Regression Modeling/Analysis of Sea Surface Temperatures over Time

Advay Kadam (advayk2), UG

STAT 429 Final Project Report, UIUC

May 9, 2025

Author Note

All research, data analysis, statistical modeling, and interpretation was done by Advay Kadam.

Abstract

Sea surface temperatures, defined as the temperature of the top few millimeters of the ocean, play a significant role in interpreting global climate conditions. We aim to effectively predict the average sea surface temperature during El Niño, periodic climate conditions associated with warmer ocean temperatures and severe weather, through regression, and determine what explanatory variables in our dataset are most significant in determining average sea surface temperatures and what variables are not as significant. In a broader context, it is important to develop statistical models with relatively accurate predictions of future sea surface temperatures, as high sea surface temperatures typically serve as a strong indicator of Hurricanes/Typhoons during El Niño. Hence, our analysis and modeling can help provide meteorologists greater insight into shifts and trends in sea surface temperatures, especially in the Pacific Ocean, where the effect of El Niño is most prevalent. In our study, we develop four diverse regression models to predict the average sea surface temperature during an El Niño event. While trying to determine suitable regression models, it was initially believed that traditional regression would be unfit, as the first three models that we trained produced high AIC/BIC values, and generally the residuals did not appear to represent white noise for the models. However, through regression on autocorrelated errors, we noticed that we were able to improve the AIC/BIC and R2 values compared to other previous regression models and see greater evidence that the residuals indicate white noise. Additionally, the correlated error regression model is shown to be best for short-range forecasting, and the other three regression models may be utilized for long-range forecasting with caution. Ultimately, these results imply that we can effectively predict the average sea surface temperatures given various parameters and determine through experimentation the best choice of features for our model.

Introduction

El Niño and La Niña are complex climate/weather patterns that take place in the Pacific Ocean and are responsible for changes in weather conditions across the world. El Niño is especially relevant to the central and eastern tropical Pacific Ocean, as El Niño years tend to have more hurricanes in their region of the Pacific. Additionally, the time frame from 1982-1983 has historically resulted in the strongest El Niño conditions, leading to the question of whether we can effectively predict the weather conditions caused during an El Niño cycle and what variables most significantly contribute to this prediction. More precisely, we hope to see which regression model is most effective at forecasting sea surface temperatures.

The dataset for this study originates from the UC Irvine Machine Learning Repository. The data was collected using the Tropical Atmosphere Ocean (TAO) Array, which is a system designed to collect more information on and understand El Niño and La Niña climate variations. This system is compressed to 70 buoys spanning the equatorial Pacific Ocean, measuring various meteorological parameters in real-time. This system of measurement was also combined with temperature probes and deep-sea level measurements down to a depth of 500 meters for further data collection. This data collection began as early as 3/7/1980 for several buoys and ended on 6/10/1998, and there are a total of 178,080 entries corresponding to approximately daily occurrences at unique longitude and latitude values. Each row of the dataset corresponds to a unique day and unique location (longitude and latitude values).

Dataset Column Information:

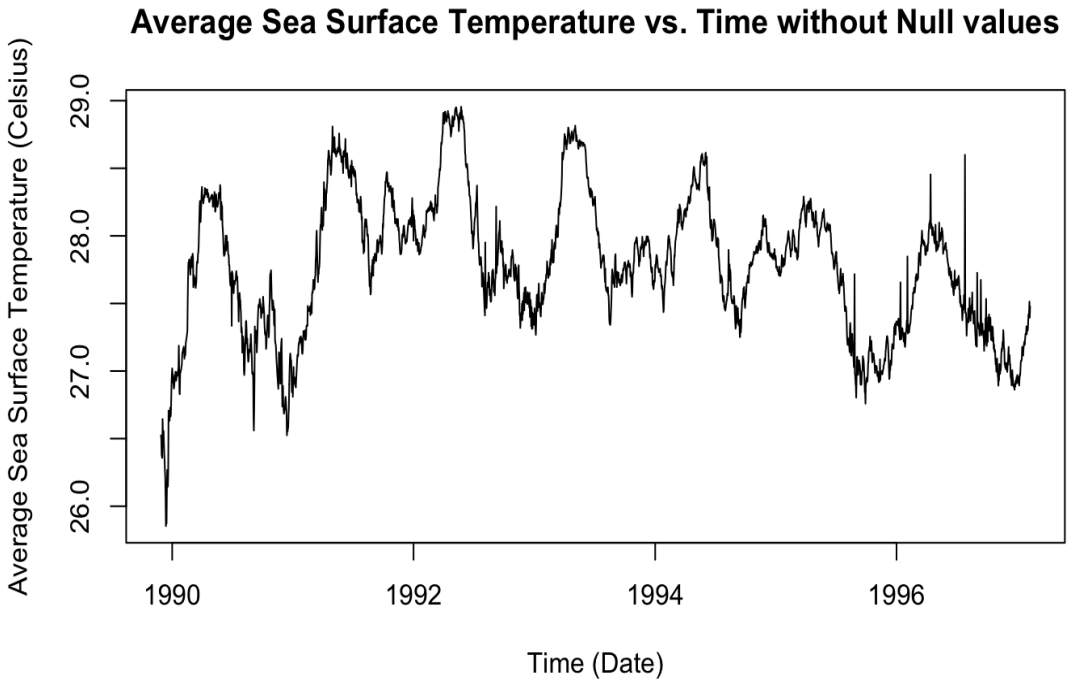
- Obs : Corresponds to entry number
- Year : 2-digit integer corresponding to year of observation in the 1900s
- Month : Month of observation (between 1 and 12)

- Day : Corresponding to day of observation (between 1 and 31)
- Date : Integer concatenating the Year, Month, Day into one number
- Latitude : Angle in degrees measuring from north-south position
- Longitude : Angle in degrees measuring from east-west position
- Zonal.Winds : Velocity (m/s) of Zonal Winds (negative if wind direction is east) in
- Meridional.Winds : Velocity (m/s) of Meridional Winds (negative if wind direction is south)
- Humidity : Percentage of humidity in the region
- Air.temp : Air temperature in celsius
- Sea.Surface.Temp : Sea surface temperature in celsius below 500 meters

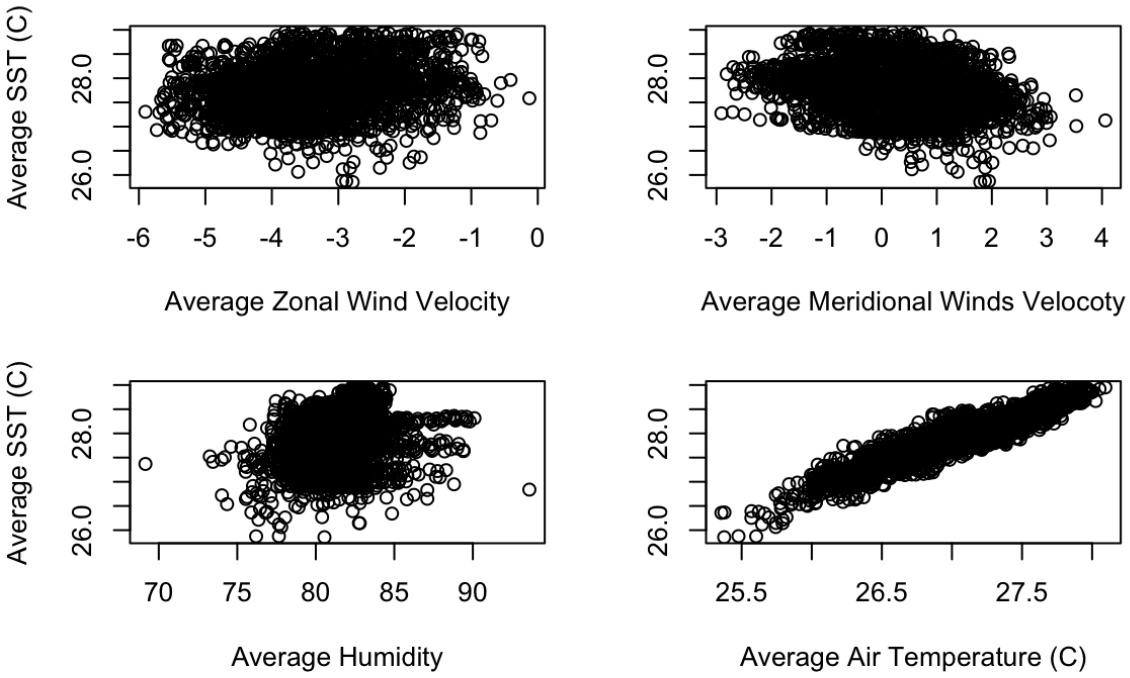
Statistical Methods

Data Cleaning

For many cases, there are numerous data points collected for a single day, so we calculate the average, and consider the average of each parameter for each date. When considering the average of each repeated row, we now have 6371 entries in our dataset, each corresponding to a unique date. However, upon further analysis, there were null values in the dataset before 11/28/1989 and after 2/8/1997, so we only consider a dataframe between the two previously mentioned dates, consisting of 2623 data points. Additionally, for further analysis, we split this dataframe into training and forecasting, where the first 2123 points are used to train the regression models, and the remaining 500 points are used for forecasting.



In the plot of the cleaned data above, we do not see an evident trend in the average sea surface temperature over time, and in general, the mean and the variance appear to be non-constant. Hence, this data does not appear to be stationary. We consider four primary variables when building our regression models: the average zonal winds, average meridional winds, average humidity, and average air temperature. While plotting each of these variables against the average sea surface temperature (the response variable), we notice a strong positive linear association between average air temperature and average sea surface temperature. However, there is no visible trend between the other predictor variables and average sea surface temperature, as seen in the plot below.



Regression Models

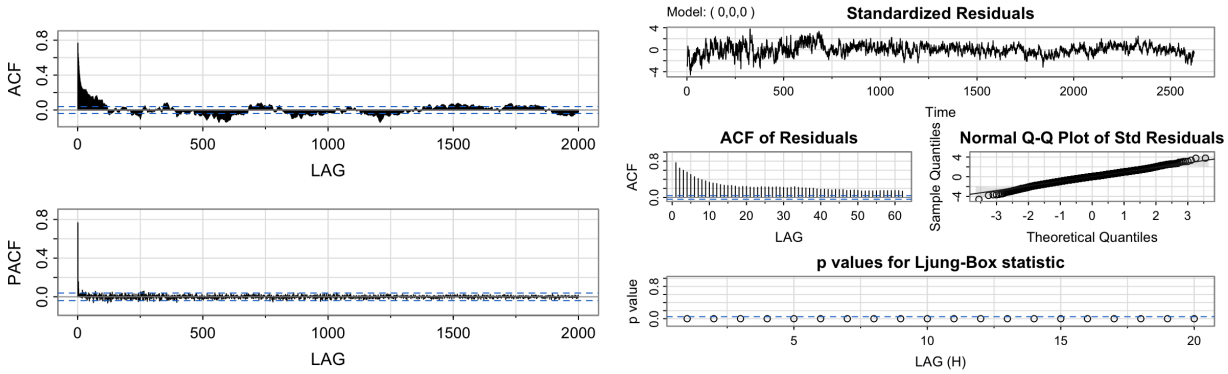
During our experimentation, we considered the four following regression models:

We define the following variables for simplicity:

- SST = Average sea surface temperature
- ZW = Average zonal wind velocity
- MW = Average meridional wind velocity
- H = Average humidity
- AT = Average air temperature
- μ_x = Represents the mean of respective parameter x
- B_i = Each beta represents a regression coefficient for the corresponding variable

Model 1:

$$SST = B_0 + B_1(time(SST)) + B_2(ZW) + B_3(MW) + B_4(H) + B_5(AT)$$

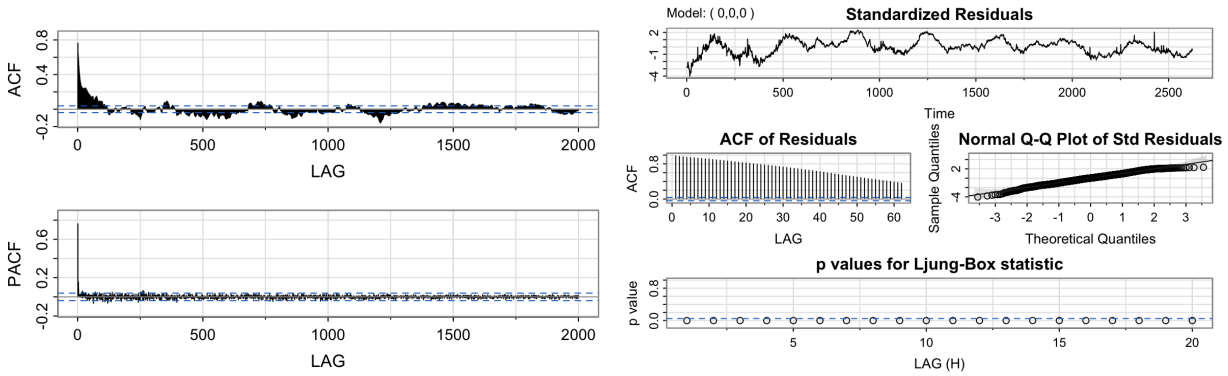


Our first model, Model 1, includes all four continuous predictor variables in our dataset as coefficients along with time. Upon analyzing the ACF and PACF of the residuals of Model 1, we notice a possible trend and consider conducting regression with autocorrelated errors, which is seen in Model 4. Additionally, we notice that in the plot of ACF of the residuals for Model 1, there are spikes at numerous lag values that fall outside of the required limits. Additionally, the p-values of the Ljung-Box statistic are generally under an alpha value of 0.05 in the plot above, and upon conducting an Ljung-Box test with a lag value of 2000, we retrieve a p-value of 2.2e-16, which implies that the residuals are not representative of white noise.

<u>Model 2</u>:

$$SST = B_0 + B_1(time(SST)) + B_2(ZW - \mu_{ZW}) + B_3(ZW - \mu_{ZW})^2 + B_4(MW - \mu_{MW})$$

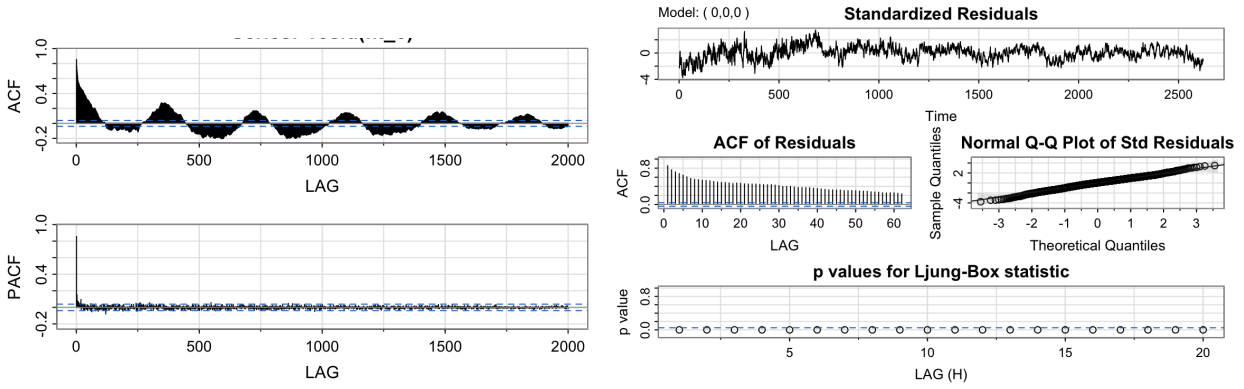
$$+ B_5(MW - \mu_{MW})^2 + B_6(H - \mu_H) + B_7(H - \mu_H)^2 + B_8(AT - \mu_{AT}) + B_9(AT - \mu_{AT})^2$$



In Model 2, we experiment by utilizing the differences of four predictor variables with their respective means and the squared differences of four predictor variables with their corresponding means as the variables for our model. We utilize these variables for our model to notice possible trends that could not be captured by simply utilizing the explanatory variables directly. Similar to Model 1, we notice that the ACF and PACF of the residuals of Model 2 fall outside the desired range. Additionally, similar to Model 1, the p-values of the Ljung-Box statistic are generally under an alpha value of 0.05 in the plot above, and upon conducting an Ljung-Box test with a lag value of 2000, we retrieve a p-value of 2.2e-16, which implies that the residuals are not representative of white noise.

Model 3:

$$SST = B_0 + B_1(time(SST)) + B_2(AT)$$

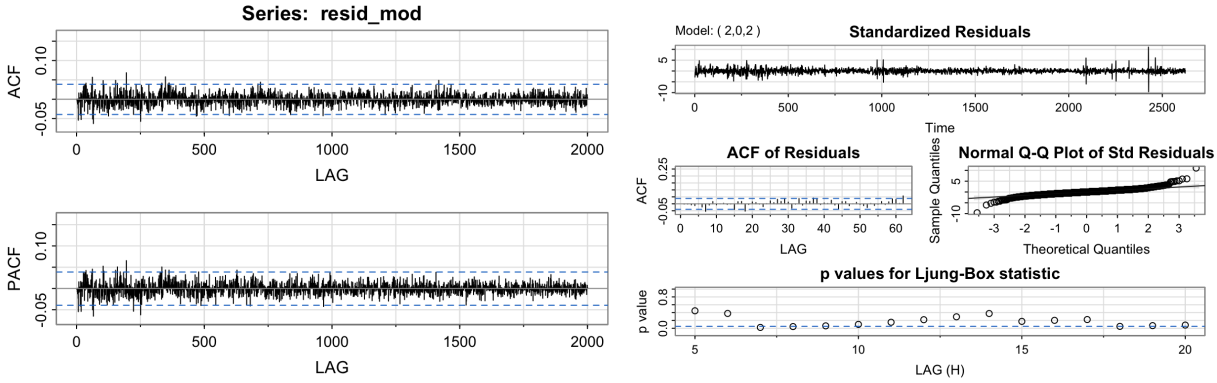


In Model 3, we only consider one predictor variable along with time, the average air temperature. This is done because our data analysis above showcases a strong linear correlation between the average air temperature and the average sea surface temperature, our response variable, a trend not visible with the other explanatory variables. However, we notice in the ACF plot of residuals that the ACF largely falls outside of the desired range, and the p-values for the Ljung-Box statistic are generally under an alpha value of 0.05 in the plot above. Upon conducting an

Ljung-Box test with lag 2000, we retrieve a p-value of 2.2e-16, which implies that the residuals are not representative of white noise.

Model 4:

$$SST = Sarima(p = 2,\ d = 0,\ q = 2,\ xreg = Model\ 1)$$



We noticed that the ACF and PACF of residuals of Model 1 cut off at approximately lag 2. Hence, we consider an ARMA(2, 2) model for the residuals for Model 1, and implement this model with autocorrelated errors in Model 4. We notice that the ACF and PACF of residuals in the plot above largely fall within the desired range. Additionally, the p-values for the Ljung-Box statistic are generally over an alpha value of 0.05 in the plot above, and after conducting an Ljung-Box test with lag 2000, we retrieve a p-value of 0.6535, which is greater than an alpha value of 0.05. Hence, we can conclude that the residuals are reminiscent of white noise, which is desired. For the given analysis, we consider Model 4 to be the most successful as the residuals represent white noise and it best fits the data based on the metrics described below in Results.

Results

Our testing data consists of 500 values, which we use to compute a short-range RMSE value (based on a 5-day forecast) and a long-range RMSE value (based on a 500-day forecast).

We notice that Model 4 has the lowest AIC\BIC values and the highest R^2 value compared to the other models. However, we notice the RMSE calculation based on a 500-day forecast is significantly worse for Model 4 compared to the other models, while the 5-day forecast is relatively accurate. Therefore, this implies that while Model 4, the correlated error regression model, is successful at forecasts over a small period but struggles when forecasting over a much larger period. Accordingly, we notice that Model 3 was most successful over a long period forecast, but struggled greatly compared to the other models in the 5-day forecast. In general, Model 3 was the least successful over a short-term forecast and had the lowest R^2 value compared to the other models. On the other hand, Model 1 and Model 2 were relatively successful in both the 5-day and 500-day forecasts and achieved a similar R^2 value over the training data. Model 1 still achieved lower AIC/BIC values and performed the best in the 5-day forecast based on RMSE.

Regression Model	R^2	**AIC**	**BIC**	**RMSE (5 days)**	**RMSE (500 days)**
Model 1	0.9141	-0.9340	-0.9183	0.0569	0.2040
Model 2	0.9170	1.478	1.494	0.1606	0.2106
Model 3	0.8756	-0.5670	-0.5581	1.050	0.1634
Model 4	0.9855	-2.706	-2.681	0.0603	0.6565

We conclude that Model 4 is generally the best performing model and successful in short-range forecasts, but other models may be explored for long-range forecasts with caution. In meteorology, long-range predictions of climate conditions are often impractical due to the sudden and dynamic nature of weather. Hence, this study and these results are most practical

over a short-range forecast during El Niño. Hence, we recommend Model 4 for such cases, utilizing an ARMA(2, 2) correlated error model with all of the explored predictor variables.

Discussion

Through testing various regression methods and narrowing down to four regression models, we were able to conclude that Model 4, the correlated error model, is our best-performing model and recommended to be used for short-range forecasts. While other models performed significantly better for long-range forecasts, in the context of meteorology, short-range forecasts are more applicable and reliable due to dynamic weather conditions, especially during El Niño. Accordingly, long-range forecasts for average sea surface temperature can still be useful, but should be treated with caution. More precisely, Model 1, Model 2, and Model 3 performed significantly better than Model 4 on long-range forecasts, but we showed that the residuals for those models did not correspond to white noise, undermining their performance. Additionally, some limitations to our study include the period of relevance. More precisely, the data used in our study range from 11/28/1989 to 2/8/1997. So, utilizing a forecast for the current time may lead to inaccurate results. However, through similar analysis of more recent data, we can experiment with similar regression strategies. For future work, we plan to explore other model options and try to improve the correlated error model. We initially found ARMA(2, 2) to be a good model for the residuals for Model 1, but we may be able to improve our current results through further experimentation.

References

Comprehensive R Archive Network (CRAN). (2025, March 18). *Sarima: Simulation and prediction with Seasonal Arima models.* The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/sarima/index.html

El Nino. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/122/el+nino

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications*. Springer.

Appendix

```{r appendix}
#Relevant R Code, df_avg represents the original dataset with null values dropped and all repeated dates being averaged, df_test is the test dataset

#Plot of Average SST vs explanatory variables
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))

plot(x = df_avg$Avg.Zonal.Winds, y = df_avg$Avg.Sea.Surface.Temp, xlab = "Average Zonal Wind Velocity", ylab = "Average SST (C)")
plot(x = df_avg$Avg.Meridional.Winds, y = df_avg$Avg.Sea.Surface.Temp, xlab = "Average Meridional Winds Velocoty", ylab = "")
plot(x = df_avg$Avg.Humidity, y = df_avg$Avg.Sea.Surface.Temp, xlab = "Average Humidity", ylab = "Average SST (C)")
plot(x = df_avg$Avg.Air.Temp, y = df_avg$Avg.Sea.Surface.Temp, xlab = "Average Air Temperature (C)", ylab = "")

trend <- time(df_avg$Avg.Sea.Surface.Temp)

#model 1
fit_1 <- lm(Avg.Sea.Surface.Temp ~ trend + Avg.Zonal.Winds + Avg.Meridional.Winds + Avg.Humidity + Avg.Air.Temp, na.action = NULL, data = df_avg)
summary(fit_1)

#residual plots and acf/pacf plots for residuals to lag 2000
checkresiduals(fit_1, test = "LB")
acf2(resid(fit_1), 2000)

#model 2
df_avg$Avg.Zonal.Winds_m = df_avg$Avg.Zonal.Winds - mean(df_avg$Avg.Zonal.Winds)
df_avg$Avg.Meridional.Winds_m = df_avg$Avg.Meridional.Winds - mean(df_avg$Avg.Meridional.Winds)
df_avg$Avg.Humidity_m = df_avg$Avg.Humidity -  mean(df_avg$Avg.Humidity)
df_avg$Avg.Air.Temp_m = df_avg$Avg.Air.Temp -mean(df_avg$Avg.Air.Temp)

fit_2 <- lm(Avg.Sea.Surface.Temp ~ trend + Avg.Zonal.Winds_m + I(Avg.Zonal.Winds_m^2) + Avg.Meridional.Winds_m + I(Avg.Meridional.Winds_m^2) + Avg.Humidity_m + I(Avg.Humidity_m^2) + Avg.Air.Temp_m + I(Avg.Air.Temp_m^2), na.action = NULL, data = df_avg)
summary(fit_2)
```

```
#residual plots and acf/pacf plots for residuals to lag 2000
checkresiduals(fit_2, test = "LB")
acf2(resid(fit_2), 2000)

#model 3
fit_3 <- lm(Avg.Sea.Surface.Temp ~ trend + Avg.Air.Temp, data = df_avg)
summary(fit_3)

checkresiduals(fit_3, test = "LB")
acf2(resid(fit_3), 2000)

#model 4: correlated errors model based on model 1
sarima_mod <- sarima(df_avg$Avg.Sea.Surface.Temp, 2,0,2,
xreg=cbind(trend,df_avg$Avg.Zonal.Winds, df_avg$Avg.Meridional.Winds,
df_avg$Avg.Humidity, df_avg$Avg.Air.Temp))
#residual plots and acf/pacf plots for residuals to lag 2000
resid_mod <- sarima_mod$fit$residuals
acf2(resid_mod, 2000)

#MSE/RMSE for models
h = 5 # change to 500 for all of df_test
pred_test = predict(fit_1, newdata = df_test[1:h,])
pred_test1 = predict(fit_2, newdata = df_test[1:h,])
pred_test2 = predict(fit_3, newdata = df_test[1:h,])

mse_1 = mean((pred_test - df_test$Avg.Sea.Surface.Temp[1:h])^2)
rmse_1 = sqrt(mse_1)
mse_2 = mean((pred_test2 - df_test$Avg.Sea.Surface.Temp[1:h])^2)
rmse_2 = sqrt(mse_2)
mse_3 = mean((pred_test3 - df_test$Avg.Sea.Surface.Temp[1:h])^2)
rmse_3 = sqrt(mse_3)

sarima_pred <- sarima.for(xdata = ts_train, n.ahead = h, p =2, d=0, q=2, xreg=as.matrix(df_avg[,
cols]), newxreg = as.matrix(df_test[,cols])[1:h, ,drop =FALSE])
mse_out <- mean((sarima_pred$pred - df_test$Avg.Sea.Surface.Temp[1:h])^2)
rmse_out <- sqrt(mse_out)

#Ljung Box Test
Box.test(resid(fit_1), type = "Ljung", lag = 2000)
```

```
Box.test(resid(fit_2), type = "Ljung", lag = 2000)
Box.test(resid(fit_3), type = "Ljung", lag = 2000)
Box.test(sarima_mod$fit$residuals, type = "Ljung", lag = 2000)
```