University of CINCINNATI

STAT6031 – Applied Regression Analysis

Prediction of Sea Surface Temperature Based on Remote Sensing Images

Final Report

By:

Xiaolei Li, Liang Zhang, Weiqi Zhao

Instructor:
Emily Lei Kang, Ph.D.

Department of Mathematical Sciences

November 24, 2023

Member Contributions

Xiaolei Li: Liang Zhang: Weiqi Zhao:

Team member	Framework	Coding	Final Report	Presentation
Xiaolei Li	Participate	Participate	Participate	Lead
Liang Zhang	Lead	Lead	Participate	Participate
Weiqi Zhao	Participate	Participate	Lead	Participate

- Xiaolei Li
  - Conducted analysis of model C
  - Wrote report Introduction and section 5.3
  - Made slides
  - Presented this project
- Liang Zhang
  - Built the flow chart for guiding the analysis
  - Conducted analysis of model A and B
  - Enhanced the framework analysis (4 section)
  - Wrote report
- Weiqi Zhao
  - Conducted analysis of framework analysis (section 4)
  - Submitted the prediction to Kaggle
  - Made slides
  - Presented this project

We agree that all team members have contributed in equal measure to this effort.

Team member:

Xiaolei Li	Xiaolei Li
(Print)	(Signature)
Liang Zhang	Liang Zhang
(Print)	(Signature)
Weiqi Zhao	Weiqi Zhao
(Print)	(Signature)

Contents

Prediction of Sea Surface Temperature Based on Remote Sensing Images

1. Introduction

Sea Surface Temperature (SST) is a critical parameter in oceanography and ocean technology, serving as a prerequisite for comprehending and predicting weather, climate dynamics, and planning diverse offshore activities (Patil et al., 2016). Extreme high sea surface temperatures have been identified as a primary contributor to global warming in the past 30 years (Yang et al. 2017). Observations of sea surface temperature are indispensable for unraveling the intricacies of the ocean's interaction with the global climate (Usharani, 2023). Therefore, recent years have seen a growing focus on predicting SST in various ocean-related domains, including fisheries, global warming, and oceanic environmental protection.

Remote sensing is an effective and efficient approach to recording the variation of the sea surface temperature (SST). For example, in the Great Barrier Reef area, Australia, the image information of the local positions (i.e., longitude and latitude) and their corresponding SST varying with years (i.e., from Year 2003 to 2017) is collected by remote sensing for the predictions of SST. Currently, these datasets can be accessed through Kaggle (https://kaggle.com/competitions/stat6031-fall-2023-final-project). A preliminary study shows that global climate models (GCMs) provide informative future projections for sea surface temperature (SST). However, these projections are presented on relatively coarse spatial resolutions due to computational limitations. Further effort is needed for the enhancement of spatial resolutions. As such, this design project takes the Great Barrier Reef area as a study area and aims to construct more refined models for the predictions of SST in the Great Barrier Reef area. Herein, multiple linear regression models are applied as an initial prior.

The remainder of the design is organized as follows. In Section 2, the design objective is highlighted. In Section 3, simple linear regression models are built to examine the linear relationship between predictors and SST, which will provide insight into the selections of the full models. In Section 4, a framework is proposed to determine the most preferable reduced regression models with a preset full multiple linear regression model. In Section 5, different full models are plugged into this framework by considering different influential factors. In Section 6, the recommended model is obtained by evaluating the accuracy of these different preferable reduced models in terms of root mean square error (RMSE). Finally, the limitations of this design project are shortly discussed as well as potential improvements.

2. Design Objective

In the dataset on Kaggle, the SST at each local position is estimated by that at the four nearest points denoted as n1, n2, n3 and n4, respectively. The location information and the corresponding SST are collected at each point. The information collected is tabulated in Table 1. The postfix ".n1", ".n2", ".n3" and ".n4" indicate that the information is originated from n1, n2, n3, n4, respectively. A total of approximately 460 million data (denoted as the training set) were recorded during the Years 2003-2017 while approximately 930 thousand local positions (denoted as the test set) were selected for the predictions of SST from the Year 2018 to 2020 in Great Barrier Reef area.

Table 1. Information collected in the dataset on Kaggle

Estimate point			Four nearest points						
lon	lat	SST	lon.n1	lat.n1	GCM.n1	…	lon.n4	lat.n4	GCM.n4

Note: lon = longitude; lat = latitude; SST = sea surface temperature; GCM = SST estimated from GCMs.

All the variables in Table 1 will be used during the model construction. The design objective is summarized as follows:

- To construct the most preferable linear models for predictions of SST from Year 2018 to 2020 in the Great Barrier Reef area
- To give a better understanding of how to construct multiple regression models
- To give a better understanding of how to evaluate the performance of multiple regression models
- To give a better understanding of how to determine the most preferable reduced linear regression models

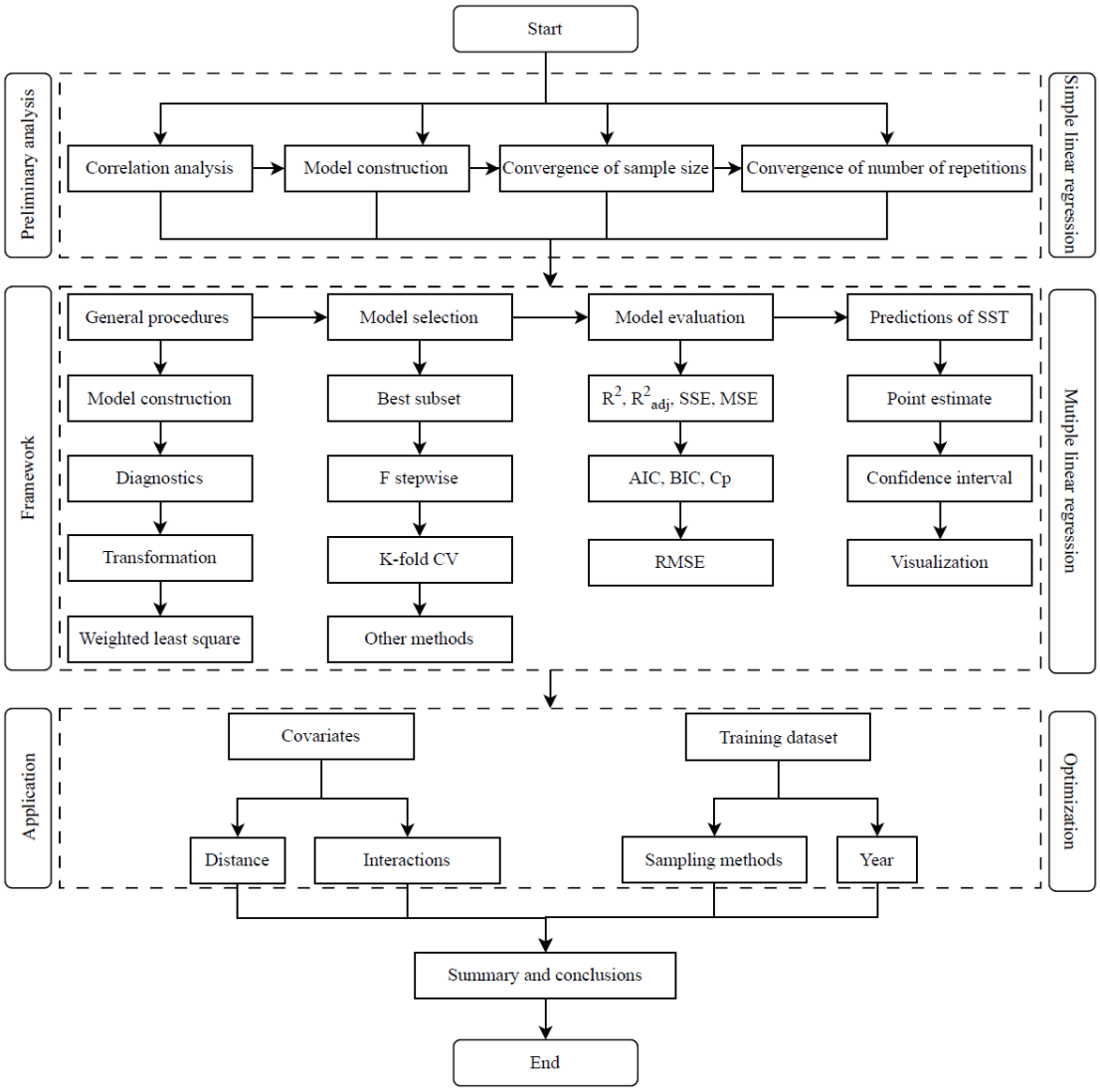


Figure 1. Flowchart in the design project.

Figure 1 shows a flowchart in the design project. The procedures are described as

follows:

Step 1: Build simple linear regression models for preliminary analysis before multiple linear regression model construction. In this step, the linear relationship between SST and all the single predictors (e.g., GCM.n1, GCM.n2) involved is first examined, which will provide insight to determine if the transformation of predictors is required or if the predictors should be added. Note that the training set has a large size. Considering a trade-off between computational burden and computational resources, only partial data from the training set will be sampled for the model construction and validation. Without loss of generosity, data from the Years 2003 to 2016 are sampled for model construction whereas the data from the Year 2017 are for model validation. The convergence of predictions and model stability (i.e., robustness) with several random sample sizes are verified. The error level will be evaluated by root mean square error (RMSE) (see Eq. (1)) during all the model construction.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{1}$$

Step 2: Propose a framework for multiple linear regression model construction. First, an arbitrary full multiple linear regression model is constructed followed by model diagnostics to check the assumptions during model construction. Transformation of predictors or SST is determined by the diagnostic results and automatically checked by its lambda (EmilyUC 2023). Note that the weighted least square (WLS) method can reach the same accuracy as ordinary least square (OLS) but is more robust. The WLS is also applied to this framework. Afterward, the best subset, F stepwise, K-fold and other methods are adopted for the model selection of the reduced models. The most preferable models are determined by different criteria such as R2, AIC, and BIC. The model accuracy in terms of RMSE is evaluated for the final model selection. The recommended model with minimum RMSE is used for the point estimate and 95% confidence interval prediction of SST.

Step 3: Apply different multiple linear regression models to the framework considering different influential factors such as covariates involved and training sets from different sampling methods and years. The different recommended linear models with different full models are obtained from the framework.

Step 4: Evaluate the model accuracy by comparing the RMSE of all the recommended models. The model yielding minimum RMSE is selected as the final recommended model. Summary and conclusions are drawn based on the results presented.

3. Simple linear regression

The position information from the four nearest points is converted into the distance to the estimated point. For example, the longitude distance from n1 is distance.lon.n1=abs(lon.n1-lon). Therefore, a total of 12 predictors are added to the full multiple linear regression models. The model for simple linear regression is:

Model for simple linear regression:
SST~GCM.n1
SST~GCM.n2

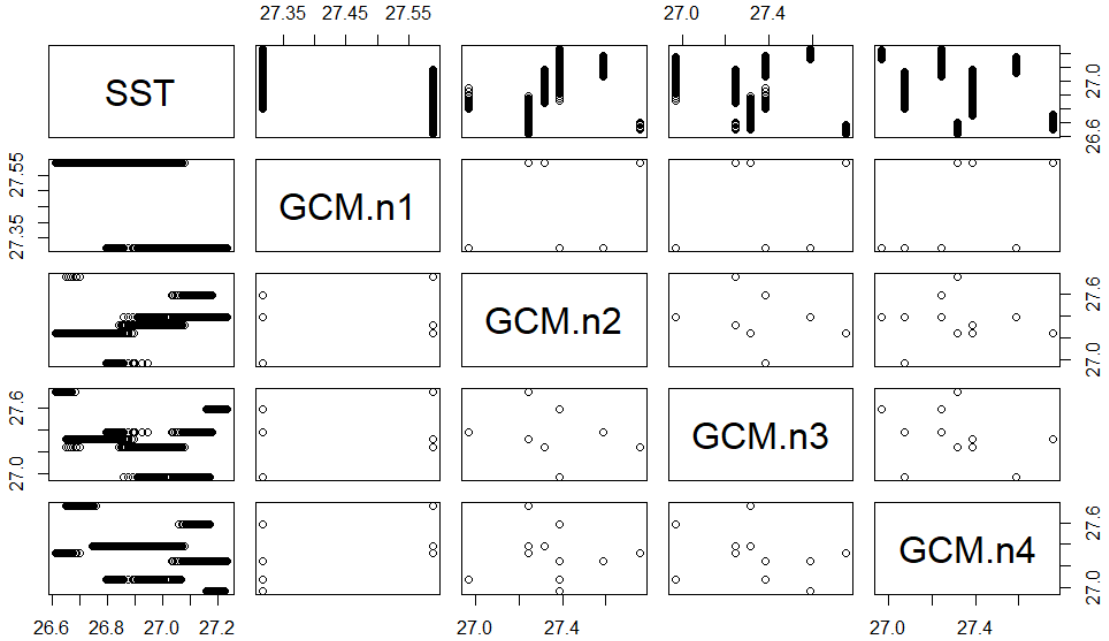
SST~GCM.n3
SST~GCM.n4
SST~distance.lon.n1
SST~distance.lon.n2
SST~distance.lon.n3
SST~distance.lon.n4
SST~distance.lat.n1
SST~distance.lat.n2
SST~distance.lat.n3
SST~distance.lat.n4

3.1 Simple linear regression model construction

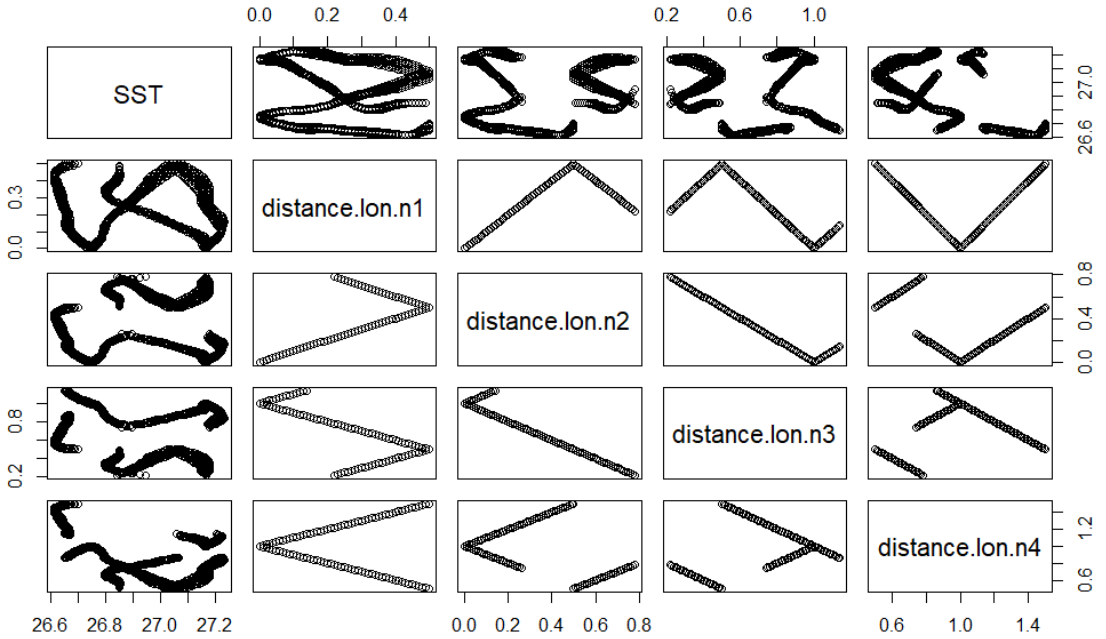
We assume the residuals ε_i satisfy:

$\varepsilon_i \sim iid, N(0,\sigma^2)$, $E(\varepsilon_i)=0, Var(\varepsilon_i)=\sigma^2, COV(\varepsilon_i,\varepsilon_j)=0$ for $i \neq j$.

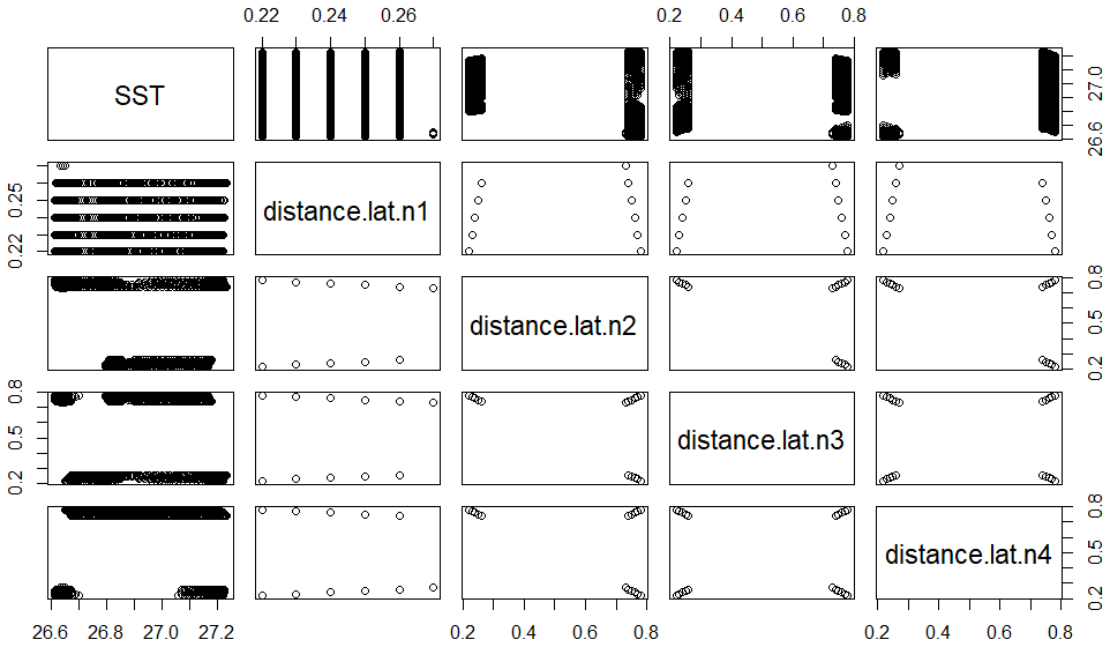
Correlation analysis results show that 12 predictors are uncorrelated to each other in Figure 2. A strong linear relationship exists between SST and GCM.n1 whereas a nonlinear relationship exists between SST and distance.lon.n1, distance.lat.n1, as illustrated in Figure 3. Diagnostics results indicate that residuals from SST~GCM.n1 are roughly normally distributed while residuals from SST~distance.lon.n1 or distance.lat.n1 do not satisfy normal distributions. Transformation might be needed for model SST~ distance.lon.n1 and SST~ distance.lat.n1. Similar results can be obtained for n2, n3 and n4. Since all the predictors are uncorrelated to each other, all of them will be applied to the multiple linear regression model.



(a) Correlation among GCM.n1, GCM.n2, GCM.n3, GCM.n4

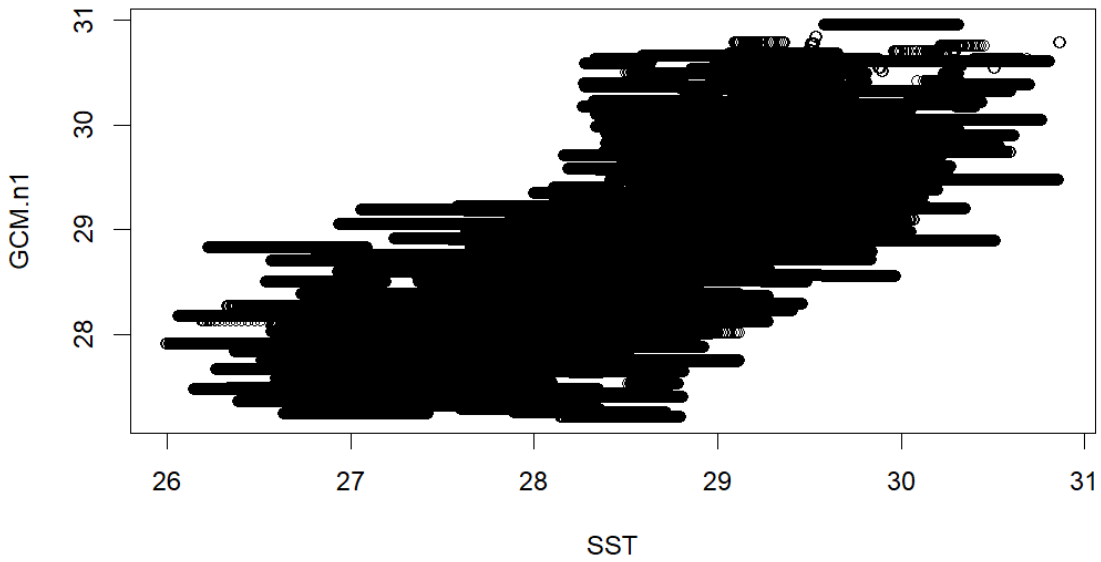


(b) Correlation among distance.lon.n1, distance.lon.n2, distance.lon.n3, distance.lon.n4

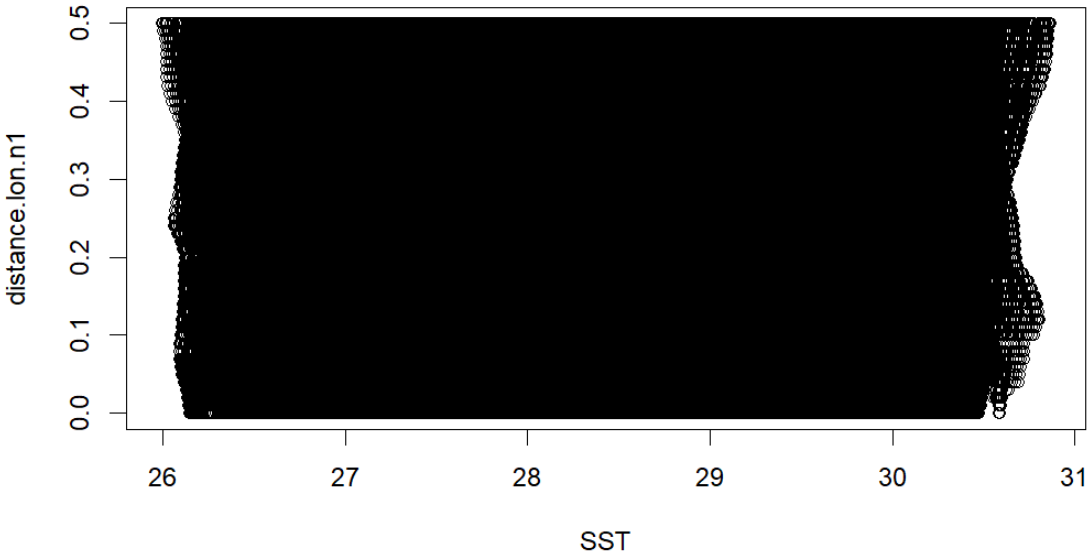


(c) Correlation among distance.lat.n1, distance.lat.n2, distance.lat.n3, distance.lat.n4

Figure 2. Correlation analysis among predictors



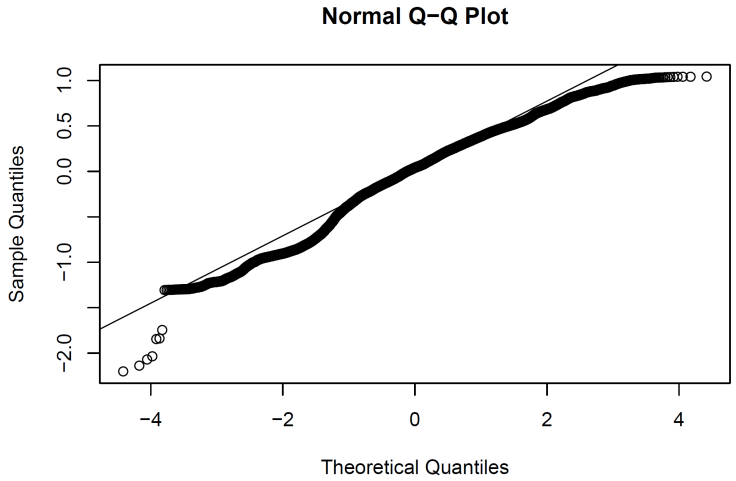
(b) Relationship between SST and GCM.n1



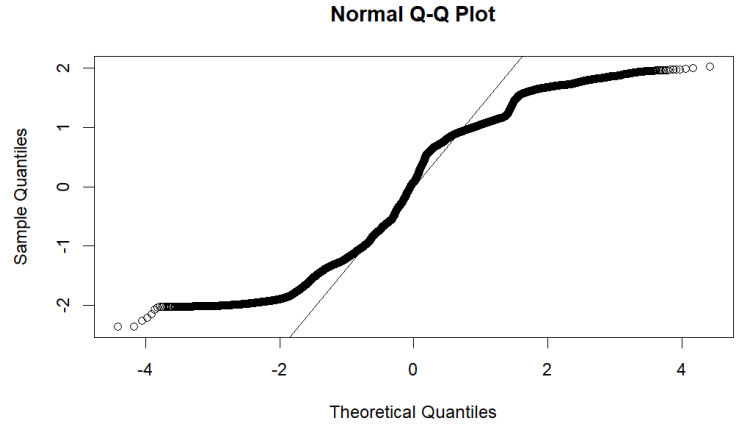
(c) Relationship between SST and distance.lon.n1

(c) Relationship between SST and distance.lat.n1

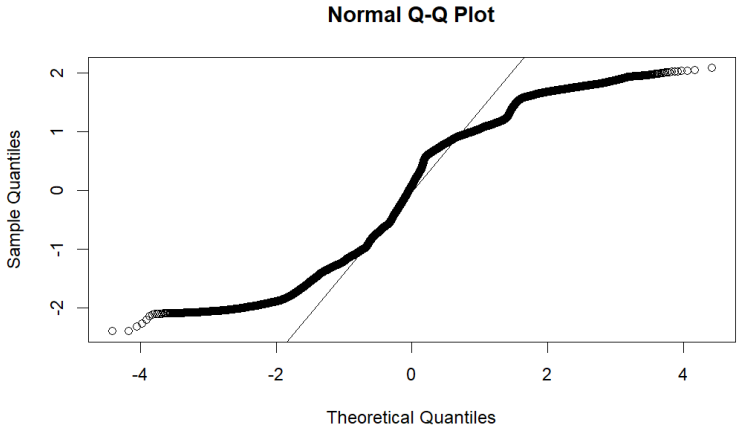
Figure 3. Simple linear regression model for n1.



(a) GCM.n1



(b) distance.lon.n1



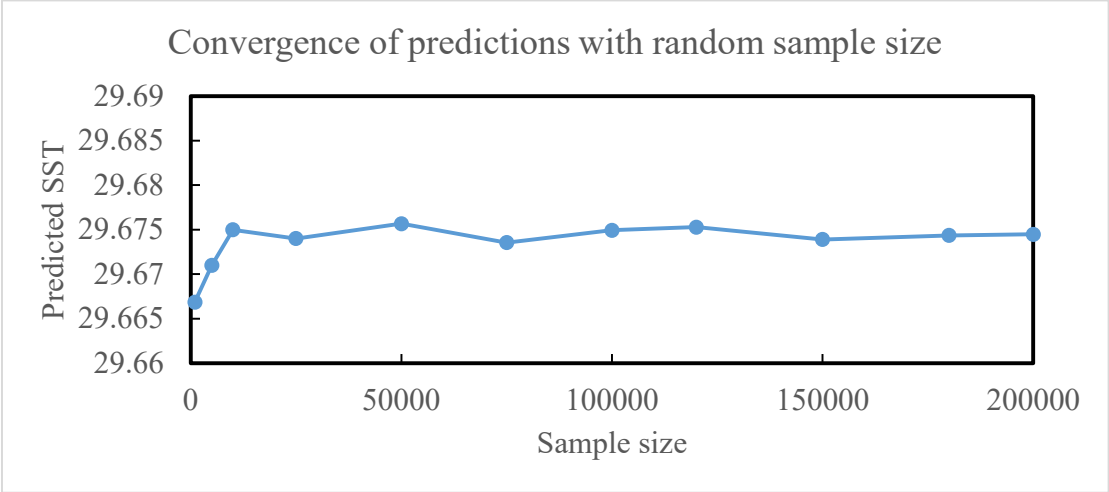
(c) distance.lat.n1

Figure 4. QQ-plot for residuals from a simple linear regression model for n1.

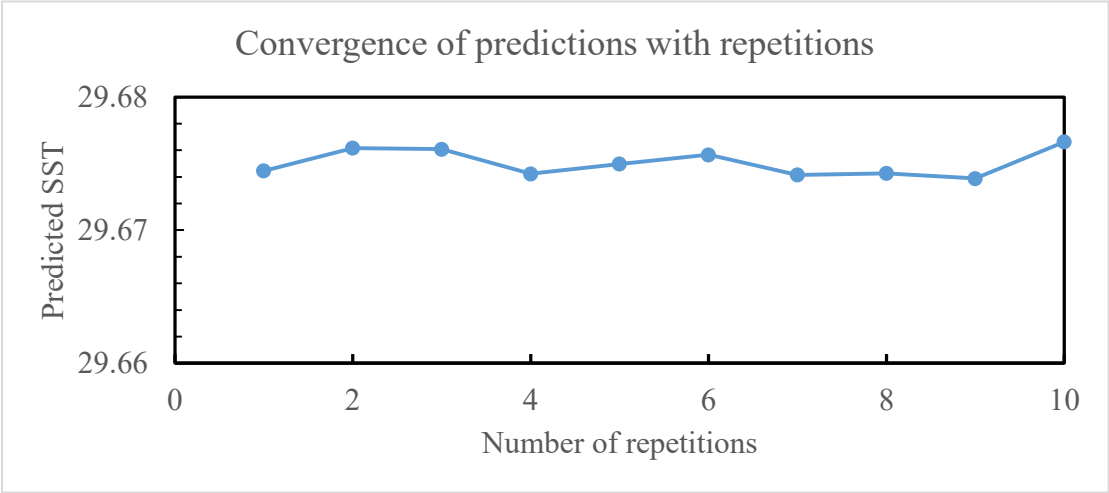
3.2 Convergence of sample size and number of repetitions

Since only partial data of the training data are used for model construction, the sample size should be large enough to capture the characteristics of the training data. The

predictions from the regressed model should be robust enough with the sampled partial data. Figures 2 and 3 show that GCM.n1, GCM.n2, GCM.n3, and GCM.n4 are uncorrelated but give considerable expiation to SST. Therefore, a full multiple regression model, SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4, is adopted to investigate the convergence of the predictions. Figure 5(a) shows that the predicted SST is converged with 50,000 data collected from each year. 100,000 data is collected from each year for the following multiple linear regression model construction. After 100,000 data in each year are randomly sampled from the training set for the model construction, the robustness is also guaranteed as illustrated in Figure 5(b).



(a) Sample size



(b) Number of repetitions

Figure 5. Convergence of predicted SST with multiple regression model: SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4

4. Framework to construct multiple linear regression models

4.1 General procedures to construct multiple linear regression models

Without loss of generosity, the following full model is adopted for the illustration of the framework. 100, 000 data are equally sampled by year from the Year 2003-2017.

Full model: SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4

4.1.1 Model construction with multiple linear regression

Figure 6 shows the multiple linear regression results from R. It can be found that R2 = 0.53. It means 53% of the variation in the response variable is explained by the four predictors.

```
Call:
lm(formula = SST ~ GCM.n1 + GCM.n2 + GCM.n3 + GCM.n4, data = sst.training.samples)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97555 -0.33810  0.03785  0.35425  1.67362

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.775564   0.017212  393.66   <2e-16 ***
GCM.n1      0.201973   0.002018  100.08   <2e-16 ***
GCM.n2      0.285351   0.001927  148.07   <2e-16 ***
GCM.n3      0.143075   0.001704   83.94   <2e-16 ***
GCM.n4      0.124579   0.001829   68.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5324 on 1399995 degrees of freedom
Multiple R-squared:  0.5357,    Adjusted R-squared:  0.5357
F-statistic: 4.038e+05 on 4 and 1399995 DF,  p-value: < 2.2e-16
```

Figure 6. Multiple linear regression results from R for full model: SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4.

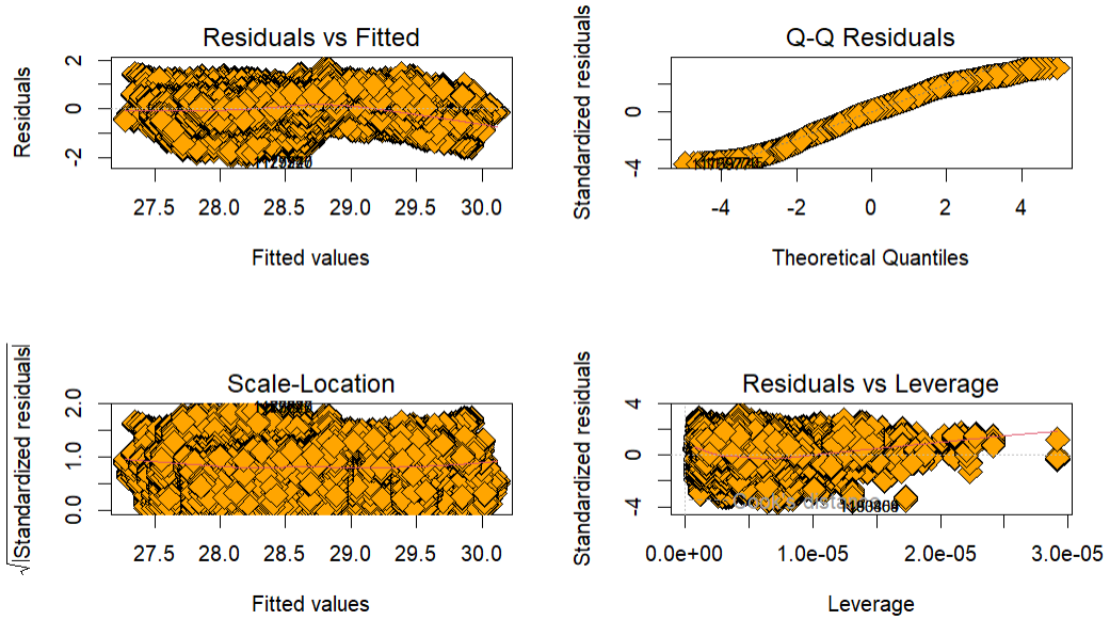


Figure 7. Diagnostics for full model: SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4.

4.1.2 Transformation

Figure 8 shows that the optimal lambda is approximately 5. However, it is found that no improvement is reached by this transformation (see Table 2). Therefore, transformation is not considered in Section 5.

Table 2. Transformation results

	R^2	RMSE
Without transformation	0.5357	0.532357
Transformation	0.5357	0.5291807

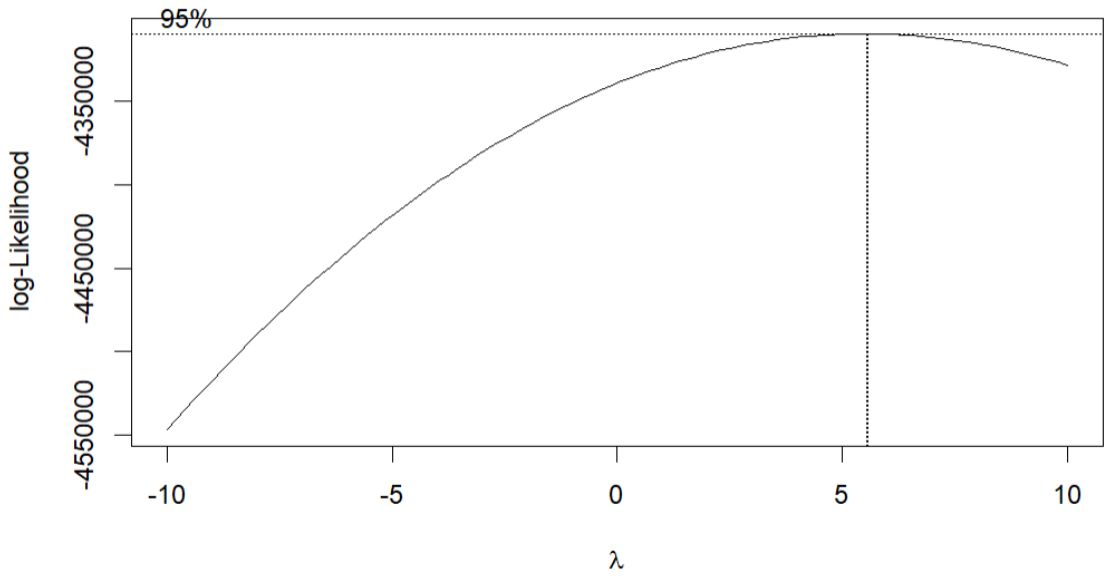


Figure 8. Best lambda for transformation.

4.1.3 Weighted Least Square (WLS)

It is found that no improvement is reached by WLS in Table 3. Therefore, WLS is not considered in Section 5.

Table 3. WLS results

	R^2	RMSE
Without transformation	0.5357	0.532357
Transformation	0.5289	0.5324281

4.2 Model selection with different criteria

In this section, the best subset, stepwise, k-fold CV, and other methods are adopted for the model selections. R^2, R^2_{adj}, AIC, BIC, Cp are evaluated for these methods. R output with each method is attached in the Appendix. It is interesting to find that all the recommended model based on different methods is SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4. Therefore, the final recommended model is the same as the preset full model SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4. This full model will be used for model validation and model prediction.

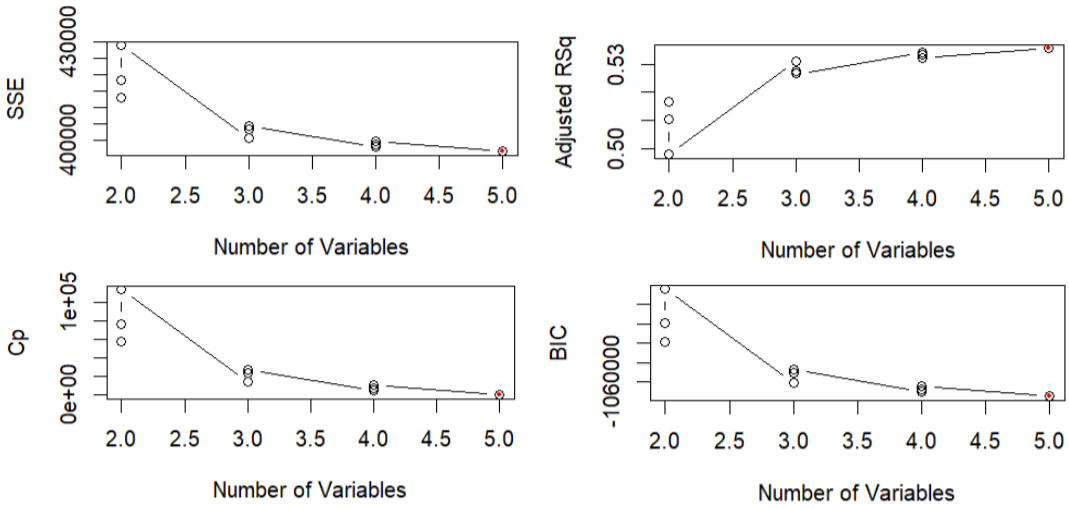


Figure 9. Model selection with the best subset.

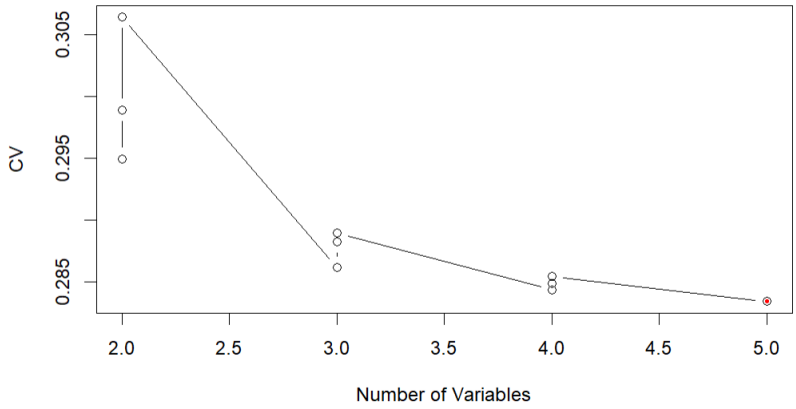


Figure 10. Model selection with K-fold CV.

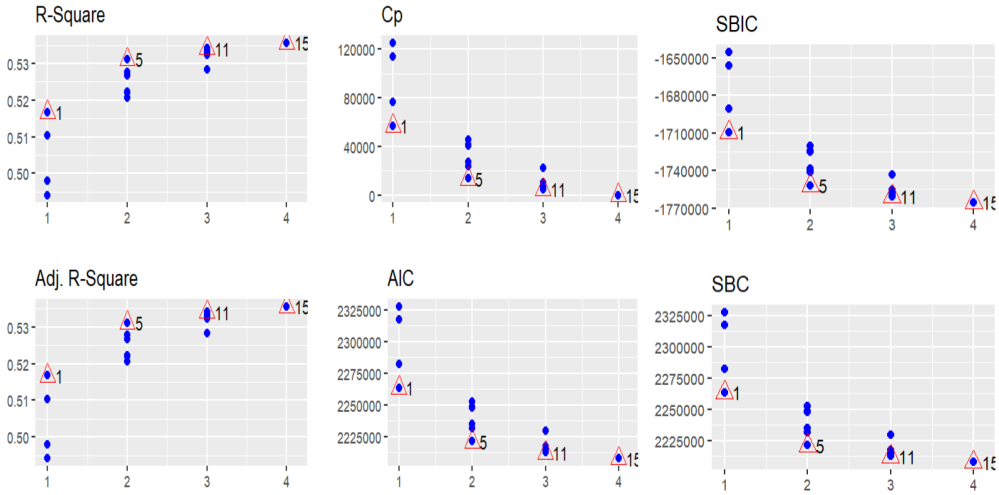


Figure 11. Model selection with other methods.

Table 4. Recommend model for selected full model: SST~ SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4

Method	Recommend model
Best subset	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4

AIC forward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4
AIC backward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4
AIC both	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4
BIC forward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4
BIC backward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4
BIC both	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4
7-fold CV	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4
Other methods	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4

4.3 Accuracy evaluation of the final model

Figure 12 shows that only a small proportion of validation samples (denoted as true data braced by a 95% confidence interval of the predictions from the model) are located at the 1:1 line. This means this recommended linear model is not good enough to predict SST in the Year 2017.

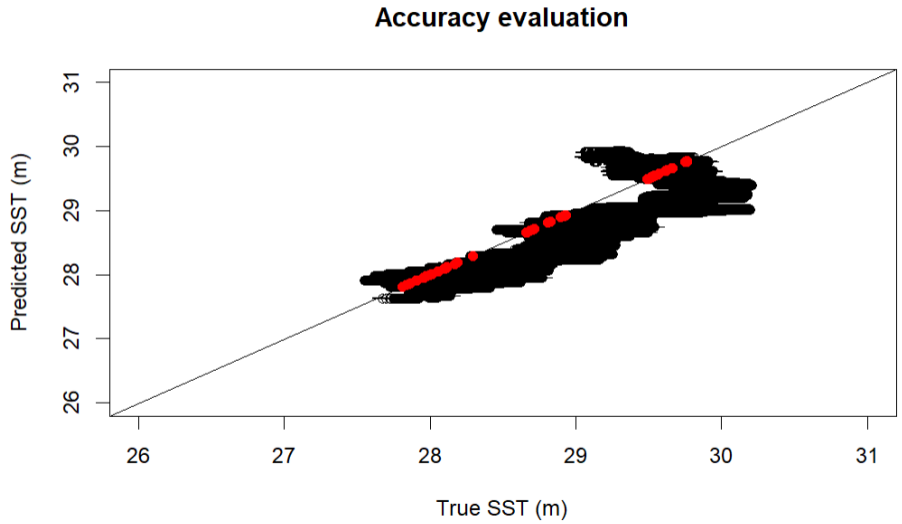


Figure 12. Accuracy evaluation for the recommended model.

4.4 Predictions of SST from Year 2018 to 2020

The point estimate and 95 confidence interval of SST from Year 2018 to 2020 are stored in the variables sst.test.predict.Y and sst.test.predictCI.Y, respectively.

4.5 Visualizations of predictions

As expected, there is a huge gap between the predicted SST and SST from GCM as shown in Figure 13. Therefore, other full models will be plugged into this framework to give better predictions of SST.

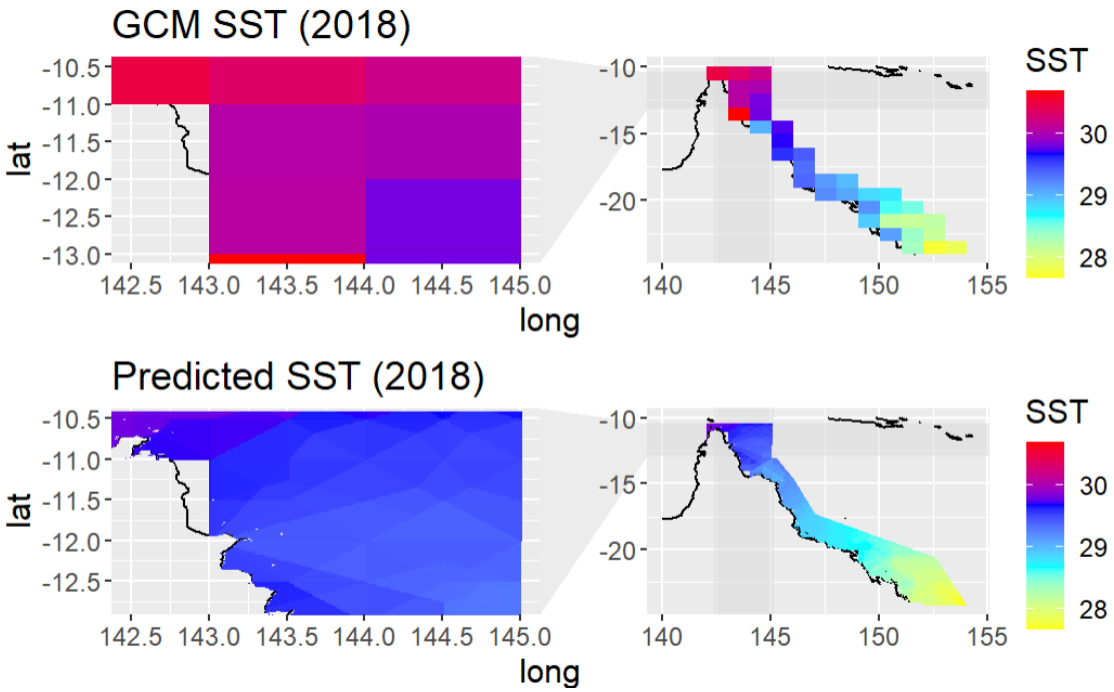


Figure 13. Visualization of prediction of SST with model: SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4

5. Application of the framework

5.1 Model A:

SST~GCM.n1+GCM.n2+GCM.n3+GCM.4+distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4, stratified sampling by year from Year 2003-2016.

Table 5 indicates that the recommended model is dropping distance.lon.n2 from the full model based on most of the criteria.

Table 5. Recommend model for selected full model: SST~GCM.n1+GCM.n2+GCM.n3+GCM.4+distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4 (stratified sampling by year)

Method	Recommend model
The best subset with SSE	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
Best subset with R^2.adj	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
Best subset with Cp	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
Best subset with BIC	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3

AIC forward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
AIC backward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
AIC both	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
BIC forward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3
BIC backward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3
BIC both	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3

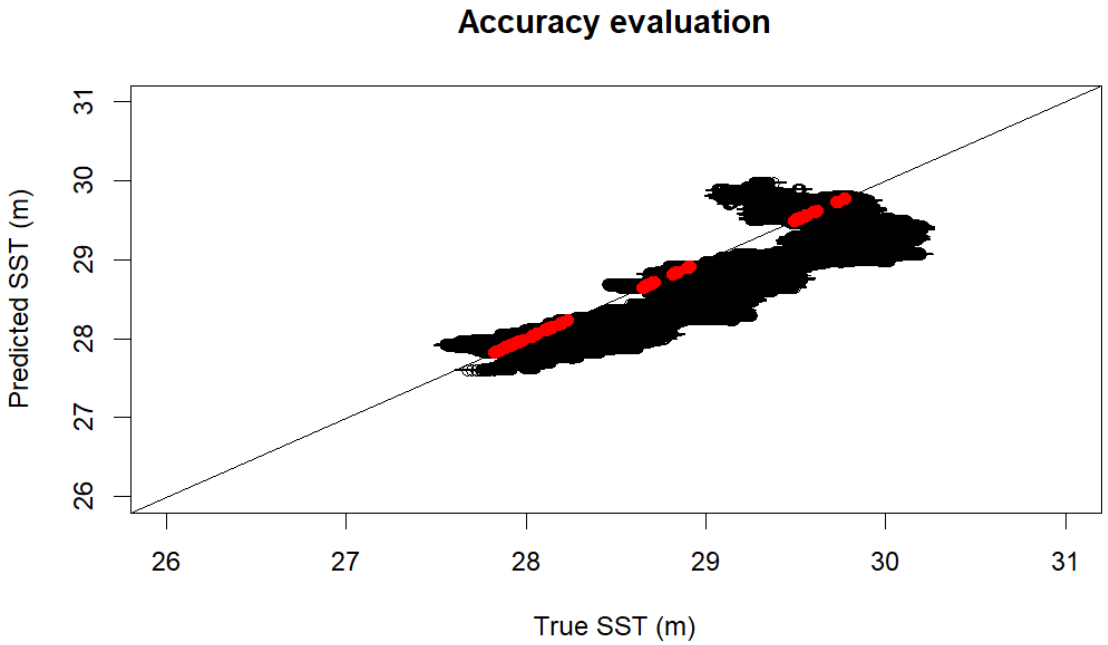


Figure 14. Accuracy evaluation for Model A (R^2= 0.5368, RMSE= 0.4776106).

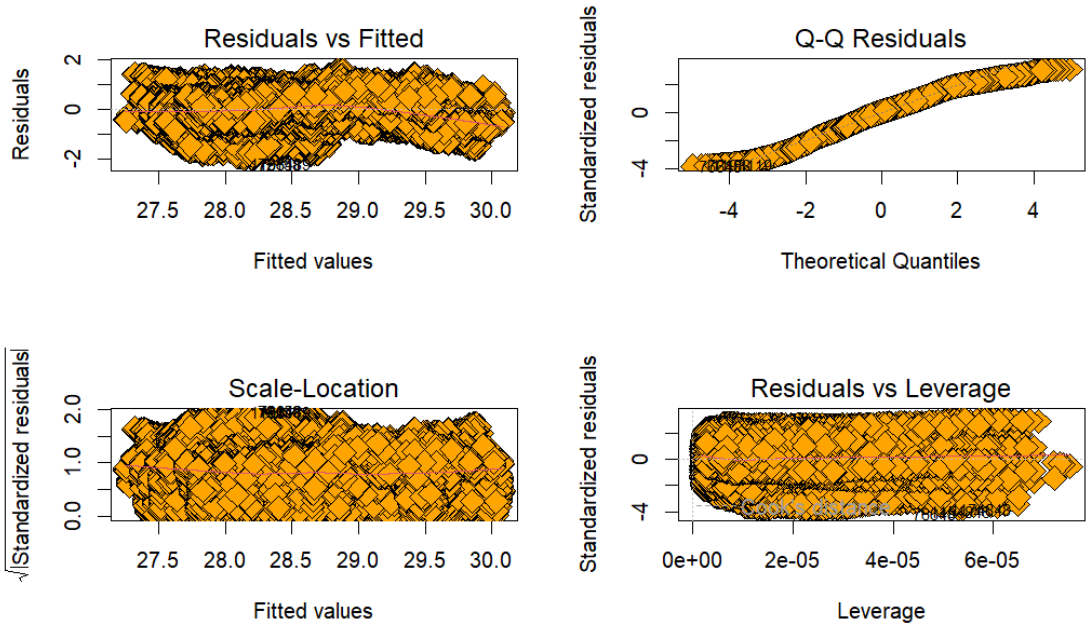


Figure 15. Diagnostics for full Model A.

5.2 Model B:
SST~GCM.n1+GCM.n2+GCM.n3+GCM.4+distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4. All the data from Year 2016 are used.

Table 6 indicates that all the methods recommend the selected full model as the recommended model. Figure 13 shows that slight improvement is reached by Model B in terms of R^2 and RMSE, compared to Model A.

Table 6. Recommend model for selected full model: SST~GCM.n1+GCM.n2+GCM.n3+GCM.4+distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4 (data from Year 2016)

Method	Recommend model
Best subset with SSE	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
Best subset with R^2.adj	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
Best subset with Cp	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
Best subset with BIC	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
AIC forward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4

AIC backward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
AIC both	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
BIC forward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
BIC backward	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4
BIC both	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4

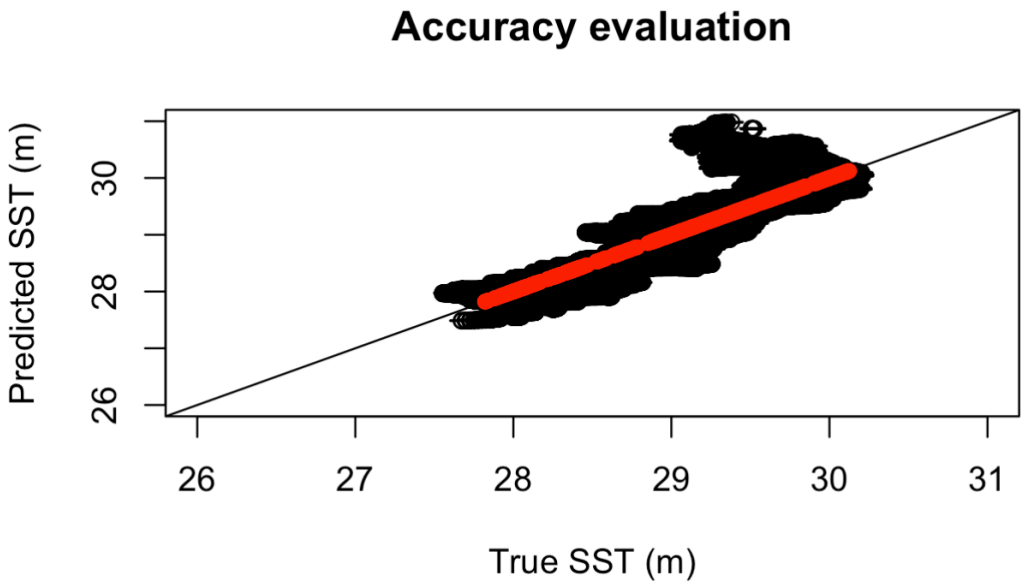


Figure 16. Accuracy evaluation for Model B (R^2= 0.8327, RMSE= 0.3267852).

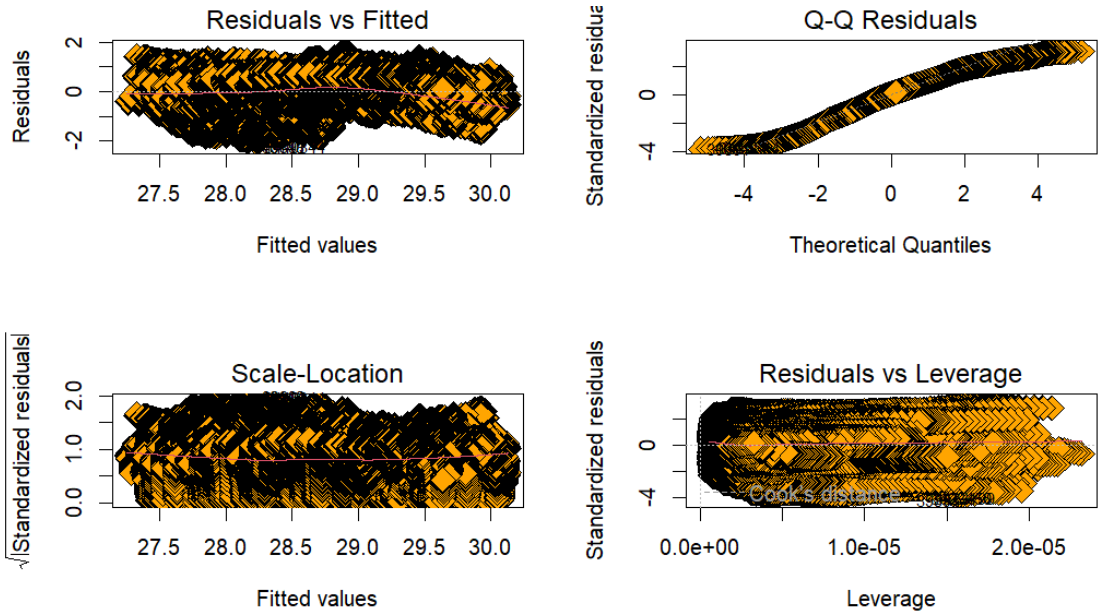
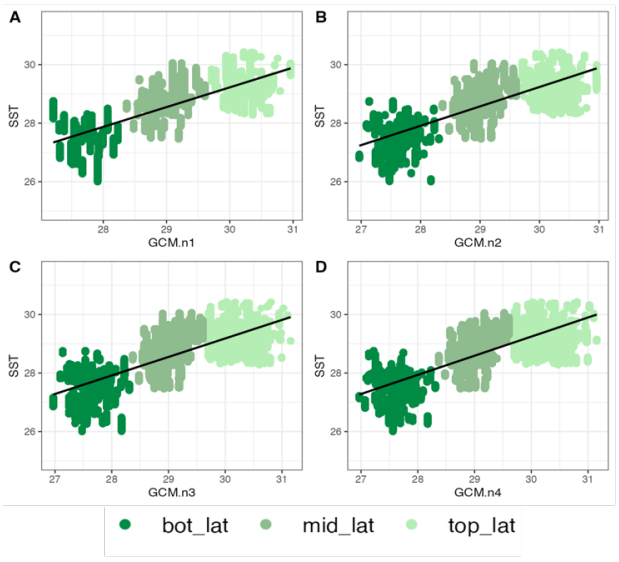


Figure 17. Diagnostics for Model B.

5.3 Model C

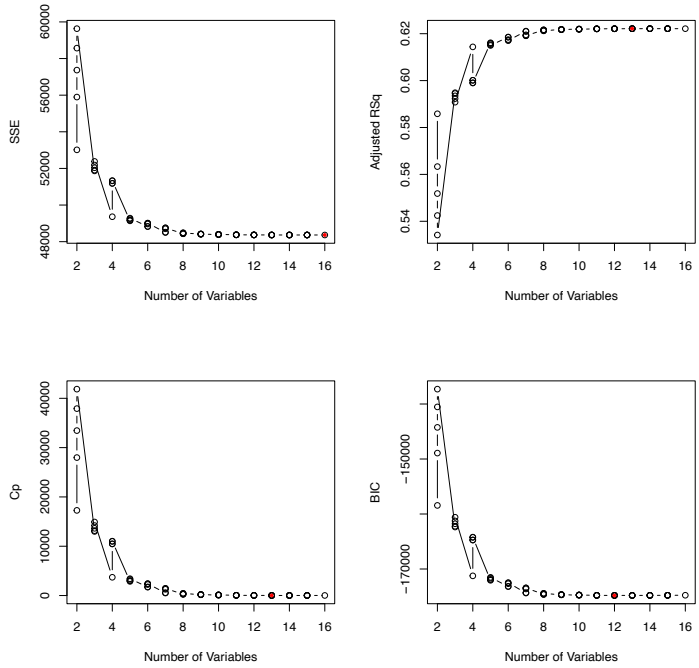
Since SST varies mainly with latitude, a sub-dataset was created based on that for efficient feature selection. 3 layers of latitudes were selected, and they are across from -10.50 to -10.41, -19.27 to -19.09, and -24.28 to -24.14. The association between SST and GCM outputs across different latitudes was shown below. The results showed that SST increases as latitude moving closer to the equator and GCM increases as well. It appears to be a linear relationship between SST and latitudes and GCM values.



Multiple linear regression models were applied on the sub-dataset and the coefficients output was listed below. Hypothesis testing was performed to see if it is useful to include “lat.n2” and “lat.n3” as predictors when others already included in the model. Based on the anova results, p-value associated with F test is very small. Thus, we concluded that it’s necessary to contain “lat.n2” and “lat.n3” in the model.

```
Call:
lm(formula = SST ~ GCM.n1 + GCM.n2 + GCM.n3 + GCM.n4 + lat.n1 +
    lat.n2 + lat.n3 + lat.n4 + lat + lon.n1 + lon.n2 + lon.n3 +
    lon.n4 + lon + year, data = combine)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84365 -0.33586  0.05843  0.34599  1.68130

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.7077595  0.6171192  44.899  < 2e-16 ***
GCM.n1      -0.0080854  0.0084052  -0.962   0.3361
GCM.n2       0.7942749  0.0082419  96.370  < 2e-16 ***
GCM.n3      -0.0598809  0.0073897  -8.103 5.38e-16 ***
GCM.n4       0.0870463  0.0079149  10.998  < 2e-16 ***
lat.n1      -1.5229759  0.0217681 -69.964  < 2e-16 ***
lat.n2       0.0089427  0.0057910   1.544   0.1225
lat.n3      -0.0052644  0.0046647  -1.129   0.2591
lat.n4       0.0140744  0.0030002   4.691 2.72e-06 ***
lat          1.5053379  0.0219979  68.431  < 2e-16 ***
lon.n1      -0.0485292  0.0071859  -6.753 1.45e-11 ***
lon.n2       0.1018160  0.0044799  22.727  < 2e-16 ***
lon.n3      -0.0247723  0.0027683  -8.948  < 2e-16 ***
lon.n4       0.0014830  0.0021101   0.703   0.4822
lon         -0.0289092  0.0127912  -2.260   0.0238 *
year        -0.0115656  0.0003614 -32.003  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5187 on 179744 degrees of freedom
Multiple R-squared:  0.6222,    Adjusted R-squared:  0.6221
F-statistic: 1.973e+04 on 15 and 179744 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Model 1: SST ~ GCM.n1 + GCM.n2 + GCM.n3 + GCM.n4 + lat.n1 + lat.n4 + lat +
    lon.n1 + lon.n2 + lon.n3 + lon.n4 + lon + year
Model 2: SST ~ GCM.n1 + GCM.n2 + GCM.n3 + GCM.n4 + lat.n1 + lat.n2 + lat.n3 +
    lat.n4 + lat + lon.n1 + lon.n2 + lon.n3 + lon.n4 + lon +
    year
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 179746 48370
2 179744 48364  2    5.9254 11.011 1.653e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

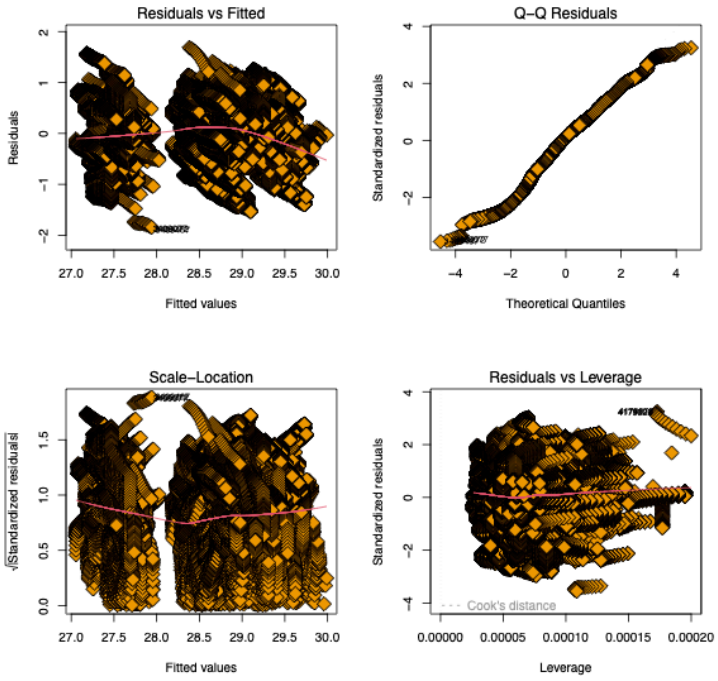
Model selection was conducted to choose a best subset from these 15 predictors based on different criteria. 5 models were chosen for each size and SSE, Ra^2, Cp, and BIC were plotted as below. Based on BIC, 12 variables were given the best results. The optimal candidate model was determined.




```
   (Intercept) GCM.n1 GCM.n2 GCM.n3 GCM.n4 lat.n1 lat.n2 lat.n3 lat.n4  lat lon.n1 lon.n2 lon.n3 lon.n4   lon year
12        TRUE  FALSE   TRUE   TRUE   TRUE   TRUE   TRUE  FALSE   TRUE TRUE   TRUE   TRUE   TRUE  FALSE  TRUE TRUE
12        TRUE  FALSE   TRUE   TRUE   TRUE   TRUE   TRUE  FALSE   TRUE TRUE   TRUE   TRUE   TRUE  FALSE  TRUE TRUE
11        TRUE  FALSE   TRUE   TRUE   TRUE   TRUE   TRUE  FALSE   TRUE TRUE   TRUE   TRUE   TRUE  FALSE FALSE TRUE
```

```
Call:
lm(formula = SST ~ GCM.n2 + GCM.n3 + GCM.n4 + lat.n1 + lat.n2 +
    lat.n4 + lat + lon.n1 + lon.n2 + lon.n3 + year, data = combine)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84279 -0.33600  0.05801  0.34633  1.68486

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.6323969  0.6041261  45.739  < 2e-16 ***
GCM.n2       0.7931463  0.0079523  99.738  < 2e-16 ***
GCM.n3      -0.0639067  0.0066093  -9.669  < 2e-16 ***
GCM.n4       0.0869837  0.0063638  13.668  < 2e-16 ***
lat.n1      -1.5268832  0.0201740 -75.686  < 2e-16 ***
lat.n2       0.0150827  0.0031270   4.823 1.41e-06 ***
lat.n4       0.0145384  0.0025263   5.755 8.68e-09 ***
lat          1.4975208  0.0217537  68.840  < 2e-16 ***
lon.n1      -0.0615911  0.0028093 -21.924  < 2e-16 ***
lon.n2       0.0927999  0.0021444  43.276  < 2e-16 ***
lon.n3      -0.0291351  0.0019742 -14.758  < 2e-16 ***
year        -0.0116427  0.0003519 -33.089  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5187 on 179748 degrees of freedom
Multiple R-squared:  0.6222,    Adjusted R-squared:  0.6221
F-statistic: 2.691e+04 on 11 and 179748 DF,  p-value: < 2.2e-16
```

In order to make sure that this model follows all the assumptions for multiple linear regression, residuals were plotted against their fitted values. The results showed that residuals are roughly uniformly distributed around the horizontal line. Although Q-Q plot did not give a perfectly straight line, it still showed that the residuals are more or less normally distributed.

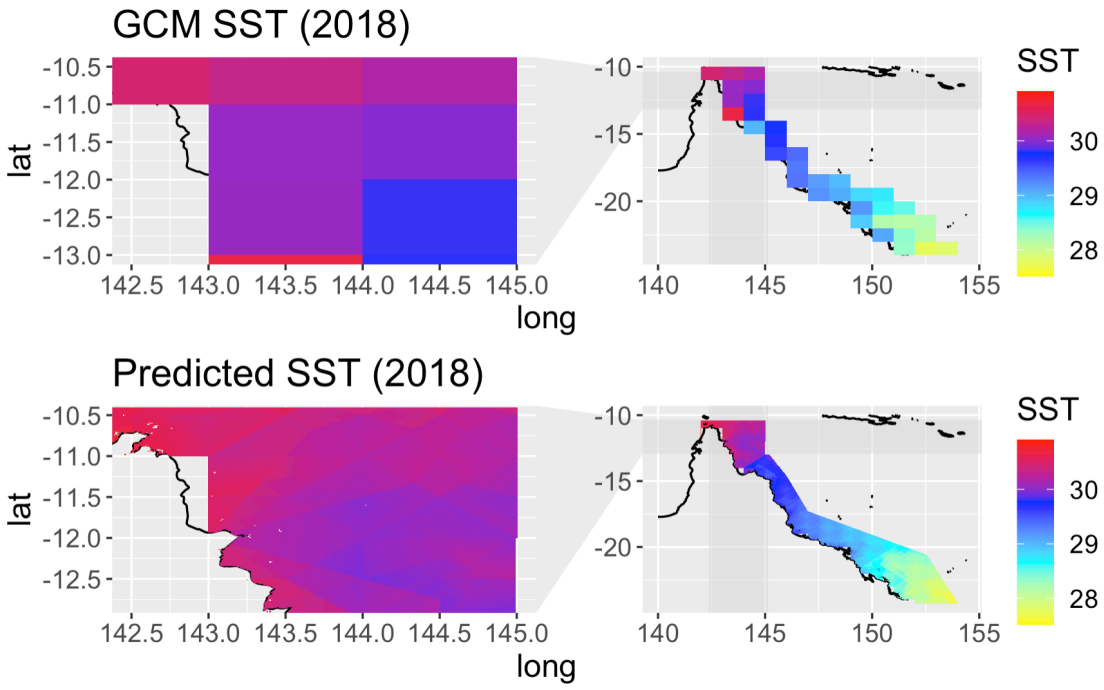


6. Summary and conclusions

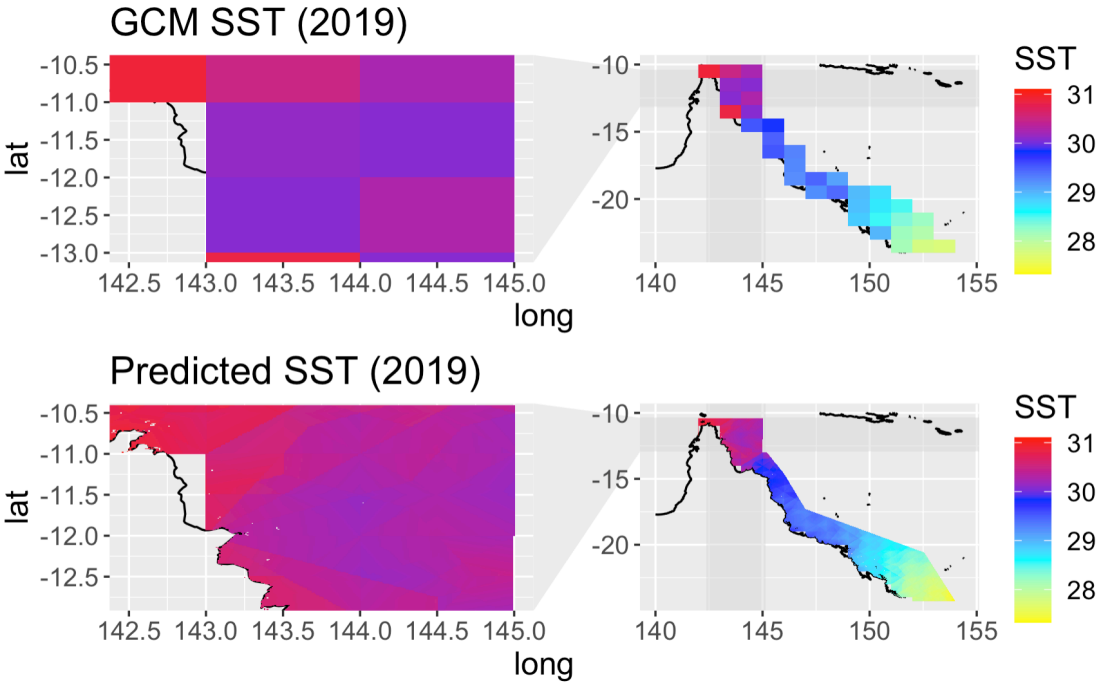
According to the RMSE from different full models, the final recommended model is:
SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+
+distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+
+distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4

Full model	Model selection results	R^2	RMSE
A	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ distance.lon.n1 +distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4	0.5368	0.4776106
B	SST~GCM.n1+GCM.n2+GCM.n3+GCM.n4+ +distance.lon.n1+distance.lon.n2+distance.lon.n3+distance.lon.n4+ +distance.lat.n1+distance.lat.n2+distance.lat.n3+distance.lat.n4	0.8327	0.3267852
C	SST~GCM.n1+GCM.n2 +GCM.n4+ lon.n1+ lon.n2+ lon.n3+ lon.n4+ +lat+lon	0.6222	0.5215781

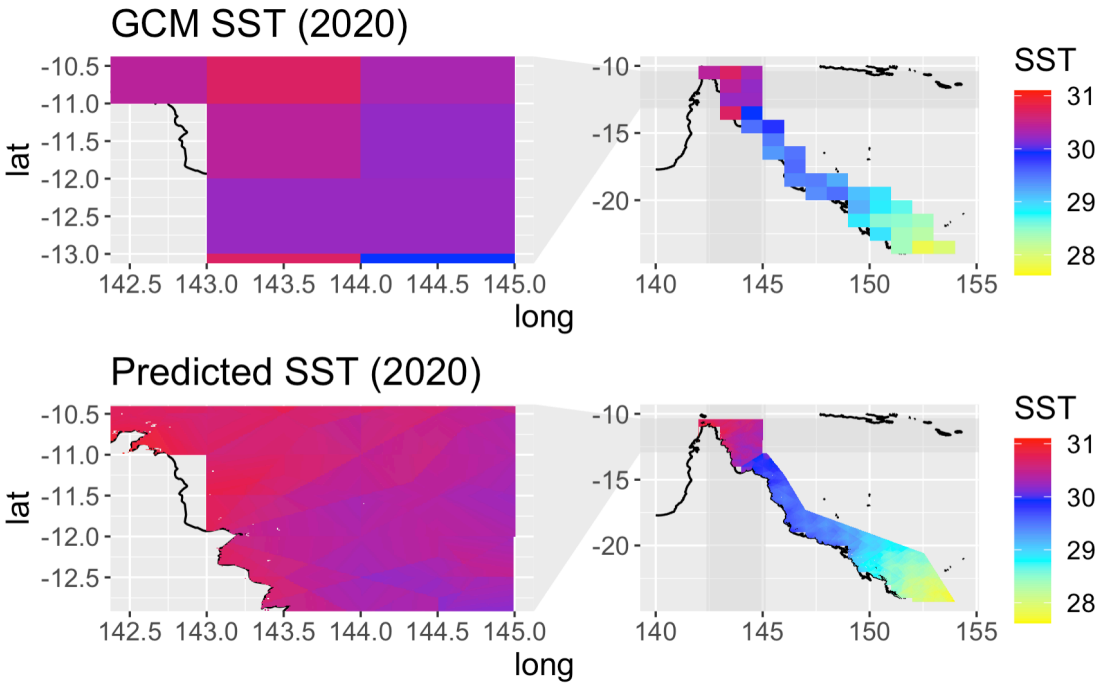
The predictions of SST from Year 2018 to 2019 are illustrated below.



(a) Year 2018



(b) Year 2019



(c) Year 2020

Figure 18. Visualization of predictions of SST from Year 2018 to 2020.

7. Limitation of this design project

In this design project, only linear relationship between SST and predictors is investigated. However, according to the results presented, there is a strong nonlinear relationship between SST and other predictors such as distance.lon.n1. Also, the effect of the interactions among the predictors are not considered.

References

EmilyUC. (2023). STAT6031 Fall 2023 Final Project. Kaggle. https://kaggle.com/competitions/stat6031-fall-2023-final-project

Patil, K., Deo, M. C., & Ravichandran, M. (2016). Prediction of sea surface temperature by combining numerical and neural techniques. Journal of Atmospheric and Oceanic Technology, 33(8), 1715-1726.

Usharani, B. (2023). ILF-LSTM: Enhanced loss function in LSTM to predict the sea surface temperature. Soft Computing, 27(18), 13129-13141.

Yang, Y., Dong, J., Sun, X., Lima, E., Mu, Q., & Wang, X. (2017). A CFCC-LSTM model for sea surface temperature prediction. IEEE Geoscience and Remote Sensing Letters, 15(2), 207-211.