# Bouys and Data Cleaning: Report

*Melody Shaff, Brian Clare, Elise Roche, and Carly Rose Willing*

*November 7, 2017*

## Choice of Buoy Locations (See Figure 1)

We examined the National Oceanic and Atmospheric Administration's National Data Buoy Center's buoys around the eastern coast of the United States and Canada and found 4 locations that met our criteria of having historical data going back to the 1980s and being fairly well-spaced geographically. From south to north, our buoys are: Mid Gulf of Mexico, Molasses Reef Florida, Cape May New Jersey and Georges Bank off the coast of Cape Cod.

## Report on Data Cleaning

Our raw buoy data was not ready for exploratory data analysis. The data had 14 columns and thousands of observations, many of which were superfluous for what we were focusing on for this project. Thus, our first step was to clean the data.

In order to clean the data, we first needed to read it into R. As we did not want to individually read in over 30 urls and we recognized the urls for each location differed by year only, we were able to create strings of the first half and second half of the url which remained consistent within a location. We used this to read in each location's data in a loop. Inside the loop we began the cleaning process.

First, we made sure all of the years were presented by 4 digit integers (as data before 1998 was represented by the last two digits, ex. 87 rather than 1987). To do this we needed to make the year a string and then change it back into a number. We then made sure all the rest of our data was in its numeric form, rather than represented as strings. Because early years were missing a minutes column, we added in the minutes for those as 00.

Second, we removed columns of no interest to us. For the purposes of this project we needed only: year (YYYY), month (MM), day (DD), hour (hh), minute (mm), air temperature (ATMP), and water temperature (WTMP). So, we selected only these columns.

Third, we needed a date column. Instead of the messy set of five columns representing time and date we transformed these (using make_datetime()) into a date object for each observation.

Fourth, we changed our temporary 'file' into the main data frame for this location. We used the temporary file to allow the code to be replicated, largely unchanged, between locations. However, at this point we are ready to establish the main data frame, unique to each location, and finish up our cleaning. Thus, we renamed file with the abbreviated notation for the location and bound that location's data into the new data frame.

Fifth, we filtered out observations so we have only one for each day. Whenever possible we chose entries at noon, minute 0, but some instances only had minutes recorded at 50 and in those cases we selected 11:50 AM instead as this is closer to noon than 12:50 PM. We filtered for one entry, at (or close to) noon, per day.

Sixth, as we now have a date object for each observation, we could make the new and necessary column for time difference (referring to the time, in number of seconds, from our data to noon). We did this simply by subtracting noon from the time at which the buoy data for each particular observation was taken.


Georges Bank
44011
Cape May
44009
Mid Gulf
42001
Molasses Reef
MLRF1

Figure 1:

Seventh, we needed to filter for the columns we care about once more. Now that we've simplified the presentation of the date we can select only this one column, instead of the initial five. We also need to include our new time difference column. Thus, our final selected columns are: date (date), time difference from noon (timediff), air temperature (ATMP), and water temperature (WTMP).

Eighth, we want to re-code our NAs to register as such. Thus, we used apply() to address entries of 999.0 and 99.0 to appear instead as NA.

Ninth, we changed our column names to make them more readable and added in necessary new columns. We changed ATMP to Air Temp, WTMP to Sea Temp, date to Date, and timediff to Time from Noon. Next, we added in a column for our group number (1), the type of data we used (buoy data), the latitude, and the longitude of the location.

Finally, we reordered the columns. As the group compiling all of the groups' data sets needs uniformity we made sure to re-order our columns last to ensure that were easy to use for the next group. Thus, the columns are now: 1) Group, 2) Type, 3) Date, 4) Time from Noon, 5) Lat, 6) Long, 7) Sea Temp, 8) Air Temp.

We used this same procedure for each of our locations' data. Our very last step was to put the cleaned data frames into an xlsx file. We did this by writing the file with the first location's data and adding the other three locations one by one as new sheets.

## Descriptive Statistics For Cape May

Table 1: Cape May Descriptive Statistics for All Years

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | -14.20 | 13.40 | 13.20 | 7.864 | 27.10 | 41.3 | 0.9228646 |
| Sea Temperature | 0.50 | 14.20 | 14.31 | 6.874 | 26.90 | 26.4 | |

Table 2: Cape May Descriptive Statistics for 1984

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | -8.60 | 14.30 | 12.91 | 7.912 | 25.20 | 33.8 | 0.9269352 |
| Sea Temperature | 3.50 | 13.50 | 13.21 | 6.737 | 25.60 | 22.1 | |

Table 3: Cape May Descriptive Statistics for 2015

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | -8.0 | 14.5 | 13.9 | 8.653 | 26.3 | 34.3 | 0.9455966 |
| Sea Temperature | 0.5 | 15.4 | 14.6 | 7.948 | 26.40 | 25.9 | |

Table 4: T-Test for Difference of Means for Cape May

| Variables of Comparison | t-value | p-value |
|---|---|---|
| Air Temperature 1984 vs 2015 | 1.5425 | 0.1234 |
| Sea Temperature 1984 vs 2015 | 2.4904 | 0.013 |

## Descriptive Statistics For Molasses Reef

Table 5: Molasses Reef Descriptive Statistics for All Years

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | 2.60 | 25.30 | 24.68 | 3.972 | 30.70 | 28.1 | 0.81659 |
| Sea Temperature | 18.00 | 26.40 | 26.49 | 2.506 | 31.20 | 13.2 | |

Table 6: Molasses Reef Descriptive Statistics for 1988

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | 9.80 | 24.50 | 24.08 | 4.077 | 29.00 | 19.2 | 0.8271741 |
| Sea Temperature | 20.5 | 26.10 | 25.98 | 2.435 | 29.60 | 9.1 | |

Table 7: Molasses Reef Descriptive Statistics for 2015

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | 8.30 | 26.50 | 25.61 | 3.516 | 29.90 | 21.6 | 0.8482789 |
| Sea Temperature | 19.9 | 27.30 | 27.20 | 2.419 | 30.90 | 11.0 | |

Table 8: T-Test for Difference of Means for Molasses Reef

| Variables of Comparison | t-value | p-value |
|---|---|---|
| Air Temperature 1988 vs 2015 | 5.41 | <.0001 |
| Sea Temperature 1988 vs 2015 | 6.71 | <.0001 |

## Descriptive Statistics for Georges Bank

Table 9: Georges Bank Descriptive Statistics for All Years

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | -11.20 | 10.70 | 10.69 | 6.148 | 25.80 | 37.0 | 0.87634 |
| Sea Temperature | 0.40 | 10.10 | 10.72 | 5.216 | 27.40 | 27.0 | |

Table 10: Georges Bank Descriptive Statistics for 1985

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | -4.10 | 11.20 | 11.02 | 5.933 | 24.80 | 28.9 | 0.8457552 |
| Sea Temperature | 3.60 | 10.35 | 10.70 | 5.038 | 23.40 | 19.7 | |

Table 11: Georges Bank Descriptive Statistics for 2015

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | 4.30 | 15.60 | 15.24 | 5.116 | 25.80 | 21.5 | 0.8811325 |
| Sea Temperature | 4.80 | 14.60 | 14.78 | 5.251 | 25.40 | 20.6 | |

Table 12: T-Test for Difference of Means for Georges Bank

| Variables of Comparison | t-value | p-value |
|---|---|---|
| Air Temperature 1985 vs 2015 | 8.87 | <.0001 |
| Sea Temperature 1985 vs 2015 | 8.948 | <.0001 |

## Descriptive Statistics for Mid Gulf

Table 13: Mid Gulf Descriptive Statistics for All Years

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | 7.50 | 25.10 | 24.71 | 3.398 | 32.20 | 24.7 | 0.84549 |
| Sea Temperature | 19.8 | 26.30 | 26.20 | 2.696 | 31.00 | 11.2 | |

Table 14: Mid Gulf Descriptive Statistics for 1984

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | 13.50 | 23.80 | 23.52 | 2.918 | 28.20 | 14.7 | 0.7219845 |
| Sea Temperature | 20.30 | 25.60 | 25.68 | 2.501 | 29.80 | 9.5 | |

Table 15: Mid Gulf Descriptive Statistics for 2016

| Variable | Minimum | Median | Mean | Standard Deviation | Maximum | Range | Correlation |
|---|---|---|---|---|---|---|---|
| Air Temperature | 16.80 | 26.80 | 25.85 | 2.905 | 29.40 | 12.6 | 0.8409042 |
| Sea Temperature | 22.60 | 27.20 | 27.46 | 1.881 | 30.50 | 7.9 | |

Table 16: T-Test for Difference of Means for Mid Gulf

| Variables of Comparison | t-value | p-value |
|---|---|---|
| Air Temperature 1984 vs 2015 | 9.79 | <.0001 |
| Sea Temperature 1984 vs 2015 | 10.18 | <.0001 |

## Report on the Descriptive Statistics

For the descriptive statistics, we decided to take a rather orthodox approach. For each buoy, the entire data set–for the air temperature variable and the sea temperature variable–was analyzed using a basic summary command. This yielded the minimum, maximum, mean, and median for each variable. The standard deviation for each variable was also calculated, as well as a correlation between the air temperature and the sea temperature. These values can be viewed in Tables 1, 5, 9, and 13.

After that, we decided that we would like to compare the temperatures from the first year of available data and the most recent year of available data. However, some of the buoy's did not have full year readings for the first year they were in service. For that reason, we used the first full year of data that was available. For example, the Molasses Reef buoy began collecting data in late 1987, so we used the first complete year, 1988. This data can be viewed in Tables 2, 6, 10, and 14. In addition, the most recent year of available data was not always complete either. For this reason, we used the most recent year of complete data. In the case of Georges Bank, this meant that instead of using the six months of data available from 2016, we used the complete data from 2015. This data can be viewed in Tables 3, 7, 11, and 15.

In order to accurately compare the changes in temperature from the first year to the most recent year, we ran t-tests for both sea temperature and air temperature. The corresponding t-values and p-values can be viewed in Tables 4, 8, 12, and 16. Interestingly enough, there was a signficiant difference in the mean temperatures for every variable, except for Cape May's air temperature. This is because climate change, particularly global warming, tends to affect ocean areas farther away from land, and closer to the equator.

As expected, the air temperatures and sea temperature for each region were highly correlated. The air and sea temperature for Molasses Reef and Mid Gulf were slightly lower. This may be due to a pervasive pattern of warm sea temperature, regardless of the air temperature. At times, a cool front may move through the region, but because it is short lived, there is not a significant impact on the corresponding sea temperature. Similarly, the range for sea temperatures is smaller than the range of temperatures for air temperature. This is because air temperature can fluctuate more rapidly before it dramatically affects the corresponding sea temperature.

# Data Visualization (See Figure 2)

## Air Temperature and Sea Temperature Correlation For Each Location

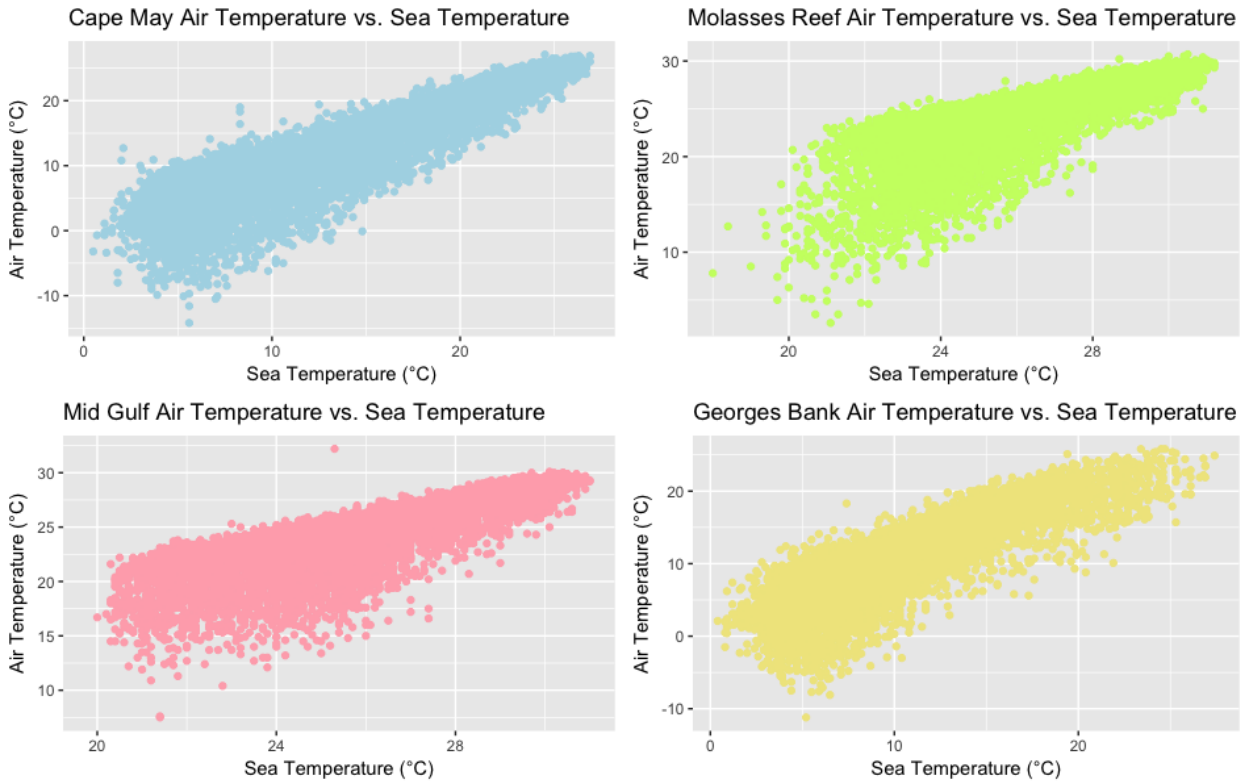


Figure 2:

Each of these plots demonstrates a positive correlation between air temperature and sea temperature. The warmer buoy locations (Mid Gulf and Molasses Reef) that are closest to the equator have a slightly weaker correlation when the air temperature is colder. This makes sense as the water in these locations is in general warmer with less variation. These results are unsurprising, but do give us a level of validation in the measurements of our buoys. If they provide data that logically makes sense in this case, we are more able to trust the air and sea temperature data supplied by these buoys. The range here is also well within what we would expect with more northern buoys having the colder sea and air temperatures and a wider range of temperatures than the buoys farther south.

## Report on Seasonal Data Visualization

We chose to compare year vs. air temperature associations for each part of the year because we wanted to narrow down the data sets in the hopes of finding some type of correlation. We also thought it would be helpful in the case that long-term temperature change might affect lower or higher temperatures differently. As can be seen in the images and in the correlation coefficients of Temperature and Year, in three of four locations, summer temperatures had a slight positive correlation over other seasons. This is in line with the idea that the earth is heating up gradually, although the correlation coefficients are small and cannot necessarily inform accurate conclusions. In the Mid Gulf, correlation between air temperature and year during three of four months was higher than all other correlations we looked at combined. This coincides with the descriptive statistics displayed previously, and the notion that regions closer to the equator will be affected more greatly by climate change. The regression lines on each graph have extremely slight slopes, in line with the correlation coefficients and the gradual nature of long-term temperature change.

## Corresponding Visualizations

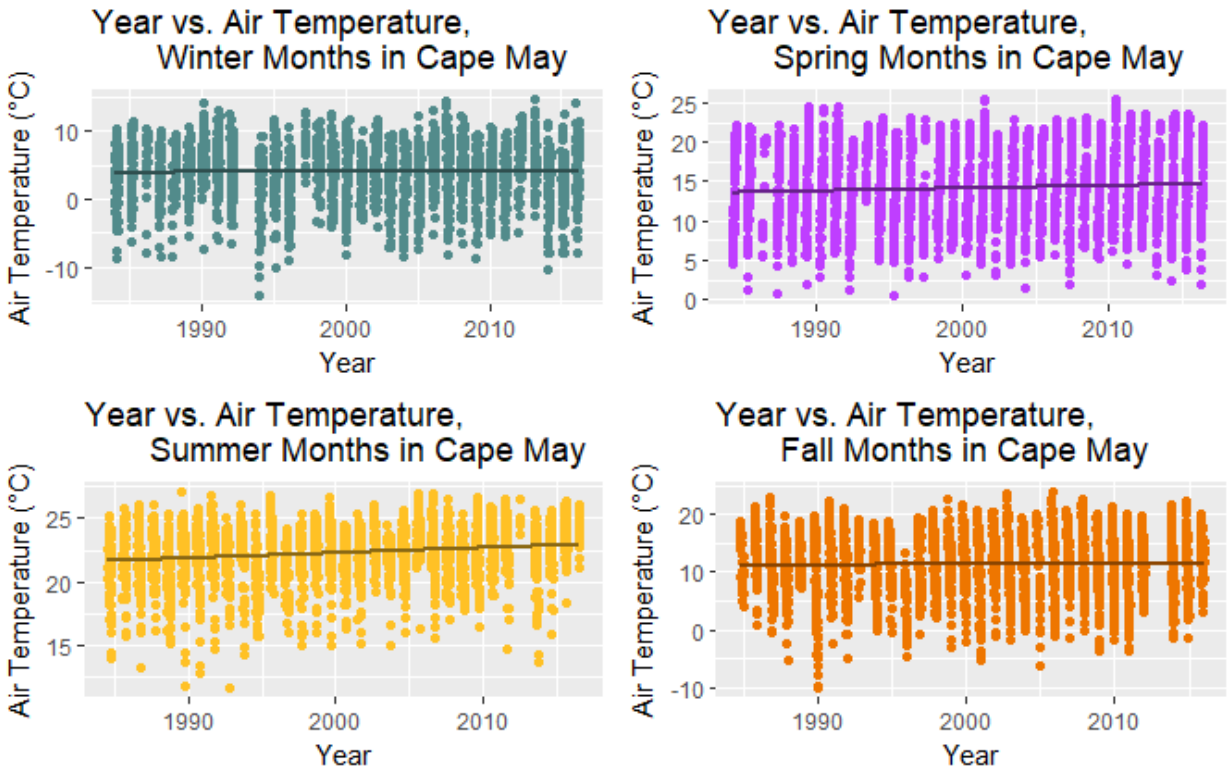


Figure 3:

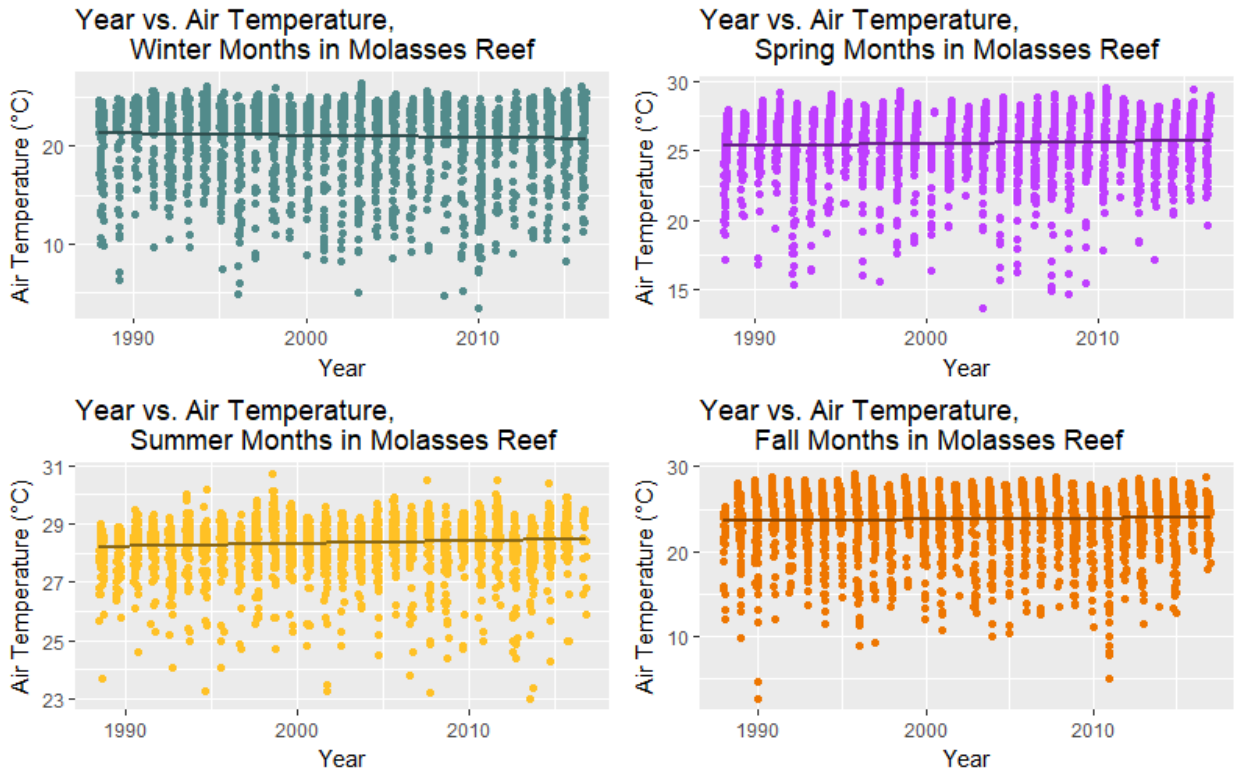


Figure 4:

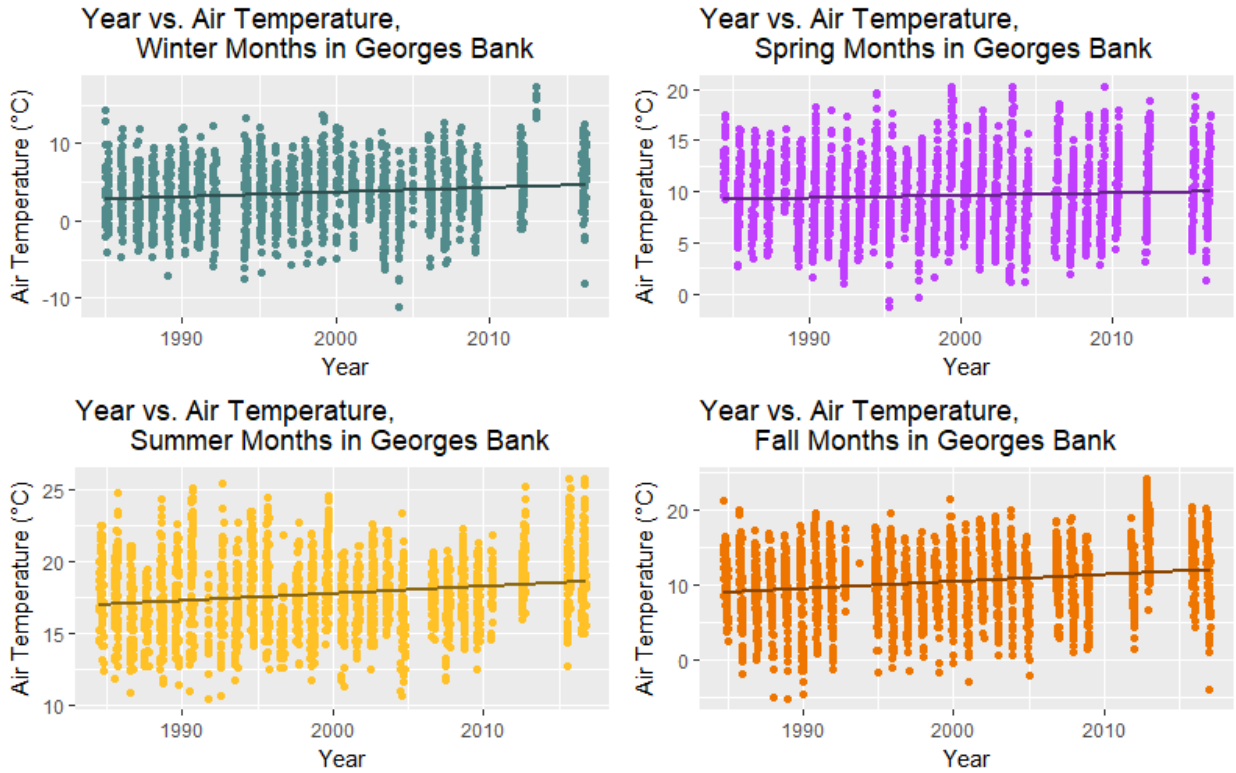


Figure 5:

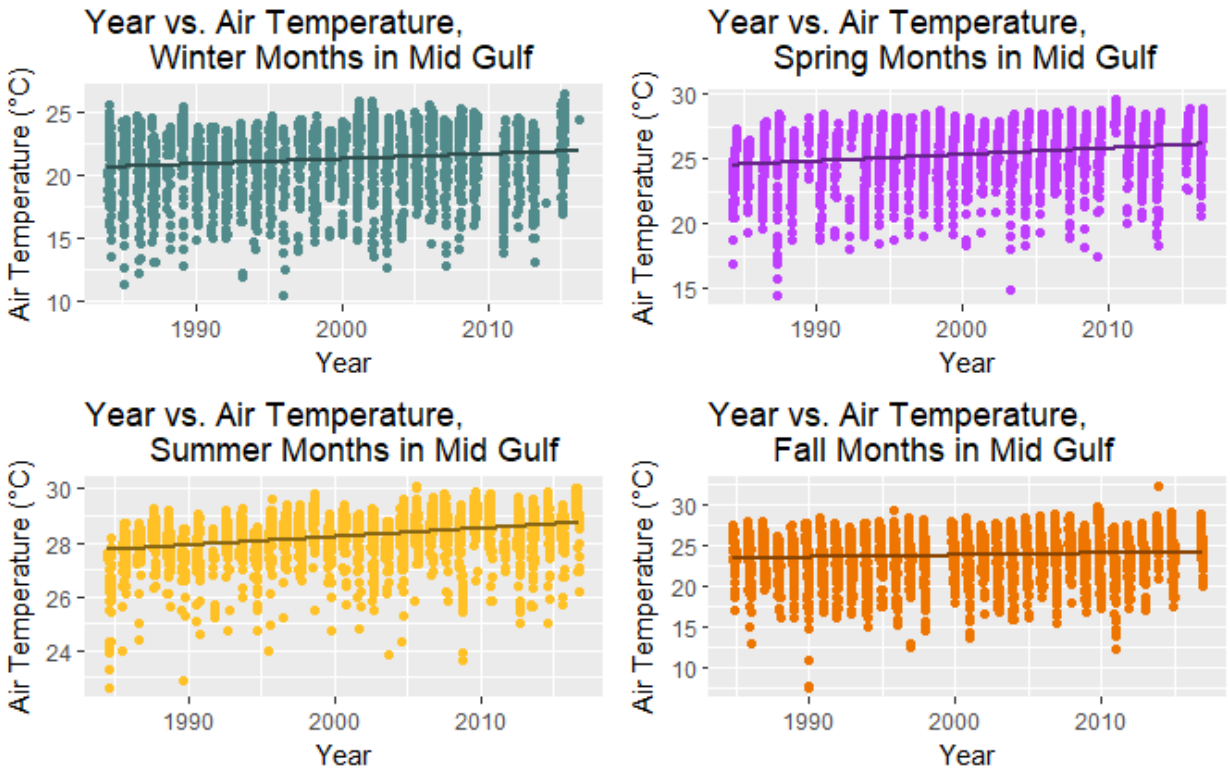
Year vs. Air Temperature,
Winter Months in Mid Gulf
Air Temperature (°C)
25
20
15
10
1990
2000
2010
Year
Year vs. Air Temperature,
Spring Months in Mid Gulf
Air Temperature (°C)
30
25
20
15
1990
2000
2010
Year
Year vs. Air Temperature,
Summer Months in Mid Gulf
Air Temperature (°C)
30
28
26
24
1990
2000
2010
Year
Year vs. Air Temperature,
Fall Months in Mid Gulf
Air Temperature (°C)
30
25
20
15
10
1990
2000
2010
Year

Figure 6: