Boston Buoy Data Analysis

Hao Shen

2020/9/22

Summary

In this report, we found evidence showing global warming in the Boston buoy data set from 1987 to 2016, with total 246,245 observations. The whole analysis process is consisted by four parts:

- Read data from URLs of NOAA and combine into a single data.frame.
- Transform separated time variables into one and prepare for analysis.
- Perform data analysis with WTMP *(Water Temperature)*
- Perform data analysis with ATMP *(Air Temperature)*

As a result, from annual perspective, the analysis revealed that both the annual mean water temperature and air temperature shows slightly upward trends (0.005413℃/year, 0.003434℃/year respectively) as indications of global warming. Then, from monthly perspective, 8 out of 12 months shows annually increasing trends on average water (Jan, Feb, Jul, Nov.) and air temperature (Jan, Feb, Mar, Nov). Besides, the upward trends of temperature in summer and downward trends of temperature in winter indicate an increase in extreme weather which is also a indicator of global warming.

Data reading and combination

Read data from NOAA

Initially, we thought this would be the simplest process, since Professor Haviland had uploaded Web crawler script. However, due to different formats and different amount of variables, the code need improvements.

At first, we choose to separate the single `for` loop into two, one for downloading data from NOAA and one to deal with formats issues, which has two advantages. First, it is helpful to deal with network issues, especially when try to connect from China, so that no timeout warning will interrupt other process. Second, it is easier to compare the difference of data formats from year to year, which is necessary for further combination.

Nonetheless, using `read.table()` function for reading data from NOAA is not a good choice, because the NOAA data sets are not totally formal. Normally there is only 1 space between 2 columns, but in some cases 2 spaces will occur. The read.table() cannot handle such things well, since It will combine these columns separated by 2 spaces into 1 column as the outcome. Luckily, we can just use `read_table2()` function to solve this problem.

As a result, in this step, we get 30 tables with total 246,245 observations and 16 to 18 variables depending on different years.

```
url1="http://www.ndbc.noaa.gov/view_text_file.php?filename=mlrf1h"
url2=".txt.gz&dir=data/historical/stdmet/"
years=c(1987:2016)
urls=paste0(url1,years,url2)
Dnames=paste0('D',years)
#Using read_table2() for table reading
for(i in years){assign(Dnames[i-years[1]+1],read_table2(urls[i-years[1]+1]))}
```

Combine data as one

As the result of previous step shows, 30 data frames do not have same dimension, which need to be transferred before combination. Here 5 main operations are performed:

- From Y2000 to Y2016, delete an additional variable of 'TIDE'.
- From Y2005 to Y2016, delete an additional variable of 'mm'.
- From Y2007 to Y2016, delete first row of units.
- Check and unify col names and set data type as 'numeric'
- Create and combine to form final data set Buoy

After combination using `rbind.data.frame()` , we get the final data frame of Buoy with 246,245 observations of 16 variables.

Note: Here, because the network connection in China to NOAA is not stable, we only show the code of this chunk and the above here to ensure the knit process will not be interrupted. So, we need to save Buoy as Rdata in work directory and load it in next chunk for further code running.

```
coln=colnames(get(Dnames[1]))
for(i in years){
  D=get(Dnames[i-years[1]+1])
  # From Y2000 to Y2016, delete an additional variable of 'TIDE'
  if(i %in% 2000:2016){D=select(D,-TIDE)}
  # From Y2005 to Y2016, delete an additional variable of 'mm'
  if(i %in% 2005:2016){D=select(D,-mm)}
  # From Y2007 to Y2016, delete first row of units
  if(i %in% 2007:2016){D=D[-1,]}
  # Check and unify col names and set data type as 'numeric'
  if(ncol(D)==length(coln)){colnames(D)=coln}
  D=sapply(D, as.numeric)
  # From Y1987 to Y1999, transfer the Year from 'XX' to '19XX'
  D[,1][D[,1]<100]=D[,1][D[,1]<100]+1900
  # Create and combine to form final data set Buoy
  if(i==years[1]){Buoy=D}
  else{Buoy=rbind.data.frame(Buoy,D)}
}
save(Buoy,file='Buoy.Rdata')
```

Transform time and prepare for analysis

Transform time

After we load Buoy data from work directory, we use Kable to display the first five rows. Then, we add a variable called DT to Buoy with `make_datetime()` function.

```
load('Buoy.Rdata')
Buoy$DT=make_datetime(Buoy$YY,Buoy$MM,Buoy$DD,Buoy$hh)
Buoy=Buoy[,-c(1:4)]
kable(Buoy[1:5, ],caption="Buoy")
```

Buoy

WD	WSPD	GST	WVHT	DPD	APD	MWD	BAR	ATMP	WTMP	DEWP	VIS	DT
275	6.2	6.9	99	99	99	999	1018.6	20.0	25.8	999	99	1987-12-04 12:00:00
273	6.5	7.3	99	99	99	999	1019.0	20.2	25.8	999	99	1987-12-04 13:00:00
271	5.6	6.2	99	99	99	999	1019.3	20.5	25.8	999	99	1987-12-04 14:00:00
285	6.1	7.3	99	99	99	999	1019.1	20.8	25.8	999	99	1987-12-04 15:00:00
283	7.2	8.1	99	99	99	999	1018.8	21.1	25.7	999	99	1987-12-04 16:00:00

Remove NA and set for analysis

Now, before move to analysis, we need to remove NA values and set some parameters for analysis. To better understand the whole step, we need to mention a bit of analysis process.

In analysis process, we will use two variables WTMP *(Water Temperature)* and ATMP *(Water Temperature)* respectively. So, here we need to create Bouy_W and Buoy_A data frames without NA values (denoted by 99.0 or 999.0) for WTMP and analysis respectively. In this step, we only delete 1,476 and 2,298 data points from Buoy_W and Buoy_A which are less than 1%.

Then, because we need to mark month of analysis results for several times, it will be clear to set it at first.

Last, in analysis, we need preset Y_W and Y_A to store annual average WTMP and ATMP data. Also, we preset M_W and M_A to store monthly average WTMP and ATMP data.

```
#Remove Display data
Buoy_W=Buoy[Buoy$WTMP<99,]
Buoy_A=Buoy[Buoy$ATMP<99,]
#Set for analysis
month=c('Jan','Feb','Mar','Apr','May','Jun',
        'Jul','Aug','Sep','Oct','Nov','Dec')
Y_W=0
Y_A=0
M_W=0
M_A=0
```

Data analysis with WTMP

Trend of annually average water temperature

As a common sense, the most direct evidence of global warming is the temperature of the Earth has a upward trend. So, from all the variables listed above, only WTMP and ATMP are used. As for other wind or tide data their relations with global warming are not clear.

Also as a common sense, the temperature varies from day to night and from winter to summer. If we want to use all the data points for analysis, moving average to eliminate cycle fluctuation is required. However, considering we just need to find evidence or a sense of global warming, which only requires rough estimates with large uncertainty, annual mean is a good choice. It can not only eliminate cycle fluctuation but also easy to calculate and it can also tolerate 1% deletion of NA data points.

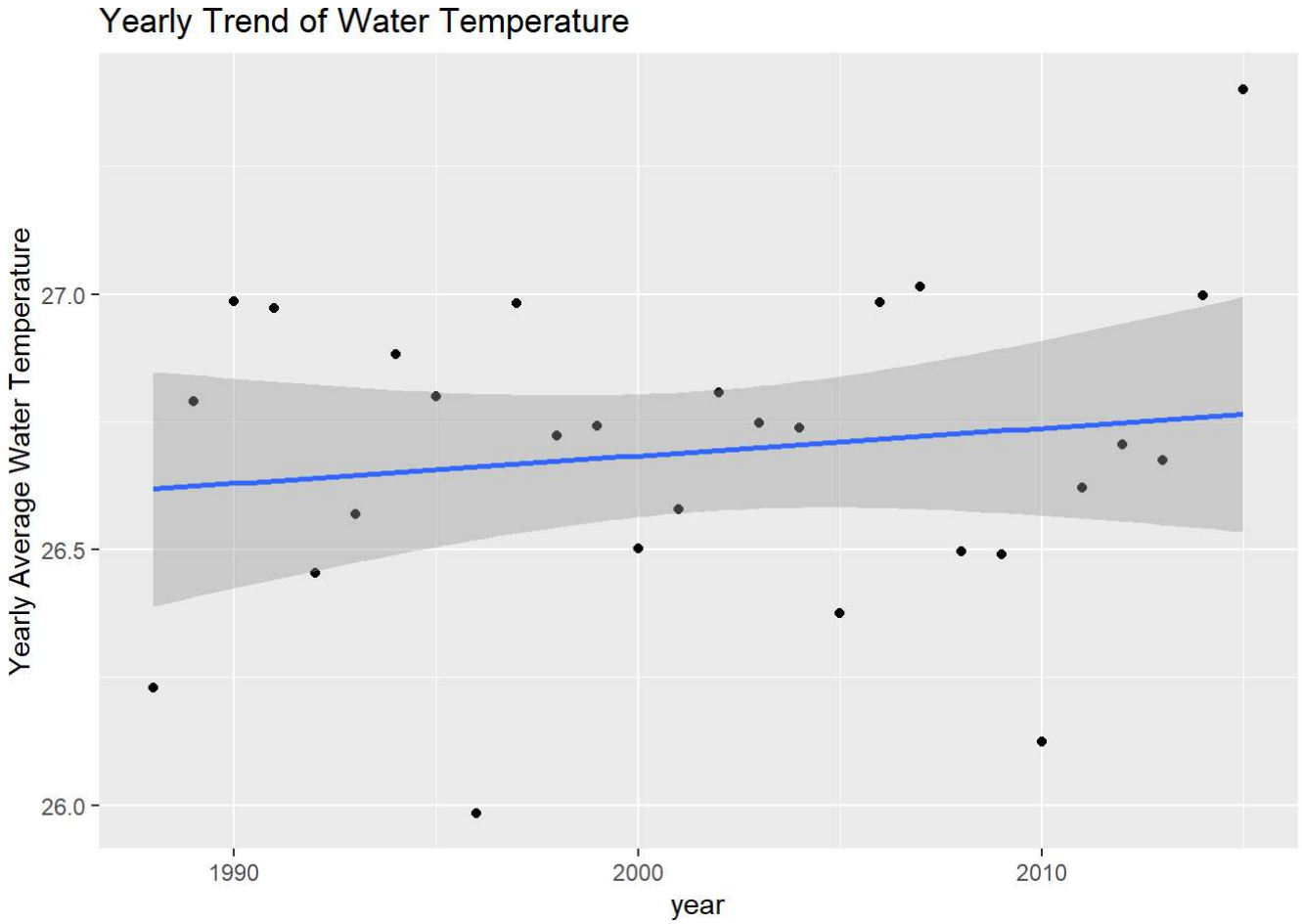
So, firstly we use `for` loop to compute annual mean temperature. Here, we use time data type for data selection and `make_date()` function to transfer loop indicator into different time points that can separate the whole 30 years data into annually. You may notice here we only choose the data from 1988 to 2015 which only

has 28 years. This is because the data of 1988 only contains data collected in winter and data of 2016 only has about 75% data points compared to other years, which are both incomplete and need to be removed.

```
#Loop for annual mean temperature calculation
for(i in 1988:2015){
  Y_W[i-1987]=mean(Buoy_W$WTMP[year(Buoy_W$DT)==i])
}
#Do regression and present results
D_W=data.frame(Time=1988:2015,TMP=Y_W)
R_W=lm(TMP~Time,data=D_W)
P_W=ggplot(D_W,aes(Time,TMP))+
  geom_point()+
  geom_smooth(method="lm",formula=y~x)+
  labs(title='Yearly Trend of Water Temperature',
       x="year", y='Yearly Average Water Temperature')
print(R_W)
```

```
##
## Call:
## lm(formula = TMP ~ Time, data = D_W)
##
## Coefficients:
## (Intercept)         Time
##   15.856940     0.005413
```

```
P_W
```



As the simple linear regression model shows the slope of fitted line is 0.005413 which represents a slightly upward trend of buoy water temperature indicating global warming and can be interpreted as buoy water temperature will increase by 0.005413℃ per year. The picture also shows the same result.

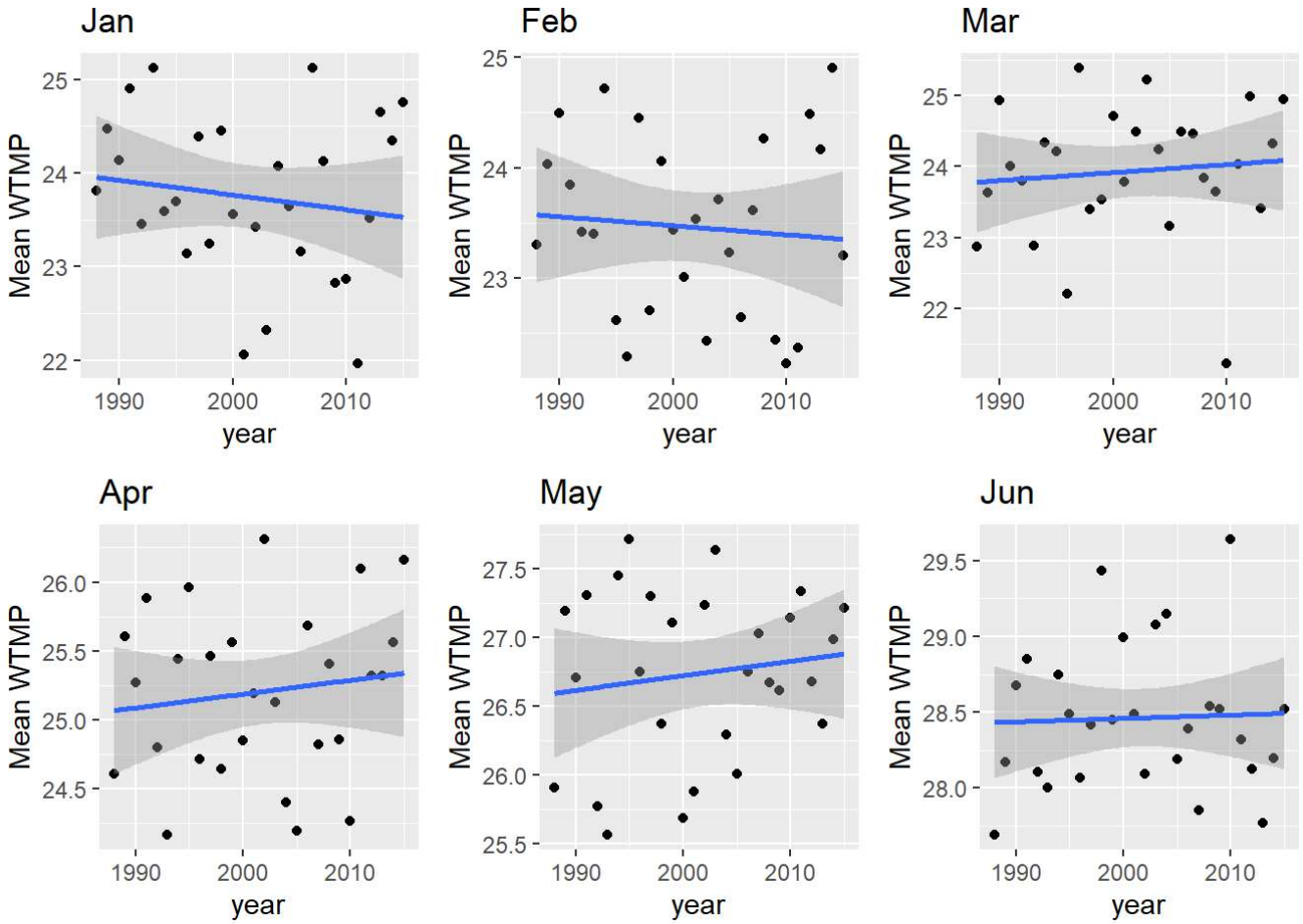
Trends of monthly average water temperature

Now we found some evidence, but the slope or annual increase of 0.005413℃ is so samll that is hard to convince us. We need to separate this slope so that may be more information will be reveal. And we choose to further analyze the data from a month perspective.

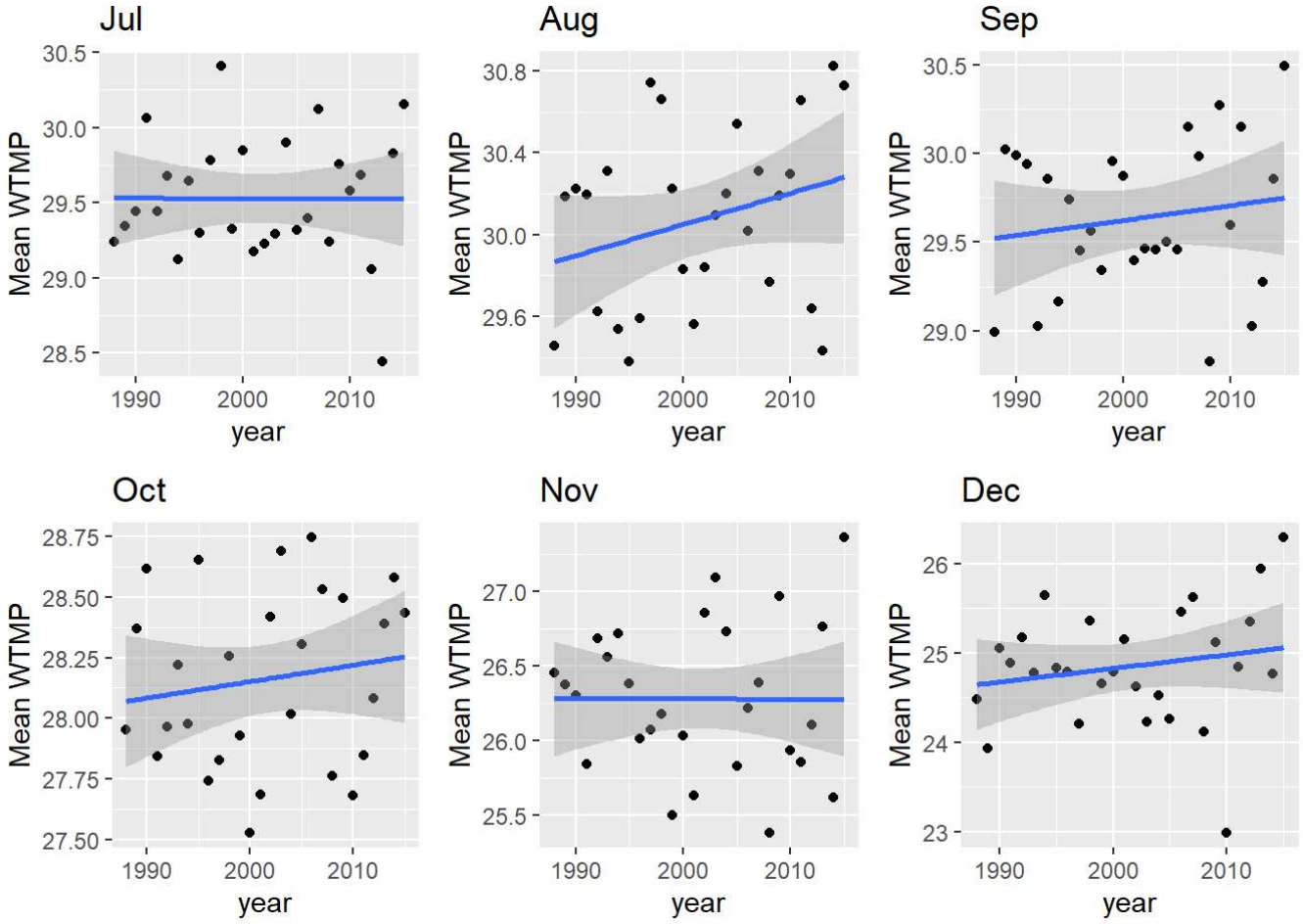
Here we consider what's the trend of temperature of each month changed during the past 30 years. For example, for June, we want to explore how the June temperature of each year changes from 1987 to 2015. So we use another `for` loop to repeat computation and use loop indicators of j, i to represent the time point of month, year respectively for data selection. Then we use `ggplot2()` to present results with month's labels and `ggarrange()` to layout.

Note: Here for convinence, we just put an outer loop on the original one which is enough for this analysis but it has a disadvantage that monthly mean temperature data expect December would be deleted and only their plots left.

```
P_WM_name=str_c('P_W',1:12,sep='')
#Outer loop for 12 months calculation and 12 plots
for(j in 1:12){
  #Inner loop for annual mean temperature of each month calculation
  for(i in 1988:2015){
    M_W[i-1987]=mean(Buoy_W$WTMP[year(Buoy_W$DT)==i&month(Buoy_W$DT)==j])
  }
  ##Do regression and present results of each month
  D_W=data.frame(Time=1988:2015,TMP=M_W)
  assign(P_WM_name[j],
         ggplot(D_W,aes(Time,TMP))+
         geom_point()+
         geom_smooth(method="lm",formula=y~x)+
         labs(title=month[j],x="year",y="Mean WTMP"))
}
#Arrange 12 plots
ggarrange(P_W1,P_W2,P_W3,P_W4,P_W5,P_W6,ncol=3,nrow=2)
```




```
ggarrange(P_W7,P_W8,P_W9,P_W10,P_W11,P_W12,ncol=3,nrow=2)
```



As the result shows, except January, February, July and November, other 8 months all display upward trends in water temperature as an indicator of global warming. More interesting is the plots show an downward trend in winter time and upward trend in summer time which means the range of annual temperature is becoming large which means more extreme weather. And extreme weather is also a phenomenon caused by global warming.

Data analysis with ATMP

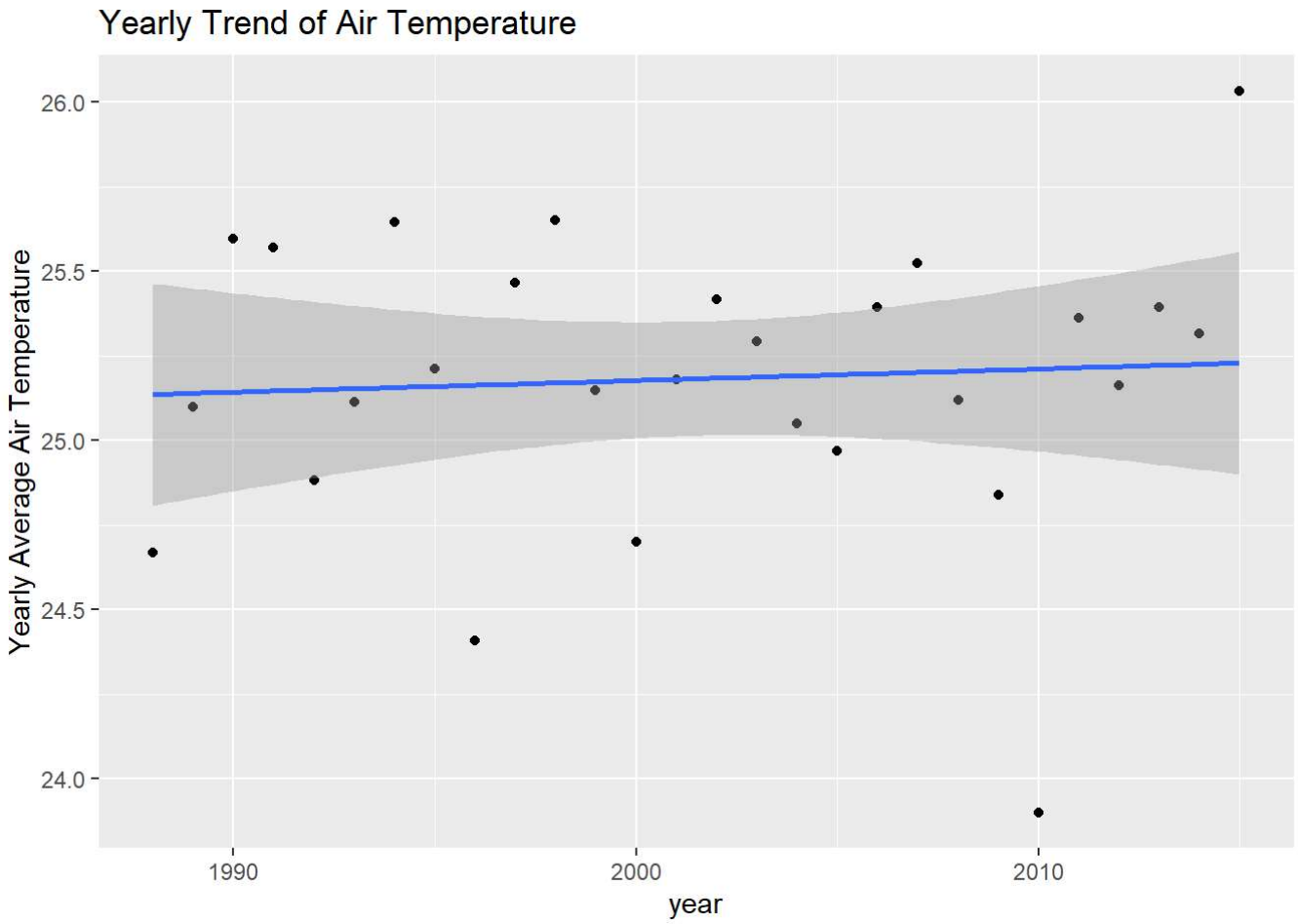
Trend of annually average air temperature

Actually, in the above part we have finished all the analysis and formed a conclusion there is evidence showing global warming. However, since there is also a variable about temperature which is about air and since we want to further confirm our conclusion, we also perform the same analysis and presentation on ATMP as we have done on WTMP.

```
#Loop for annual mean temperature calculation
for(i in 1988:2015){
  Y_A[i-1987]=mean(Buoy_A$ATMP[year(Buoy_A$DT)==i])
}
#Do regression and present results
D_A=data.frame(Time=1988:2015,TMP=Y_A)
R_A=lm(TMP~Time,data=D_A)
P_A=ggplot(D_A,aes(Time,TMP))+
  geom_point()+
  geom_smooth(method="lm",formula=y~x)+
  labs(title='Yearly Trend of Air Temperature',
       x="year", y='Yearly Average Air Temperature')
print(R_A)
```

```
##
## Call:
## lm(formula = TMP ~ Time, data = D_A)
##
## Coefficients:
## (Intercept)         Time
##   18.309780     0.003434
```

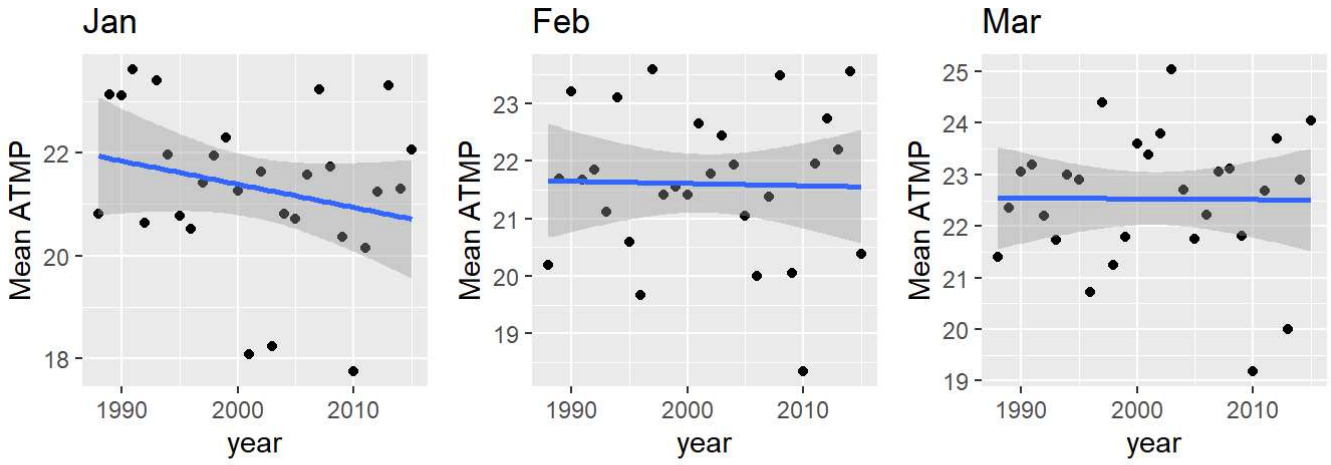
```
P_A
```

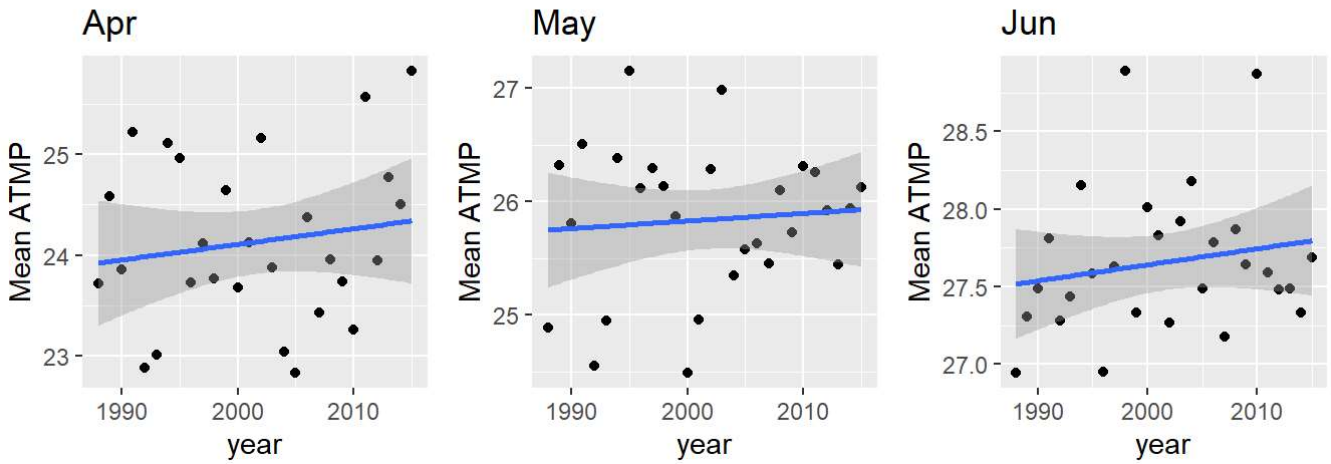


As the simple linear regression model shows the slope of fitted line is 0.003434 which also represents a slightly upward trend of buoy air temperature indicating global warming and can also be interpreted as buoy air temperature will increase by 0.003434℃ per year. The picture no doubly shows the same result.

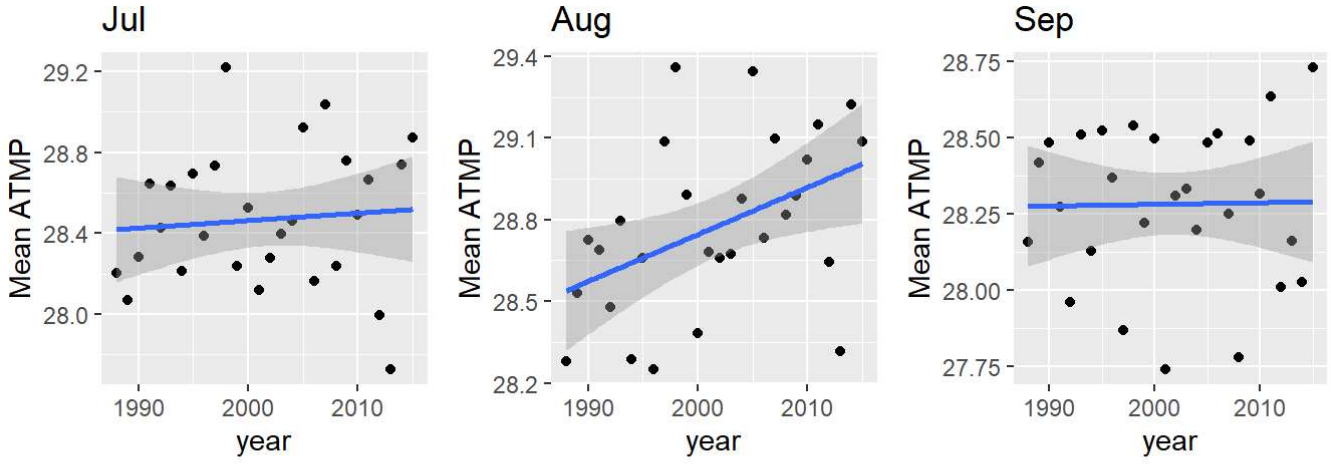
Trends of monthly average air temperature

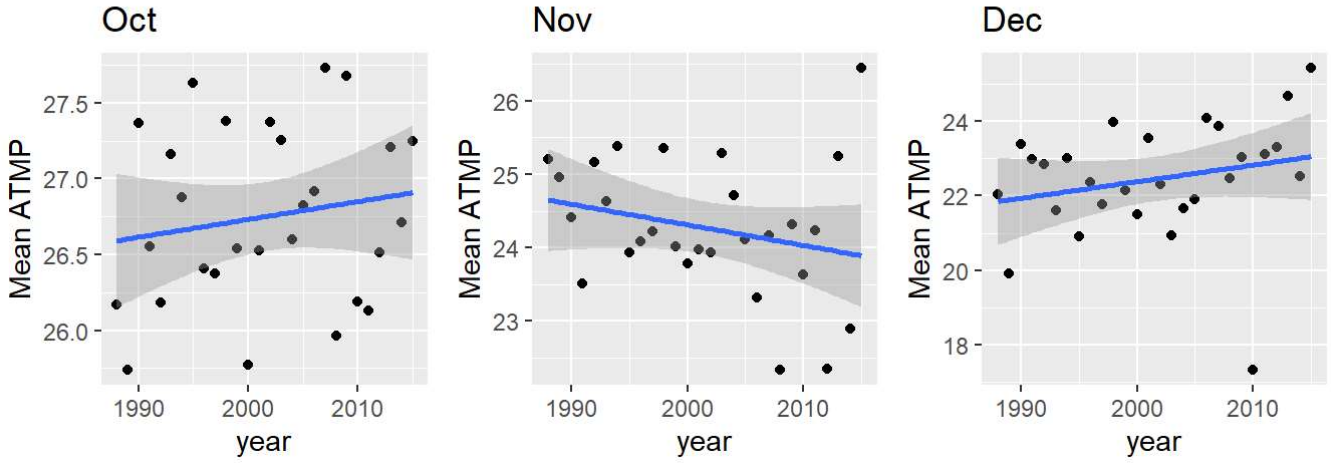
```
P_AM_name=str_c('P_A',1:12,sep='')
#Outer loop for 12 months calculation and 12 plots
for(j in 1:12){
  #Inner loop for annual mean temperature of each month calculation
  for(i in 1988:2015){
    M_A[i-1987]=mean(Buoy_A$ATMP[year(Buoy_A$DT)==i&month(Buoy_A$DT)==j])
  }
  ##Do regression and present results of each month
  D_A=data.frame(Time=1988:2015,TMP=M_A)
  assign(P_AM_name[j],
         ggplot(D_A,aes(Time,TMP))+
         geom_point()+
         geom_smooth(method="lm",formula=y~x)+
         labs(title=month[j],x="year",y="Mean ATMP"))
}
#Arrange 12 plots
ggarrange(P_A1,P_A2,P_A3,P_A4,P_A5,P_A6,ncol=3,nrow=2)
```






```
ggarrange(P_A7, P_A8, P_A9, P_A10, P_A11, P_A12, ncol=3, nrow=2)
```





Also, as the result shows, except January, February, March and November, other 8 months all display upward trends in air temperature as an indicator of global warming. The plots also show an downward trend in winter time and upward trend in summer time which means the range of annual temperature is becoming large which means more extreme weather, a phenomenon caused by global warming.

In conclusion, there are three evidence both from water temperature and air temperature showing the existence of global warming:

- Slightly positive slopes indicate temperature increased each year slightly.
- Most months in the past 28 years showed increases in temperature.
- Upward and downward trends of temperature in summer and winter indicate more extreme weather, a phenomenon caused by global warming.