Sea Surface Temperatures

An Application of Time Series Methods and Integration of Fuctional Modeling

Ann Marie Matheny, Amanda Suleski, Evan Tiffany

4/1/2020

Abstract

...can fill this in last...

Division of Labor

...can fill this in last...

Table 1: Missing Data

	Winter	Spring	Summer	Fall	
1950	5	7	0	6	18
1960	4	3	8	6	21
1970	3	6	5	7	21
1980	2	2	9	4	17
1990	4	4	10	8	26
2000	4	9	4	5	22
2010	2	4	1	5	12
	24	35	37	41	137

Introduction

The purpose of this report is to inform readers on our team's collaborative efforts in examining sea surface temperatures (SST). The objective of this project was to explore modern methologies in dealing with missing data as well as forcasting techniques. Overall, we analyzed the performace of three statistical models; a benchmark model, a tradiational time series model, and a functional model. We graded each models performance based on minimized variance in the residuals it produced. Before these steps could even be taken however, our team worked together to address the initial problem of missing data. Many avenues were explored on how to handle this obstacle. Since we are working with climatological data that produces a seasonal trend, we concluded that interpolation was the appropriate choice. We at first receuved only a portion of the data to train our models and then the full data set to then test how well our models performed. This report consisely documents our thought process and statistical procedures.

The Data

The data used in this analysis was provided by the National Oceanic and Atmospheric Agency (NOAA) and was believd to be measured within the region of the south Pacific. The data received at first consisted simply of two variables; a date variable (month, year) and a monthly average sea surface temperature given in degrees Celcius. The first record was taken in January 1950 and the data spans to December of 2014. It includes 780 observations. However, 137 of these observations were missing the sea surface temperature for a given time stamp. As we delved into our assignment and our analysis structure matured we later received the complete data set which contained all data points. This new data was applied to our 3 modeling techniques.

Examining Missing Data

After researching different missing data techniques applied in time series and consulting our project advisor, we concluded that seasonal interplation was the most applicable and appropriate for this data sense given the cyclical and seasonal nature of the series data.

We also took several steps in cleaning, wrangling and preparing the data in order to run any sort of preliminary analysis. First, we created a season variable (winter, spring, summer, fall) to look at any seasonal trends that appeared relevant in the data over time. We also created a decade variable to indicate what decade each monthly avergae was recorded in to compare any trends in the data across the different decades. We repeated this same procedure for creating a separate month and year varibale (this is apart from the month-year variable provided).

Missing Data Interpolations

We had 3 basic approaches for the data interpolation: using the original data, a linear interpoliation, and a cubic spline interpolation. The nonparametric approach was the most appropriate; a cubic spline is a spline constructed of piecewise third-order polynomials which pass through a set of m control points (it should be mentioned that natural cubic spline interpolation does have a tendency of overfitting). (WE NEED TO

CITE THIS!!) This method produced minimal variance an was added as a new numeric varibale to the SST data as an interpolation metric. The linear approach did not perform as well, mostly due to the fact that the data violates linearity.

Original Data

The variance in the original data is 1.582. We calculated the mean and standard deviation of the sea surface temperatures of all years per month in order to further examine the data.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   22.70   24.79   25.81   25.78   26.68   29.14     137
```

	Mean	SD
Jan	25.45804	1.0178743
Feb	26.25393	0.6038061
Mar	26.97385	0.3676477
Apr	27.38055	0.3222460
May	26.86132	0.6098501
June	26.35696	0.5711306
July	25.61456	0.6493074
Aug	24.83622	0.7230877
Sep	24.69250	0.7643564
Oct	24.80926	0.7890523
Nov	24.78312	1.0351496
Dec	25.05949	1.4309290

Linear Interpolation

The variance after the linear interpolation is slightly lower than the variance of the original data at 1.542. This interpolation keeps the quartiles approximately the same as the original data. While the mean sea surface temperature per month remains approximately the same as the original data, the standard deviation increased for each month.

	Mean	SD
Jan	25.51151	1.0873482
Feb	26.22781	0.6355249
Mar	26.90738	0.4354900
Apr	27.25372	0.4750281
May	26.86164	0.5446654
June	26.31829	0.5516208
July	25.56521	0.5913527
Aug	24.98372	0.7400923
Sep	24.74205	0.7957403
Oct	24.79395	0.9244672
Nov	24.85823	1.2203654
Dec	25.07110	1.3724462

Cubic Spline Interpolation

The variance after the cubic spline interpolation is larger than the variance of the original data at 1.605. The quartiles are just barely lower than the original data. The standard deviation of sea surface temperatures per month is larger than the original data. The standard deviation is slightly larger than the linear interpolation in the winter months and about the same in the summer months.

	Mean	SD
Jan	25.49712	1.1043799
Feb	26.23090	0.6346559
Mar	26.96230	0.3974344
Apr	27.31279	0.4197625
May	26.91878	0.5503460
June	26.32420	0.5552533
July	25.54950	0.6129347
Aug	24.91547	0.7182987
Sep	24.70874	0.7933589
Oct	24.77165	0.9502289
Nov	24.82955	1.2543366
Dec	25.05953	1.3796583

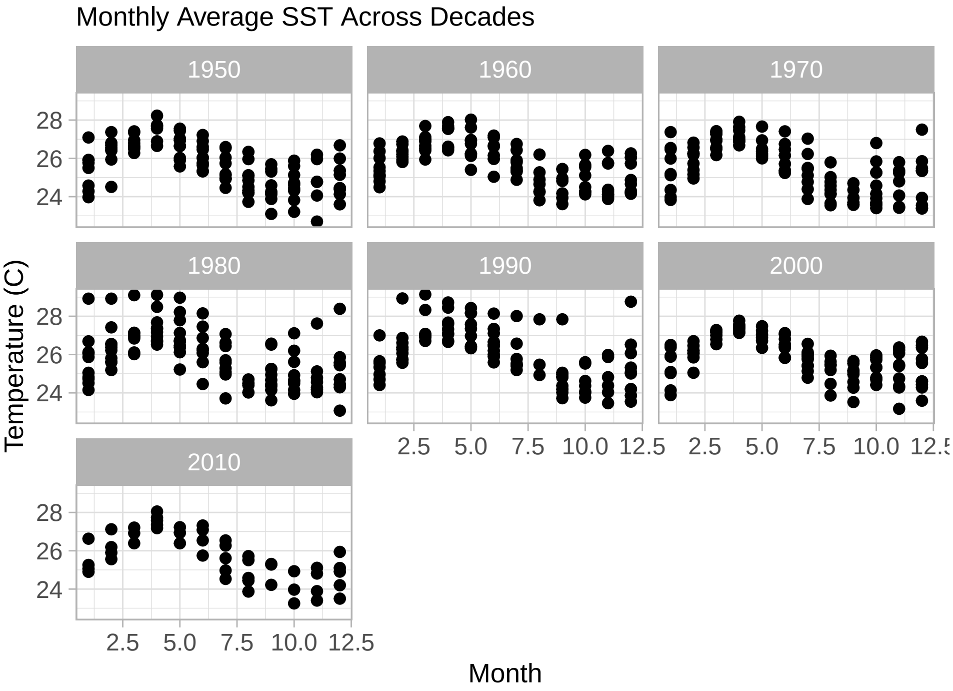
Analysis of Missing Data Interpolations

Squared Error

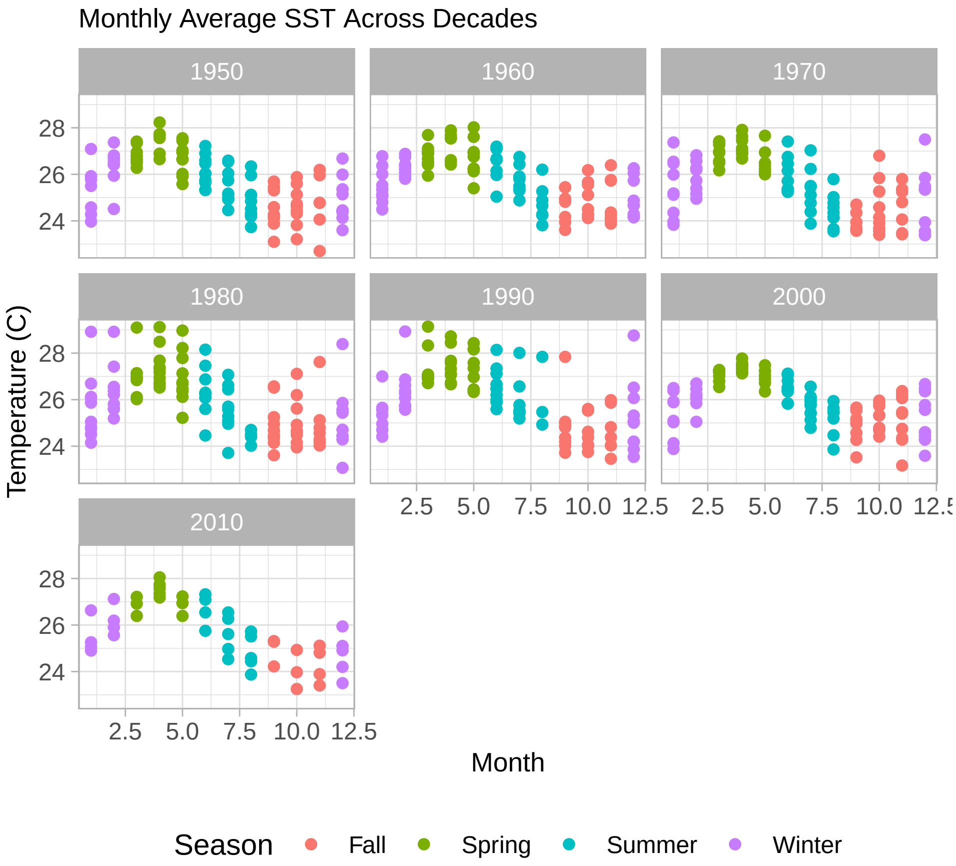
As explained above, the cubic spline interpolation should yield the best results since the data is not linear. Once we received the complete data set, we analyzed the accuracy of our interpolations. The mean squared errors of our linear interpolation and cubic spline interpolation is 0.0254 and 0.0144 respectively. Thus confirming our expectations that the cubic spline interpolation is more accurate.

Preliminary Exploratory Data Analysis

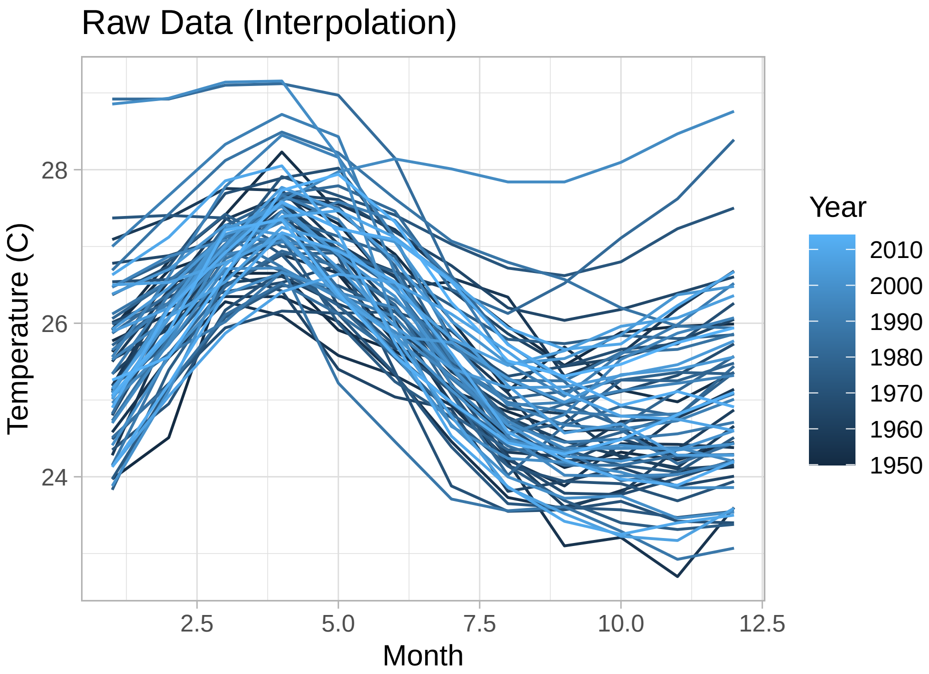
As previously stated, our team wanted to explore seasonal and time trends that may or may not have been present in the data once all wrangling and cleaning was complete.



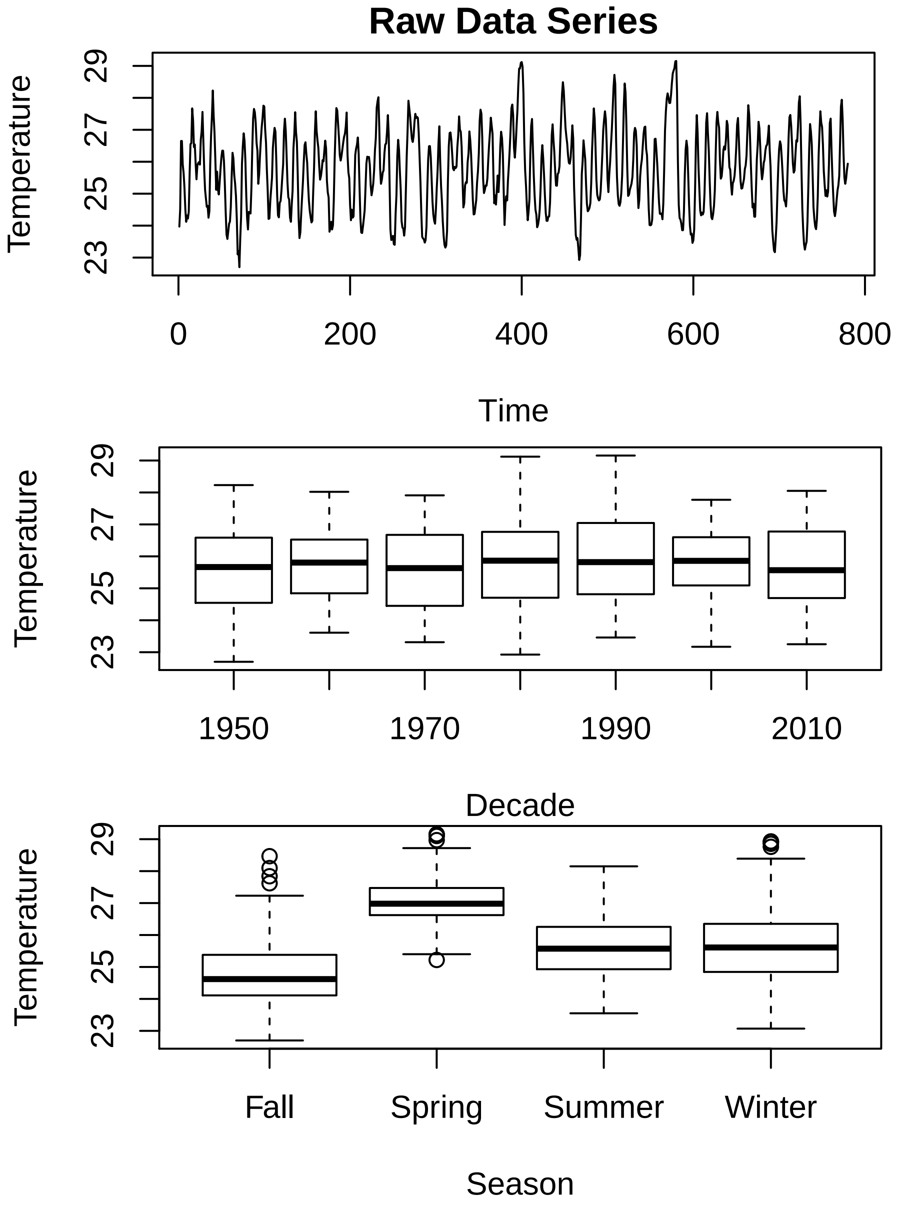
This first plot breaks up the overall times series across the 7 decades. The most inconsistent decacde is 1980, where we see high averages in comparison to the other deacdes as well as a wider, more variant monthly averages. Likewise the 1990 plot also has notably higher SST, especailly whne coming the decade that follows. We see much less varaition in the later decaded of the SST, which could be due to confounding factors such as climate change, but that is beyond the scope of this report.



This graph presents the same inforamtion as the previous one, but this time we added a color aesthetic to indicate the four seasons.



These two graphics show the flucuate of SST over time. The first plot (pictured above) shows the time series of the cubic spline interpolation over time, month 0 being the first data point month (January 1950). This second one (pictured below) is a cluster of lines, each on indicating a different year as a fucntion of time. The 65-year lines are all plotted against each other and an overall sinusoidal pattern is evident.



These boxplots were constructed to again visualize the temperature distributions across the seasons and decades parameters. In the first plot illustrating the decade data, it is interesting that the medians are constistant and constant over the 60 or so year time span, apart from what one might have hypothesized with rising sea temperatures due to global warming. In the second box plot figure, it is esay to see that spring overall has the warmest sea surface temperatures, and given that this data was collected in the southern hemisphere, this is to be expected. Interestingly, there seems to be little differnce in the temperature distribution between the summer and winter months; they are practically identical.

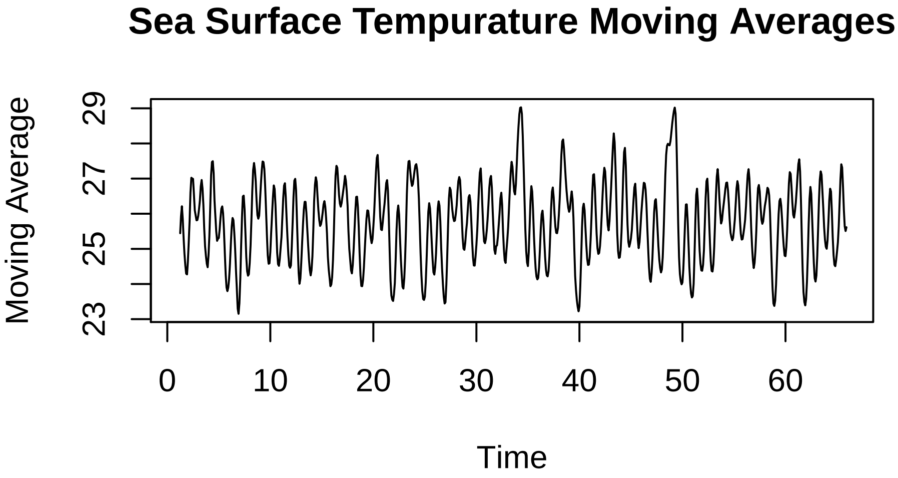
Differencing the Data

Time series datasets may contain trends and seasonality, which may need to be removed prior to modeling. Trends can result in a varying mean over time, whereas seasonality can result in a changing variance over time, both which define a time series as being non-stationary. Stationary datasets are those that have a stable mean and variance, and are in turn much easier to model. Differencing is a widely used data transform for making time series data stationary. For example, when modeling, there are assumptions that the summary statistics of observations are consistent. In time series terminology, we refer to this expectation as the time series being stationary. These assumptions can be easily violated in time series by the addition of a trend, seasonality, and other time-dependent structures. The observations in a stationary time series are not dependent on time. Our team checked if our time series is stationary by looking at a line plot of the series over time. Sign of obvious trends, seasonality, or other systematic structures in the series are indicators of a non-stationary series. A more accurate method would be to use a statistical test, such as the Dickey-Fuller test, which we performed on our data using the `adf.test()` R function. It determines whether a unit root, a feature that can cause issues in statistical inference, is present in a time-series sample. The main idea is the bigger the negative value, the stronger the confiramtion of stationarity. In our results, we have a test statistic of -5.128 and a small p-value of 0.01 indicating significance.

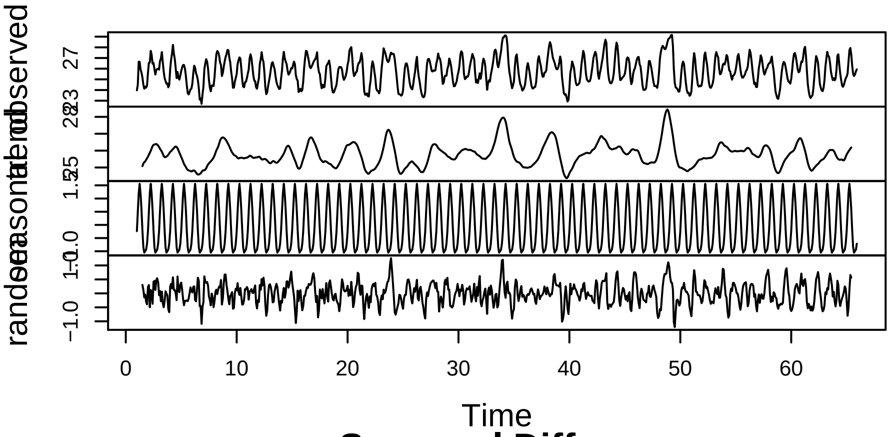
One of the time series models we hoped to examie was the Auto-Regressive (AR) Moving Average (MA) model and in particular the ARIMA (p,d,q) model. In order to construct this we needed to determine the parts of the model notion by calculating the autocorrelation and partial autocorrelation. These results are displayed in what are called correlgrams, which indicate how the data is related to itself over time based on the number of periods apart, or lags. The autocorrelation function (ACF) displays the correlation between series and lags for the Moving Average (q) of the ARIMA model, and the partial autocorrelation function (PACF) displays the correlation between returns and lags for the auto-regression (p) of the ARIMA model.

In order to carry out these statistical procedure, we decomposed the time series and ran an ACF and PACF on the series' seasonal components. We also differenced the original time series (with a time lag of 12 for each month) and ran the data throught the same functions.

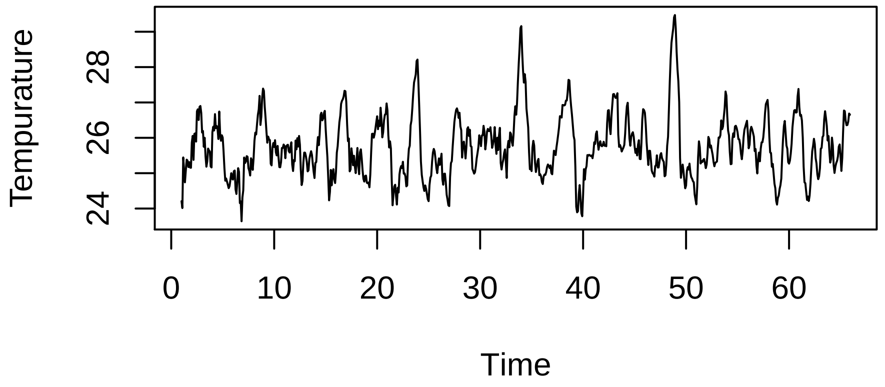
```
##
##  Augmented Dickey-Fuller Test
##
## data:  series
## Dickey-Fuller = -5.1282, Lag order = 9, p-value = 0.01
## alternative hypothesis: stationary
```


Sea Surface Tempurature Moving Averages

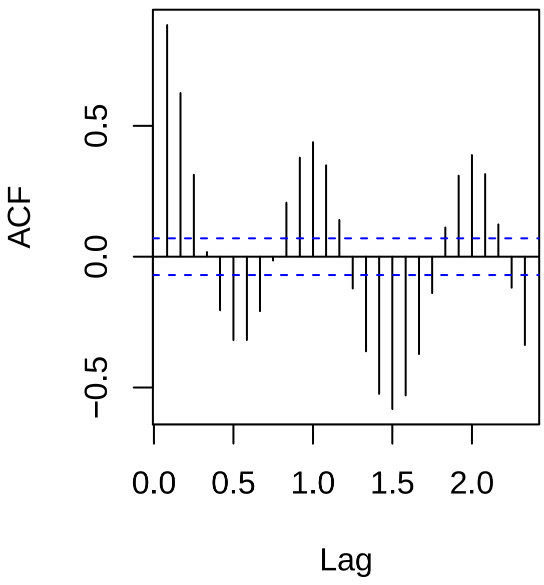
Decomposition of additive time series



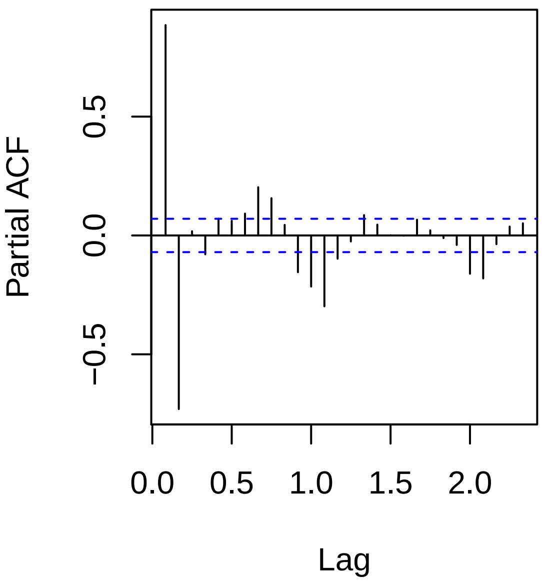
Seasonal Difference

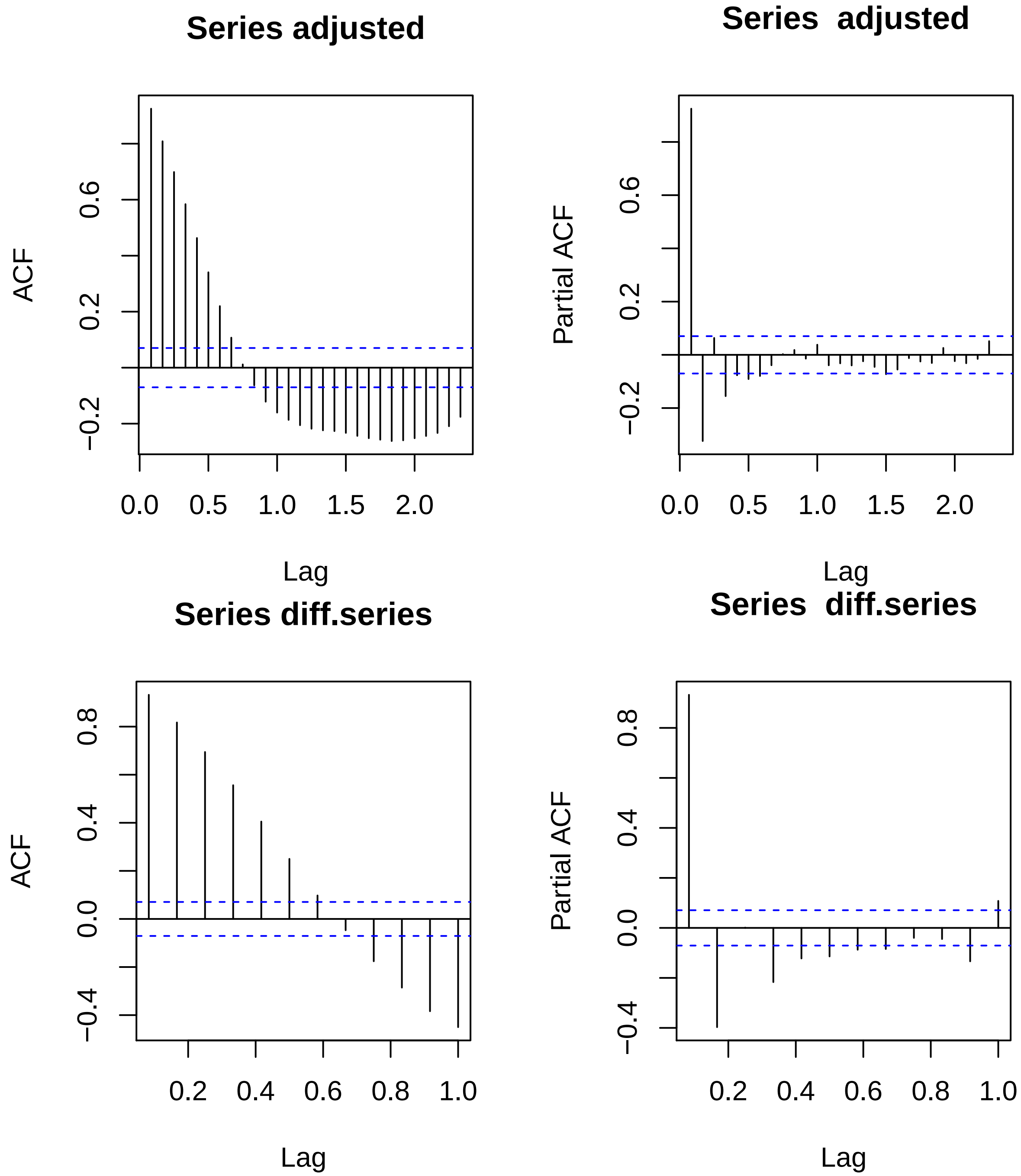


Series series



Series series



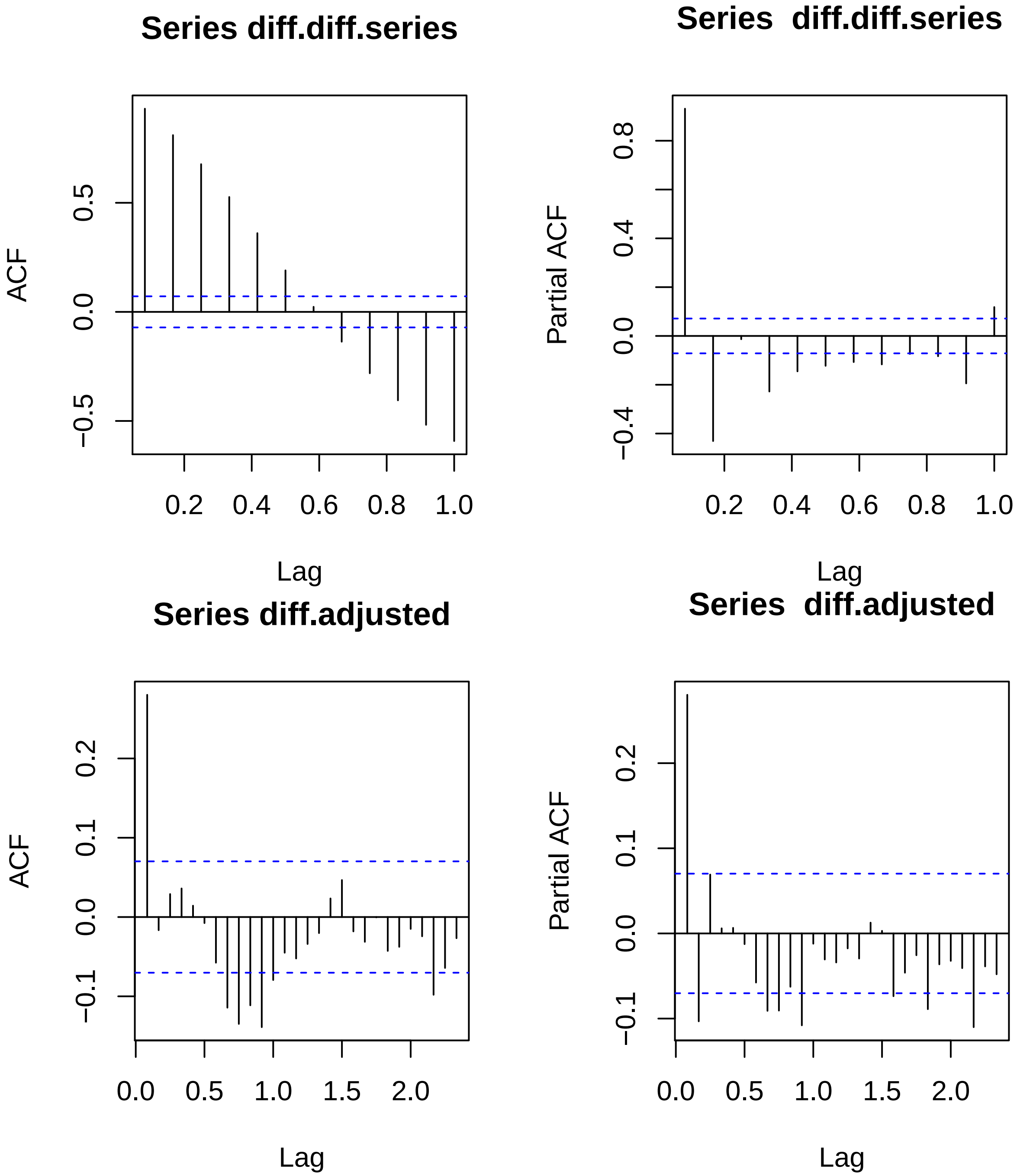


Series adjusted

Series adjusted

Series diff.series

Series diff.series



The blue-dotted line determines the strength of correlations. The values that cross the blue-dotted lines determine the notation for the ARIMA model for each dataset.

Modeling Process

A Random Walk, Linear Model

The first model we included in our analysis was a benchmark model to gauge as a criterion before continuing with our analysis. The first one we chose to look into was a Random Walk (RW) model. A random walk can be expressed by the following:

$$x_t = x_{t-1} + \omega_t$$

The time series is purely predicted as a stochastic model with time dependency based entirely on the previous time point $t-1$. Note that a random walk time series is not stationary (as the AR polynomial root is not greater than 1). Applying first differencing would result in a white noise time series (which would be *stationary*), and would have minimal autocorrelation.

$$\nabla x_t = x_t - x_{t-1} = \omega_t$$

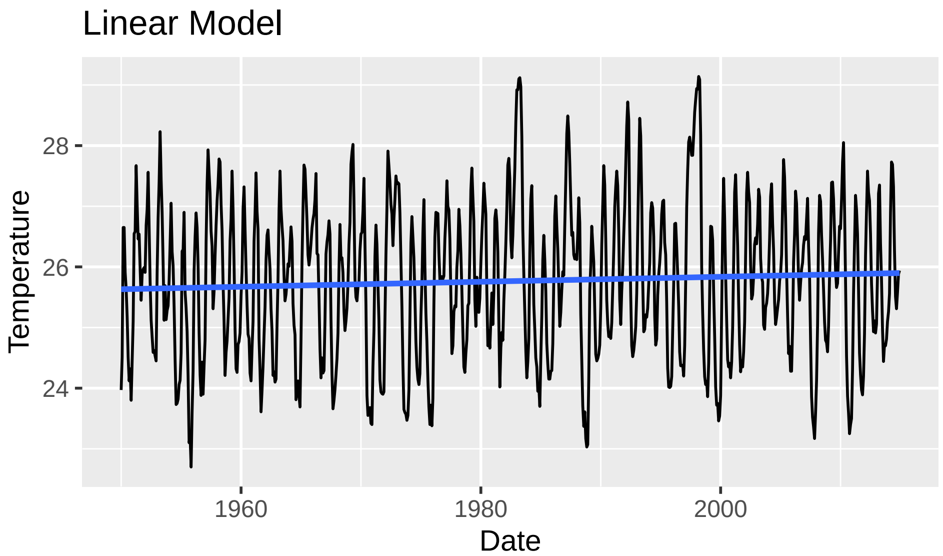
Since we are in particular working with cyclical data a seasonal random walk is more appropriate to apply. If the seasonal difference (the season-to-season change) of a time series looks like stationary noise, this suggests that the mean (constant) forecasting model should be applied to the seasonal difference. For monthly data, whose seasonal period is 12, the seasonal difference at period t is $X(t) - X(t-12)$. Applying the mean model to this series yields the equation:

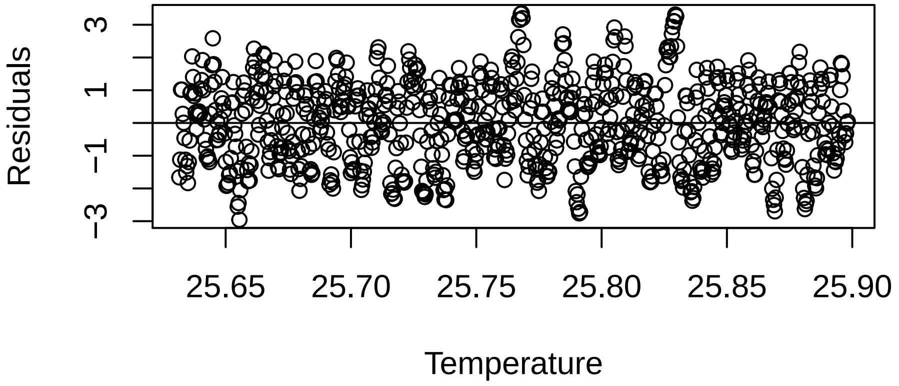
$$X(t) - X(t-12) = \alpha$$

This forecasting model will be called the seasonal random walk model, because it assumes that each season's values form an independent random walk. Thus, the model assumes that January's temperature for 2020 is a random step away from January's temperature for 2019, February's tempeature for 2020 is a random step away from February's temperature for 2019, etc., and the mean temperature of every step is equal to the same constant (denoted here as alpha). The forecast for Jan 2020 ignores all data after Jan 2019; it is based entirely on what happened exactly one year ago.

The second benchmark model we briefly looked into was a linear model. As shown by the plots below, the linear model does not fit the data well. Instead of a seasonal pattern, this forecast is constantly slightly increasing. There is a very wide range of residuals from -3 to 3; however, the plot also shows the homogeneity of error variance.

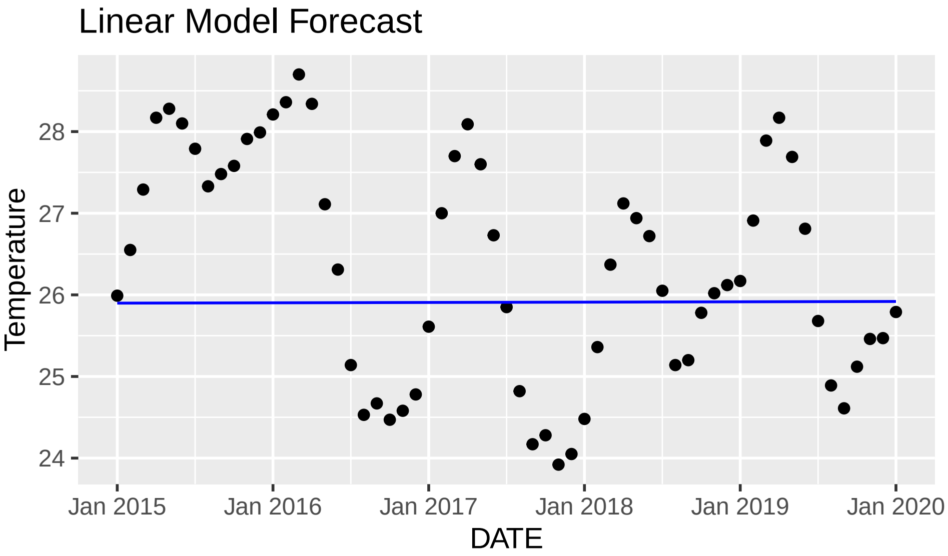
```
## integer(0)

##
## Call:
## lm(formula = SST ~ as.numeric(DATE), data = sst.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9555 -1.0079  0.0255  0.9472  3.3519
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      17.620332   4.797942   3.672 0.000257 ***
## as.numeric(DATE)  0.004108   0.002420   1.698 0.089986 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.268 on 778 degrees of freedom
## Multiple R-squared:  0.00369,    Adjusted R-squared:  0.00241
## F-statistic: 2.882 on 1 and 778 DF,  p-value: 0.08999
```





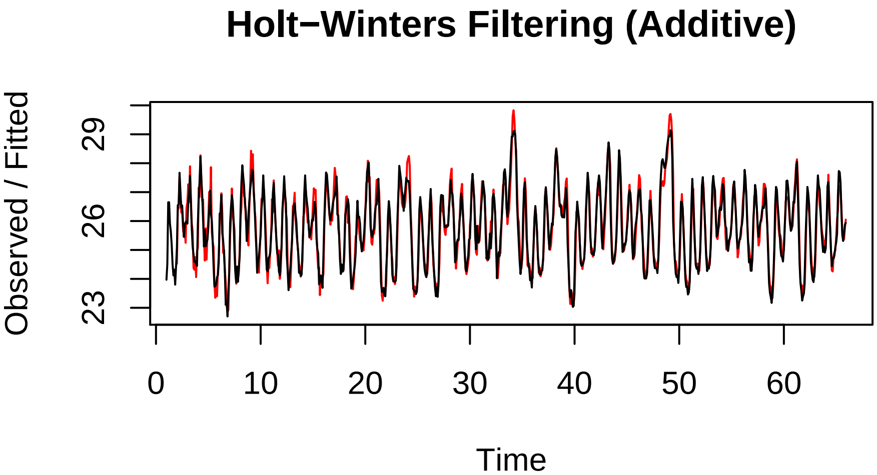
Linear Model Analysis

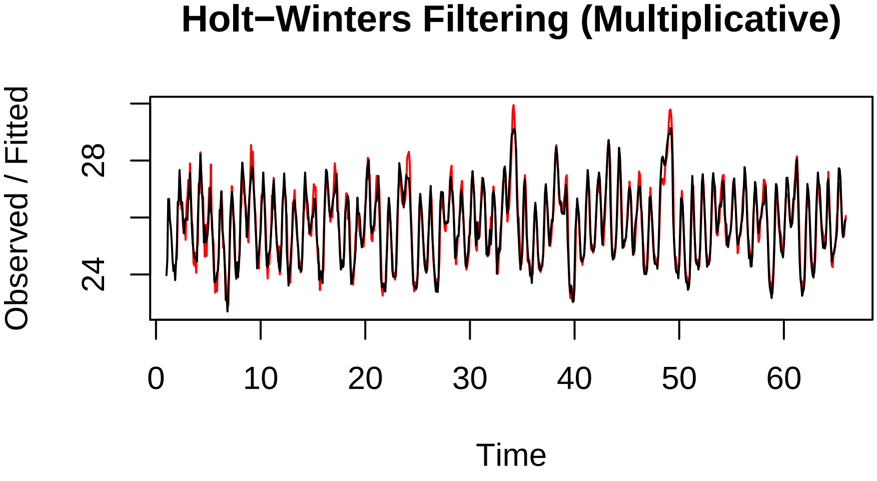
The mean squared error of the linear model forecast is 2.065. The plot below shows the linear model forecast (blue line) and the actual sea surface temperature values as points.



The Holt-Winters Model

The second type of modeling we explored was a traditional time series model. For this, we decided the Holt-Winters Forecast was most applicable to our data. Holt (1957) and Winters (1960) extended Holt's method to capture seasonality. The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations — one for the level l_t,trend b_t, a seasonal component s_t, and corresponding smoothing paramters α, β, and γ and m is used to denote seasonality. There are two variations to this model, the additive and multiplicative methods. The additive method is favored when the seasonal variations are roughly constant through the time series. The multiplicative method is favored when the seasonal variations are changing proportional to the level of the series. For the purposes of this analysis we used both modeling methods, but forecasted the time series using the additive method. While there did not appear to be a significant difference between the two fits, the additive was deemed more appropriate since we are working with seasonal data which typically has little variation. The sum of squared error for the additive model is 99.797 compared to the multiplicative model which is 102.742.

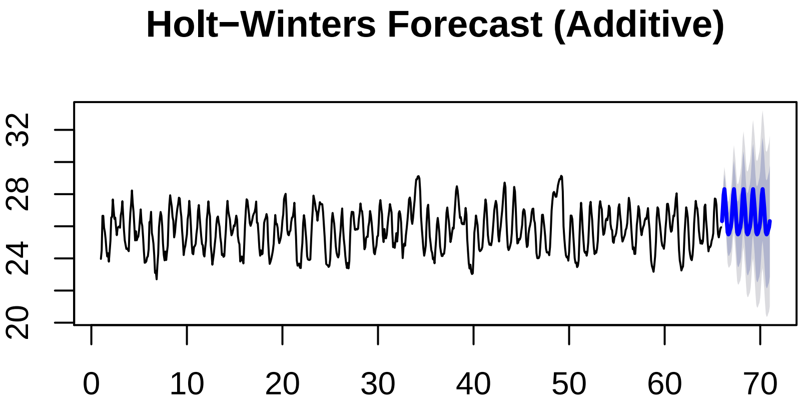


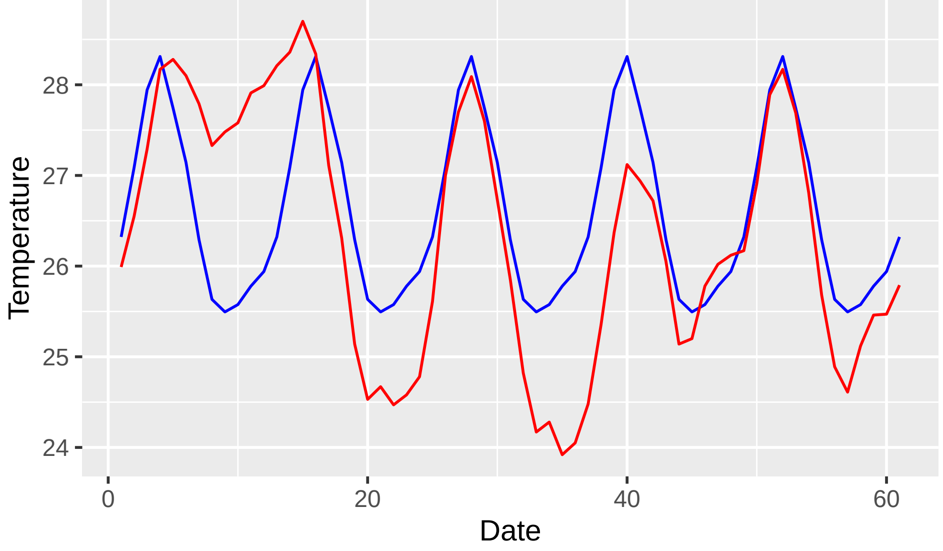


Holt-Winters Analysis

Using the additive model we forecasted the sea surface temperature for the next 5 years. The plots below show the forecasted temperatures. The second plot compares the Holt-Winters forecast (blue) to the actual sea surface temperature (red). The mean squared error of the forecast is 1.066.

Prediction intervals?? https://www.sciencedirect.com/science/article/abs/pii/016920709090103I

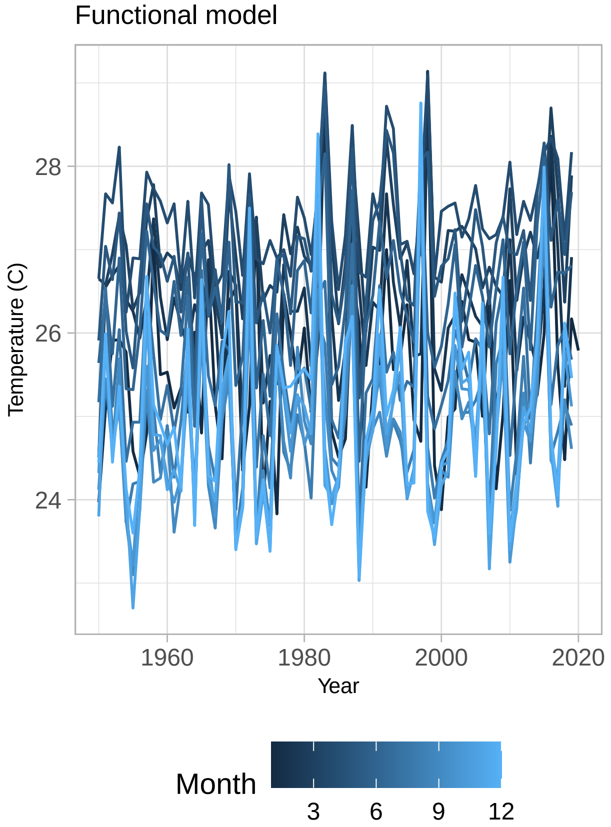


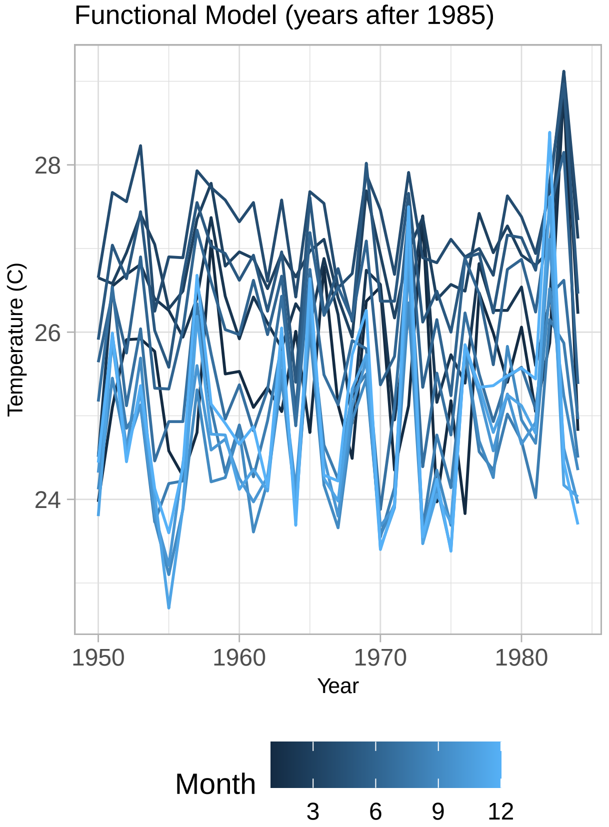


Functional Model

The last model we explored was a Functional Model. For the functional model, we considered the data as representing a time series of curves, particularly, sinusodal curves, as expected from seasonal, cyclical flucuations. We can use this information to specify that for each year 1950-2020, there is a functional specification for the annual cycle. In particular the function we found to best fit the data is as follows (where t denotes the temperature, y is the year and m indicates month)

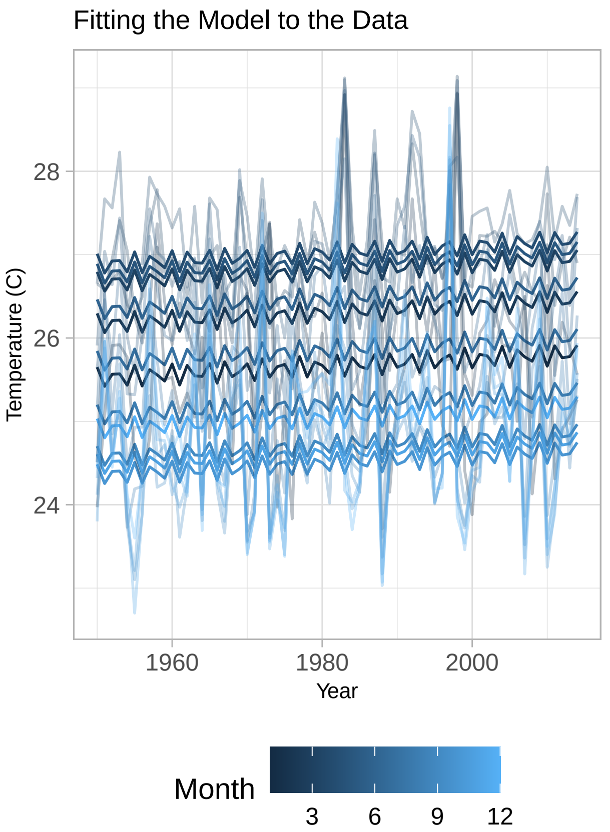
$$t_{ij} = y_i + sin(2\pi m_{ij})/(12 - 0.6) + cos(y_i * 10)$$

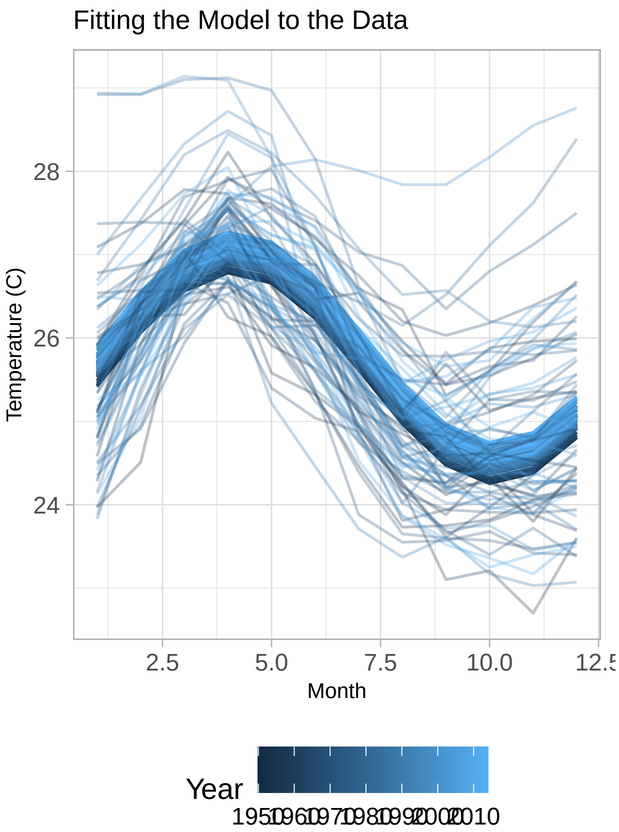


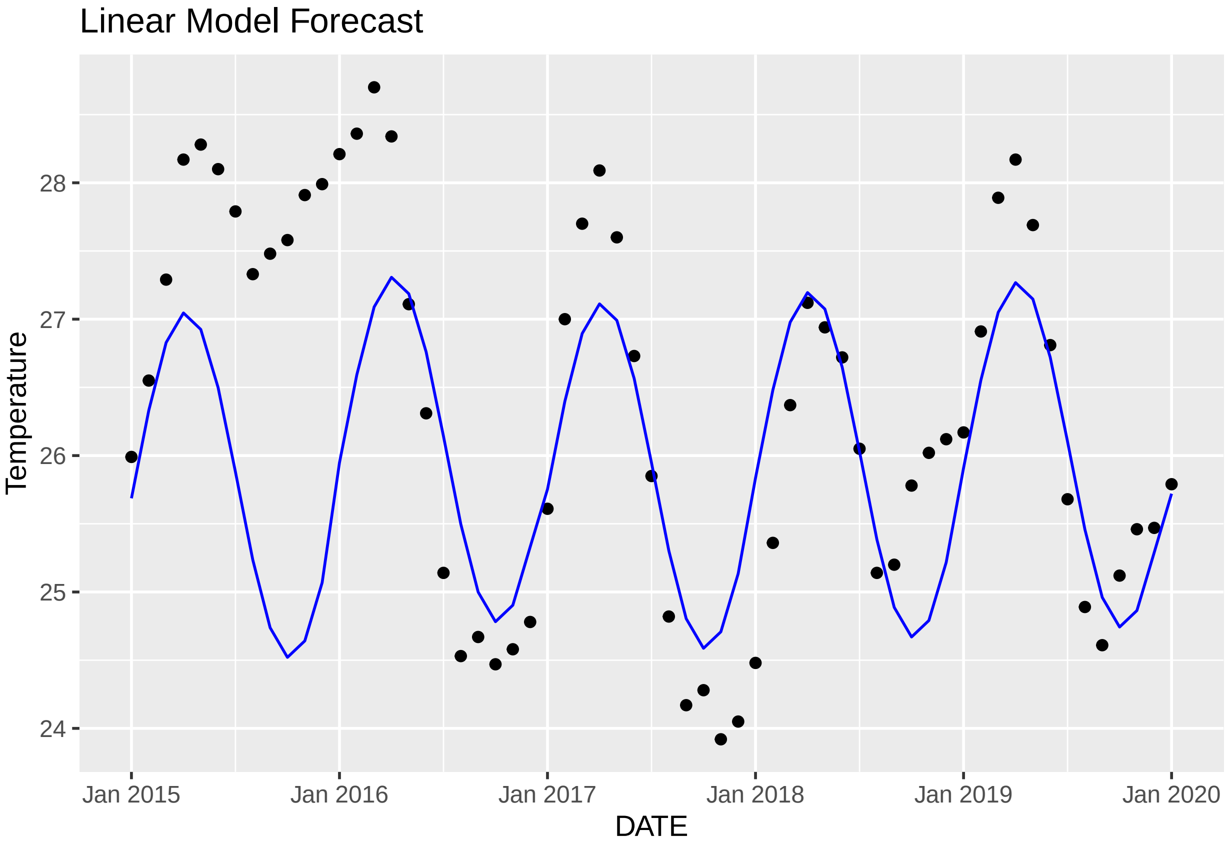


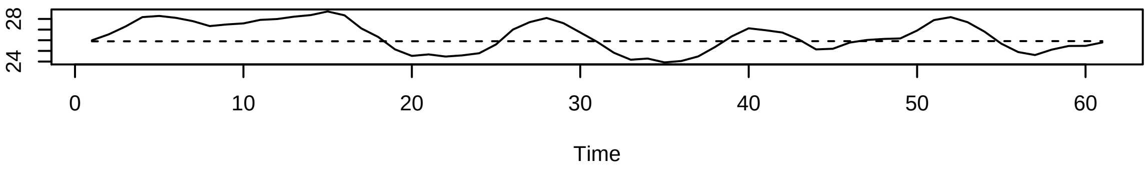
The plots above aim to show the generic functional pattern of the data. We can see a sinusoidal function each year with the first plot, showing how the sea surface temperature changes monthly with a color change over years. The second and third plot aim to show the yearly pattern changes. The relationship in this case is much less obvious, but we can see generally the time between peaks and troughs to fit a sin relationship to the yearly component.

```
##
## Call:
## lm(formula = SST ~ YEAR + sin(2 * pi * MONTH/12 - 0.6) + cos(YEAR *
##     10), data = sst.dat)
##
## Coefficients:
##                  (Intercept)                          YEAR
##                    16.567381                      0.004639
## sin(2 * pi * MONTH/12 - 0.6)                cos(YEAR * 10)
##                     1.266221                     -0.134710
```

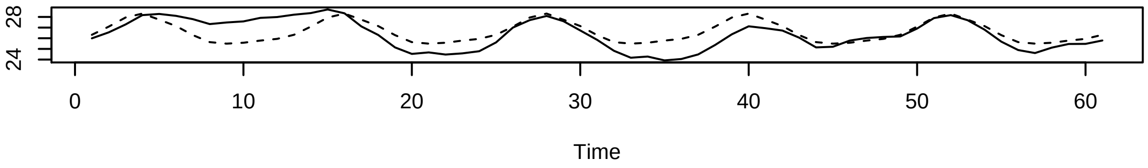

Fitting the Model to the Data
Temperature (C)
28
26
24
1960
1980
2000
Year
Month
3
6
9
12


Fitting the Model to the Data
Temperature (C)
28
26
24
2.5
5.0
7.5
10.0
12.5
Month
Year
1950
1960
1970
1980
1990
2000
2010

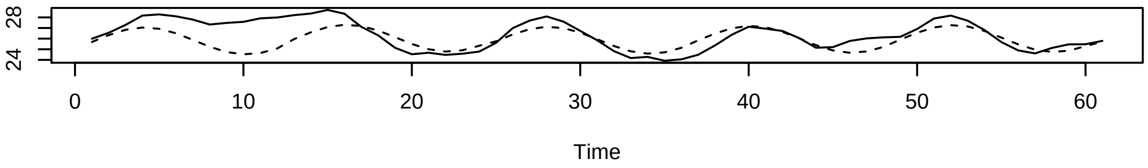

Linear Model Forecast
Temperature
28
27
26
25
24
Jan 2015
Jan 2016
Jan 2017
Jan 2018
Jan 2019
Jan 2020
DATE



Linear



Holt−Winters



Functional

	Linear	Holt-Winters	Functional
Mean Squared Error	2.049878	1.056047	1.313298
Absolute Error	74.877035	50.063925	51.177666

Discussion and Conclusion

discuss pros and cons of each method, amount of effort that went in and was it worth it for marginal improvements

Appendix

Work Cited

Holt, C. E. (1957). *Forecasting seasonals and trends by exponentially weighted averages* (O.N.R. Memorandum No. 52). Carnegie Institute of Technology, Pittsburgh USA.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324–342.