Buoy-Project

Ruxin Liu

9/24/2020

Introduction

As people are now more and more concerned about our ecosystem, we can easily acquire information about global warming all over the internet. However, even there is clear scientific evidence about this current climate change, a large proportion of the general public may still not believe in this or not realize the severity of this issue (NASA 2020). In this study, the data collected by the National Data Buoy Center (NOAA) from the year 1999 to the year 2018 was explored and analyzed to show evidence of global warming (NDBC 2000).

Methodology

Data cleaning and variable selection

The general approach of this study is to use the 20-year buoy data to perform linear regression models. The reason why the year 1999 to the year 2018 was selected is that the last decade was the warmest since the year 1880 and these were the latest available data on the NDBC website (Climate Hot Map). Since the buoy data was collected over years, there were updates and changes on how many variables are measured. During the exploring process, it is found that the data of the year 1999 only contained 16 columns, from the year 2000 to 2004 there were 17 columns while starting from the year 2005, the data included one more column mm, which was the minute variable. In order to keep consistency throughout different years, the new variables that were added in later years are deleted, since these data were not complete and should not be used in the analysis.

After merging the 20 data sets into a single data frame, only variables YYYY (year), MM (month), DD (day), ATMP (Air Temperature) and WTMP (Water Temperature) are kept, since these are the factors that can show direct evidence of global warming. There were temperature variables in the data with values of 999, which is impossible for a temperature to be. These were assumed to be missing values and were converted to NA first and then were deleted from the data to avoid confusion and miscalculation. All the date-related variables were formatted into POSIX numbers with lubridate package (Hadley Wickham & Garrett Grolemund 2016).

Sampling frequency

In order to reduce the number of total observations and at the same time keeping all the important information, the sampling frequency was filtered down to one data point per day, which was the maximum temperature of the day. The reason why using maximum temperature instead of the average temperature was that average values could possibly hide some outliers in the data while the maximum value could capture them. Global warming itself is an unusual speeding event on the temperature, so it is better to look for extreme values in the data. Then, using the daily maximum temperature, the yearly average of maximum temperature was computed and plotted to see any potential general trend for the purpose of Exploratory Data Analysis. Lastly, since the temperatures change a lot in different seasons. The maximum temperature data were further separated into seasons and linear regressions were performed based on the seasonal sub-datasets. The data was collected in Boston, and therefore spring is from March, April and May; summer is from June, July and August; fall is from September, October and November; and the winter is from December, January

and February (Kathryn Cirrone 2019). The year variables used in linear regressions were all centered by subtracting the value of 1999, therefore the intercepts can then be interpreted as the average temperature for the first year of the data (year 1999). All the detailed codes and references can be found in the Appendix.

Results and Discussion

From Fig.1 shown in the Appendix, we can see that the average maximum temperatures fluctuated up and down over the years. However, in the year 2005, the average annual max temperature peaked and after that, the temperature kept staying at a relatively high level. This plot suggested that there may be an increasing trend for the temperature, so linear regressions are fitted to confirm with this conjecture. From Table 1 & 2, the slope coefficients for both air temperature (slope = 0.02090) and water temperature (slope = 0.02090) were positive. However, they are not statistically significant. For the spring, from Table 3 & 4 the slope coefficients for air temperature (slope = 0.015552) and water temperature (slope = 0.012890) were both positive and significant. For the summer, from Table 5 the slope was positive and significant, which was 0.008529, and from Table 6 the slope was negative but not significant, which was -0.005415. For the fall, from Table 7 the slope was positive and significant, which was 0.018076, and from Table 8 the slope was positive but not significant, which was 0.008436. Lastly for the winter, from Table 9 & 10 the slope coefficients for air temperature (slope = 0.049556) and water temperature (slope = 0.040290) were both positive and very significant. Since the intercept of the regression models all corresponded to the average maximum temperature in the first year of this data, which didn't provide much useful information. However, the slope coefficients were very useful.The positive slopes suggested that with 1 unit of increase in years, the annual maximum temperature on average will increase by the amount of the coefficient. Out of the 10 linear models, there was only one producing a negative slope. By looking at linear regressions only, the results showed that the temperature will increase with the increase of years. Although the values of slopes seem to be very small, it actually meant a lot when putting back to the content. For example, the slope coefficient for maximum air temperature in spring is only 0.015552. This means that after 100 years, the maximum temperature in spring is predicted to increase by 1.5 degree celsius. This number is scary because all living organisms on the earth are very sensitive to temperature and 100-years is a very short period for the ecosystem. Therefore, this amount of increase can be detrimental.

Conclusion

In conclusion, all linear regression models showed a positive slope coefficient, except for one. And many of the coefficients also showed significance, which made the results more reliable. These results concluded that both air and water maximum temperature will increase through time. Although there could be better models than simple linear regression and there might be many assumptions and insights into the data that were not taking account in this study, it could suggest that there is evidence of global warming in the data only collected by a single weather buoy at a single location. Therefore, even given the inevitable issues of this data, the results still showed pretty good evidence and confirmed that global warming is happening. For future studies, a large dataset and a longer time span could be used to fit a more accurate and convincing model. Also, the time series could be performed to forecast future temperatures. Global warming is actually happening and everyone is responsible for trying to alleviate it.

Appendix

Code

```
# Loading required packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
```

```
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
```

Importing the data

```
# The first part of the data URL
url_1 <- "http://www.ndbc.noaa.gov/view_text_file.php?filename=mlrf1h"
# The second part of the data URL
url_2 <- ".txt.gz&dir=data/historical/stdmet/"
# Read the latest available 20 years of data
years <- c(1999:2018)
# The complete 20 URLs
urls <- str_c(url_1, years, url_2, sep = "")
filenames <- str_c("mr", years, sep = "")
```

```
# Load in the 20-year data from the NDBC website
# Year 1999 - 2006
for(i in 1:8){
  suppressMessages(
    # Fill any missing values with NA:
    assign(filenames[i], read.table(urls[i], header = TRUE, fill = TRUE))
  )

}
# Year 2007 - 2018: Column names started with #, which will cause issue
for(i in 9:20){
  suppressMessages(
    # Fill any missing values with NA and use the same column names as year 2006
    assign(filenames[i], read.table(urls[i], header = FALSE,
                                    fill = TRUE, col.names = colnames(mr2006))),
  )

}
```

Since the buoy data was collected over years, there were updates and changes on how many variables are measured. During the exploring process, it is found that the data from the year 1999 only contained 16 columns(without Tide), the year 2000 to 2004 contained 17 columns, while starting from the year 2005, the data included one extra column mm, which is the minute variable. Since the mm variable only contained values of 0 and the earlier years didn't have this information available, the mm variable will not be used in the analysis and can be removed from the data.

```
# Add Column Tide to the 1999 data
mr1999$TIDE <- NA
# Delete the mm column
n <- length(urls)
for (i in 1:n){
```

```
  file <- get(filenames[i])
  colnames(file)[1] <-"YYYY"
  if(ncol(file) == 18){
    file <- subset(file, select = -mm )
  }
  if(i == 1){
    MR <- file
  }else{
    MR <- rbind.data.frame(MR, file)
  }

}
```

Cleaning the data

The goal of this project is to find possible evidence of global warming, therefore, the possibly related variables are Air Temperature (ATMP) and Water Temperature (WTMP). So, the other variables could be removed for now.

```
MR_temp <- subset(MR, select = c(YYYY, MM, DD, ATMP, WTMP))
```

For the temperature variable, there are many missing values indicated by 999.0 (temperature could not go up to 999), which will cause issues when calculating averages and findind maximun values. Therefore, these points should be substituted by NA.

```
MR_temp[MR_temp == 999] <- NA
# Delete the rows with NA
MR_temp <- na.omit(MR_temp)
```

Formating the data into posix numbers with lubridate

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
# Format the year, month and time variables into one single date variable
MR_temp <- MR_temp %>%
  select(YYYY, MM, DD, ATMP, WTMP) %>%
  mutate(date = make_datetime(YYYY, MM, DD))
# The new variable is POSIX number
is.POSIXt(MR_temp$date)
```

```
## [1] TRUE
```

Filtering the data down to one data point per day (the maximum temperature)

```
library(dplyr)
MR_max <- MR_temp %>%
```

```
  select(YYYY, MM, date, ATMP, WTMP) %>%
  group_by(date) %>%
  # Finding the max and exclude NA
  summarize(max_ATMP = max(ATMP),
            max_WTMP = max(WTMP))
```

```
MR_year_mean <- MR_max %>%
  group_by(year(date)) %>%
  # Finding the yearly average max temperature and exclude NA
  summarize(mean_ATMP = mean(max_ATMP),
            mean_WTMP = mean(max_WTMP))
```

```
MR_spring <- MR_max %>%
  filter(month(date) %in% c(3, 4, 5))
```

```
MR_summer <- MR_max %>%
  filter(month(date) %in% c(6, 7, 8))
```

```
MR_fall <- MR_max %>%
  filter(month(date) %in% c(9, 10, 11))
```

```
MR_winter <- MR_max %>%
  filter(month(date) %in% c(12, 1, 2))
```

EDA

```
library(ggplot2)
ggplot(MR_year_mean, aes(`year(date)`, mean_ATMP)) +
  geom_point(colour = "orange") +
  geom_point(data = MR_year_mean, aes(y = mean_WTMP), colour = 'purple') +
  ggtitle("Yearly Mean Temperature Trend") +
  labs(y ="Mean Temperature (C)", x = "Year",
       caption = "Fig.1 The yearly changing trend for maximum temperature")
```

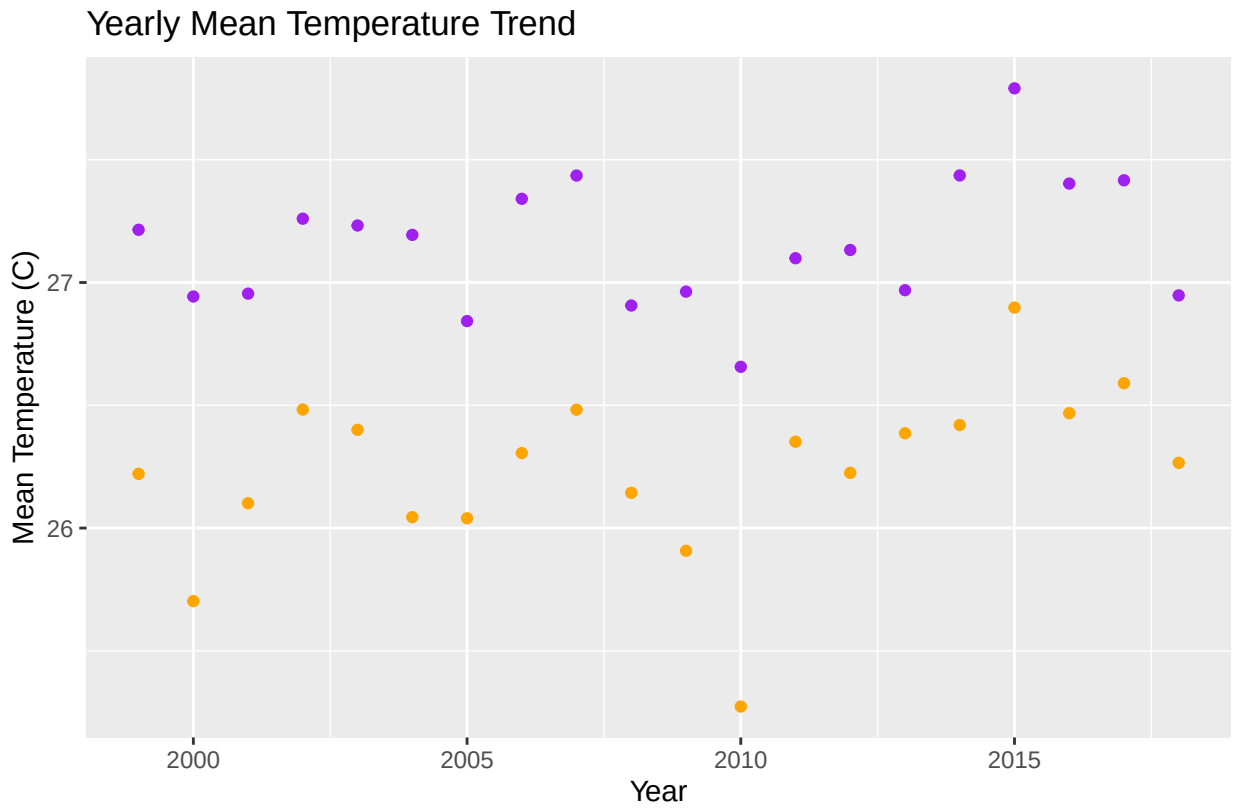

Fig.1 The yearly changing trend for maximum temperature

Analysis

Table 1

```
# Linear regression Year vs ATMP
center_year <- MR_year_mean$`year(date)` - 1999
summary(lm(MR_year_mean$mean_ATMP ~ center_year))
```

```
##
## Call:
## lm(formula = MR_year_mean$mean_ATMP ~ center_year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99358 -0.10322  0.06038  0.17875  0.52687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.03687    0.14267 182.498   <2e-16 ***
## center_year  0.02090    0.01284   1.628    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3311 on 18 degrees of freedom
## Multiple R-squared:  0.1283, Adjusted R-squared:  0.07988
## F-statistic: 2.649 on 1 and 18 DF,  p-value: 0.121
```

Table 2

```
# Linear regression Year vs WTMP
summary(lm(MR_year_mean$mean_WTMP ~ center_year))
```

```
##
## Call:
## lm(formula = MR_year_mean$mean_WTMP ~ center_year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51754 -0.20988  0.01208  0.17014  0.55901
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.04730    0.11451  236.20   <2e-16 ***
## center_year  0.01154    0.01030    1.12    0.277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2657 on 18 degrees of freedom
## Multiple R-squared:  0.06515,    Adjusted R-squared:  0.01322
## F-statistic: 1.254 on 1 and 18 DF,  p-value: 0.2774
```

Table 3

```
# Linear regression Year vs spring ATMP
year_spring <- year(MR_spring$date) - 1999
summary(lm(max_ATMP ~ year_spring, data = MR_spring))
```

```
##
## Call:
## lm(formula = max_ATMP ~ year_spring, data = MR_spring)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0373 -1.0012  0.1694  1.2904  4.9538
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.066245   0.085645 292.676   <2e-16 ***
## year_spring  0.015552   0.007746   2.008   0.0448 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.836 on 1762 degrees of freedom
## Multiple R-squared:  0.002283,   Adjusted R-squared:  0.001717
## F-statistic: 4.032 on 1 and 1762 DF,  p-value: 0.04481
```

Table 4

```
# Linear regression Year vs spring WTMP
summary(lm(max_WTMP ~ year_spring, data = MR_spring))
```

```
##
```

```
## Call:
## lm(formula = max_WTMP ~ year_spring, data = MR_spring)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6979 -0.9205  0.0264  1.0280  3.3764
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.756067   0.067911 379.263   <2e-16 ***
## year_spring  0.012890   0.006142   2.099    0.036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.456 on 1762 degrees of freedom
## Multiple R-squared:  0.002494,   Adjusted R-squared:  0.001928
## F-statistic: 4.405 on 1 and 1762 DF,  p-value: 0.03598
```

Table 5

```
year_summer <- year(MR_summer$date) - 1999
# Linear regression Year vs summer ATMP
summary(lm(max_ATMP ~ year_summer, data = MR_summer))
```

```
##
## Call:
## lm(formula = max_ATMP ~ year_summer, data = MR_summer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2693 -0.4546  0.0880  0.5307  3.1477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.001095   0.036507 794.408   <2e-16 ***
## year_summer  0.008529   0.003337   2.556   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7865 on 1762 degrees of freedom
## Multiple R-squared:  0.003693,   Adjusted R-squared:  0.003127
## F-statistic: 6.531 on 1 and 1762 DF,  p-value: 0.01068
```

Table 6

```
# Linear regression Year vs summer WTMP
summary(lm(max_WTMP ~ year_summer, data = MR_summer))
```

```
##
## Call:
## lm(formula = max_WTMP ~ year_summer, data = MR_summer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.2667 -0.6127  0.1171  0.6791  2.3008
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.842510   0.043251 689.984   <2e-16 ***
## year_summer -0.005415   0.003954  -1.369    0.171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9319 on 1762 degrees of freedom
## Multiple R-squared:  0.001063,   Adjusted R-squared:  0.0004963
## F-statistic: 1.875 on 1 and 1762 DF,  p-value: 0.171
```

Table 7

```
year_fall<- year(MR_fall$date) - 1999
# Linear regression Year vs fall ATMP
summary(lm(max_ATMP ~ year_fall, data = MR_fall))
```

```
##
## Call:
## lm(formula = max_ATMP ~ year_fall, data = MR_fall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1165  -1.3135   0.5196   1.5356   3.7100
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.189997   0.089056 305.312   <2e-16 ***
## year_fall    0.018076   0.008006   2.258   0.0241 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.957 on 1764 degrees of freedom
## Multiple R-squared:  0.002881,   Adjusted R-squared:  0.002316
## F-statistic: 5.097 on 1 and 1764 DF,  p-value: 0.02408
```

Table 8

```
# Linear regression Year vs fall WTMP
summary(lm(max_WTMP ~ year_fall, data = MR_fall))
```

```
##
## Call:
## lm(formula = max_WTMP ~ year_fall, data = MR_fall)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.272 -1.306  0.176  1.277  3.493
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.372163   0.069939 405.667   <2e-16 ***
```

```
## year_fall    0.008436   0.006288   1.342     0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.537 on 1764 degrees of freedom
## Multiple R-squared:  0.00102,    Adjusted R-squared:  0.0004532
## F-statistic:   1.8 on 1 and 1764 DF,  p-value: 0.1798
```

Table 9

```
year_winter <- year(MR_winter$date) - 1999
# Linear regression Year vs winter ATMP
summary(lm(max_ATMP ~ year_winter, data = MR_winter))
```

```
##
## Call:
## lm(formula = max_ATMP ~ year_winter, data = MR_winter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7866  -1.0388   0.5665   1.5625   3.7656
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.841484   0.106014 215.457  < 2e-16 ***
## year_winter  0.049556   0.009566   5.181 2.46e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.302 on 1769 degrees of freedom
## Multiple R-squared:  0.01494,    Adjusted R-squared:  0.01439
## F-statistic: 26.84 on 1 and 1769 DF,  p-value: 2.465e-07
```

Table 10

```
# Linear regression Year vs winter WTMP
summary(lm(max_WTMP ~ year_winter, data = MR_winter))
```

```
##
## Call:
## lm(formula = max_WTMP ~ year_winter, data = MR_winter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -0.7319  0.0523  0.8702  2.6076
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.147735   0.052323 461.511   <2e-16 ***
## year_winter  0.040290   0.004721   8.534   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.136 on 1769 degrees of freedom
```

```
## Multiple R-squared:  0.03954,    Adjusted R-squared:  0.039
## F-statistic: 72.83 on 1 and 1769 DF,  p-value: < 2.2e-16
```

Reference of R Packages

1. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
2. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
3. Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr
4. Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.
5. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Bibliography

1. NASA. (2020). *GLOBAL CLIMATE CHANGE* [online]. Available from: https://climate.nasa.gov/evidence/ [accessed 25 September 2020].
2. NDBC. (2000). *Station 44013* [online]. Available from: https://www.ndbc.noaa.gov/station_page.php?station=44013 [accessed 25 September 2020].
3. Climate Hot Map. *GLOBAL WARMING EFFECTS AROUND THE WORLD* Available from: https://www.climatehotmap.org/global-warming-effects/air-temperature.html [accessed 25 September 2020].
4. Kathryn Cirrone. (2019). *Trip Savvy* [online]. Available from: https://www.tripsavvy.com/weather-in-boston-climate-seasons-average-monthly-temperature-4177628 [accessed 25 September 2020].
5. Wickham,H. and Grolemund,G. (2016). *R for Data Science*. 1st ed. O'Reilly Media.