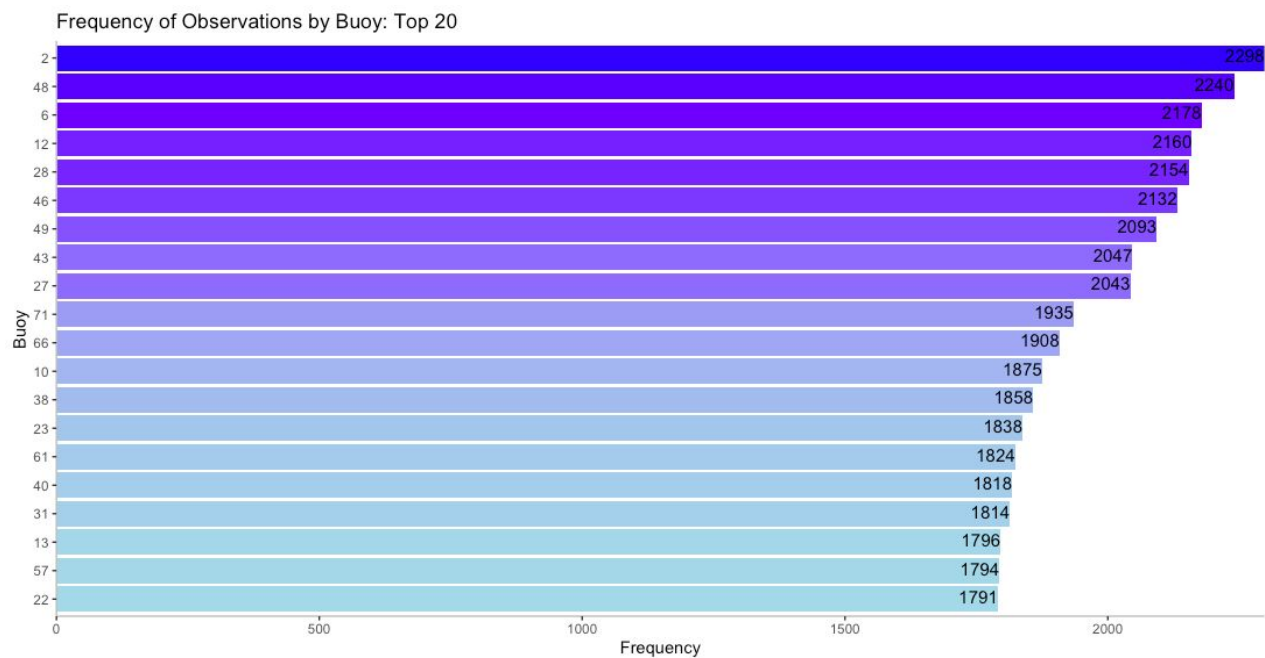


# A2: Regression -- Invincibles

## Part 1: Descriptive statistics

The “El Nino” data we analyzed in Assignment 1 was measured by 72 buoys distributed in significantly different time periods. The different properties of each buoy would introduce omitted variable bias if we would regress all of them. As a result of this, we chose only one buoy to conduct the regression analysis. We use buoy 2 for our analysis since it contains the most observations (2298 entries) as shown in the barplot below:



The cleaned data looks like the table below:

	Obser vatio n	New. Date	Lati tude	Lon gitude	Zonal .Wind s	Meridi onal. Winds	Humid ity	Air. Te mp	Sea.Surf ace.Tem p
1	12401	19900501	0.02	-139.93	-5.9	2.1	87.1	27.21	851
2	12402	19900502	0.01	-139.92	-4.4	-1.6	87.9	27.12	831

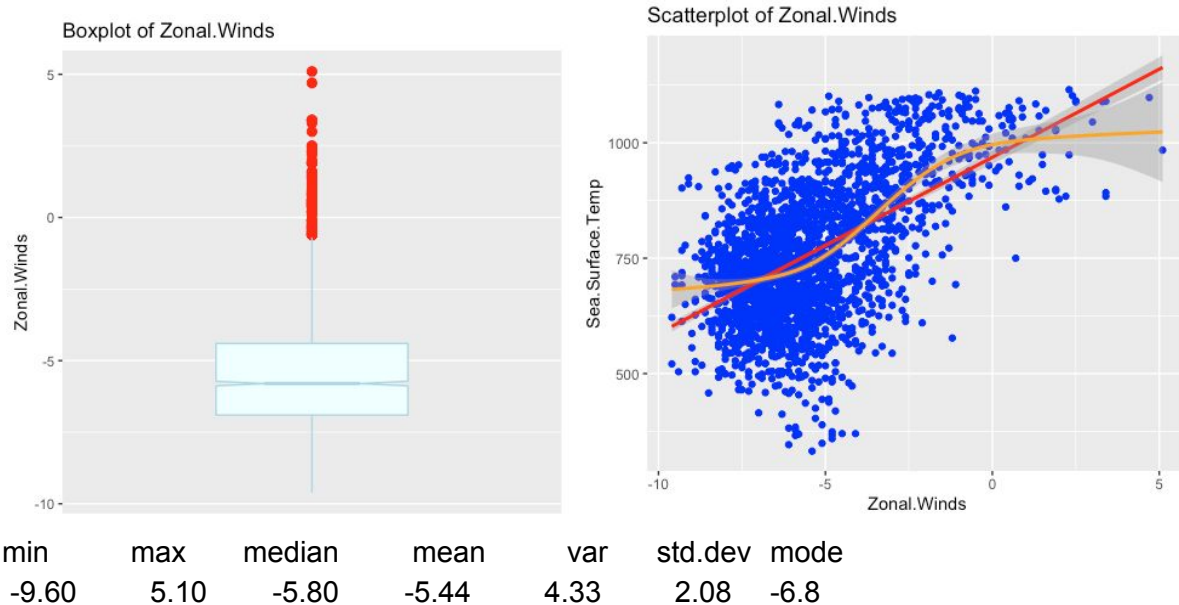
<b>3</b>	12403	19900503	0.02	-139.92	-5.8	1.1	86.5	27.56	830
<b>4</b>	12404	19900504	0.01	-139.92	-4.1	0.8	86.3	27.32	832
<b>5</b>	12405	19900505	0.01	-139.92	-3.3	-0.2	84.1	27.23	844

We are specifically interested in relevant and important variables, including Zonal.Winds, Meridional Winds, Humidity, Air Temp and Sea.Surface.Temp. The target variable is Sea.Surface.Temp, because this El Nino data set studies the El Nino effect in tropical pacific, and Sea.Surface.Temp indicates the change of the El Nino effect. As a result, we want to investigate the association between Sea.Surface.Temp and the rest of the variables, so that we can know whether the rest of the variables showcase any significant sign of the El Nino effect. If those variables are not statistically significantly associated with Sea.Surface.Temp, in future, we can remove those measures from the buoys.

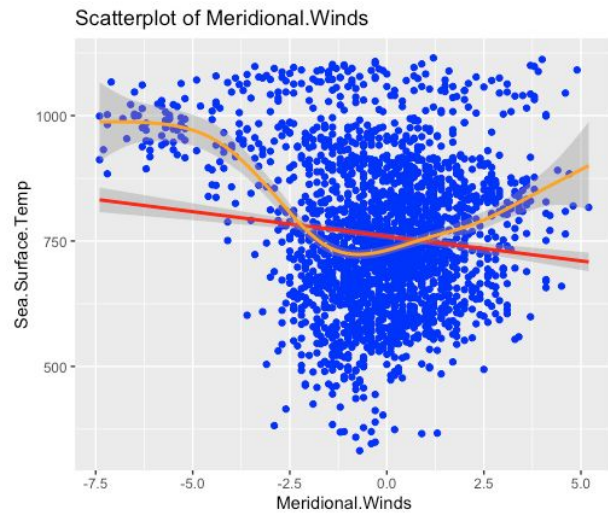
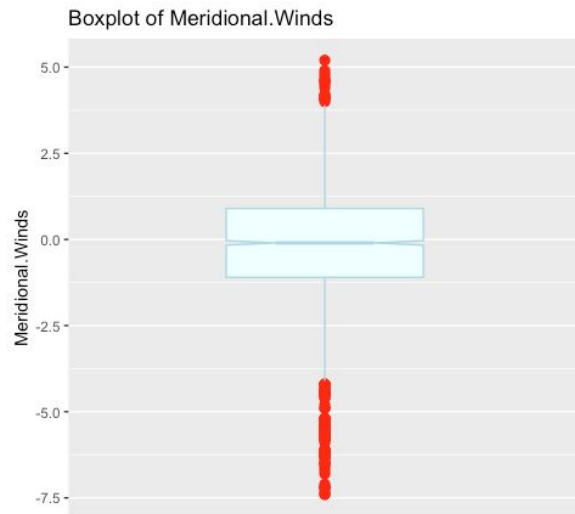
The predictor variables are Zonal.Winds, Meridional Winds, Humidity and Air Temp.

Here are the plots and descriptive statistics below. In the box-and-whisker plots, the outliers are plotted as red dots. In the scatterplots, both linear(red) and curvilinear(orange) regression are applied to the plots.

## Descriptive statistics of Zonal.Winds:

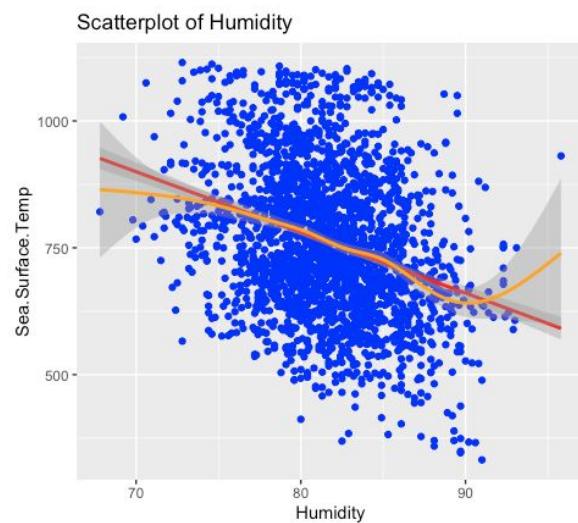
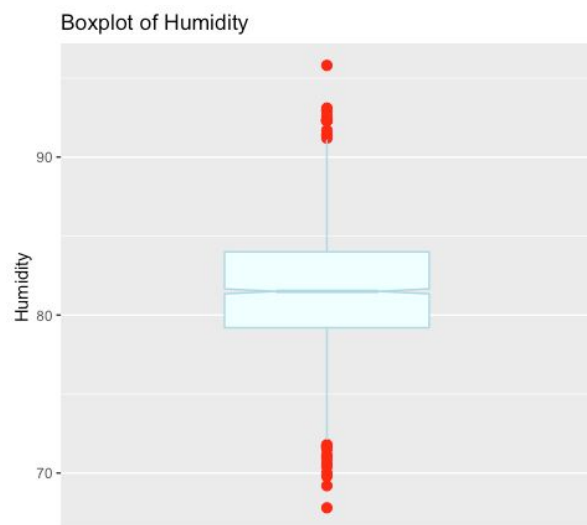


## Descriptive statistics of Meridional.Winds:



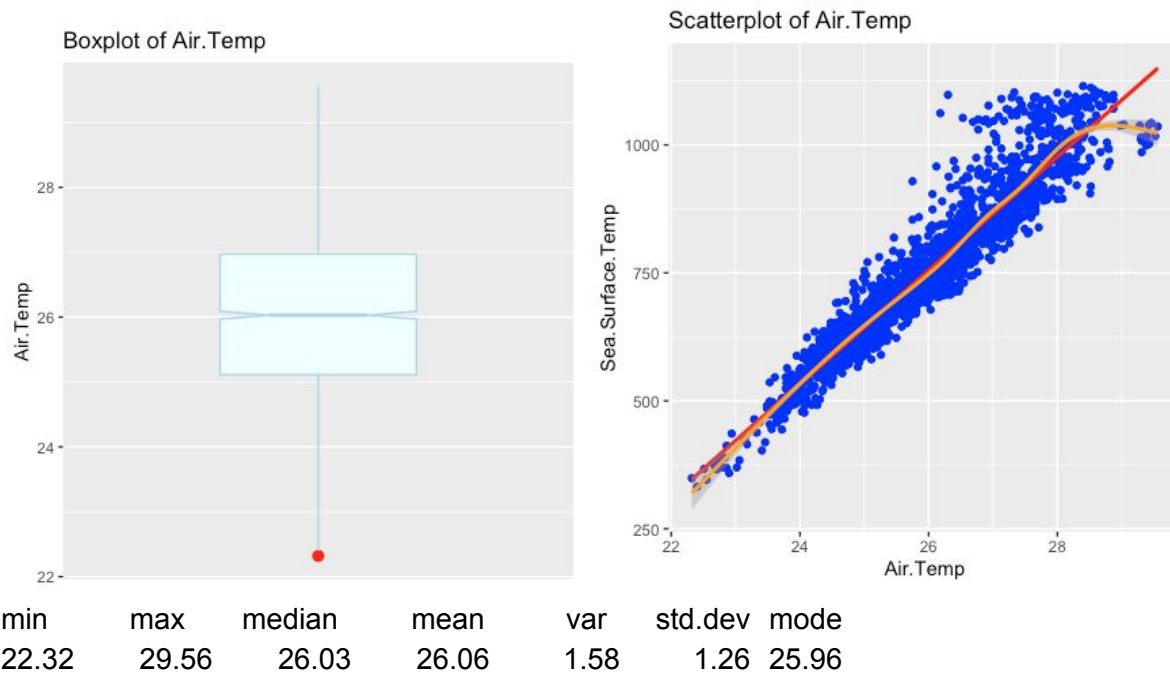
min	max	median	mean	var	std.dev	mode
-7.40	5.20	-0.10	-0.17	3.28	1.81	0.4

## Descriptive statistics of Humidity:

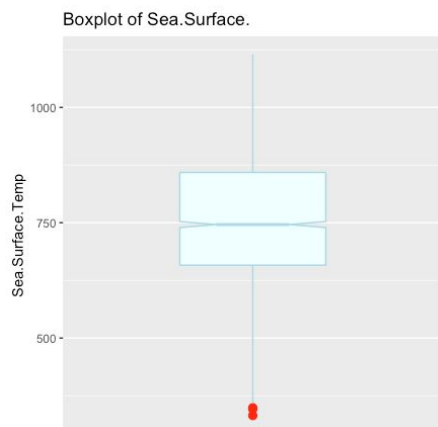


min	max	median	mean	var	std.dev	mode
67.80	95.80	81.50	81.61	14.48	3.81	79.8

## Descriptive statistics of Air.Temp:



## Descriptive statistics of Sea.Surface.Temp:



Because Sea.Surface.Temp is the target variable, a scatterplot cannot be applied to this variable.

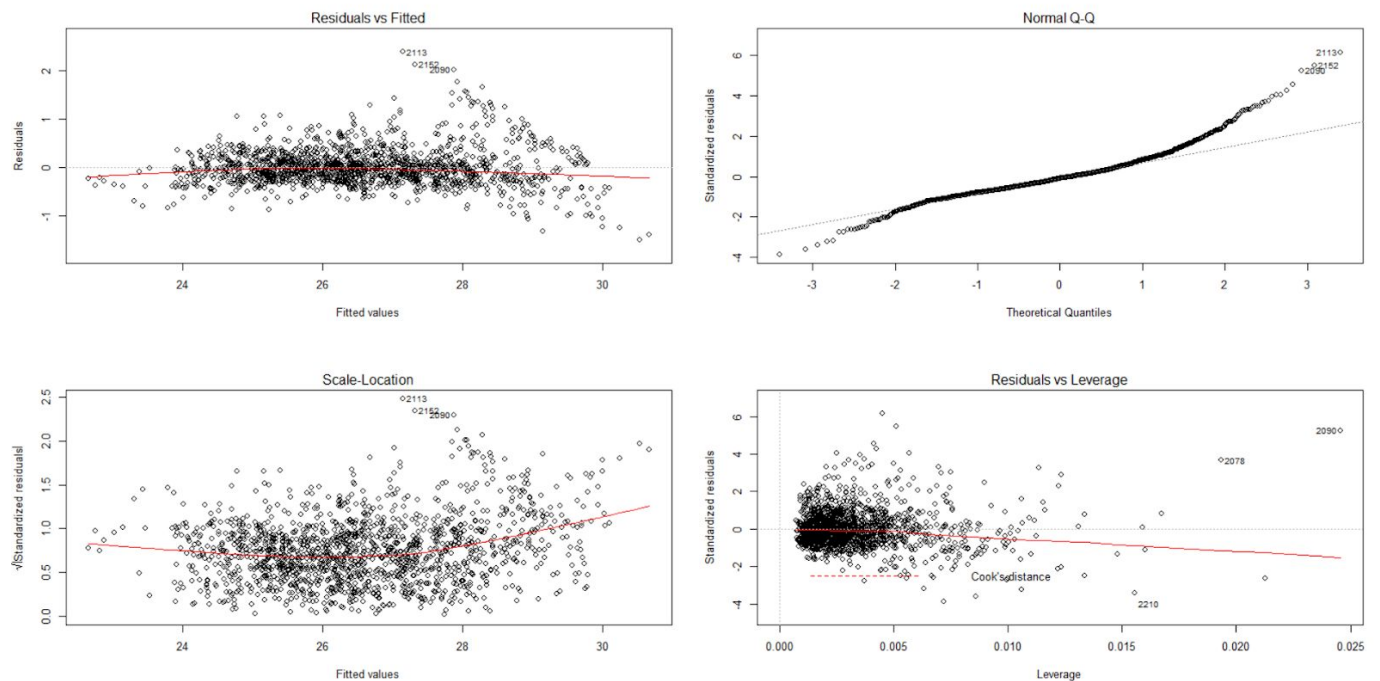
min	max	median	mean	var	std.dev	mode
332.00	1115.00	746.00	761.09	21557.46	146.82	713

## Part 2: Multiple Linear Regression

### Simple Linear Regression:

We randomly parted the training and validation sets. The training set originally included 1500 data points and the validation set included 798 data points. All four variables' p-values are close to 0. They can be regressed as predictors because they are statistically significant. Thereby, we regressed "Sea.Surface.Temp" on four independent variables: "Zonal.Winds", "Meridional.Winds", "Humidity", and "Air.Temp." By plotting the residuals, we discovered numerous residuals and eventually removed 87 data from the training set.

## Residual Plotting (Original)

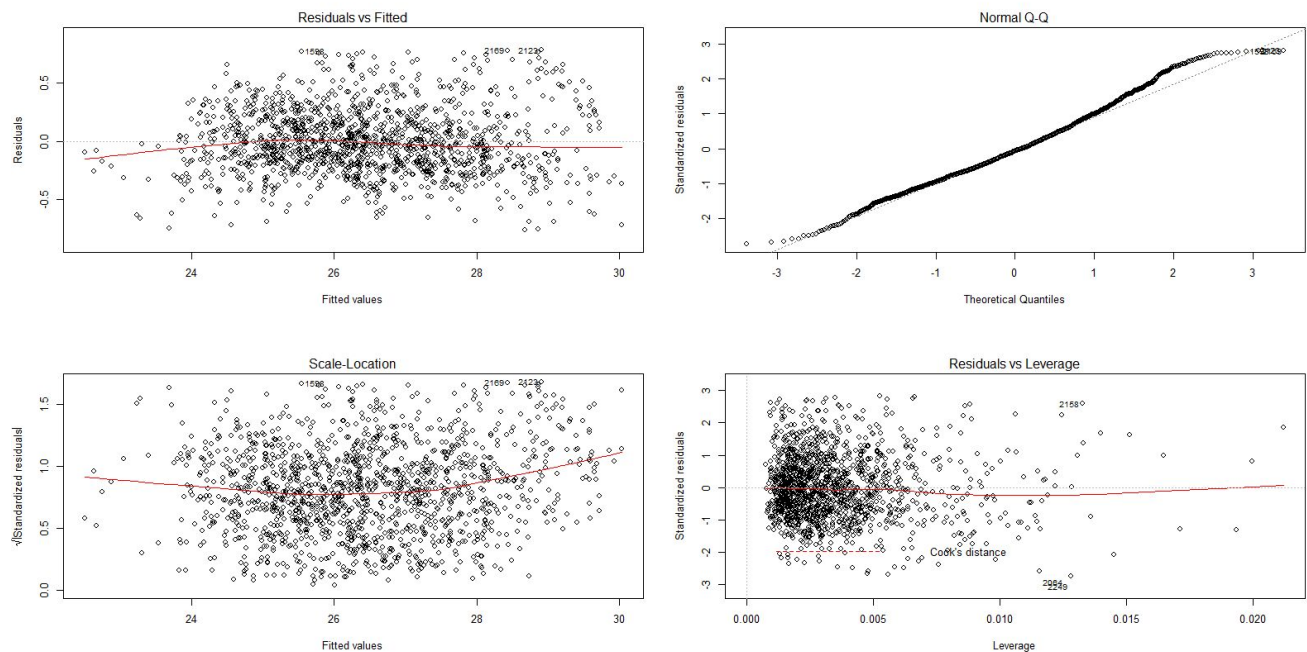


VIF (Original): all smaller than 5

Zonal.Winds	Meridional.Winds
1.247775	1.019644

Humidity	Air.Temp
1.160142	1.364073

## Residual Plotting (After Removing Outliers)



## Summary of Multiple Regression (After Removing Outliers)

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	3.656375	0.304712	11.999	< 0.0000000000000002	***
Zonal.Winds	0.099559	0.004345	22.913	< 0.0000000000000002	***
Meridional.Winds	-0.025868	0.004151	-6.232	0.0000000000606	***
Humidity	-0.033804	0.002123	-15.922	< 0.0000000000000002	***
Air.Temp	1.002654	0.007028	142.672	< 0.0000000000000002	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2786 on 1408 degrees of freedom

Multiple R-squared: 0.9603, Adjusted R-squared: 0.9602

F-statistic: 8511 on 4 and 1408 DF, p-value: < 0.00000000000000022

## Best Model Statistics (After Removing Outliers)

Adjusted R-squared: 0.9602

RMSE: 0.4092914



# Linear Regression With Interaction And Squared Terms:

The above model is the simplest model that we explored. We proceeded to explore a more complex model with the addition of squared and interaction terms for each of the variables. We then used the step AIC algorithm with “backward” selection in order to find the most relevant variables. With “backward”, we start with all the variables and then the variable that gives the minimum AIC when dropped, is dropped for the next iteration until there is no significant drop in AIC is noticed. Note that the data was standardized before carrying out this regression. Below are the results from this model:

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-0.0007311	0.0060743	-0.120	0.904217	
zonal_winds	-0.9249636	0.3345587	-2.765	0.005758	**
meridional_winds	1.5688520	0.2768976	5.666	1.71e-08	***
Humidity	-0.7434930	0.1653662	-4.496	7.39e-06	***
air_temp	1.3594144	0.4090398	3.323	0.000908	***
zonal_winds.meridional_winds	0.0741958	0.0136837	5.422	6.73e-08	***
zonal_winds.Humidity	0.5710270	0.1567442	3.643	0.000277	***
zonal_winds.air_temp	0.5782606	0.2109067	2.742	0.006174	**
meridional_winds.Humidity	-0.5986896	0.1470536	-4.071	4.89e-05	***
meridional_winds.air_temp	-0.9390529	0.1724628	-5.445	5.93e-08	***
Humidity.air_temp	0.9083694	0.1887142	4.813	1.61e-06	***
zonal_winds_squared	0.0546582	0.0204813	2.669	0.007687	**
air_temp_squared	-1.2212723	0.2793506	-4.372	1.31e-05	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.252 on 1711 degrees of freedom

Multiple R-squared: 0.9368, Adjusted R-squared: 0.9363

F-statistic: 2112 on 12 and 1711 DF, p-value: < 2.2e-16

## Model Statistics

Adjusted R-squared: 0.9363

RMSE: 0.2673344

The above model achieves lower adjusted R squared as compared to the simpler model. All the variables are also statistically significant at 10 percent statistical significance as a result of applying the AIC criterion. However, this outcome is not a guarantee that this is the most parsimonious model. We proceeded to check for multicollinearity among the variables by calculating the Variance Inflation Factors(VIFs).

Below is table that shows the VIFs for the variables in the model above:

Variable	VIF
meridional_winds_squared	1.651
zonal_winds.meridional_winds	4.869
zonal_winds_squared	7.680
zonal_winds.Humidity	7.799
Humidity	538.930
meridional_winds.Humidity	623.315
Humidity.air_temp	797.862
meridional_winds.air_temp	827.671
air_temp_squared	1700.301
meridional_winds	2118.720
air_temp	3295.693

The table above suggests that we should only include four predictors in the model whose VIFs are below 10. These are *meridional\_winds\_squared* , *zonal\_winds:meridional\_winds*, *zonal\_winds\_squared* and *zonal\_winds.Humidity*. High values of the VIF by themselves do not discount the results of regression analyses nor call for the elimination of one or more independent variables from the analysis. Rather they suggest the need to use alternative regression approaches or require combining independent variables into a single index.<sup>1</sup> With this weather data, we know that weather statistics of the previous days will affect today's weather. In the next section of this report, we explore seasonality in the data and apply autoregression techniques.

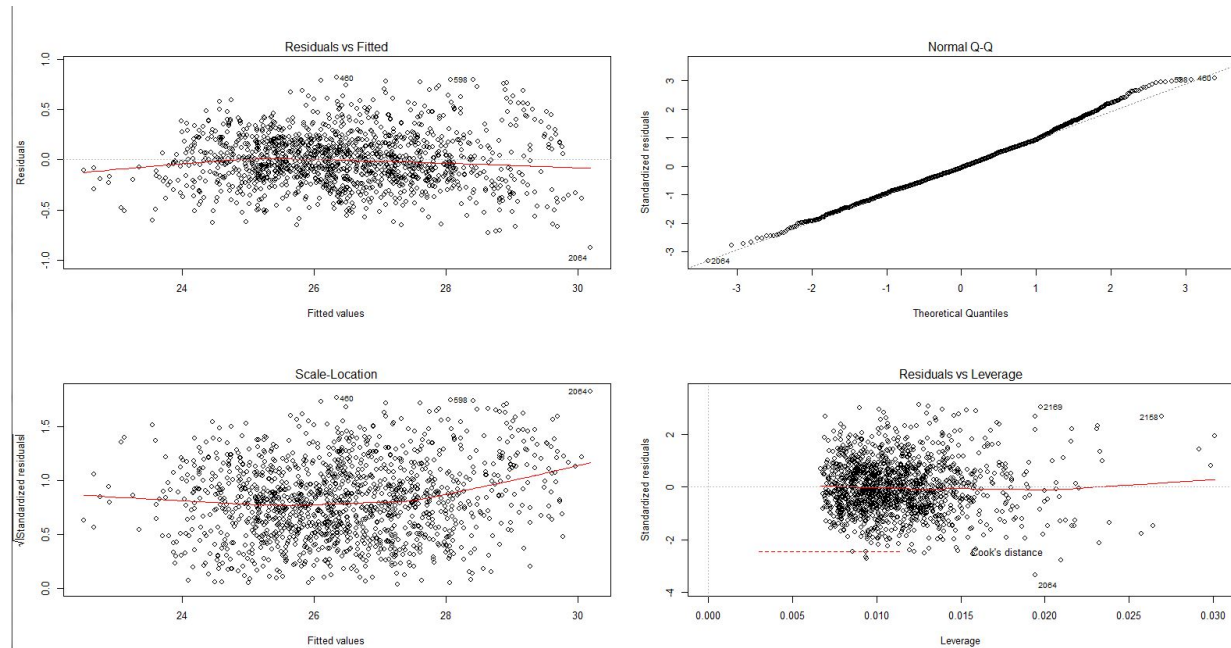
## Adding Dummies to the Model

Although the data was not collected evenly every year, there might be seasonality in the series. We added monthly dummy variables in our dataset and ran the regression with dummies. We removed the same 73 observations from the training set. The removed data accounts for only 5% of the total training data. It's reasonable because we have not neglected too much data and we have still captured most of the information.

---

<sup>1</sup> <https://link.springer.com/article/10.1007/s11135-006-9018-6>





## Dummies Model Statistics

Adjusted R-squared: 0.9638

RMSE: 0.39707

The main four variables' coefficients differ from the best model without dummy variables, but mostly they are aligned with the previous best model's coefficients. We removed the "January dummy" to avoid the dummy variable trap. The coefficients of dummies in April and May have negative values. These months may be relatively cooler in a year. Also, it seems that the sea surface temperature was still quite low just before the summer. This may be caused by the water temperature lag compared to the continent temperature. In these months, the value of dependent variable, sea surface temperature, should be reduced. August, September and October were the months with the highest temperature. The dummy coefficients in these months have the highest positive values. Note that the buoy traveled around the equatorial area. The temperature seasonal variance should be significantly lower than other places on the earth. Below are the results for this model:

```

Residuals:
    Min       1Q   Median       3Q      Max
-0.87536 -0.17805 -0.01438  0.16685  0.81973

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.631786    0.331615   7.936 0.00000000000000424 ***
Zonal.Winds     0.101032    0.004357  23.188 < 0.00000000000000002 ***
Meridional.Winds -0.032439    0.004274  -7.590 0.000000000000005810 ***
Humidity        -0.033577    0.002239 -14.995 < 0.00000000000000002 ***
Air.Temp        1.037693    0.007745 133.981 < 0.00000000000000002 ***
new_el_nino_2    0.085548    0.037935   2.255    0.0243 *
new_el_nino_3    0.092892    0.040885   2.272    0.0232 *
new_el_nino_4   -0.010277    0.039823  -0.258    0.7964
new_el_nino_5   -0.025454    0.035106  -0.725    0.4685
new_el_nino_6    0.059474    0.034464   1.726    0.0846 .
new_el_nino_7    0.162135    0.034211   4.739 0.00000236393006306 ***
new_el_nino_8    0.252033    0.034946   7.212 0.00000000000090056 ***
new_el_nino_9    0.258279    0.036967   6.987 0.00000000000434092 ***
new_el_nino_10   0.239278    0.037483   6.384 0.00000000023485734 ***
new_el_nino_11   0.047402    0.035090   1.351    0.1770
new_el_nino_12   0.053885    0.035300   1.526    0.1271
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2654 on 1397 degrees of freedom
Multiple R-squared:  0.9642,    Adjusted R-squared:  0.9638
F-statistic: 2511 on 15 and 1397 DF,  p-value: < 0.00000000000000022

```

## Dummies, Interaction and Squared Terms Regression:

In the end, we decided to wrap all variables in the same model, and the backward selection suggests the following variables for regression. Their respective VIFs are shown as well.

Residual standard error: 0.2386 on 1702 degrees of freedom  
Multiple R-squared: 0.9437, Adjusted R-squared: 0.943  
F-statistic: 1359 on 21 and 1702 DF, p-value: < 2.2e-16

new_el_nino_1	new_el_nino_2
1.242129	1.270765
new_el_nino_3	new_el_nino_4
1.333398	1.377991
new_el_nino_5	new_el_nino_6
1.569691	1.377927
new_el_nino_8	new_el_nino_9
1.319415	1.324338
new_el_nino_10	zonal_winds
1.251818	2.042560
meridional_winds	humidity
1.772421	1.343289
air_temp	`zonal_winds:meridional_winds`
1.990387	1.874905
`zonal_winds:humidity`	`zonal_winds:air_temp`
1.431302	2.140489
`meridional_winds:humidity`	`meridional_winds:air_temp`
1.282815	2.420577
`humidity:air_temp`	meridional_winds_squared
1.761553	1.600211
air_temp_squared	
1.769854	

[1] 0.2476846

However, as the last line (RMSE=0.2476) in the above table suggests, this model actually predicts the test data well, despite the fact that it fits the training data well. The R squared on training data is 0.94, similar to models run above. The following table shows the result of the regression.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.091578	0.013490	6.789	1.56e-11	***
new_el_nino_1	-0.058596	0.022861	-2.563	0.010459	*
new_el_nino_2	-0.057566	0.026504	-2.172	0.029997	*
new_el_nino_3	-0.119121	0.029079	-4.096	4.39e-05	***
new_el_nino_4	-0.165593	0.027837	-5.949	3.27e-09	***
new_el_nino_5	-0.269319	0.023719	-11.354	< 2e-16	***
new_el_nino_6	-0.163272	0.022336	-7.310	4.09e-13	***
new_el_nino_8	0.111484	0.022316	4.996	6.46e-07	***
new_el_nino_9	0.078669	0.024210	3.250	0.001179	**
new_el_nino_10	0.110303	0.023771	4.640	3.75e-06	***
zonal_winds	0.203069	0.008116	25.021	< 2e-16	***
meridional_winds	-0.059216	0.007652	-7.739	1.71e-14	***
humidity	-0.044051	0.006695	-6.580	6.26e-11	***
air_temp	0.919944	0.008117	113.334	< 2e-16	***
`zonal_winds:meridional_winds`	0.029997	0.005558	5.398	7.71e-08	***
`zonal_winds:humidity`	0.025460	0.006979	3.648	0.000272	***
`zonal_winds:air_temp`	0.013942	0.008068	1.728	0.084160	.
`meridional_winds:humidity`	-0.023701	0.006782	-3.495	0.000486	***
`meridional_winds:air_temp`	-0.031395	0.007900	-3.974	7.36e-05	***
`humidity:air_temp`	0.059119	0.007194	8.218	4.06e-16	***
meridional_winds_squared	-0.015787	0.003891	-4.057	5.19e-05	***
air_temp_squared	-0.021801	0.006228	-3.501	0.000476	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2386 on 1702 degrees of freedom

Multiple R-squared: 0.9437, Adjusted R-squared: 0.943

F-statistic: 1359 on 21 and 1702 DF, p-value: < 2.2e-16

## Model interpretations and reflections

Results from the multiple regression suggests that we need a simpler model for this data.

We have run four regressions above, including simple linear regression, regression with squared and interaction terms, simple linear regression with monthly dummy variable and regression with all variables. All four regressions have similar R-squared values above 0.9, however, regression with squared and interaction terms and regression with all variables have significantly lower RMSE values, i.e. 0.26 and 0.24 respectively. It makes sense, since the temperature varies from time to time in a year and there are some interactive effects among different variables, e.g. zonal winds and humidity as the above table suggests.