp. 7–16, 2006.

[95] D. Cook and D. F. Swayne, *Interactive and Dynamic Graphics for Data Analysis*, vol. 188. Springer, 2007.

[96] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[97] C. Cortes and V. N. Vapnik, "Support vector network," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[98] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[99] N. S. Raghavendra and P. C. Deka, "Support vector machine applications in the field of hydrology: A review," *Applied Soft Computing*, vol. 19, pp. 372–386, 2014.

[100] M. Espinoza, J. A. Suykens, and B. De Moor, "LS-SVM regression with autocorrelated errors," in *Proc. of the 14th IFAC Symposium on System Identification (SYSID)*, pp. 582–587, 2006.

[101] D. Mendes and J. A. Marengo, "Temporal downscaling: a comparison between artificial neural network and autocorrelation techniques over the amazon basin in present and future climate change scenarios," *Journal of Theoretical and Applied Climatology*, vol. 100, no. 3-4, pp. 413–421, 2010.

[102] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[103] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augusin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 307, pp. 1–11, 2008.

[104] R. Genuer, J. M. Poggi, and C. T. Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[105] B. Hofner, A. Mayr, and M. Schmid, "gamboostLSS: An R package for model building and variable selection in the GAMLSS framework," *Journal of Statistical Software*, vol. 23, no. 7, 2015.

[106] W. Y. Loh, "Fifty years of classification and regression trees," *International Statistical Review*, vol. 82, no. 3, pp. 329–348, 2014.

[107] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, vol. 81. World Scientific, 2014.

[108] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[109] C. Y. Wenwen Li and D. Sun, "Mining geophysical parameters through decision-tree analysis to determine correlation with tropical cyclone development," *Journal of Computer & Geosciences*, vol. 35, pp. 309–316, 2009.

[110] S. D. Silliman, "The numerical evaluation by spline of fourier transforms," *Journal of Approximation Theory*, vol. 12, pp. 32–51, 1974.

[111] B. D. Marx, P. H. C. Eilers, J. Gampe, and R. Rau, "Bilinear modulation models for seasonal tables of counts," *The Journal Statistics Computing*, vol. 20, pp. 191–202, 2010.

[112] G. H. J. O. Ramsay and S. Graves, *Functional Data Analysis with R and MATLAB*, vol. 1–207. Springer, 2009.

Appendix A

Generalized Additive Models

Algorithm 3 : GAM models with P-splines basis for fitting SST data

```
maxi ← 8
maxj ← 8
maxk ← 8
maxl ← 8
maxm ← 8
rownum ← maxi ∗ maxj ∗ maxk ∗ maxl ∗ maxm
minAIC ← matrix(, rownum, 6)
o ← 1
  for i = 2 → maxi do
      for j = 2 → maxj do
          for k = 2 → maxk do
              for l = 2 → maxl do
                  for m = 2 → maxm do
                      G = gam(SST ∼ s(Temp, df = i) + s(Humd, df = j) + s(Rain, df = k) + s(Nrdays, df = l) + s(Doy, df = m), data)
                      minAIC[o, 1] ← i
                      minAIC[o, 2] ← j
                      minAIC[o, 3] ← k
                      minAIC[o, 4] ← l
                      minAIC[o, 5] ← m
                      minAIC[o, 6] ← AIC(G)
                      o ← o + 1
                  end for
              end for
          end for
      end for
  end for
  Select the Best GAM Models by minimum AIC
  idx ← which.min(minAIC[, 6])
  return (minim ← minAIC[idx, ])
print(minim)
```

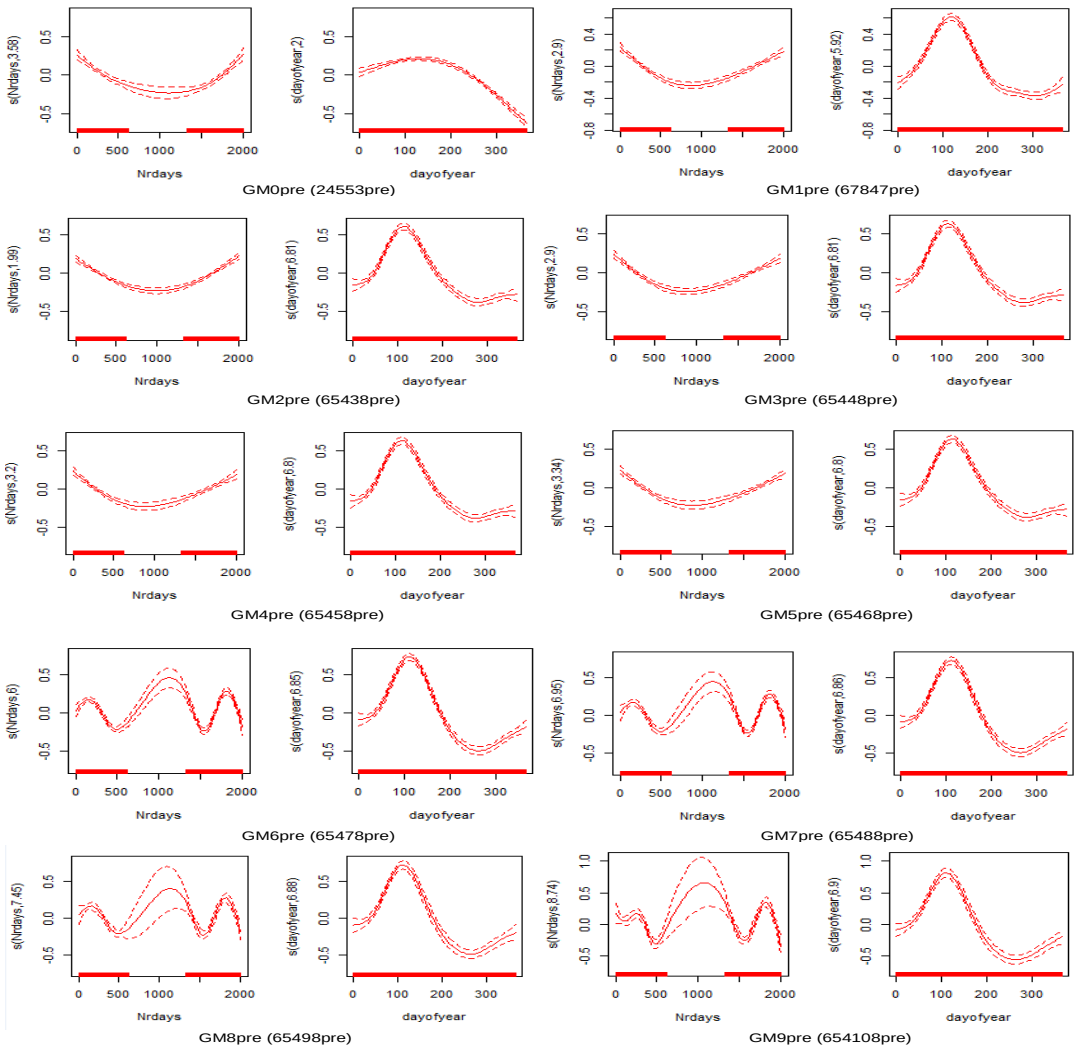


Figure A.1: *The smoothing spline for time covariates pre-transformation rainfall of the SST data by various degree compositions of GAM models. The pattern of time variability as shown in the models GM0pre to GM9pre.*

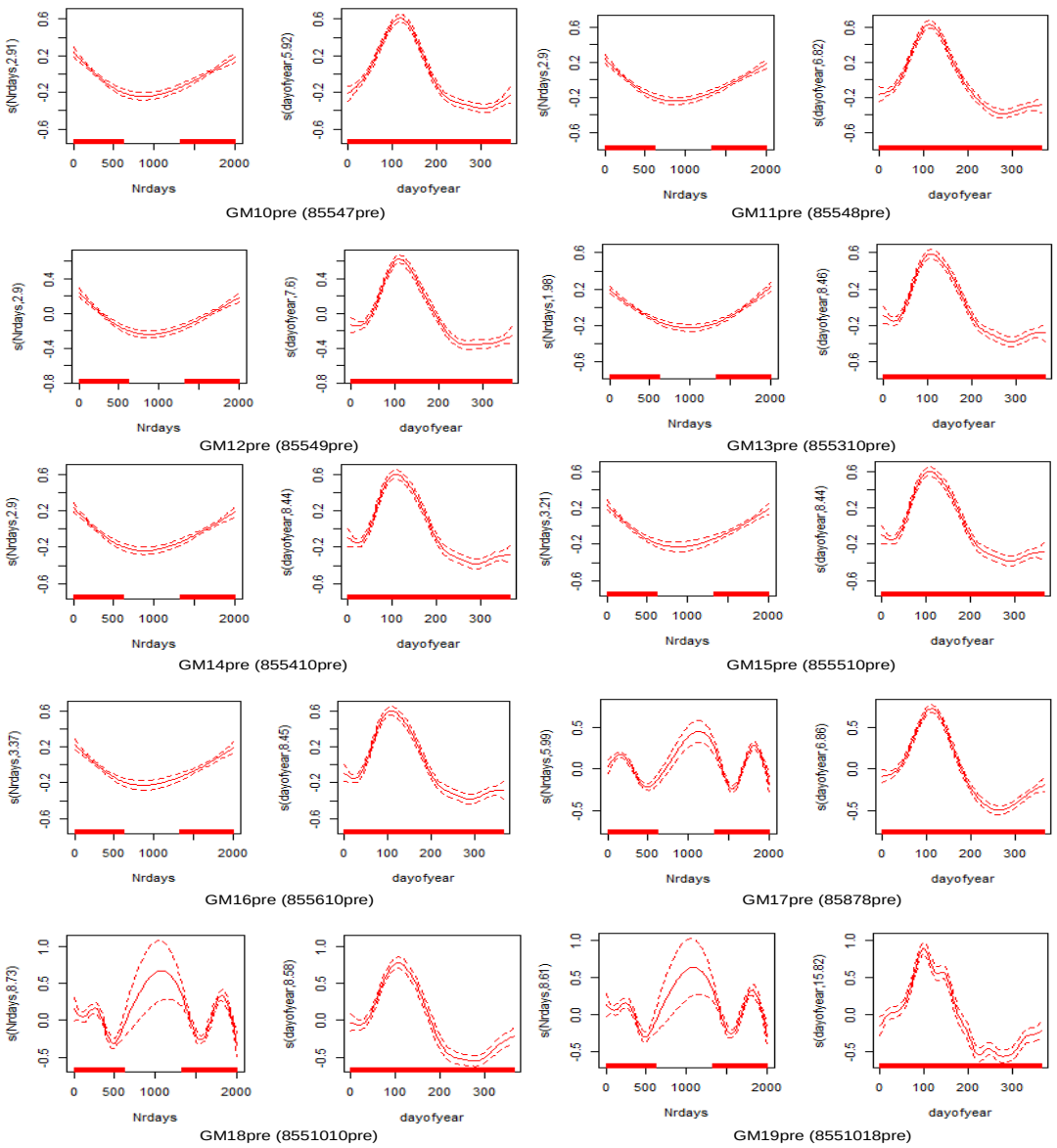


Figure A.2: *The pattern of time variability as shown in the models GM10pre to GM19pre. A gap observation has many patterns depending on the chosen edf values in the structure GAM models, i.e., degree of compositions of its covariates which is mainly for time covariates.*

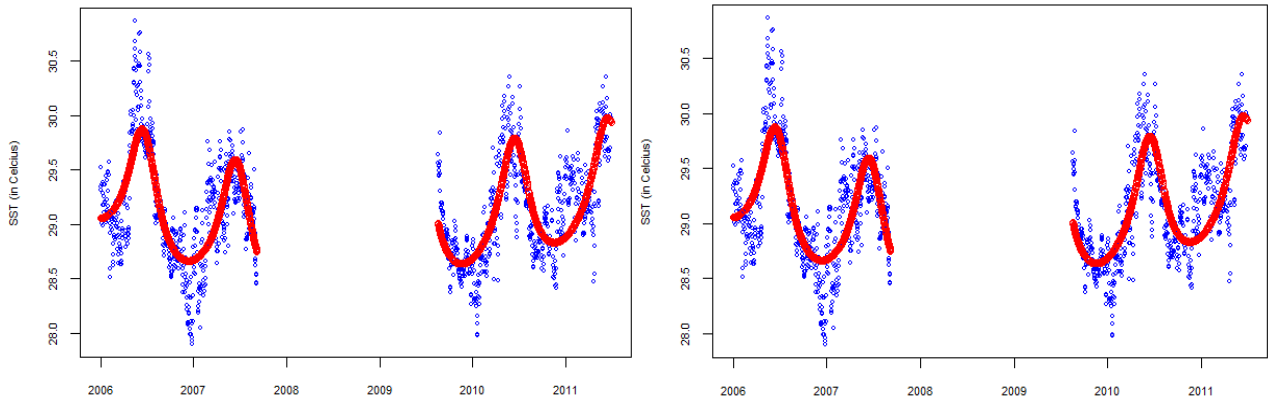


Figure A.3: *Model fitting with time covariate effects, where the model (0,0,0,5,7) is with 5 and 7 df's (left), and model (0,0,0,8,7) with 8 and 7 df's (right).*

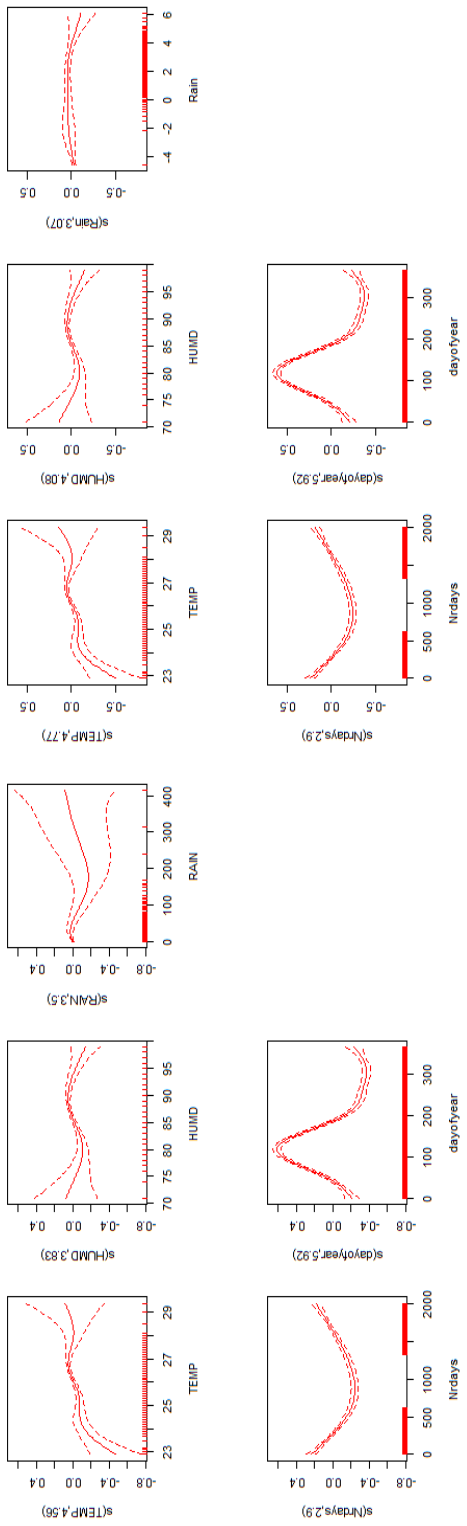


Figure A.4: *The figures of GM1pre-67847 and GM1post-67847 models.*

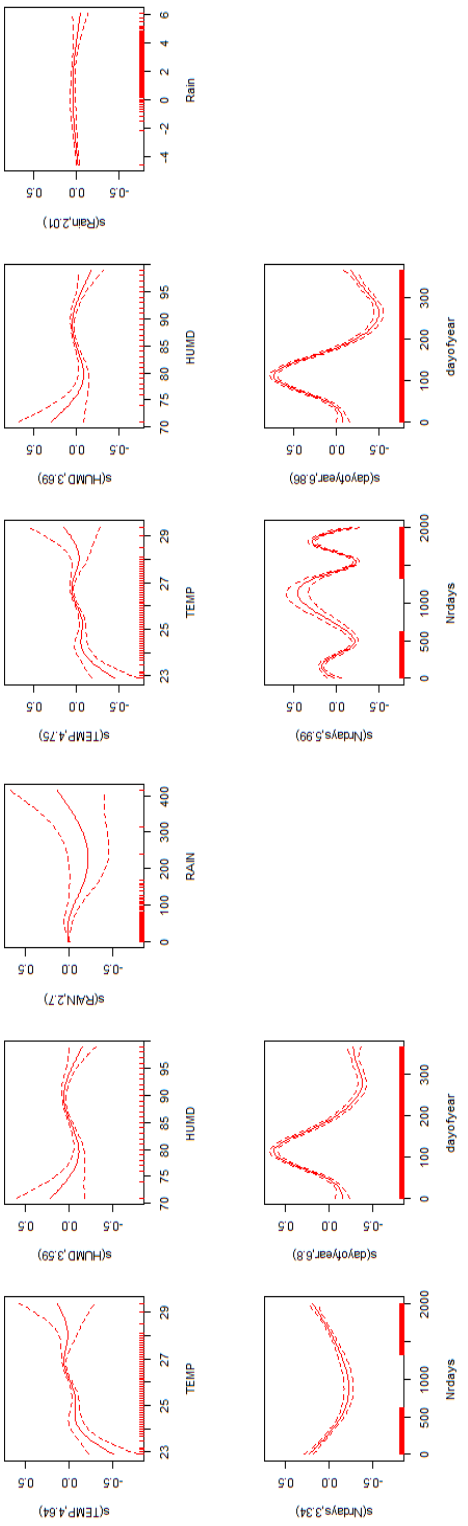


Figure A.5: *The figures of GM6pre-65478 and GM6post-65478 models.*

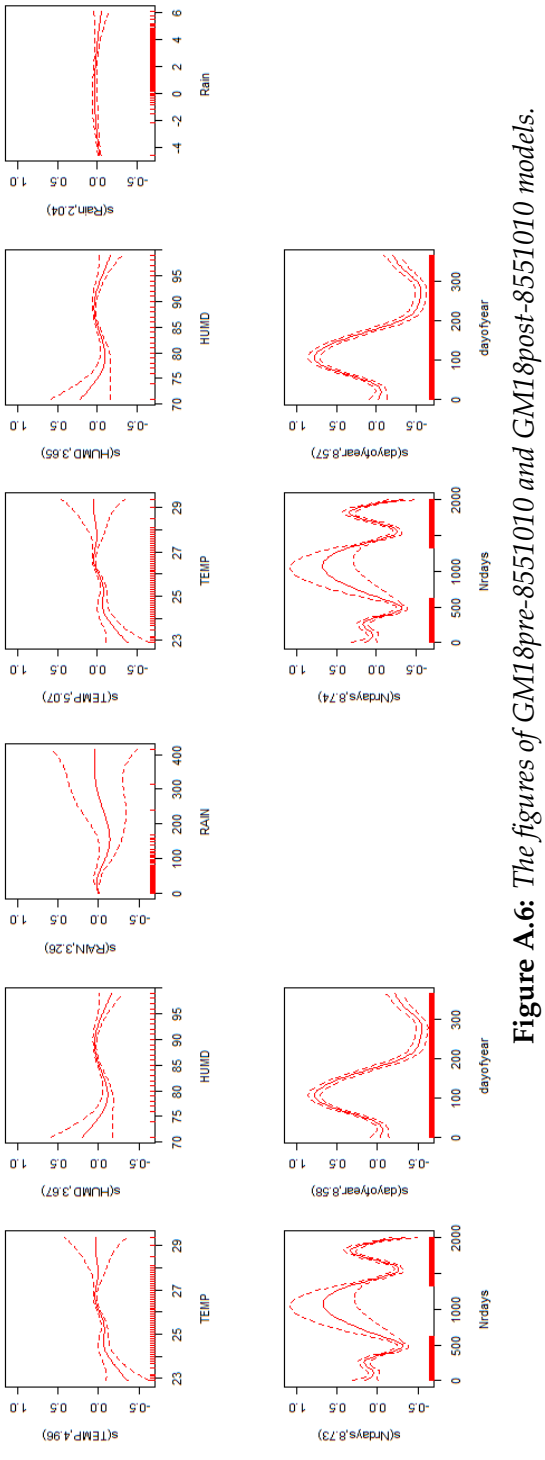


Figure A.6: *The figures of GM18pre-8551010 and GM18post-8551010 models.*

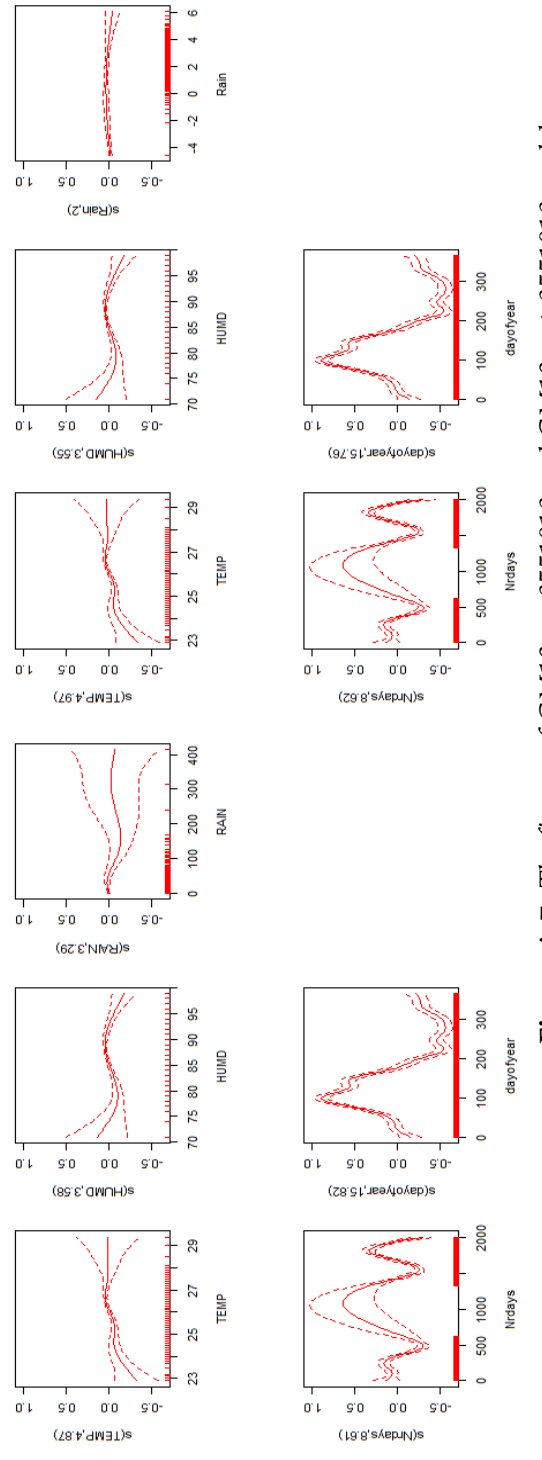


Figure A.7: *The figures of GM19pre-8551018 and GM19post-8551018 models.*

Appendix B

Gamboost Models

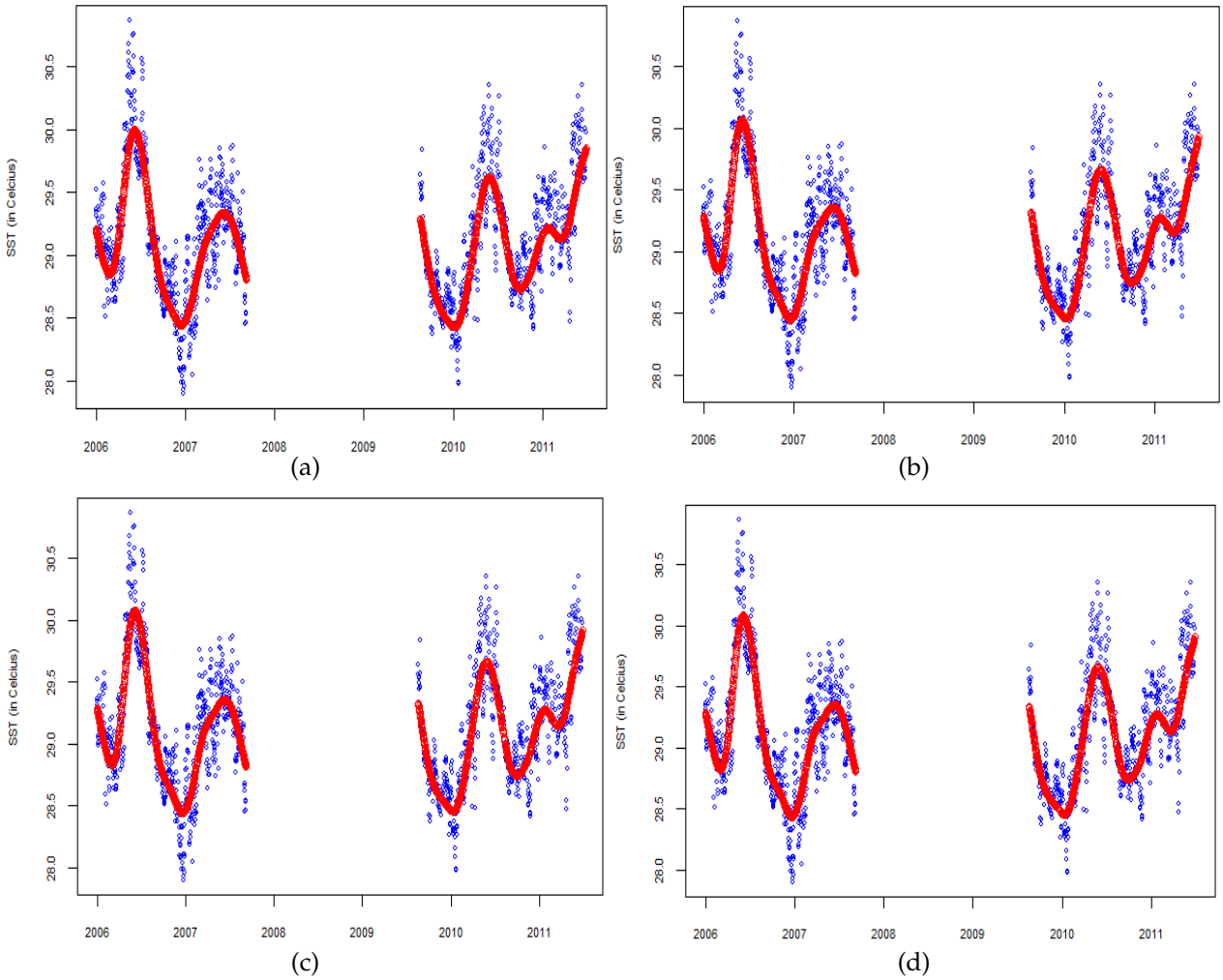


Figure B.1: *Illustration of the SST data fitting by GMboost26 to GMboost29 models for (a) to (d) respectively. The plots show the appropriate models on global fitting with similar patterns, which can be seen in detail in Table 4.16.*

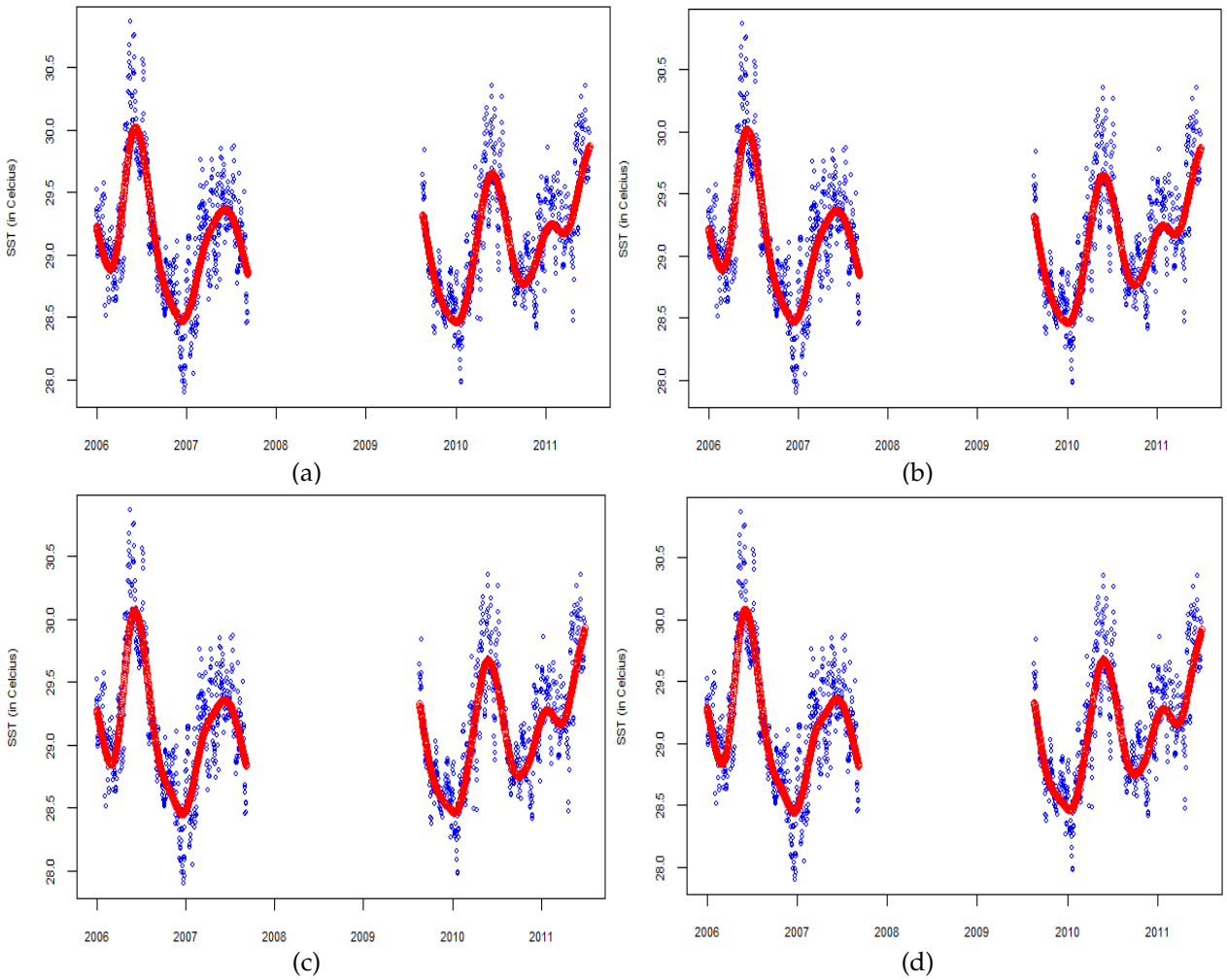


Figure B.2: *Illustration of the SST model fitting for GMboost25post to GMboost28post models with transformed rainfall covariate, (a) to (d) respectively. The models have different df and AIC, see Table 4.17.*

Appendix C

GAMLSS Models

Algorithm 4 GAMLSS by P-splines code for minimum AIC

$maxi \leftarrow 10$
$maxj \leftarrow 10$
$maxk \leftarrow 10$
$maxl \leftarrow 10$
$maxm \leftarrow 10$
$rownum \leftarrow maxi * maxj * maxk * maxl * maxm$
$minAIC \leftarrow matrix(, rownum, 6)$
$o \leftarrow 1$
for $i = 2 \rightarrow maxi$ **do**
for $j = 2 \rightarrow maxj$ **do**
for $k = 2 \rightarrow maxk$ **do**
for $l = 2 \rightarrow maxl$ **do**
for $m = 2 \rightarrow maxm$ **do**
$G = gamlss(SST \sim ps(\textbf{Temp}, df = i) + ps(\textbf{Humd}, df = j) + ps(\textbf{Rain}, df = k) + ps(\textbf{Nrdays}, df = l) + ps(\textbf{Doy}, df = m), data)$
$minAIC[o, 1] \leftarrow i$
$minAIC[o, 2] \leftarrow j$
$minAIC[o, 3] \leftarrow k$
$minAIC[o, 4] \leftarrow l$
$minAIC[o, 5] \leftarrow m$
$minAIC[o, 6] \leftarrow GAIC(G)$
$o \leftarrow o + 1$
end for
end for
end for
end for
end for*Select the Best GAMLSS Models by minimum AIC*
$idx \leftarrow which.min(minAIC[, 6])$
return (*minim* $\leftarrow minAIC[idx,]$)

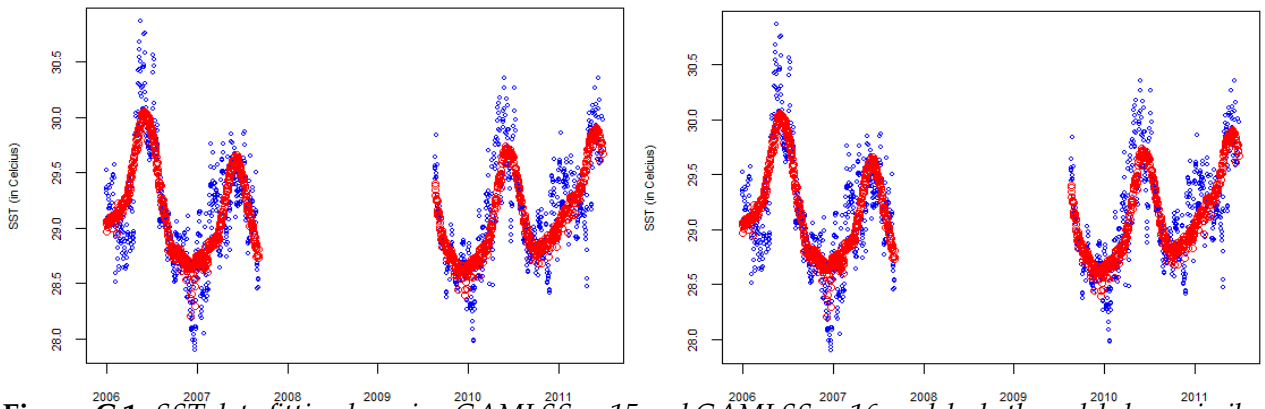


Figure C.1: *SST data fitting by using GAMLSSpre15 and GAMLSSpre16 models, both models have similar patterns in global fitting. However, the specification of Nrdays covariate of both models are different, to see in detail refer to Tables 4.19 and 4.20.*

Appendix D

GamboostLSS Models

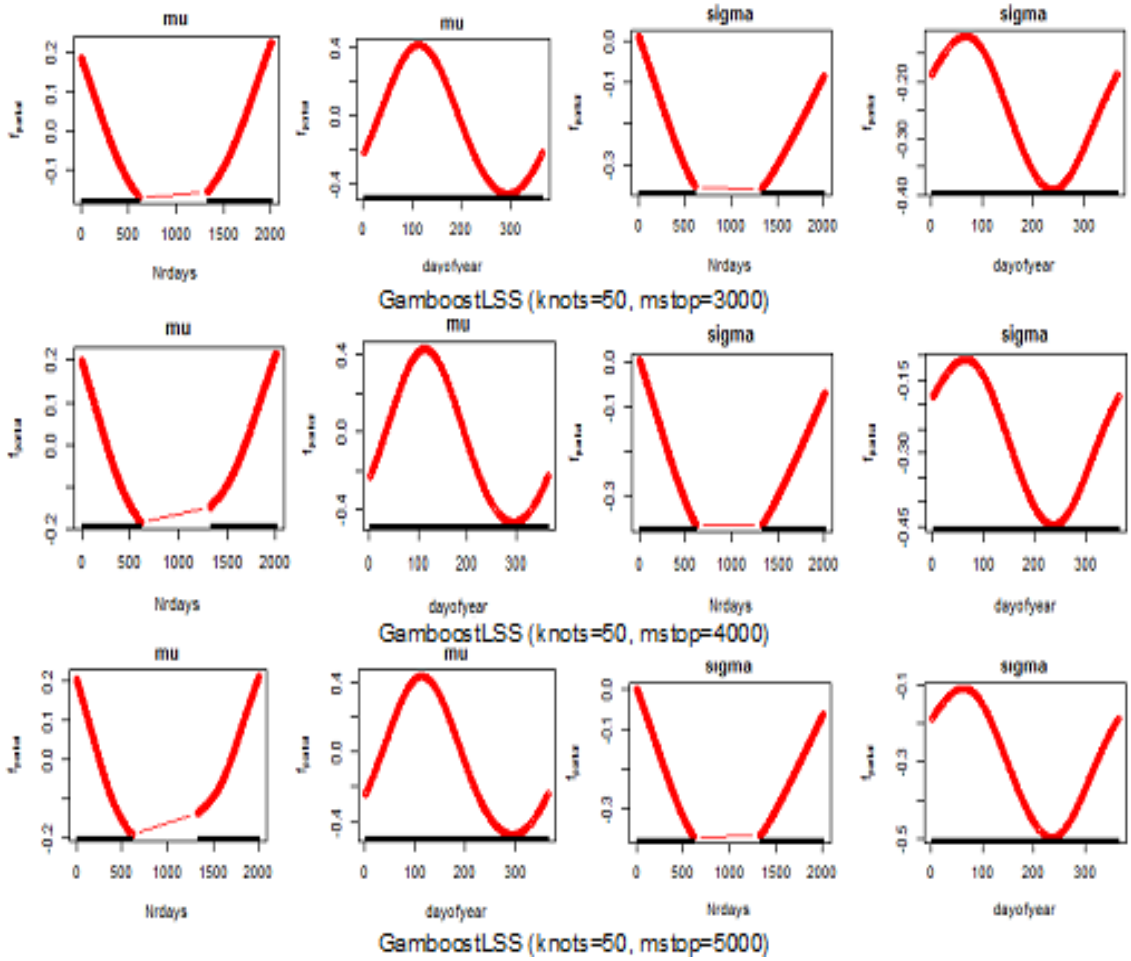


Figure D.1: *Time covariates effects of gamboostLSS models show similar patterns for location and scale of annual and seasonal effects. For annual effects before and after the gap shows similar trends for each step of* m_{stop}*= 3000-5000.*

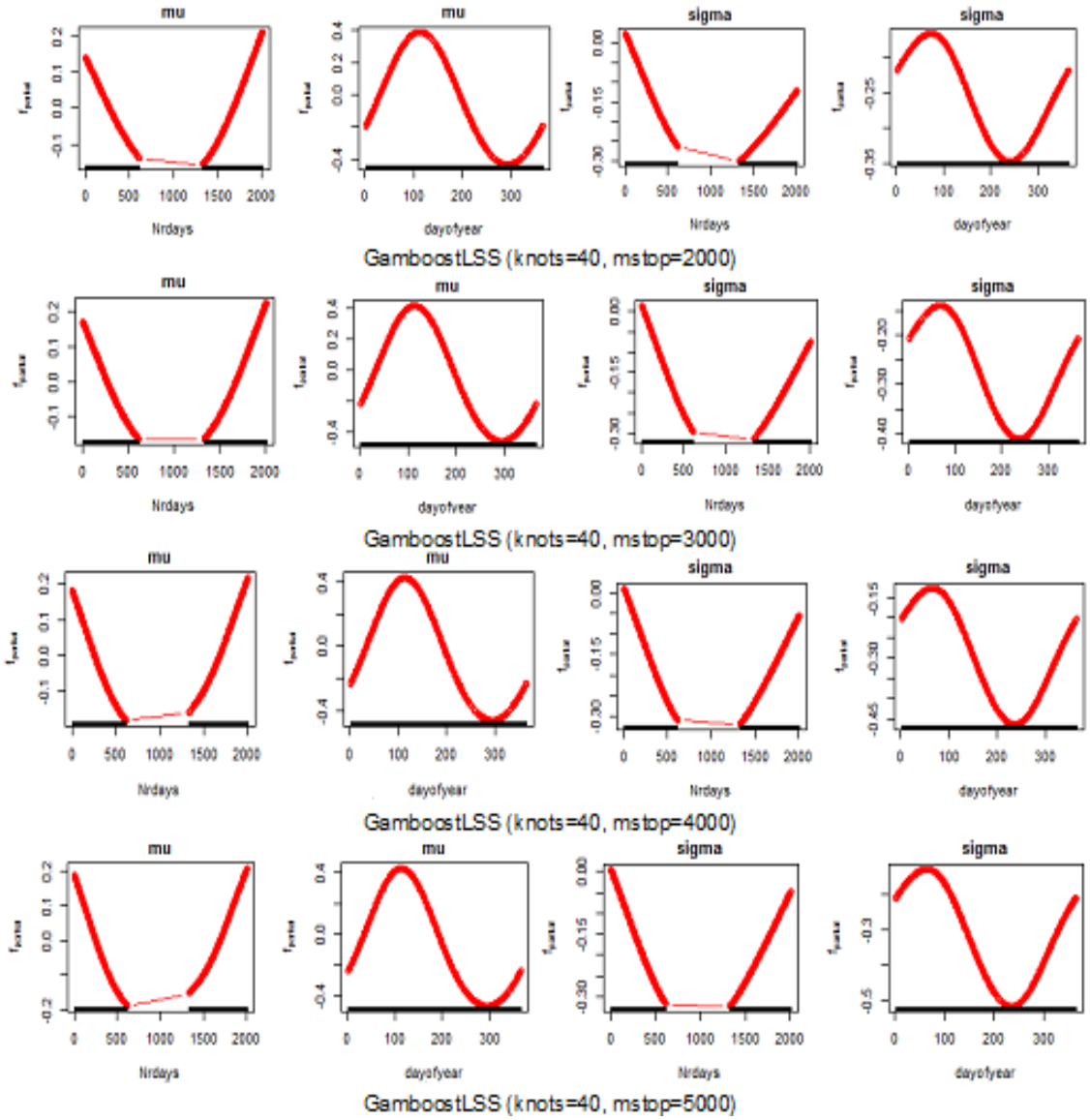


Figure D.2: *Local fitting of gamboostLSS models with different m_{stop}= 2000-5000 and fixed knots= 40 for time covariates.*

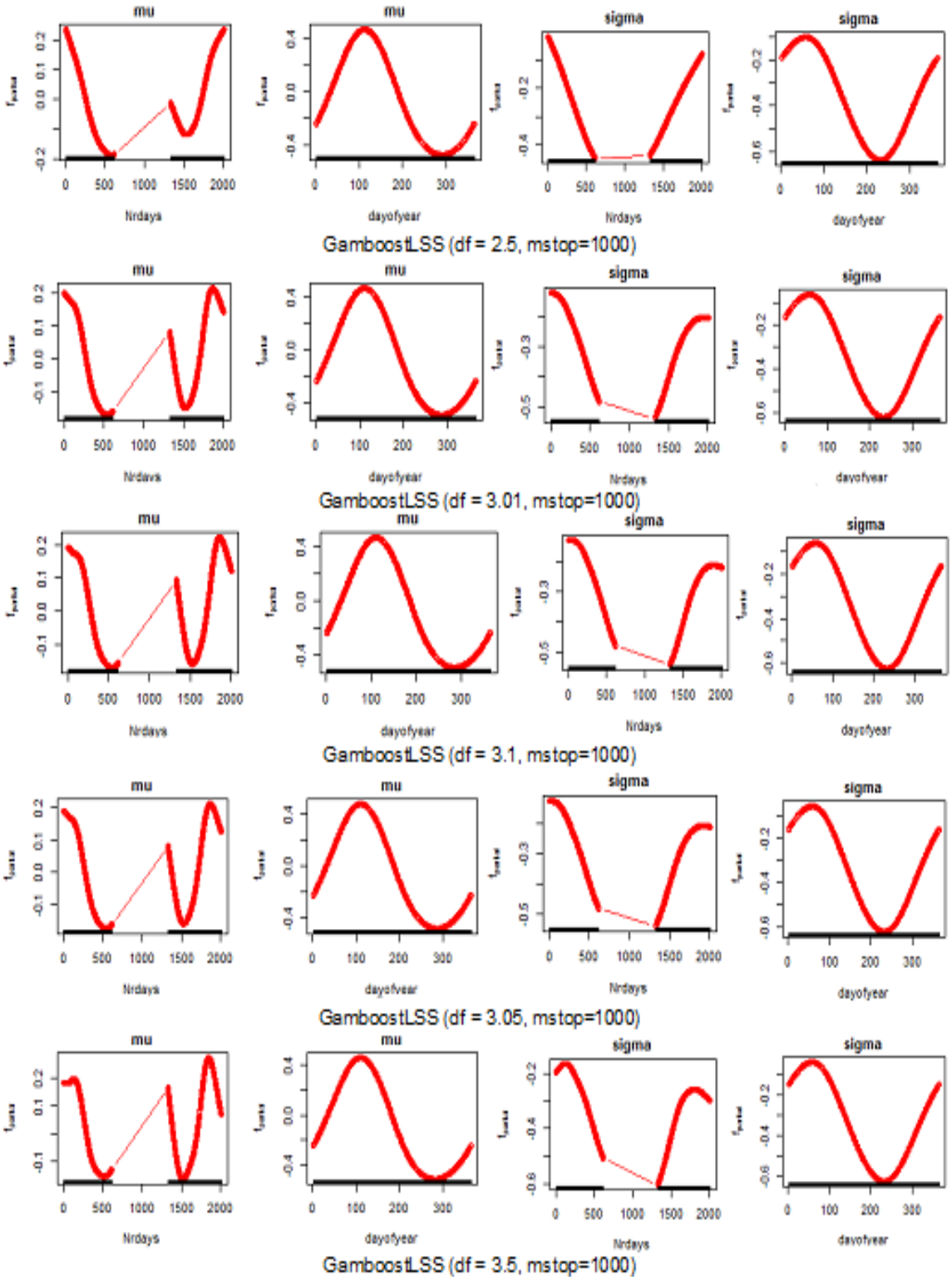


Figure D.3: *Illustration of local fitting with different degrees of freedom df = 2.5-3.5 for time covariates of the SST data fitting with transformation of rainfall.*

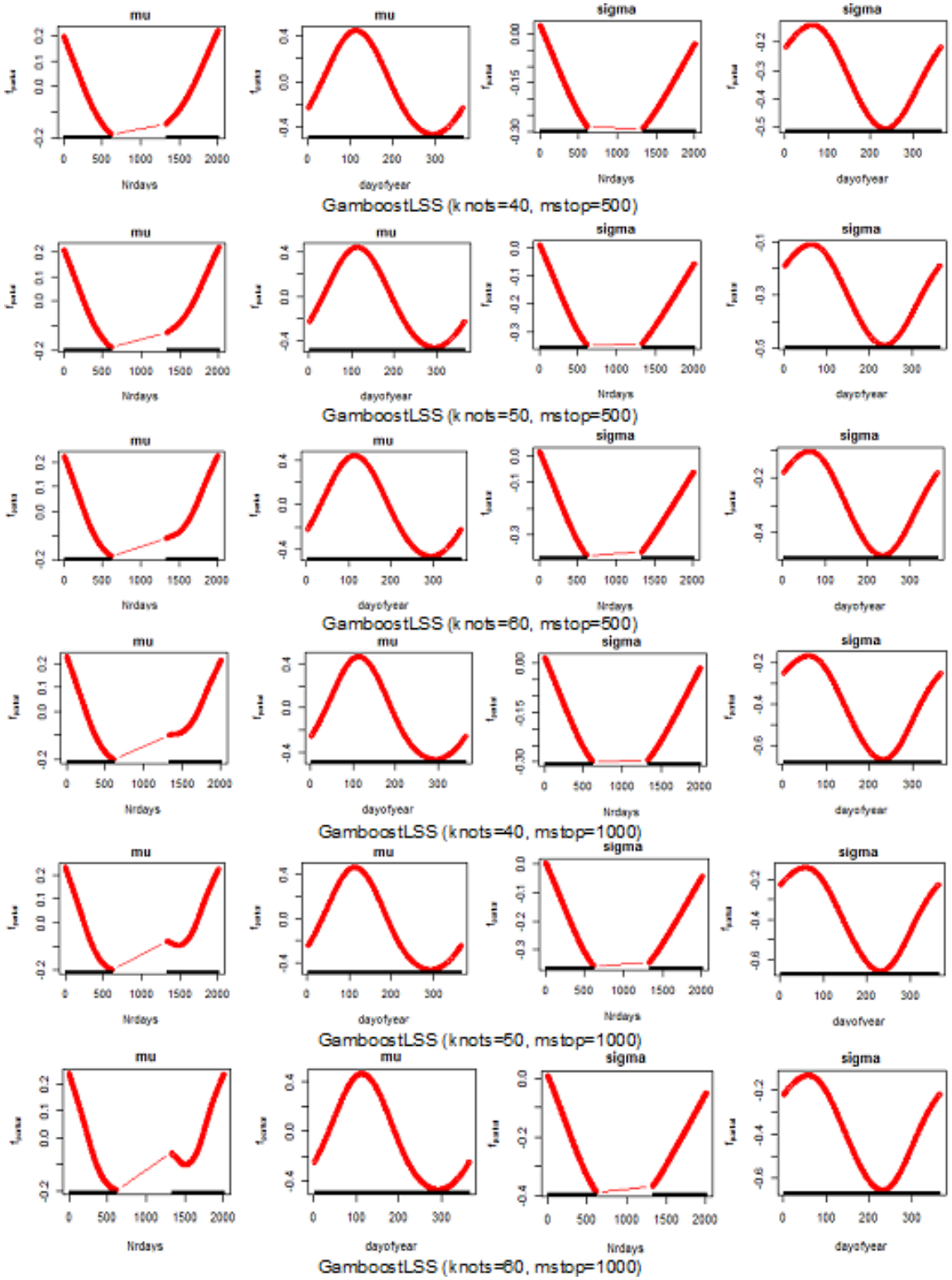


Figure D.4: *Time covariates effects of gamboostLSS models (40 to 60 knots) show similar patterns for location* (μ) *of annual effects and for scale* (σ) *of seasonal effects. For the annual effects before and after the gap shows a slight change for each increase in every 10 knots.*

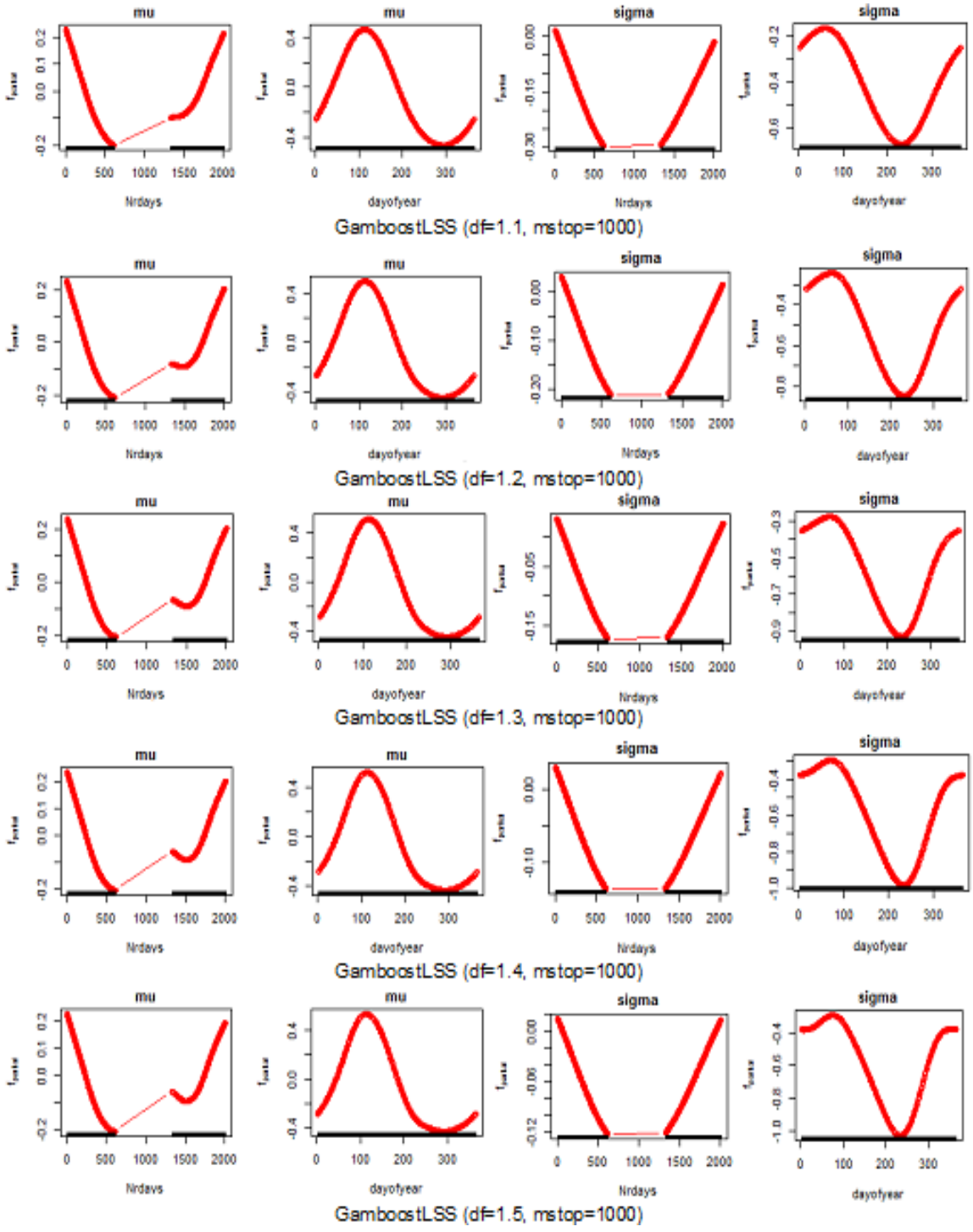


Figure D.5: *Local fitting of time covariates with different degrees of freedom df using the gamboostLSS model of the SST data. The local fitting produces the similar patterns of time covariates.*

Appendix E

GamboostLSS-AR(1) Models

E.1 Autocorrelation of the Gamboost Models

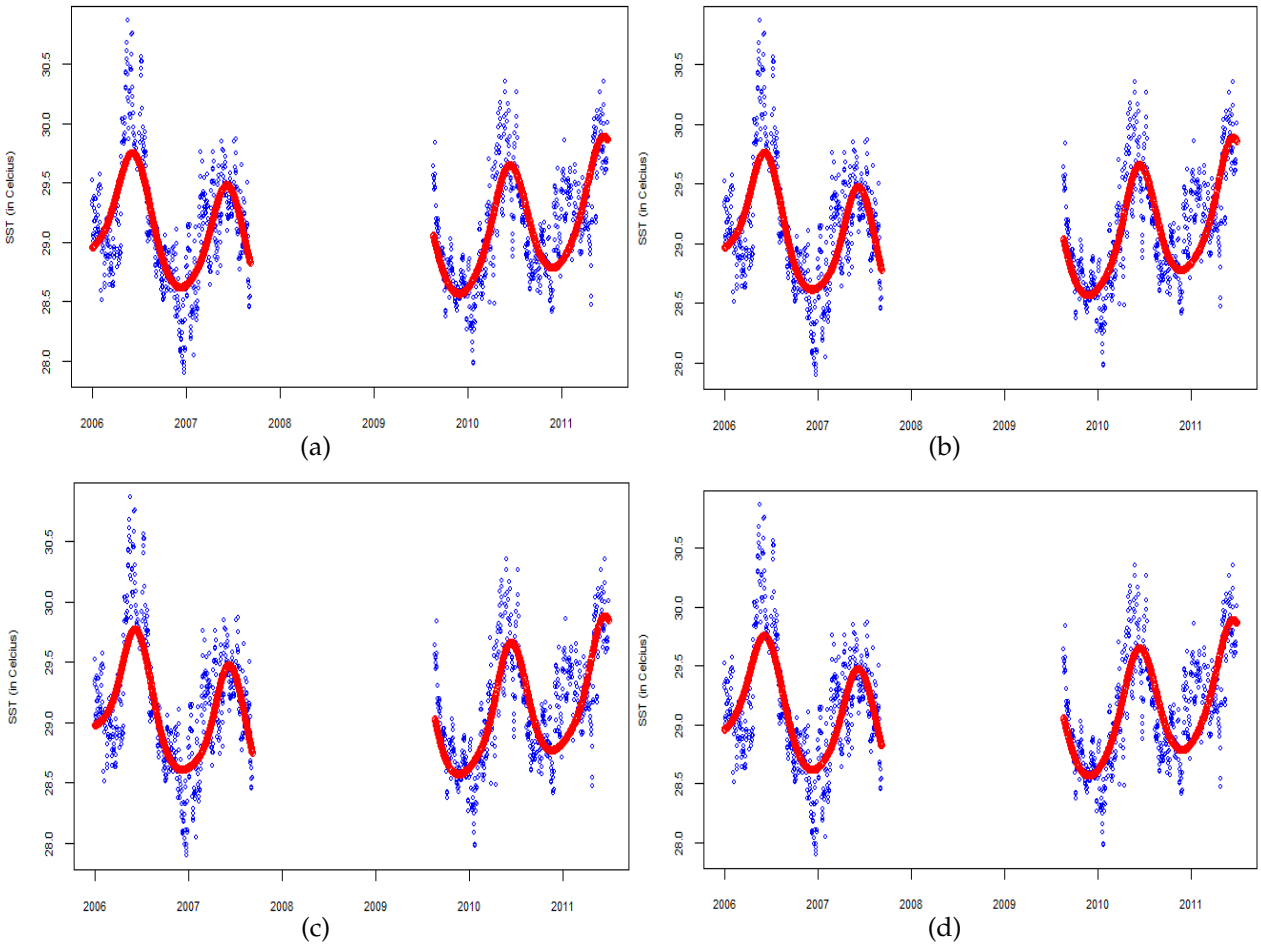


Figure E.1: *An illustration of the appropriate gamboost-AR(1) models fitting of the SST data: GMb1-AR(1) to GMb4-AR(1) models for (a) to (d) respectively, with fixed df=2.5 for Nrdays and df=1.5 for Doy covariates, to see in detail refer to Tables 5.1 and 5.2.*

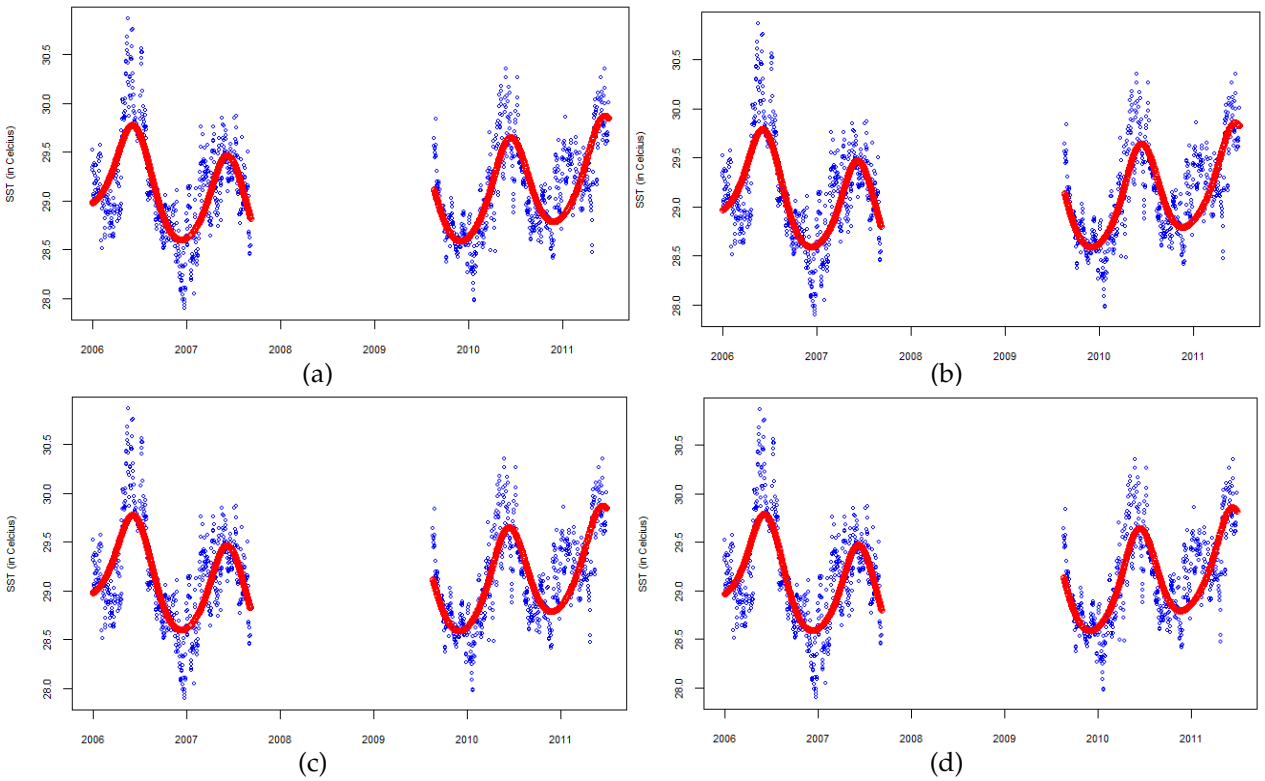


Figure E.2: *An illustration of the appropriate gamboost-AR(1) models: GMb5-AR(1) to GMb8-AR(1) models for (a) to (d) respectively, with df=3.5 for Nrdays and df=1.5 for Doy, to see in detail refer to Tables 5.1 and 5.2.*

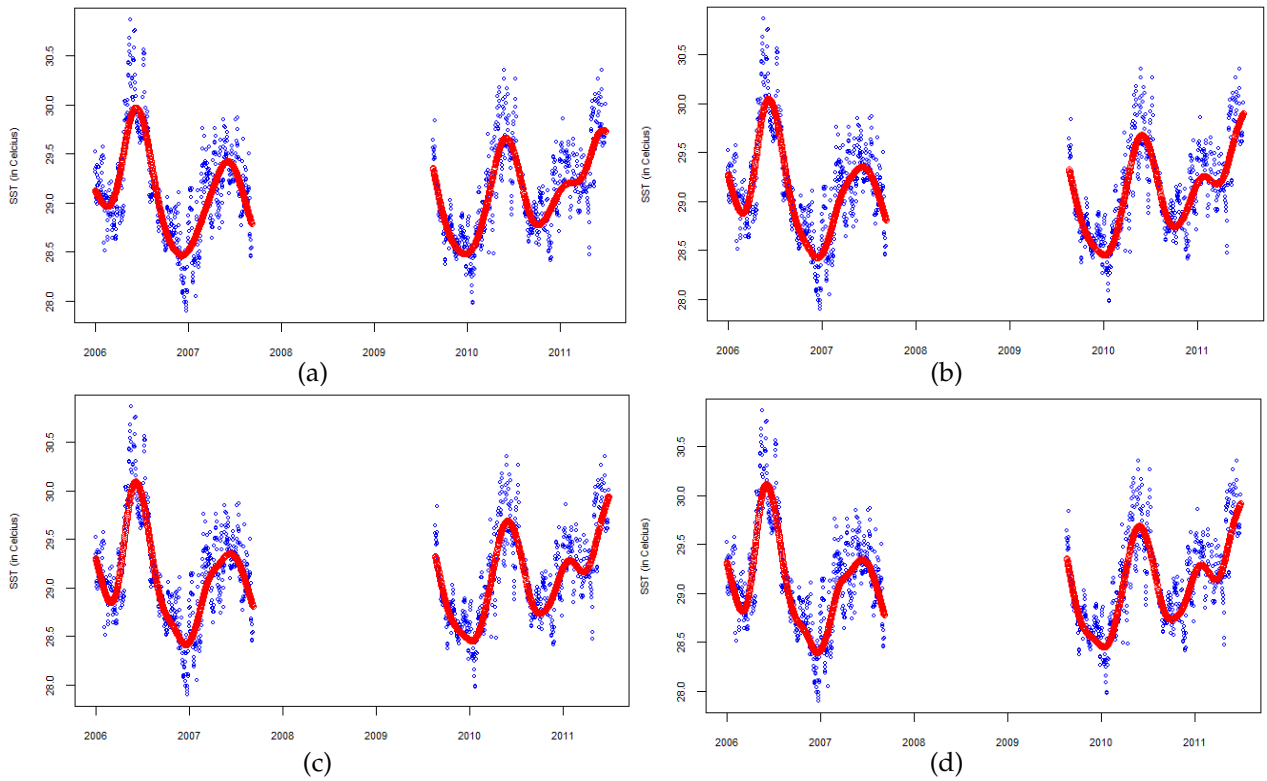


Figure E.3: *The SST data fitting by GMboost20-AR(1) to GMboost30-AR(1) models with (a) to (d) respectively. The models show similar patterns of global fitting.*

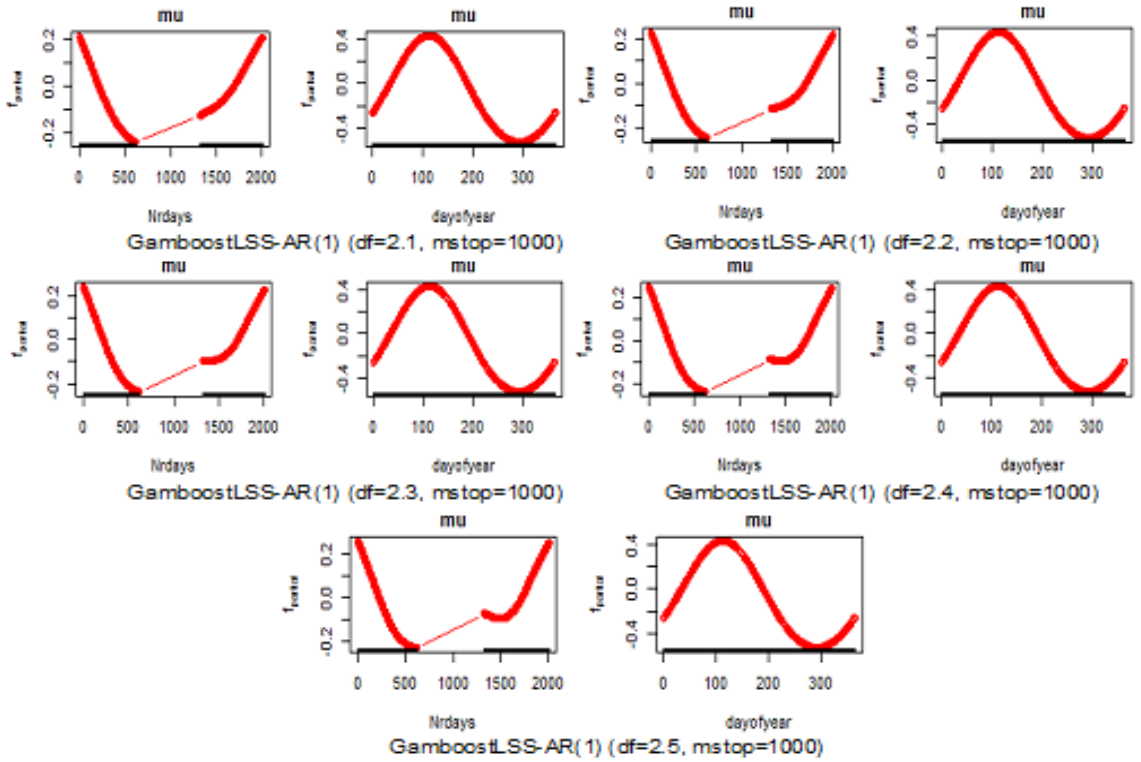


Figure E.4: *Local fitting of time covariate using gamboostLSS-AR(1) models with m_{stop} = 1000 and different df. In local fitting it shows similar patterns, excluding slight changes after the gap for df= 2.5.*

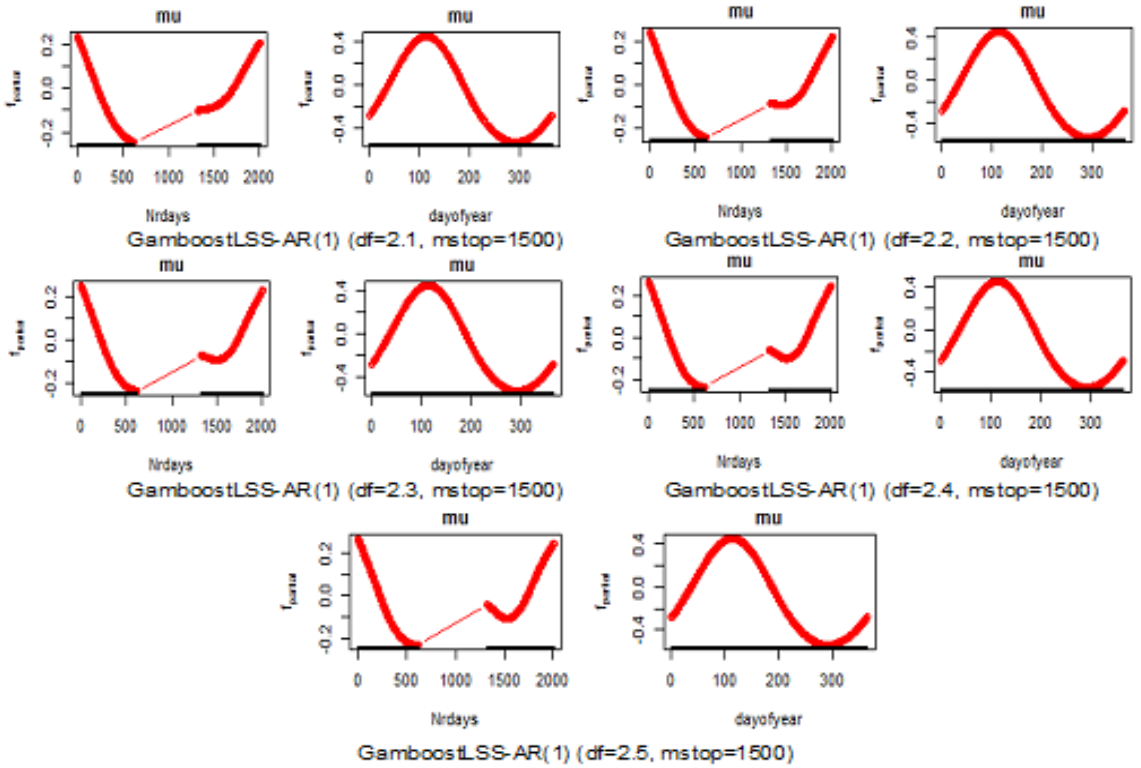


Figure E.5: *The patterns of time covariates in gamboostLSS-AR(1) models fitting with m_{stop}= 1500 and different df. The patterns show a decrease before the gap and an increase after the gap for Nrdays effect and the same pattern for Doy effect, excluding slight changes after the gap for df= 2.4 and 2.5.*

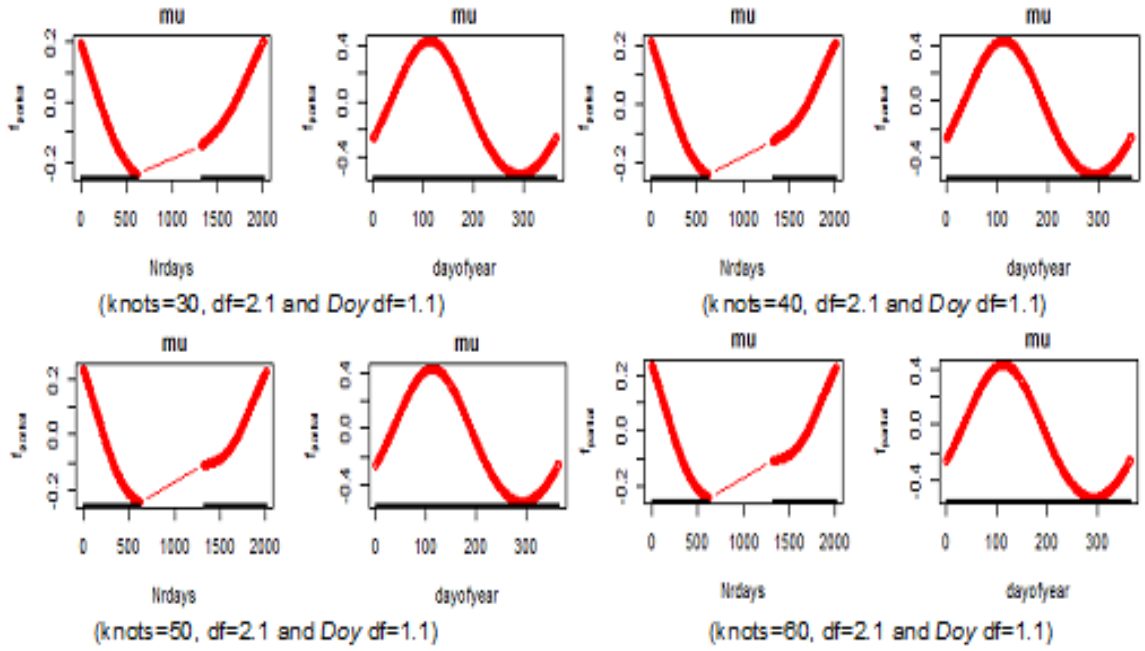


Figure E.6: *Local fitting of time covariates using gamboostLSS-AR(1) models with df= 2.1 and different knots of the Nrdays covariate and df= 1.1 at the Doy covariate show similar patterns.*

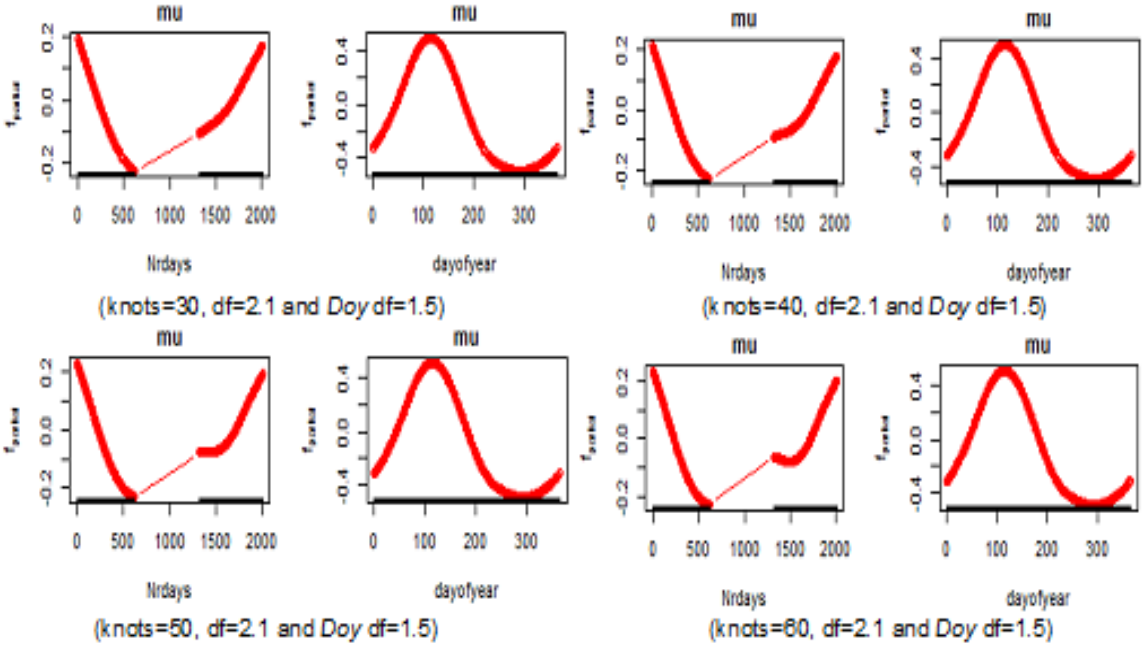


Figure E.7: *GamboostLSS-AR(1) models fitting with fixed df = 2.1 and different knots of the Nrdays covariate and df= 1.5 at the Doy covariate show the similar patterns.*

E.2 Gamboost-AR(1) Models with Transformation

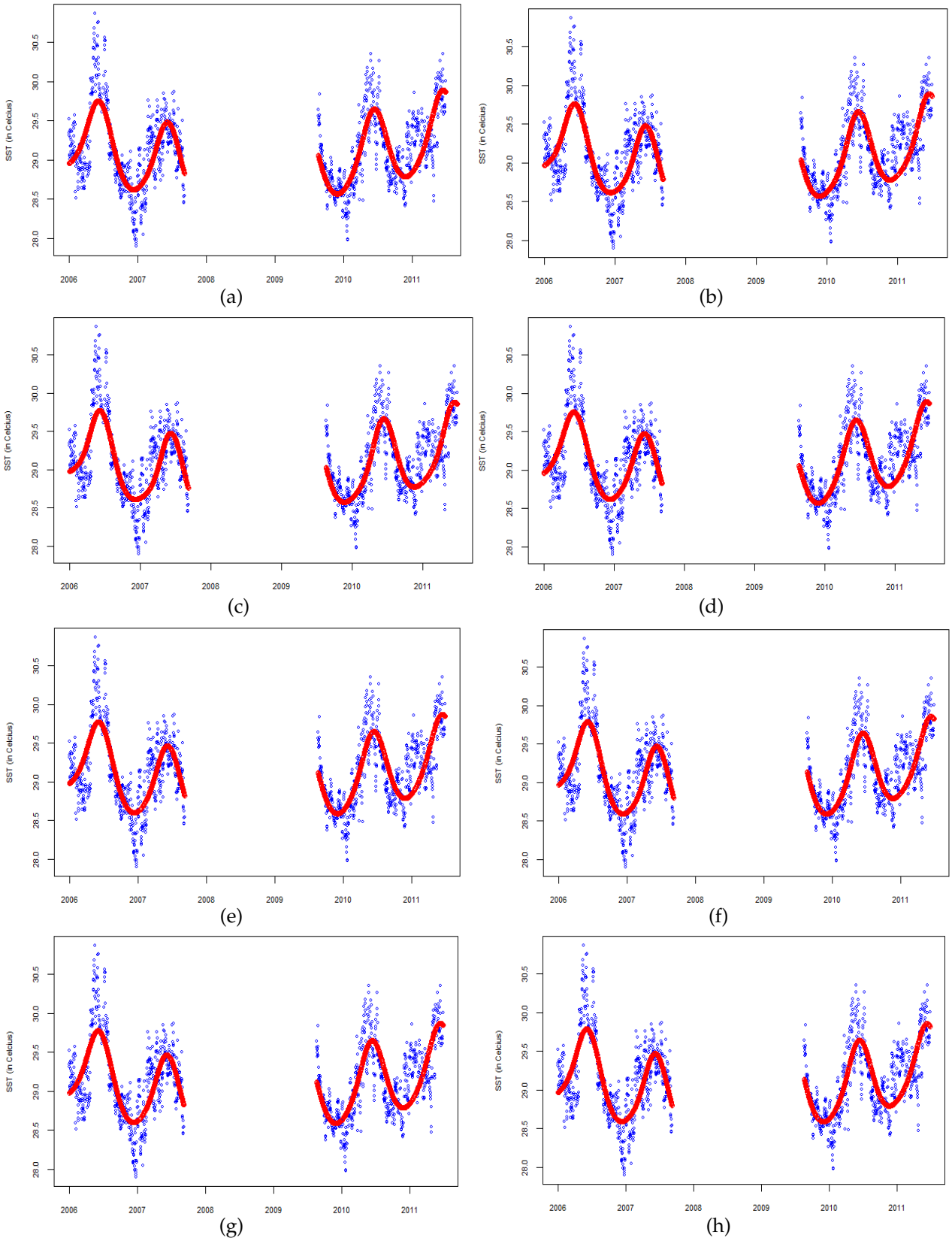


Figure E.8: *The SST data fitting by GMboost1-AR(1) to GMboost8-AR(1) models with transformation of rainfall. The models show similar patterns, to see in detail refer to Tables 5.1 and 5.3.*

E.3 GamboostLSS-AR(1) Models with Transformation

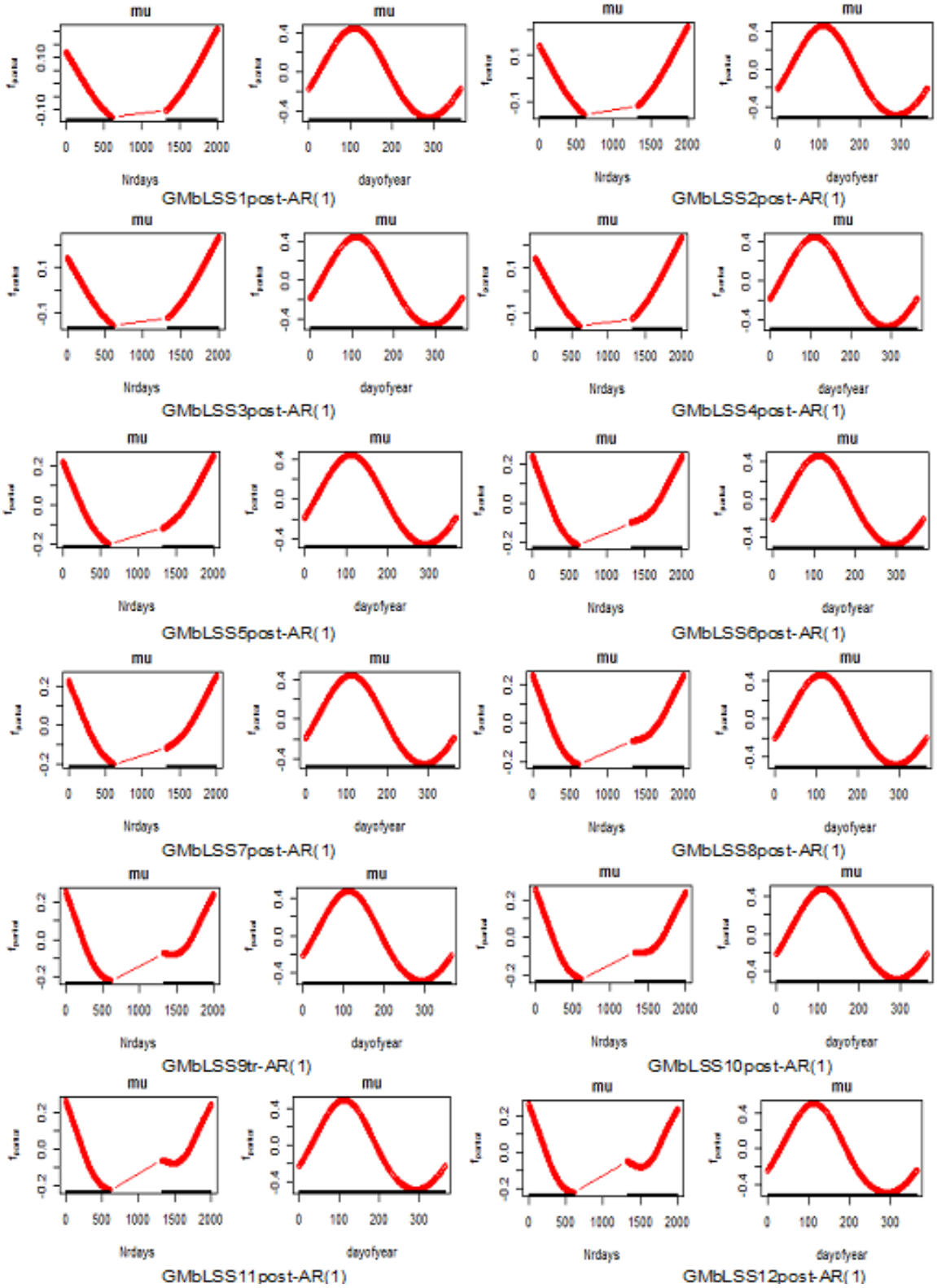


Figure E.9: *The similar patterns of time-covariates on local fitting for the SST data by gamboostLSS-AR(1) models with transformation of rainfall, to see in detail refer to Table 5.8.*

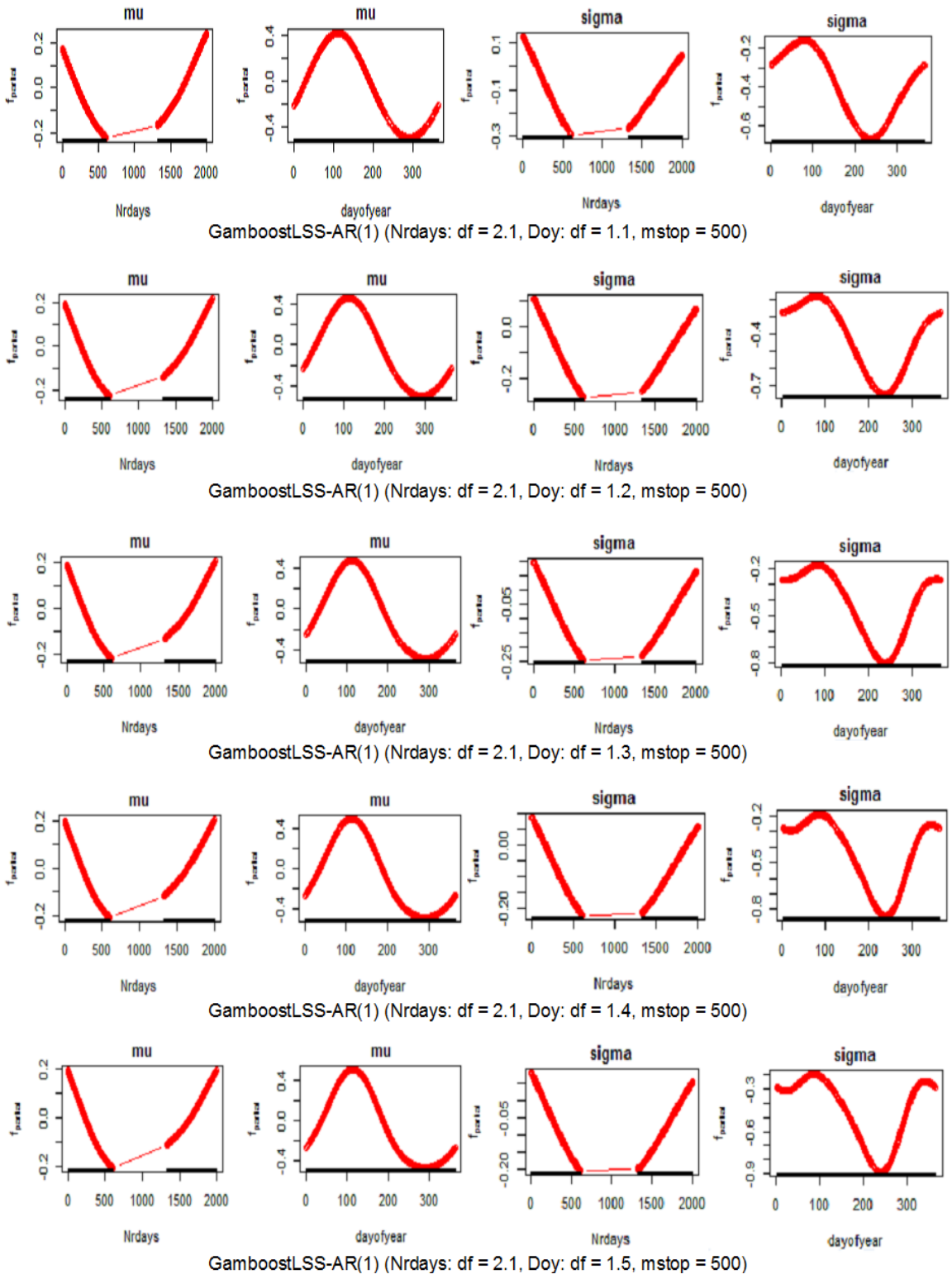


Figure E.10: *The patterns of time covariates in local fitting use gamboostLSS-AR(1) models with transformation. The patterns show a decrease before the gap and an increase after the gap for the Nrdays effect and a similar pattern for the Doy effect. However, in the beginning fitting for the Doy covariate, it shows a slight difference for df= 1.2 - 1.5 with fixed m_{stop}= 500.*

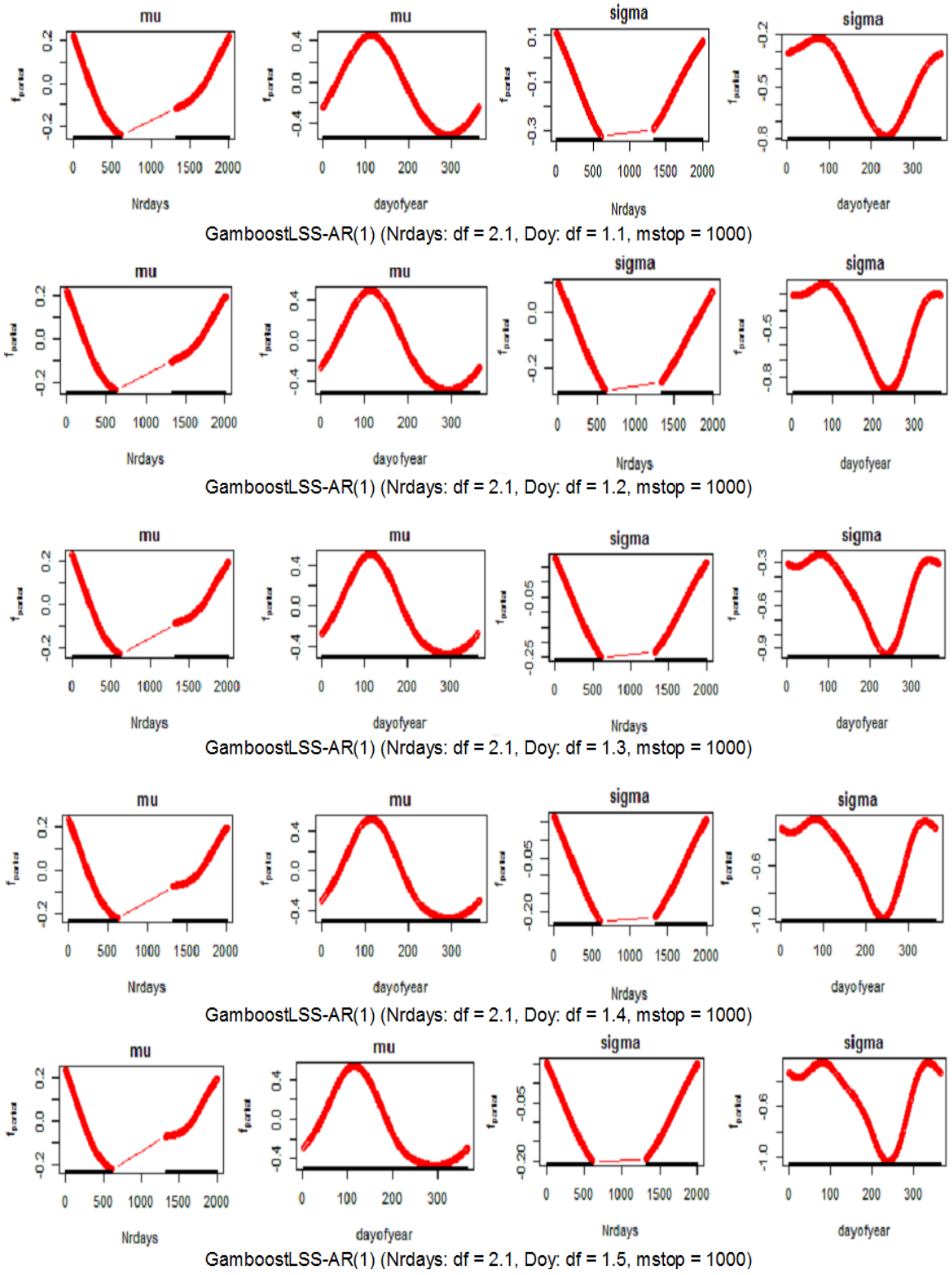


Figure E.11: *The patterns of time covariates in local fitting using gamboostLSS-AR(1) models with transformation. The patterns show a decrease before the gap and an increase after the gap for the Nrdays effect and a similar pattern for the Doy effect. However, in the beginning fitting for the Doy covariate, it shows a slight difference for df= 1.2 - 1.5 with fixed m_{stop}= 1000.*

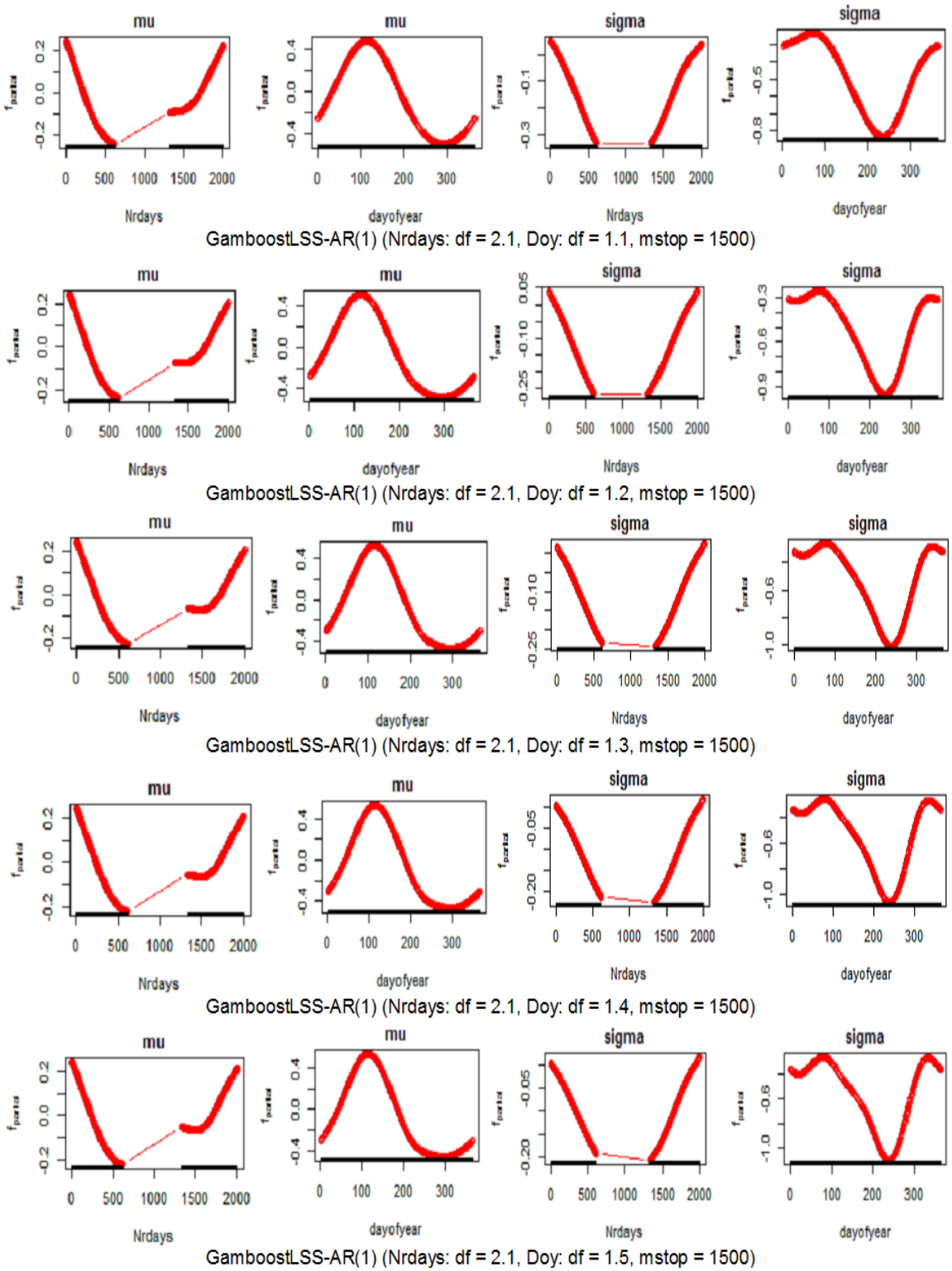


Figure E.12: *The patterns of time covariates in local fitting using gamboostLSS-AR(1) models with transformation. The patterns show a decrease before the gap and an increase after the gap for the Nrdays effect and a similar pattern for the Doy effect. However, in the beginning fitting for the Doy covariate, it shows a slight difference for df= 1.2 - 1.5 with fixed m_{stop}= 1500.*

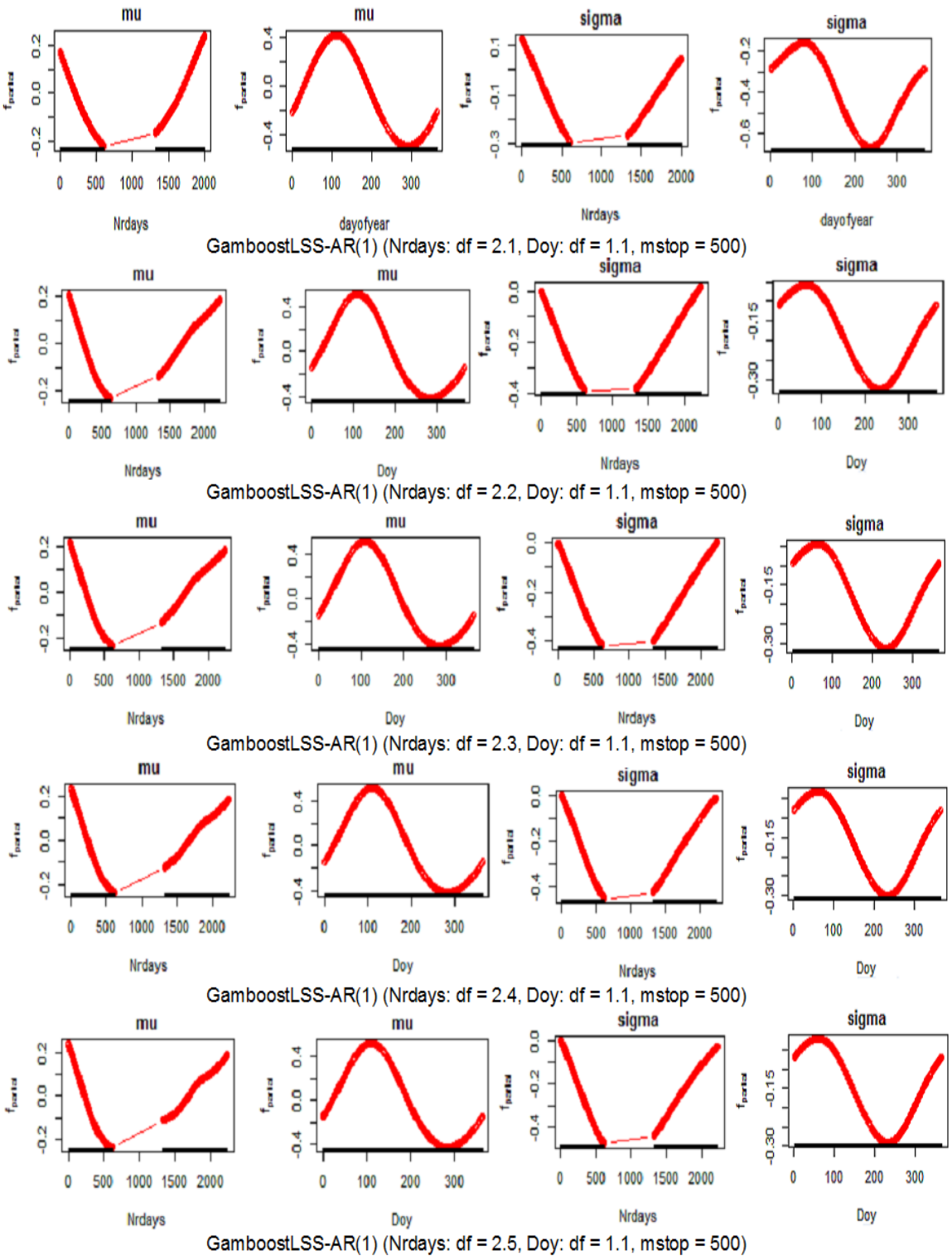


Figure E.13: *The patterns of time covariates in local fitting using gamboostLSS-AR(1) models with transformation. The patterns show a decrease before the gap and an increase after the gap for the Nrdays effect and a similar pattern for the Doy effect.*

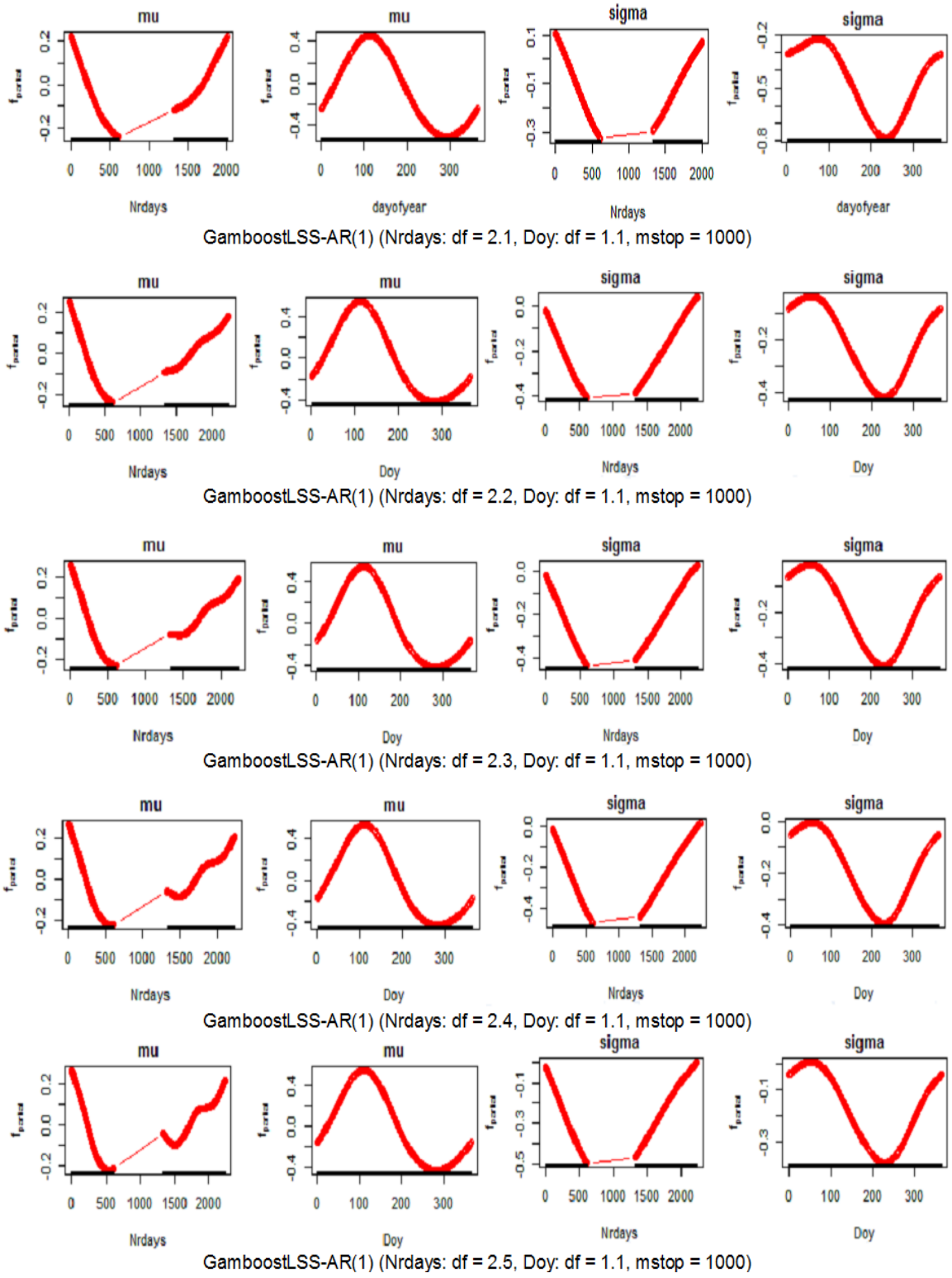


Figure E.14: *The patterns of time covariates in local fitting using gamboostLSS-AR(1) models with transformation. The patterns show a decrease before the gap and an increase after the gap for the Nrdays effect and a similar pattern for the Doy effect. However, after the gap for the Nrdays covariate, it shows a slight difference for df= 2.2 - 2.5 with fixed m_{stop}= 1000.*

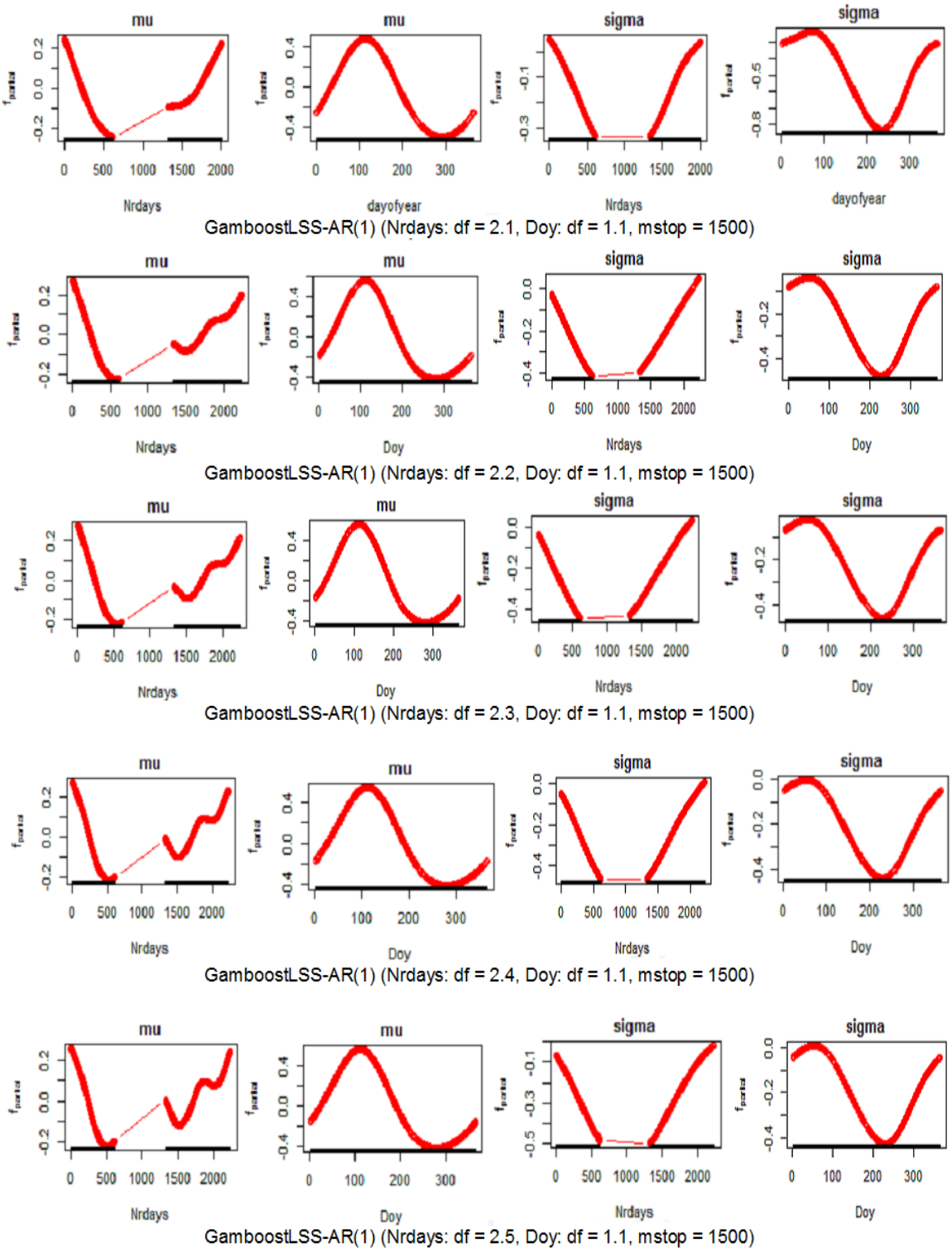


Figure E.15: *The patterns of time covariates in local fitting using gamboostLSS-AR(1) models with transformation. The patterns show a decrease before the gap and an increase after the gap for annual effect and a similar pattern for the Doy effect. However, after the gap for the Nrdays covariate, it shows a slight fluctuation for df= 2.2 - 2.5 with fixed* m_{stop}*= 1500.*

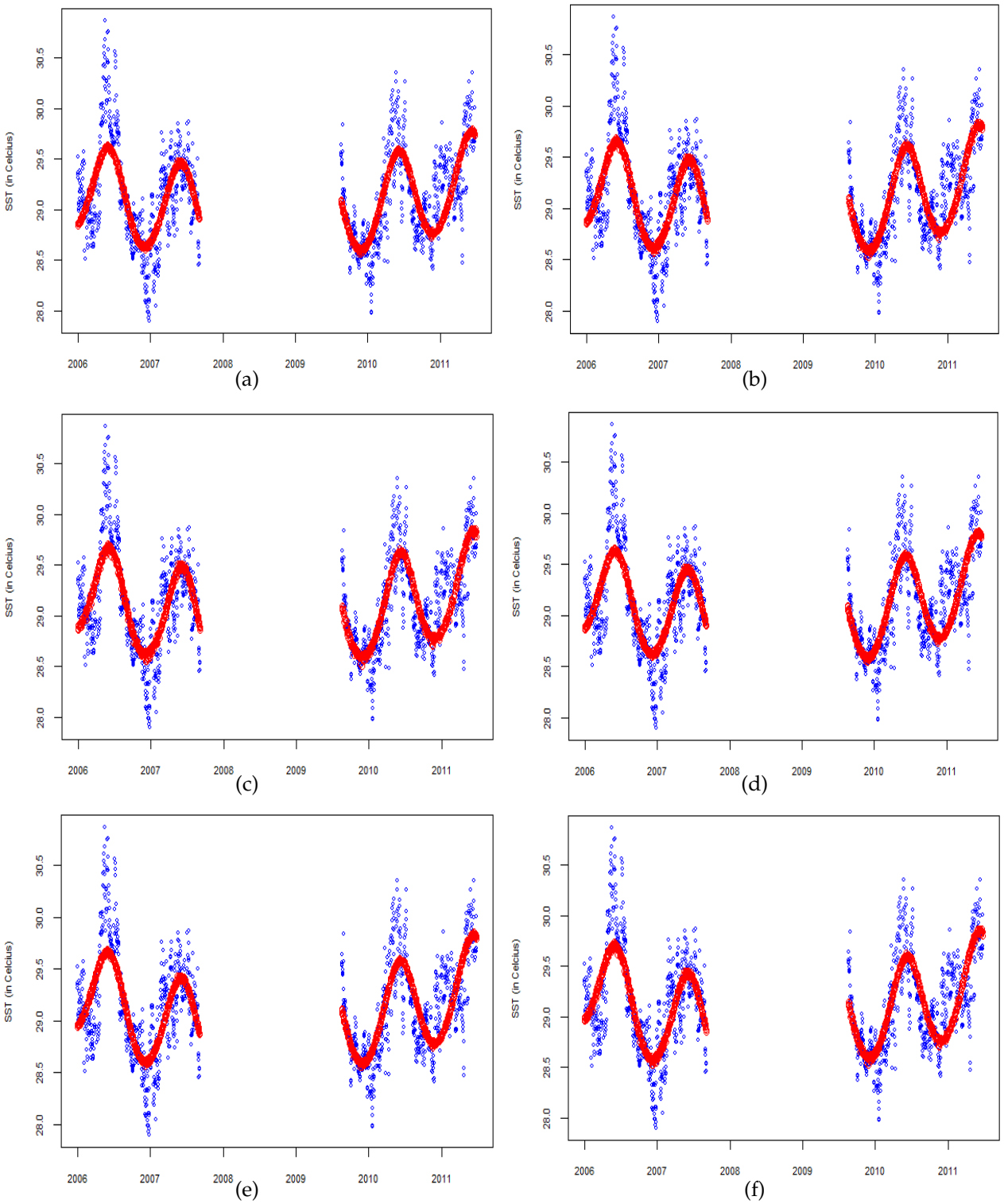


Figure E.16: *The SST data fitting by GMbLSS1post-AR(1) - GMbLSS6post-AR(1) models with different knots, df and m_{stop}. Although the models have different hyper-parameters specifications, they will all have similar patterns in the global fitting, to see in further detail refer to Table 5.8.*

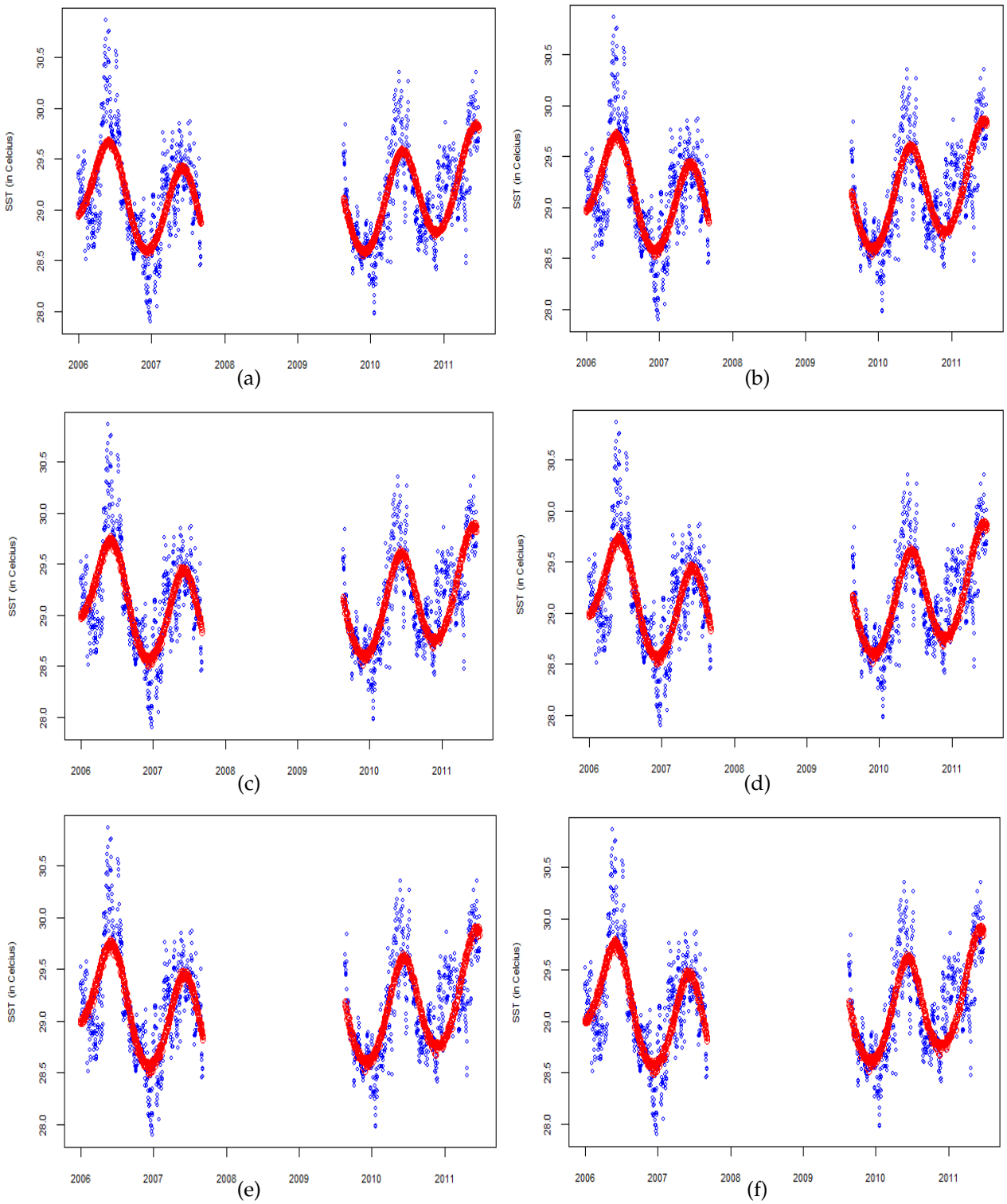


Figure E.17: *The SST data fitting by GMbLSS7post-AR(1) - GMbLSS12post-AR(1) models with different hyper-parameters specifications. The models have similar patterns in the global fitting, to see in detail refer to Table 5.8.*

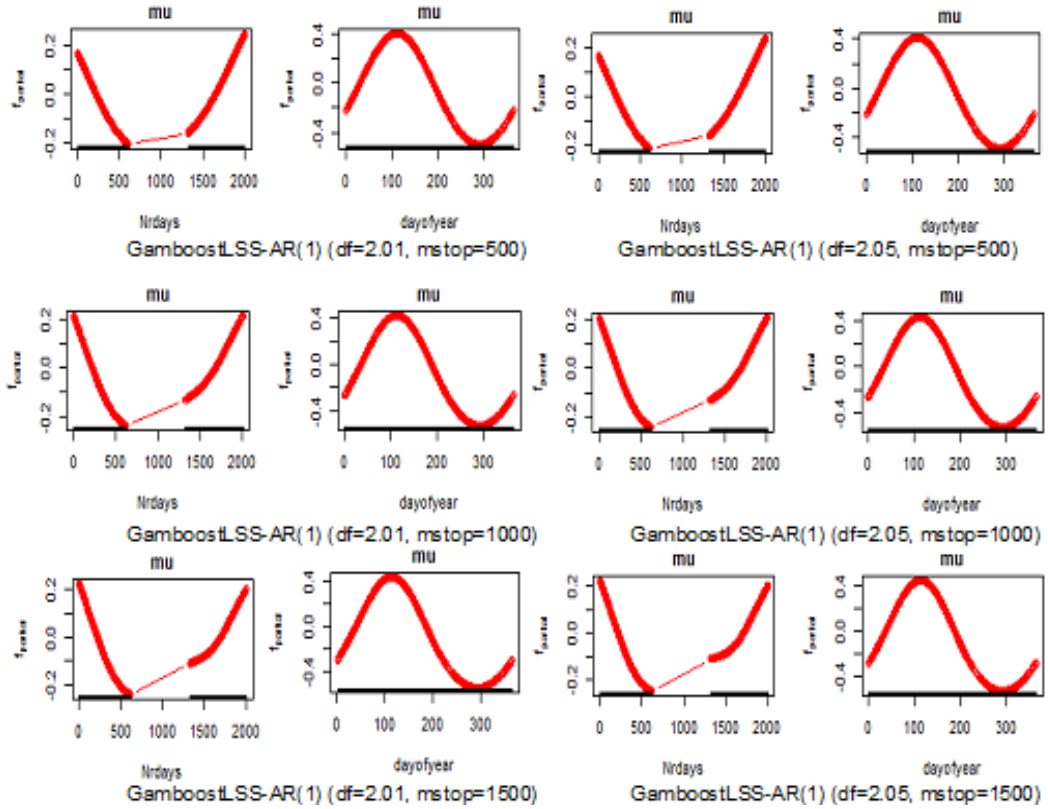


Figure E.18: *Local fitting of time covariates using gamboostLSS-AR(1) models with different m_{stop} and df. In local fitting, it shows a slight difference of df's unchanged patterns of time covariates.*

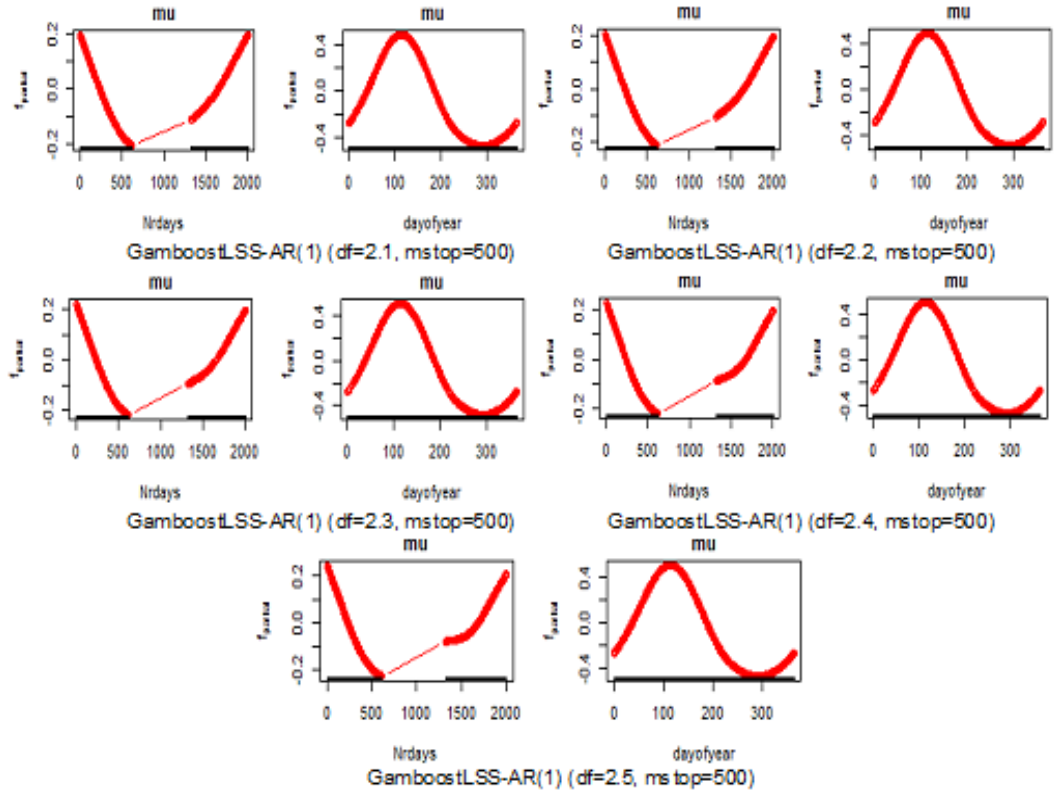


Figure E.19: *GamboostLSS-AR(1) models fitting with different df= 2.1-2.5 and fixed m_{stop}= 500 for the Nrdays and Doy covariates.*

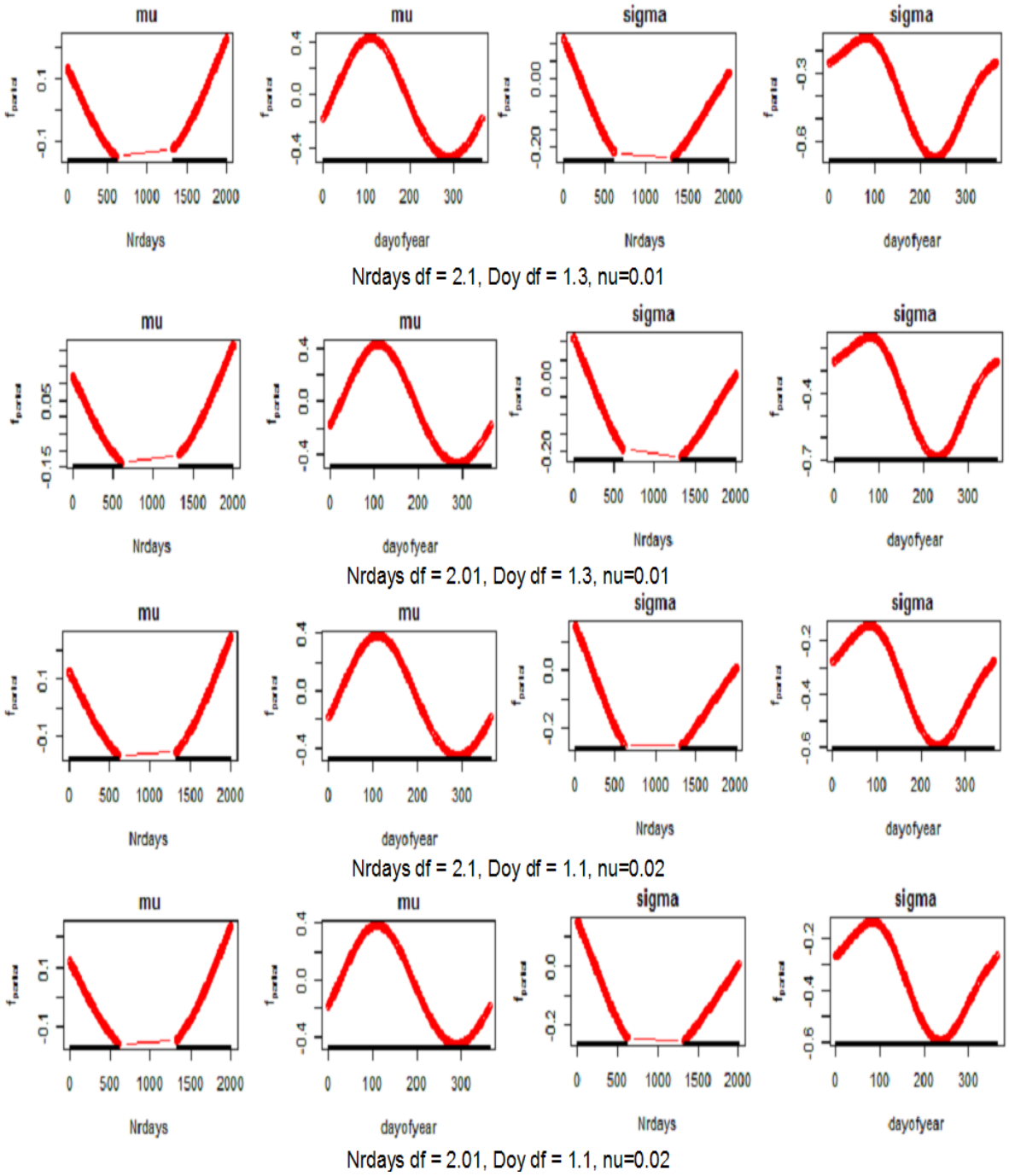


Figure E.20: *The patterns of time covariates in local fitting use gamboostLSS-AR(1) models with transformation. The patterns show a decrease before the gap and an increase after the gap for the Nrdays effect and almost the same pattern for the Doy effect.*