Buoy Project

Haoran Su, Haoyu Li, Xijia Luo, Yifan Liu

2020/9/24

```
library(tidyverse)
library(stringr)
library(tidyr)
library(lubridate)
library(arm)
library(ggplot2)
library(dplyr)
library(forcats)
library(lme4)
library(purrr)
library(readr)
library(tibble)
```

Introduction

Over the past 50 years, the average global temperature has increased at the fastest rate in recorded history. Scientists agree that the earth's rising temperatures are fueling longer and hotter heat waves, more frequent droughts, heavier rainfall, and more powerful hurricanes. On the other hand, the earth's ocean temperatures are getting warmer, too. Thus, it is important for us to monitor the ocean temperatures with buoys. A buoy is a floating device equipped to measure weather parameters. It might be anchored or allowed to drift in the open ocean currents. This report using data gathered from the National Oceanic and Atmospheric Administration's National Data Buoy Center. The purpose of this report is to explore the factors that might influence the air temperature (ATMP) above the sea level.

Analysis

Import Data

```
#1999
Buoy1999 = read.csv("1999.txt", sep = "", header = T)
Buoy1999$TIDE = 0
Buoy1999$mm = 0
colnames(Buoy1999)[1] = "X.YY"
colnames(Buoy1999)[5] = "WDIR"
colnames(Buoy1999)[12] = "PRES"

#2000
Buoy2000 = read.csv("2000.txt", sep = "", header = T)
Buoy2000$TIDE = 0
Buoy2000$mm = 0
colnames(Buoy2000)[1] = "X.YY"
```

```
colnames(Buoy2000)[5] = "WDIR"
colnames(Buoy2000)[12] = "PRES"

#2001
Buoy2001 = read.csv("2001.txt", sep = "", header = T)
Buoy2001$mm = 0
colnames(Buoy2001)[1] = "X.YY"
colnames(Buoy2001)[5] = "WDIR"
colnames(Buoy2001)[12] = "PRES"


#2002
Buoy2002 = read.csv("2002.txt", sep = "", header = T)
Buoy2002$mm = 0
colnames(Buoy2002)[1] = "X.YY"
colnames(Buoy2002)[5] = "WDIR"
colnames(Buoy2002)[12] = "PRES"

#2003
Buoy2003 = read.csv("2003.txt", sep = "", header = T)
Buoy2003$mm = 0
colnames(Buoy2003)[1] = "X.YY"
colnames(Buoy2003)[5] = "WDIR"
colnames(Buoy2003)[12] = "PRES"

#2004
Buoy2004 = read.csv("2004.txt", sep = "", header = T)
Buoy2004$mm = 0
colnames(Buoy2004)[1] = "X.YY"
colnames(Buoy2004)[5] = "WDIR"
colnames(Buoy2004)[12] = "PRES"

#2005
Buoy2005 = read.csv("2005.txt", sep = "", header = T)
mm2005 = Buoy2005$mm
Buoy2005$mm = NULL
Buoy2005$mm = mm2005
colnames(Buoy2005)[1] = "X.YY"
colnames(Buoy2005)[5] = "WDIR"
colnames(Buoy2005)[12] = "PRES"

#2006
Buoy2006 = read.csv("2006.txt", sep = "", header = T)
mm2006 = Buoy2006$mm
Buoy2006$mm = NULL
Buoy2006$mm = mm2006
colnames(Buoy2006)[1] = "X.YY"
colnames(Buoy2006)[5] = "WDIR"
colnames(Buoy2006)[12] = "PRES"

#2007
Buoy2007 = read.csv("2007.txt", sep = "", header = T)
mm2007 = Buoy2007$mm
```

```
Buoy2007$mm = NULL
Buoy2007$mm = mm2007
Buoy2007 = Buoy2007[-c(1), ]

#2008
Buoy2008 = read.csv("2008.txt", sep = "", header = T)
mm2008 = Buoy2008$mm
Buoy2008$mm = NULL
Buoy2008$mm = mm2008
Buoy2008 = Buoy2008[-c(1), ]


#2009
Buoy2009 = read.csv("2009.txt", sep = "", header = T)
mm2009 = Buoy2009$mm
Buoy2009$mm = NULL
Buoy2009$mm = mm2009
Buoy2009 = Buoy2009[-c(1), ]

#2010
Buoy2010 = read.csv("2010.txt", sep = "", header = T)
mm2010 = Buoy2010$mm
Buoy2010$mm = NULL
Buoy2010$mm = mm2010
Buoy2010 = Buoy2010[-c(1), ]

#2011
Buoy2011 = read.csv("2011.txt", sep = "", header = T)
mm2011 = Buoy2011$mm
Buoy2011$mm = NULL
Buoy2011$mm = mm2011
Buoy2011 = Buoy2011[-c(1), ]

#2012
Buoy2012 = read.csv("2012.txt", sep = "", header = T)
mm2012 = Buoy2012$mm
Buoy2012$mm = NULL
Buoy2012$mm = mm2012
Buoy2012 = Buoy2012[-c(1), ]

#2013
Buoy2013 = read.csv("2013.txt", sep = "", header = T)
mm2013 = Buoy2013$mm
Buoy2013$mm = NULL
Buoy2013$mm = mm2013
Buoy2013 = Buoy2013[-c(1), ]

#2014
Buoy2014 = read.csv("2014.txt", sep = "", header = T)
mm2014 = Buoy2014$mm
Buoy2014$mm = NULL
Buoy2014$mm = mm2014
Buoy2014 = Buoy2014[-c(1), ]
```

```
#2015
Buoy2015 = read.csv("2015.txt", sep = "", header = T)
mm2015 = Buoy2015$mm
Buoy2015$mm = NULL
Buoy2015$mm = mm2015
Buoy2015 = Buoy2015[-c(1), ]

#2016
Buoy2016 = read.csv("2016.txt", sep = "", header = T)
mm2016 = Buoy2016$mm
Buoy2016$mm = NULL
Buoy2016$mm = mm2016
Buoy2016 = Buoy2016[-c(1), ]

#2017
Buoy2017 = read.csv("2017.txt", sep = "", header = T)
mm2017 = Buoy2017$mm
Buoy2017$mm = NULL
Buoy2017$mm = mm2017
Buoy2017 = Buoy2017[-c(1), ]

#2018
Buoy2018 = read.csv("2018.txt", sep = "", header = T)
mm2018 = Buoy2018$mm
Buoy2018$mm = NULL
Buoy2018$mm = mm2018
Buoy2018 = Buoy2018[-c(1), ]

#2019
Buoy2019 = read.csv("2019.txt", sep = "", header = T)
mm2019 = Buoy2019$mm
Buoy2019$mm = NULL
Buoy2019$mm = mm2019
Buoy2019 = Buoy2019[-c(1), ]

#Combine

Buoy =
  rbind(
    Buoy1999,
    Buoy2000,
    Buoy2001,
    Buoy2002,
    Buoy2003,
    Buoy2004,
    Buoy2005,
    Buoy2006,
    Buoy2007,
    Buoy2008,
    Buoy2009,
    Buoy2010,
    Buoy2011,
    Buoy2012,
```

```
    Buoy2013,
    Buoy2014,
    Buoy2015,
    Buoy2016,
    Buoy2017,
    Buoy2018,
    Buoy2019
  )
```

Data Clean

```
Buoy2 = filter(Buoy, hh == 10)
Buoy2 = data.frame(sapply(Buoy2, as.numeric))
Buoy2[Buoy2 == 99 | Buoy2 == 999 | Buoy2 == 9999] = 0
Buoy2$date = make_datetime(Buoy2$X.YY, Buoy2$MM, Buoy2$DD)
```

##Methodology How we combine different data set and how we clean it using certain methodology

For our dataset, we choose from year 1999-2019, which are 20 year. When we first look at 20 years data of the Boston buoy, we noticed that different years have different column names. For example, for the year of 1999 and 2000, the "Tide" column has no value. Then from 2005-2019, each dataset has one more column called "mm" which means the minutes.

What we use to solve these problems is that we decide to make these dataset exact the same, meaning that the column name and numbers of columns must be the same. Then it will be convenient for us to combine the data without considering the slight differences that will cause error.

Another problem for these datasets is that there are plenty of "NA" data that use 99, 999 to cover it. For those data, we also need to deal with them and eliminate the error that may cause in the future analysis.

Model

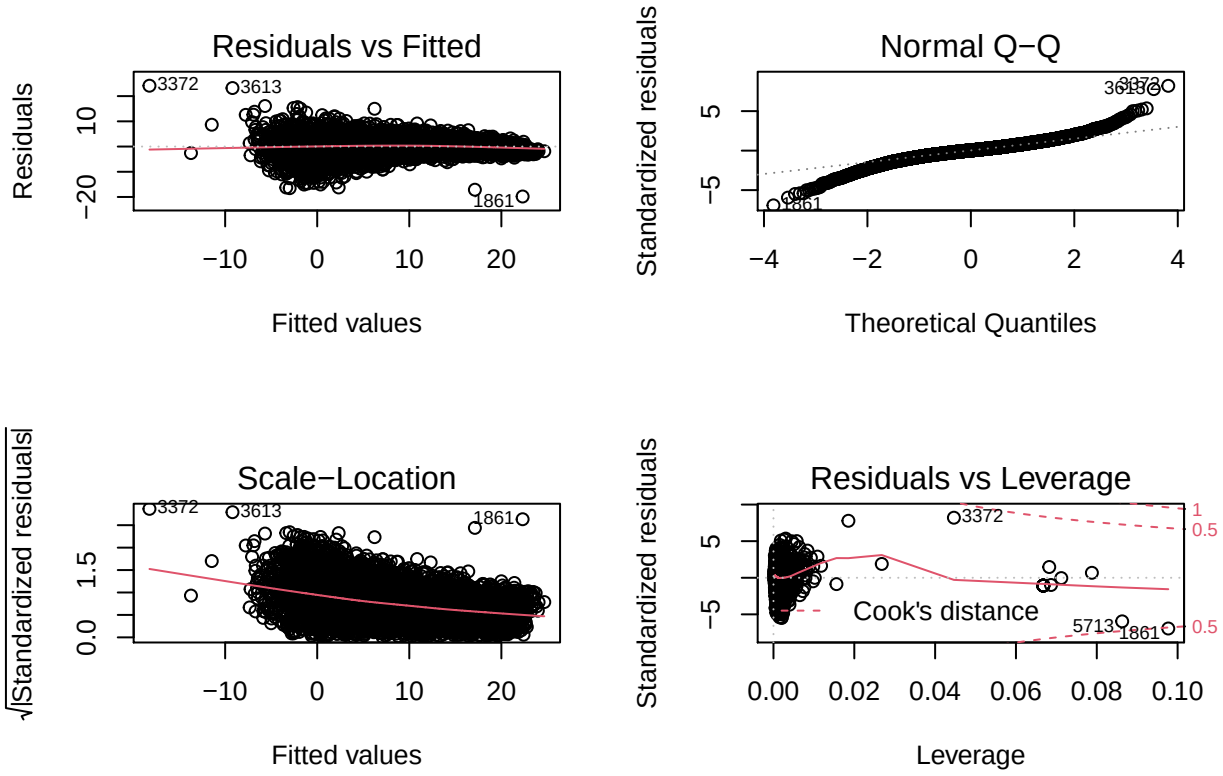
Model 1

```
set.seed(10000)
fit = lm(ATMP ~ WDIR + WSPD + GST + WVHT + DPD + APD + MWD + PRES + WTMP, data = Buoy2)
summary(fit)
```

```
##
## Call:
## lm(formula = ATMP ~ WDIR + WSPD + GST + WVHT + DPD + APD + MWD +
##     PRES + WTMP, data = Buoy2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8089  -1.4352   0.0941   1.5874  24.1204
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.1173281  0.7773541   4.010 6.13e-05 ***
## WDIR        -0.0068754  0.0003433 -20.026  < 2e-16 ***
## WSPD         2.7080940  0.0906586  29.871  < 2e-16 ***
## GST         -2.4889193  0.0746491 -33.342  < 2e-16 ***
## WVHT        -0.1856673  0.0847440  -2.191  0.02849 *
## DPD          0.0246329  0.0143055   1.722  0.08513 .
```

```
## APD          0.1486093  0.0404421   3.675  0.00024 ***
## MWD         -0.0050706  0.0004647 -10.913  < 2e-16 ***
## PRES        -0.0031337  0.0007734  -4.052 5.14e-05 ***
## WTMP         1.1104439  0.0068777 161.456  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.009 on 7381 degrees of freedom
## Multiple R-squared:  0.8539, Adjusted R-squared:  0.8537
## F-statistic:  4792 on 9 and 7381 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(fit)
```



Model 2

We assume that there are interactions between the valuables, so we try to see if the interactions would make the fit better. We add the interactive variable of WDIR(Wind direction during the same period) and MWD(The direction from which the waves at the dominant period). We add the interactive variable of WSPD(average wind speed) and GST(Peak gust speed). We do a log transformation on the variable PRES(sea level pressure) because the data of PRES is relatively large compared to other variables.

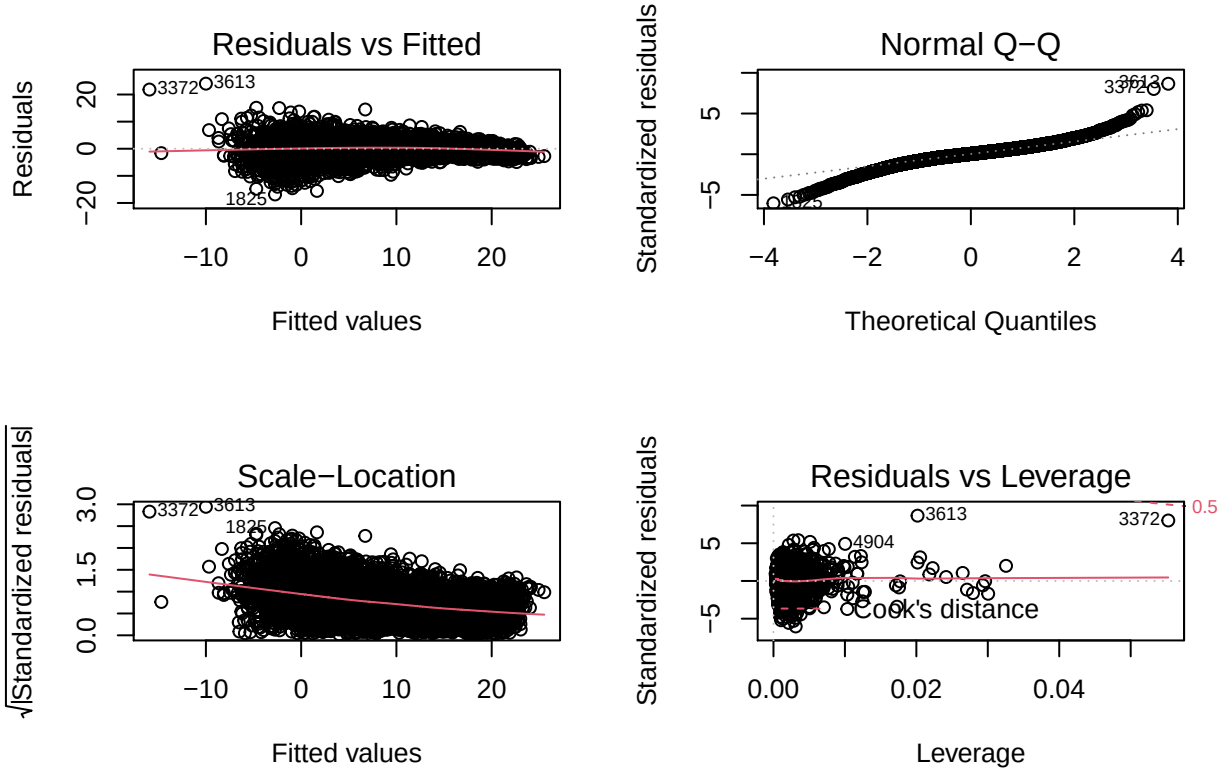
Model 2 with interaction and logarithm

```
set.seed(10000)
fit2 = lm(
  ATMP ~ WDIR + WSPD + GST + WVHT + DPD + APD + MWD + log(PRES) + WTMP + WDIR *
    MWD + WSPD * GST,
```

```
  data = Buoy2,
  subset = Buoy2$PRES > 0
)
summary(fit2)
```

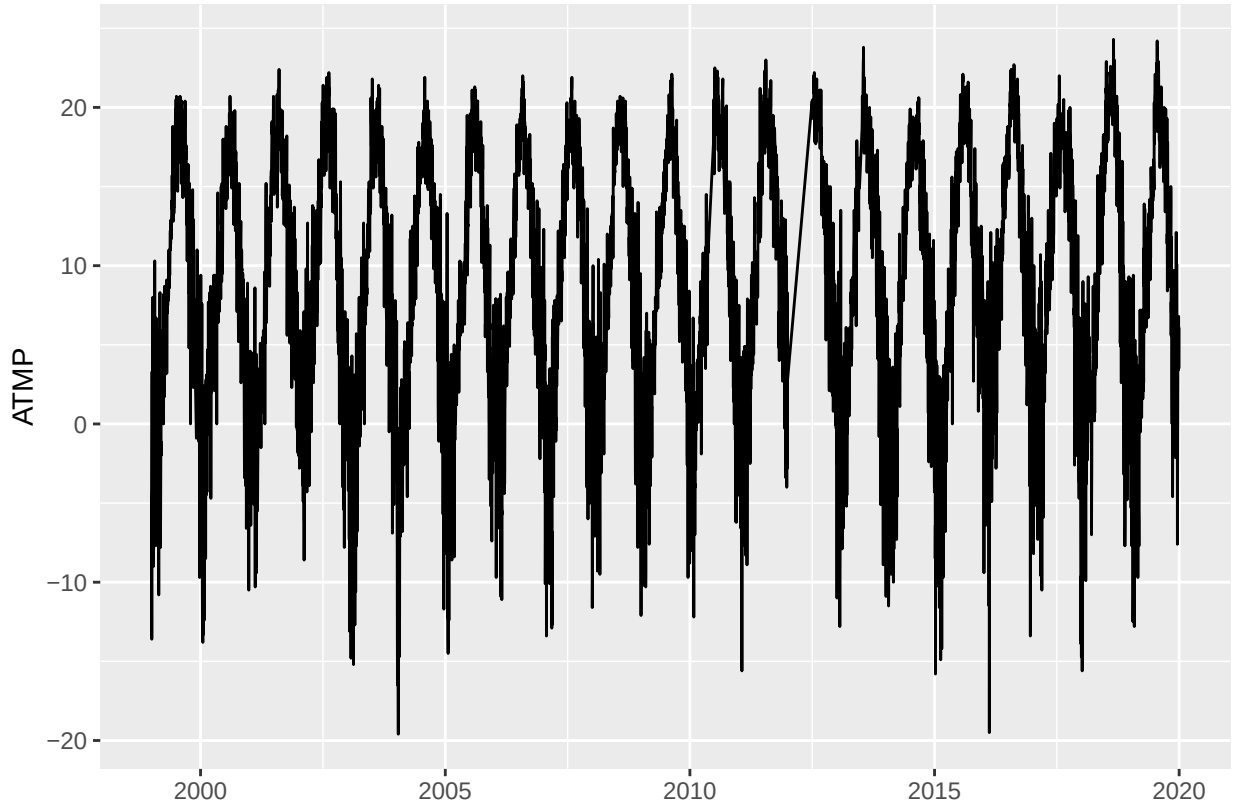
```
##
## Call:
## lm(formula = ATMP ~ WDIR + WSPD + GST + WVHT + DPD + APD + MWD +
##     log(PRES) + WTMP + WDIR * MWD + WSPD * GST, data = Buoy2,
##     subset = Buoy2$PRES > 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8605  -1.3308   0.1026   1.5426  23.9979
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.232e+02  2.814e+01  32.809  < 2e-16 ***
## WDIR        -7.118e-03  3.570e-04 -19.938  < 2e-16 ***
## WSPD         2.678e+00  8.863e-02  30.217  < 2e-16 ***
## GST         -2.387e+00  8.032e-02 -29.720  < 2e-16 ***
## WVHT        -2.219e-01  7.937e-02  -2.795  0.00520 **
## DPD          4.958e-03  1.339e-02   0.370  0.71109
## APD          1.119e-01  3.838e-02   2.915  0.00357 **
## MWD          1.209e-03  1.429e-03   0.846  0.39776
## log(PRES)   -1.333e+02  4.062e+00 -32.809  < 2e-16 ***
## WTMP         1.087e+00  6.486e-03 167.540  < 2e-16 ***
## WDIR:MWD    -2.092e-05  5.332e-06  -3.924 8.79e-05 ***
## WSPD:GST    -1.041e-02  2.007e-03  -5.189 2.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.8 on 7364 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8732
## F-statistic:  4618 on 11 and 7364 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(fit2)
```

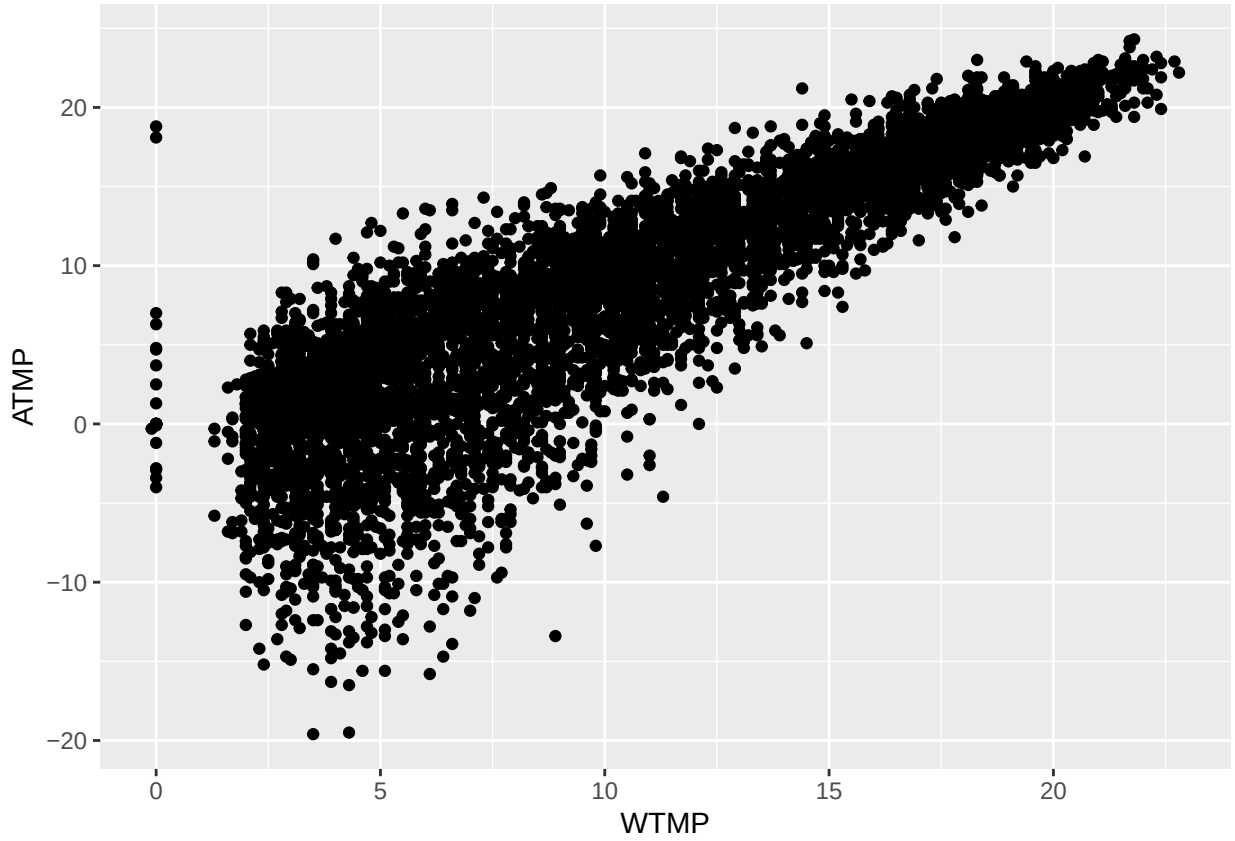


Plot

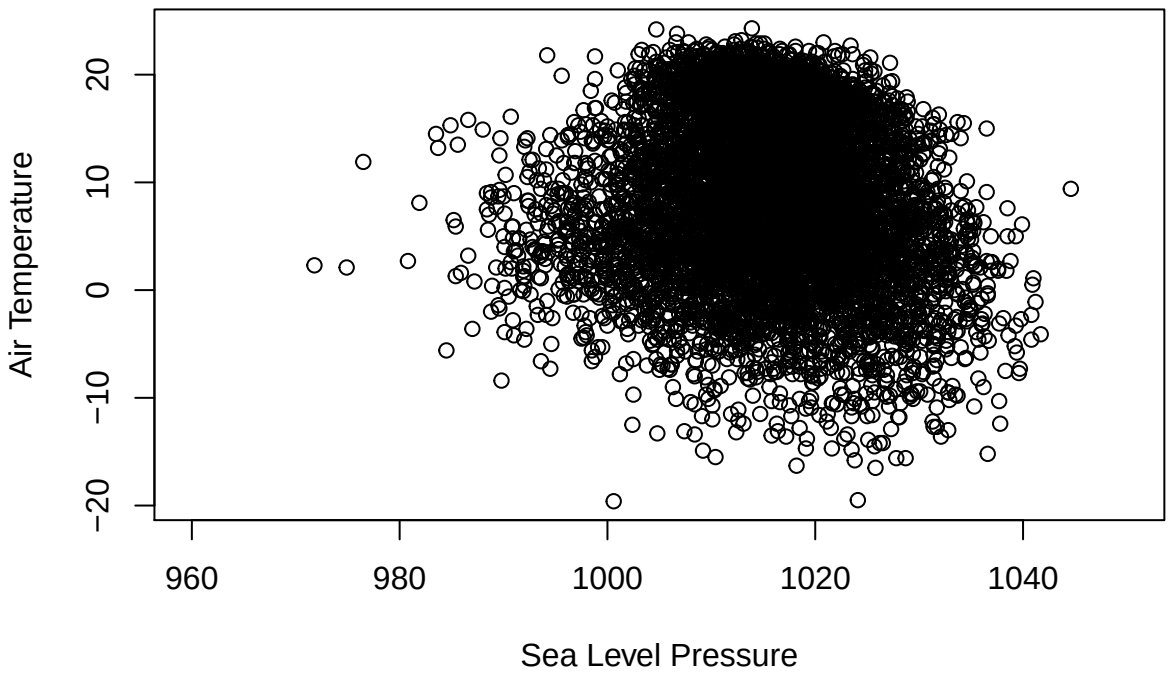
```
ggplot(data = Buoy2, aes(x = date, y = ATMP)) + geom_line() + xlab("")
```




```
ggplot(data = Buoy2, aes(x = WTMP, y = ATMP)) + geom_point()
```




```
plot(
  Buoy2$PRES,
  Buoy2$ATMP,
  xlim = c(960, 1050),
  xlab = "Sea Level Pressure",
  ylab = "Air Temperature"
)
```



Summary

According to the results of our models, the air temperature(ATMP) is heavily influenced by various variables such as wind direction, wind speed, guest speed, significant wave height, average wave period, the nature logarithm of sea level pressure, the sea surface temperature, the interaction between wind direction and wave direction, and the interaction between average wind speed and peak gust speed.

From the plot of "ATMP" and "date", we can clearly see the seasonal fluctuations from the waveform. Over time, each year's maximum temperature point tends to rise which indicates the trend of global warming. From the plot of air temperature and sea surface temperature, we can clearly see that there is a positive relationship between these two variables.

Considerations(curiosity)

There are several additional considerations to help to fully understand this study. Firstly, our data sources are very limited since all these data come from the same observation station(44013). To make our conclusion and model more reliable, we need to collect data from more stations in different locations such as in the west American or in Eurasia. Secondly, we don't have not enough time span on our data. We think 20 year is not enough to make an accurate conclusion or prediction on ATMP above sea level. If possible, we think 200 years will be a good choice for us to optimize our model.

Reference

https://www.ndbc.noaa.gov/station_page.php?station=44013