Analysis of Climate Factors in Natural Disaster Patterns

Samuel Van Gorden

3/13/2022

Introduction

Recent upward trends in natural disasters have caused incalculable loss of property, economic well-being, and life to millions all across the globe, especially in poor and underdeveloped nations. From fires in California and tornadoes in Kentucky, to flash floods in India and Nepal and a devastating earthquake in Haiti, natural disasters have become one of the greatest threats to a peaceful existence on earth[4]. It is natural to question the role of climate change and its precursors such as rising temperatures and emissions in the preponderance of such maleficent events. An understanding of the causes of the recent influx of natural disasters as well as potential solutions necessitate an investigation into recent trends in climate data.

The motivation for this report is to provide a better understanding of the link between drivers and correlating factors of climate change and natural disaster patterns. We will observe data from 1900 to the present in order to detect trends and predict future outcomes. Descriptive and inferential analyses will be performed on various types and quantities of natural disasters over the factors of time, global average land and ocean temperatures, and CO_2 and methane emissions. In performing these analyses we hope to discover any correlation between natural disasters and high vs. low temperatures and emission levels and the establish a trend that will help us know what to expect in the future. The scope of this paper is to assess factors that cause or are correlated with the characteristics of natural disasters. Future studies should make use of the information here to establish best solutions to the problem of increasing natural disaster occurrences.

Description of Data

This report utilizes three datasets. The first contains information about natural disasters occurring globally from 1900 to 2021. We utilize the following variables:

- Disaster.Type - Categorical (15 levels) - Divides types of natural disasters into 15 categories
- Year - Continuous - Year in which each natural disaster occurred

The second dataset contains land and ocean average temperatures from 1899 to 2015. We utilize the following variables:

- LandAverageTemperature - Continuous - Average global temperature as measured on land
- dt - Continuous - Date at which each temperature was measured (monthly granularity)

We average the temperature data over 12-month periods to get data at yearly granularity to match our other datasets.

The third dataset contains emissions data for each year from 1750 to 2020. We utilize the following variables:

- country - Categorical (244 levels) - Country in which emissions were measured
- year - Continuous - Year in which emissions were measured
- co2 - Continuous - CO_2 level measured in million tonnes
- methane - Continuous - Methane level measured in million tonnes of CO_2-equivalents

Methods

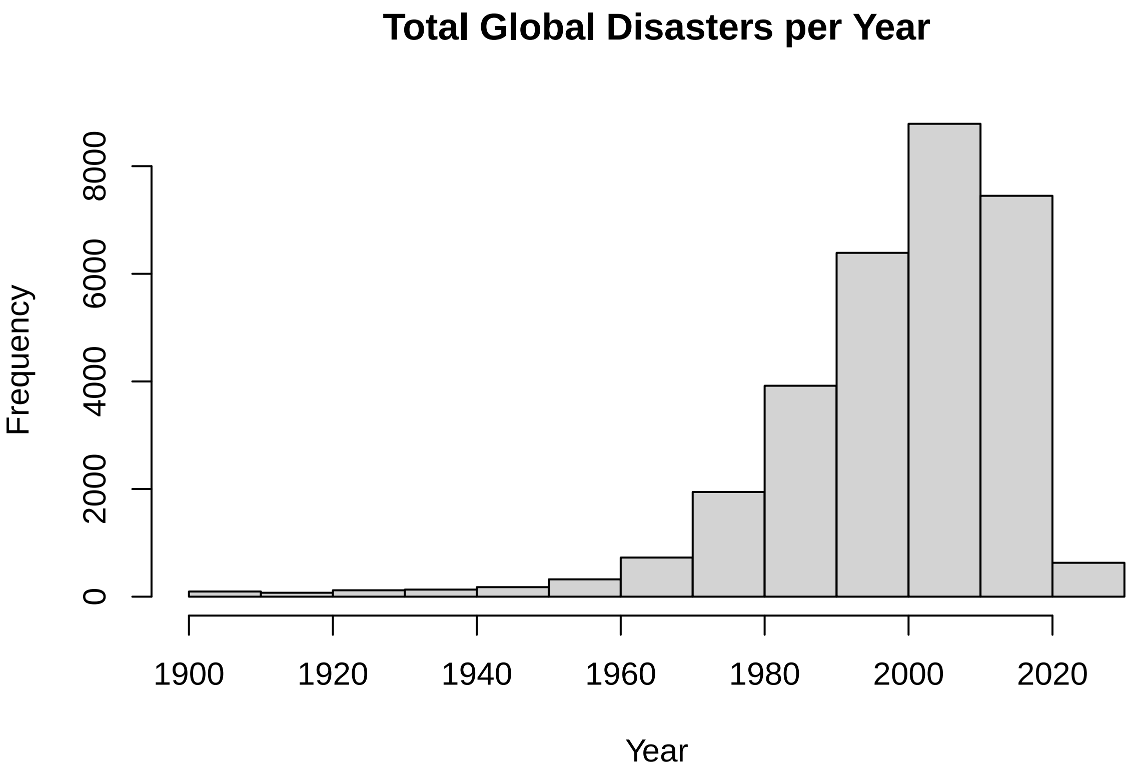
Various statistical methods were used to present our data in an informative way, as well as perform analysis on historical data and make future predictions.

Graphical representations of our data, such as histograms, scatterplots, and curvilinear regression plots are used throughout the report. We used linear regression modeling to fit a relationship between average land temperatures and quantity of natural disasters. Time series analyses were performed to observe trends in temperature over time and make predictions about the future. We performed ANOVA to determine whether or not different types of natural disasters occurred at different temperatures or different levels of emissions. Finally, we performed a QDA analysis to see if we could accurately predict which type of disasters would be most likely to occur under which temperature/emissions conditions.

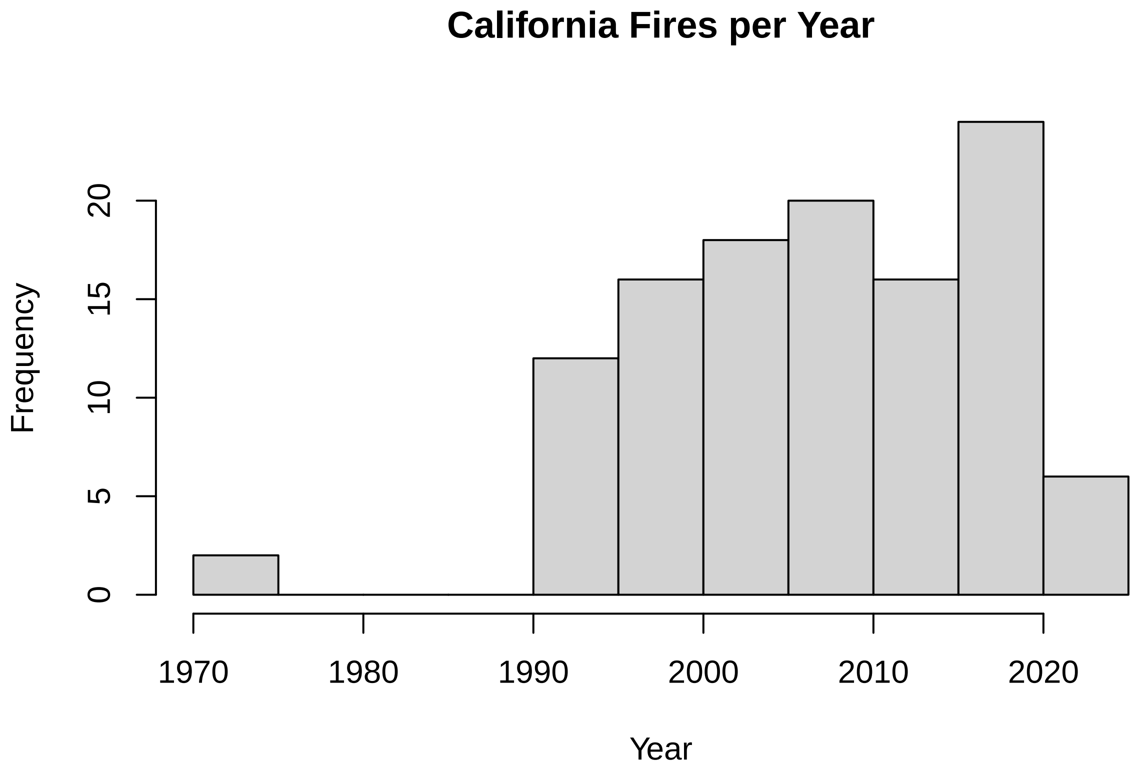
All of the coding for the project was done in R using RStudio. Plots were made using Base R, ggplot2, and plotly. We used several packages including dplyr for data manipulations, forecast for time series analysis, and MASS and caret for supervised classification modeling.

Results and Analysis

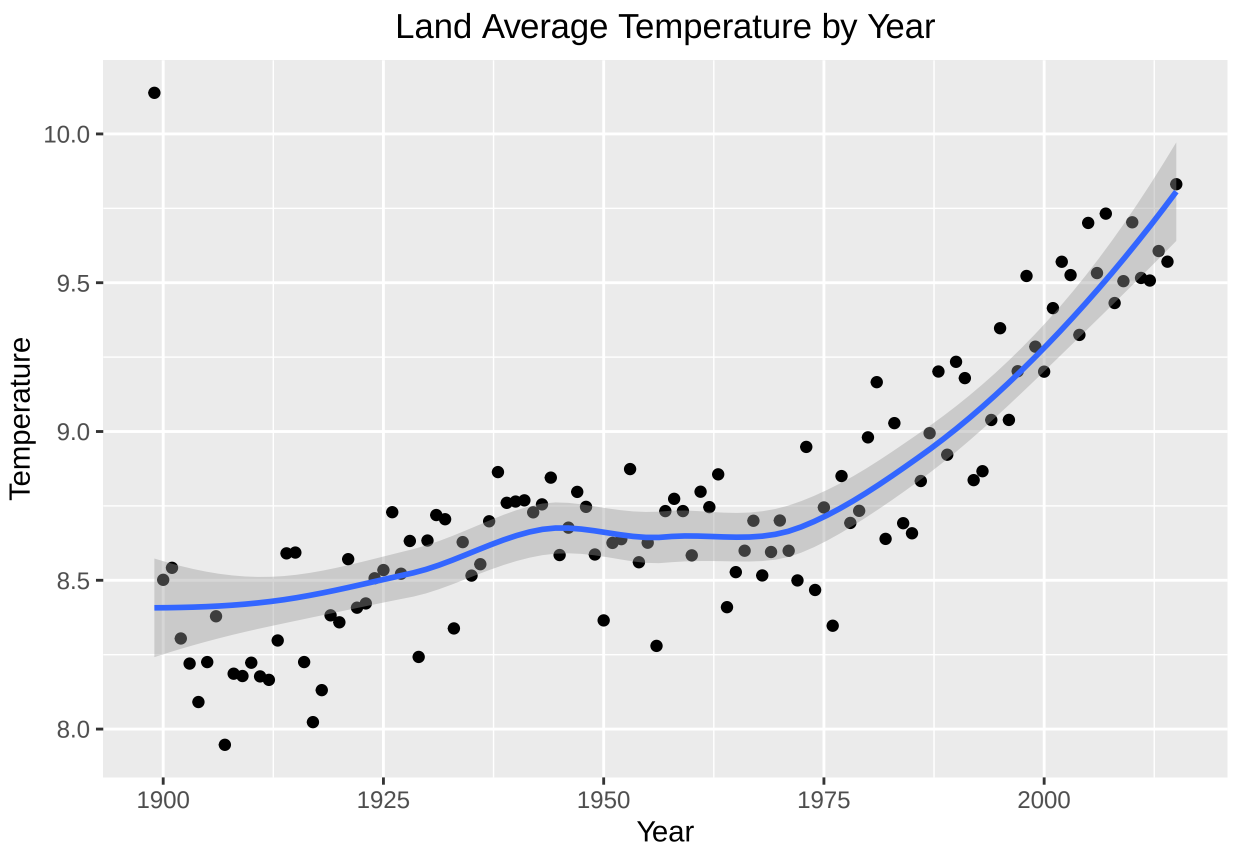
Relationship between Land Average Temperature and Quantity of Natural Disasters



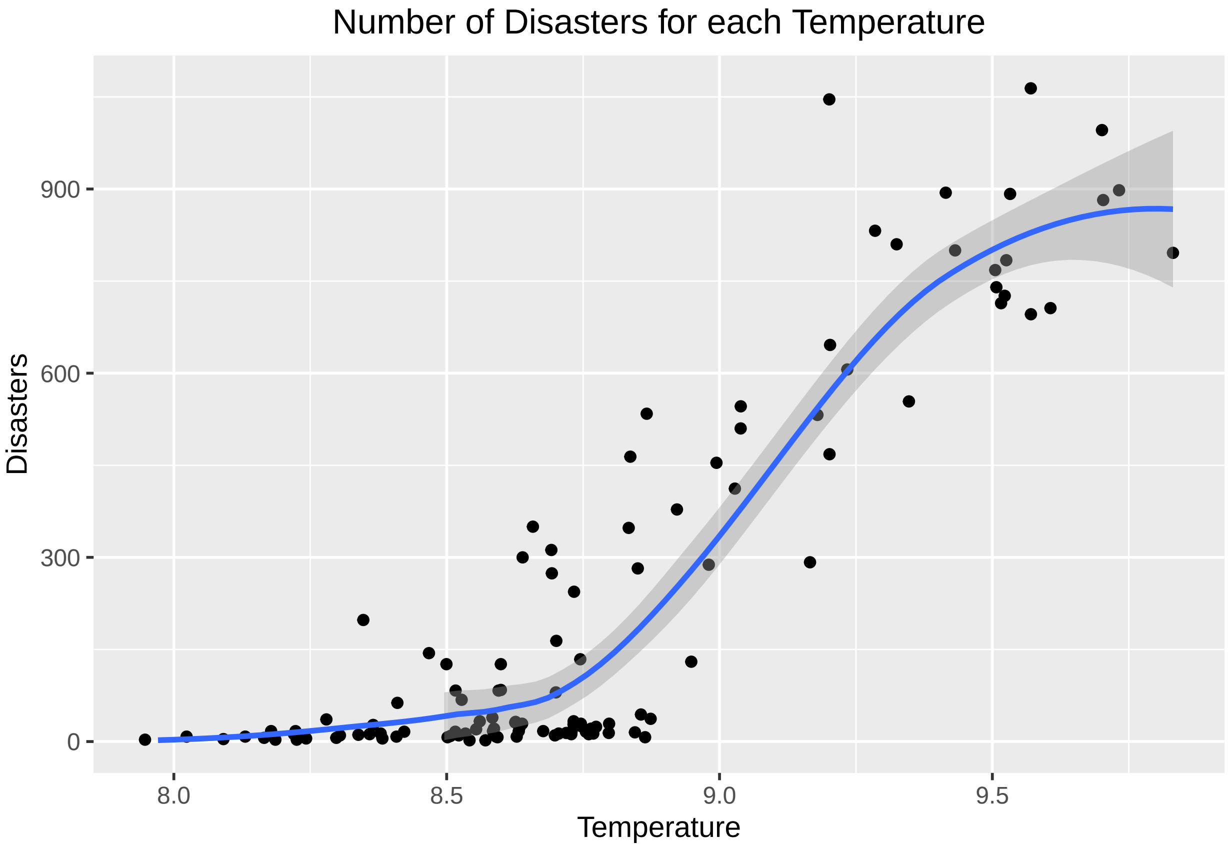
We see that the yearly number of global natural disasters increases roughly exponentially from 1910 to 2010 before finally falling down a bit in the 2010s. The number appears low for the 2020s but we are only two years into this period, and the data captured only goes through 2021.



Looking at data more locally, we see that California fires have mostly been on an upward trend but not at an exponential rate like total global natural disasters. It is noteable that there are no data on fires in California from before 1970 and from 1975 to 1990. Since this is obviously not the case in actuality, we can postulate that data on fires in California is incomplete.



This graphic includes a scatterplot of the global land average temperatures for each year as well as a regression line drawn through them with the shaded region representing a 95% confidence interval. There is a clear upward trend in average temperature over time since 1900, which probably comes with little surprise. One interesting thing to point out is the massive outlier around the year 1900. It is worth looking into whether this is due to some anomolous event or if it is just a data entry error.



This plot shows the trend between total number of disasters and average land temperature. As in the previous plot, the points represent actual data points whereas the curve is a fitted regression model. It appears that the relationship between temperature and number of disasters follows a logistic trend defined by the following function:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{1}$$

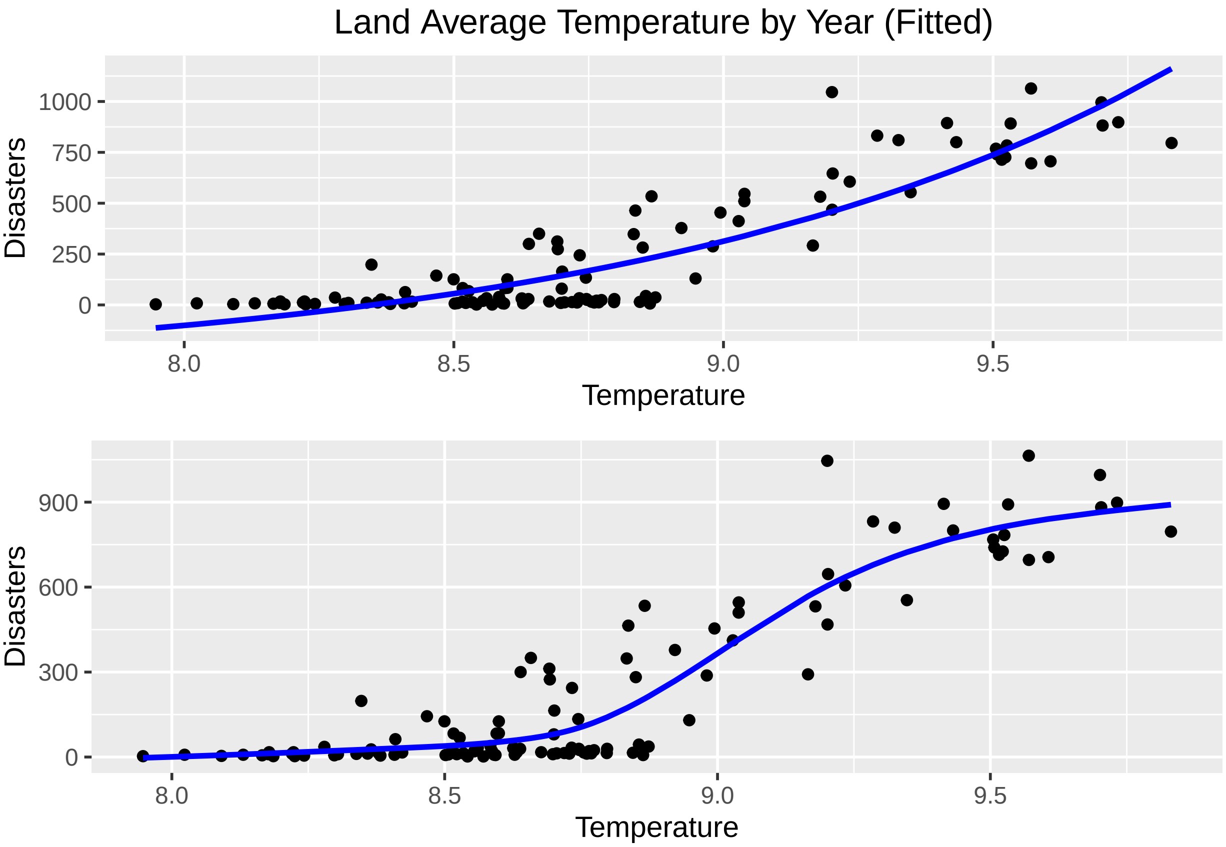
where

L is the curve's maximum value

x_0 is the x value of the function's midpoint

k is the logistic growth rate[5]

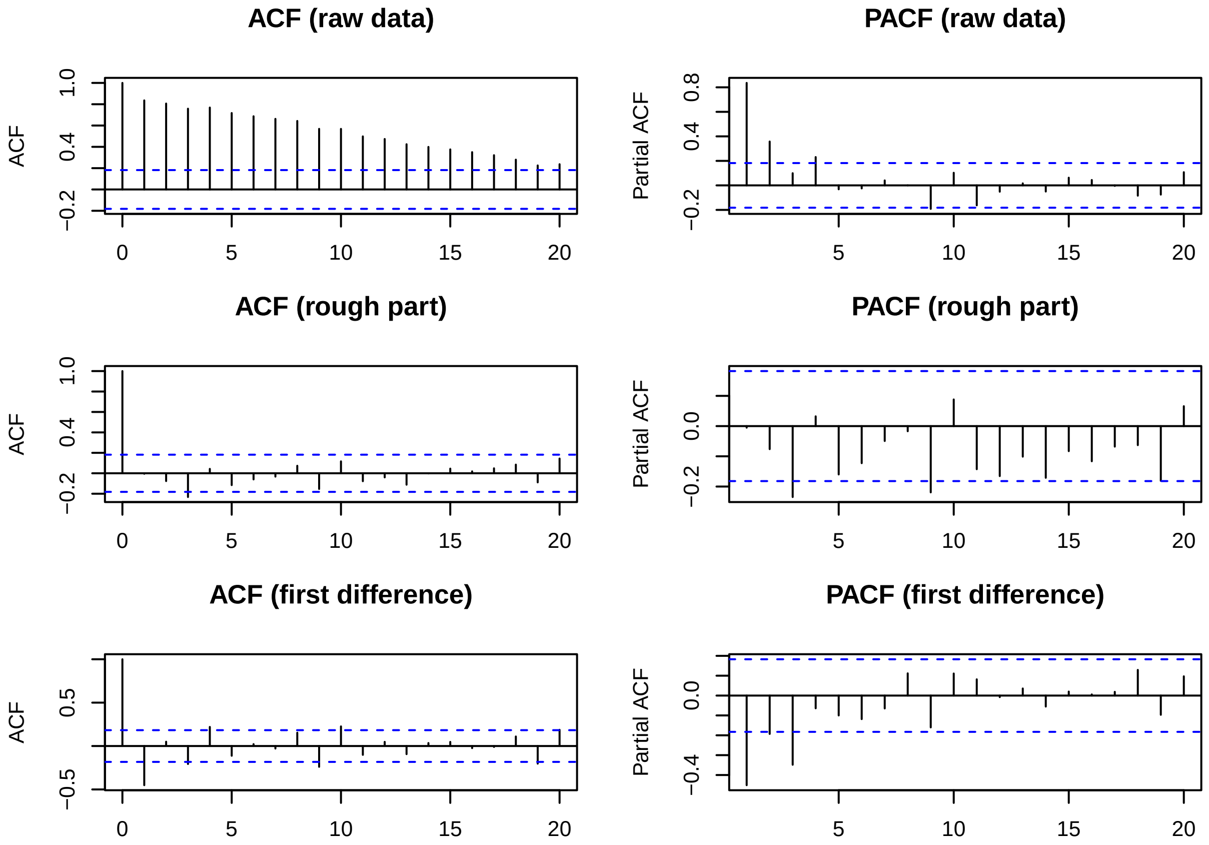
In essence, the number of disasters is roughly constantly low for temperatures on the lower end, increases roughly linearly for temperatures in the middle, and is roughly constantly high for temperatures on the higher end.



We attempt to model the data using two methods. The first plot above shows an attempt at using the sigmoid generating function in R. However, it does not generate the S-shaped curve that we are looking for. Nonetheless, it is not a bad fit for our data.

The second plot utilizes splines, which are locally generated cubics. We experimented with the parameter *nknots* for the R function smooth.spline and determined that a value of 6 produced the best fit. While the first plot was not bad, the method of splines generated a model that much more closely matches the S-shaped curve we were looking for.

Time Series Analysis of Land Average Temperature

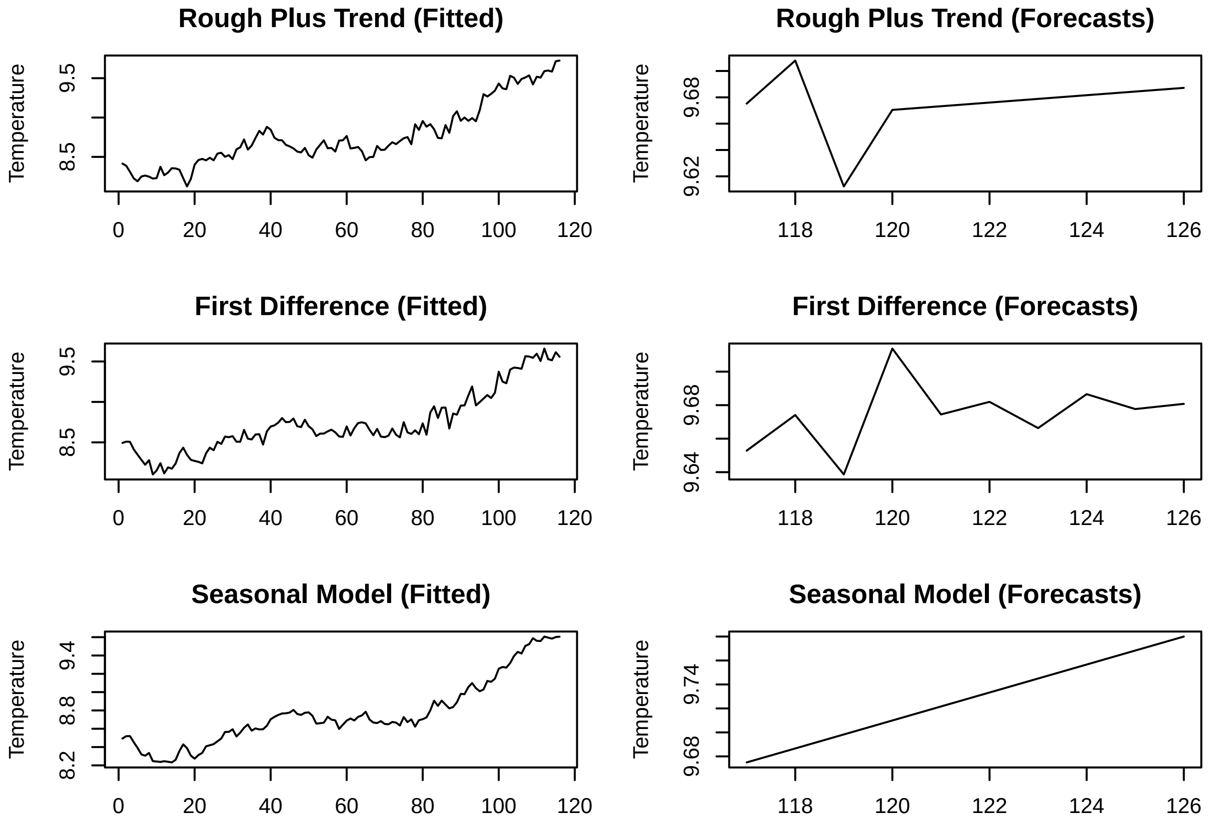


We obtain the ACF and PACF plots for the raw data and two different methods of time series generation. ACF and PACF plots can be used to determine the Auto-Regressive (AR) and Moving Average (MA) components, respectively, of a time series model. The ACF and PACF plots for the original data, "rough" part of the data, and first difference of the data (i.e. $Y_t = X_t - X_{t-1}$, where X_t is the original data series) are displayed above. The first ACF/PACF plots for the original series show no discernable MA component and a possible AR component of order 2.

However, We can probably improve upon this model by analyzing three different methods of handling time series data. The first method involves subtracting off the "trend" part of the data and analyzing the remaining "rough" part of the data. From the ACF plot, it can be determined that the rough part has an MA component of order 3. The PACF plot gives no relevant information.

The second method involves taking the first difference of the data and then analyzing Y_t. In this case, the PACF plot shows that this first difference series probably has an AR component of order 3 while the ACF plot gives no relevant information.

Finally, for the third method we utilize the R function auto.arima with parameter *seasonal* set to TRUE. This will automatically generate the "best" model according to the algorithm used in auto.arima, and will check for any seasonal components. Since we don't need to calculate the order of the AR and MA components, as this is handled by R, we do not need the ACF and PACF plots.

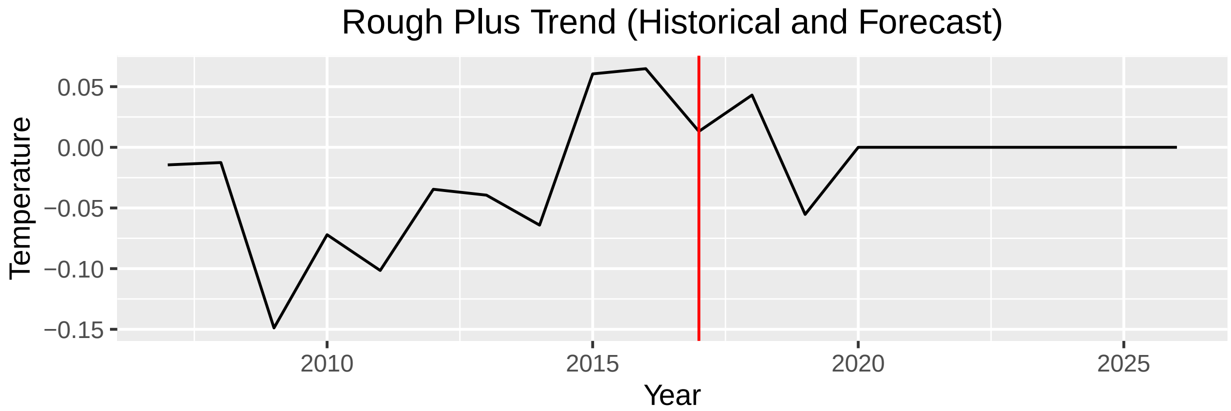


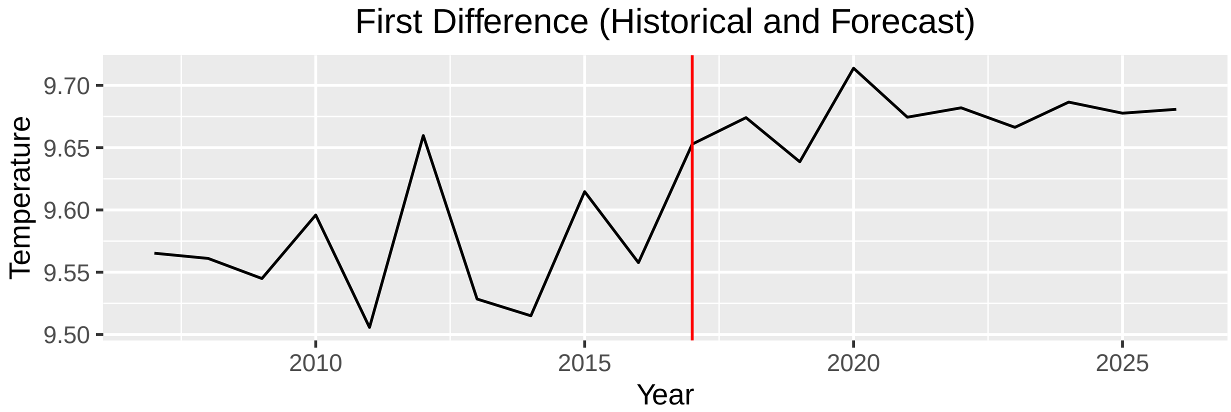
Here we display the results of fitting the three previously described models against the historical data in the left column, and the forecasted data obtained from each model in the right column.

The auto-generated seasonal model produces a smoother plot than the rough plus trend and first difference models. This is not necessarily a good thing, however, because we would like to keep the individual fluctuations present in the original data when making new predictions. Looking at AIC values may help us decide which model is best for forecasting. The AIC values for the rough plus trend, first difference, and seasonal models are -115.8025208, -58.0940606, and -61.4049308, respectively. We choose the best model by choosing the model with the lowest AIC value - in this case the rough plus trend model.

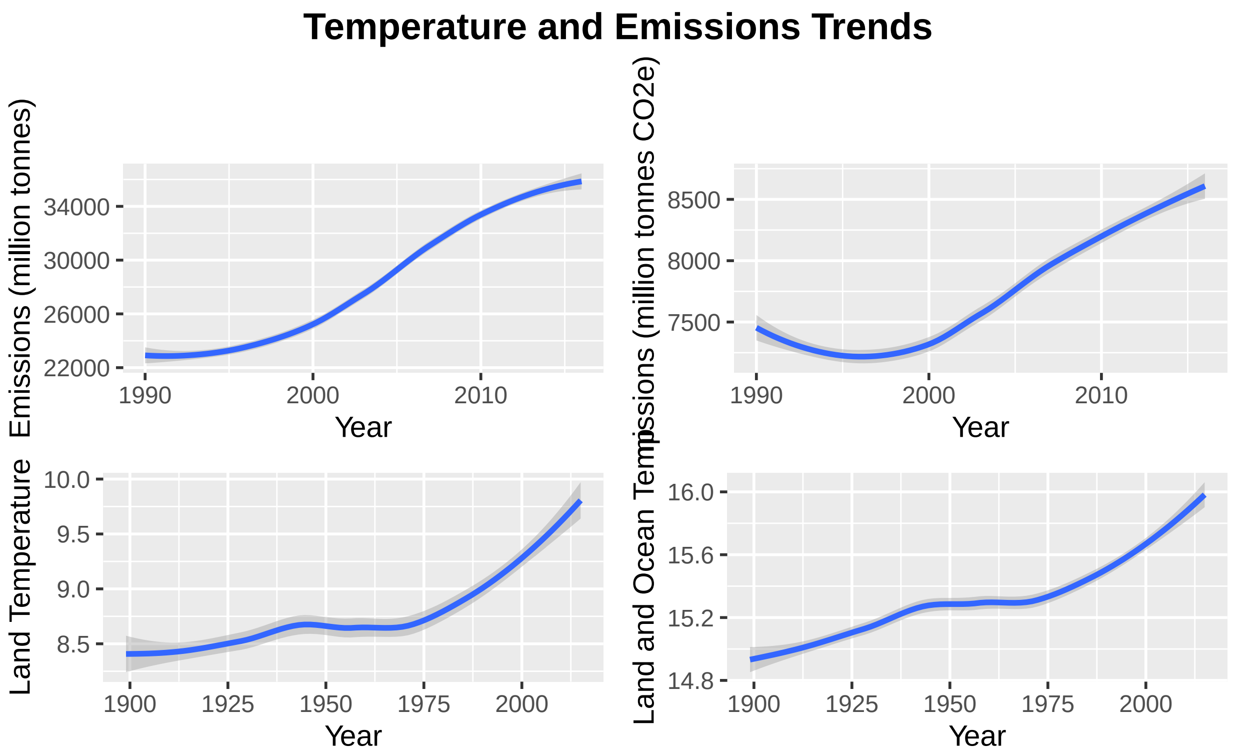
The ten-year forecasts for the three models look significantly different. Rough plus trend shows variations only up to the third year and then simply follows a straight upward-sloping line, corresponding to the trend. This is due to the fact that MA models of order q only provide random variance up to the q_{th} forecasted year. The first difference forecast shows some amount of variance through all ten forecasted years while also seeming to follow the trend. This makes it seem promising in its predictive value. The seasonal model forecast is a constantly upward-sloping line which follows the trend but provides no random variance.

Based on the AIC criterion and the forecast plots, it seems that we can rule out the seasonal model. The first difference model appears to be best if we need a model that provides accurate yearly fluctuations for more than three years, whereas the rough plus trend model seems to be the most likely to give the most accurate average prediction.



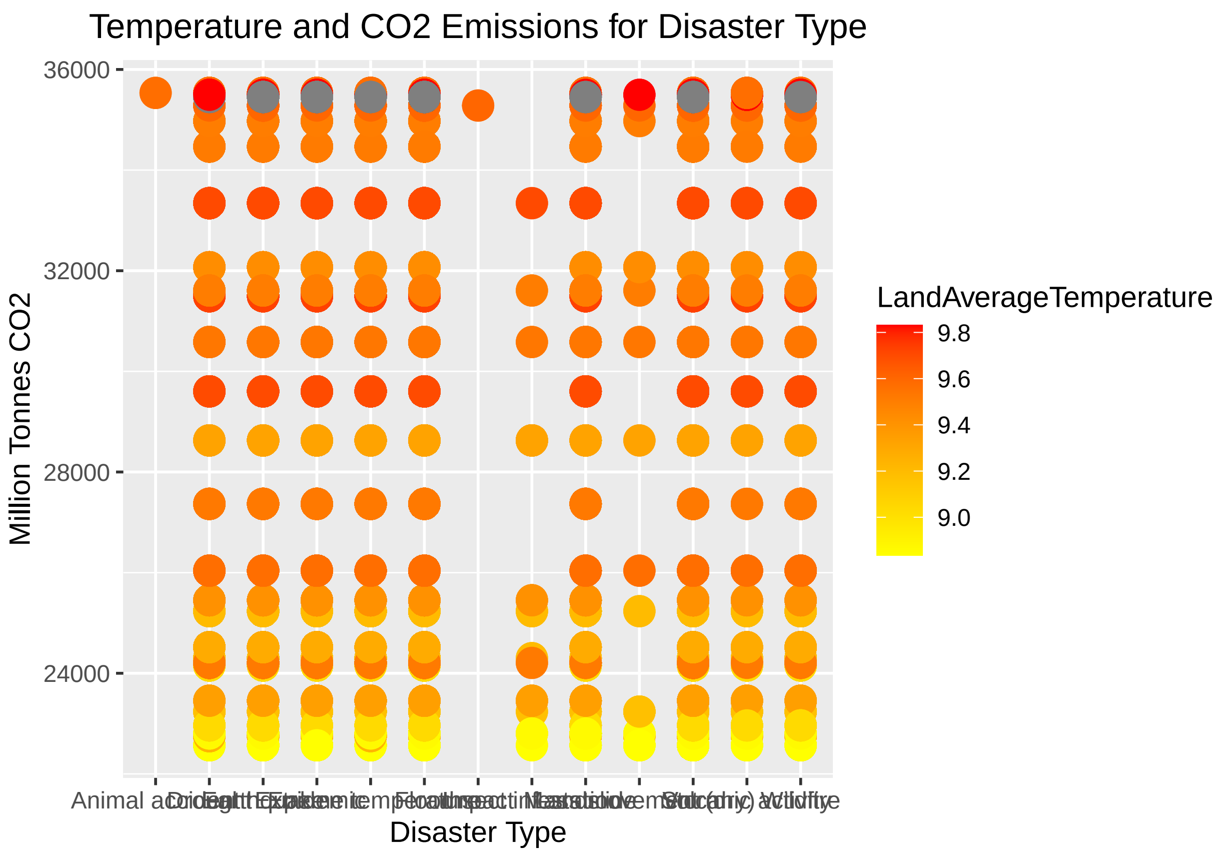


We plot the last ten years of historical data and the ten forecasted years for the rough plus trend and first difference models above. The red line indicates the point at which we go from fitting historical data to forecasting future years. By zooming in on the data, we can see that both the fitted historical values and the forecasted values differ for the two models.



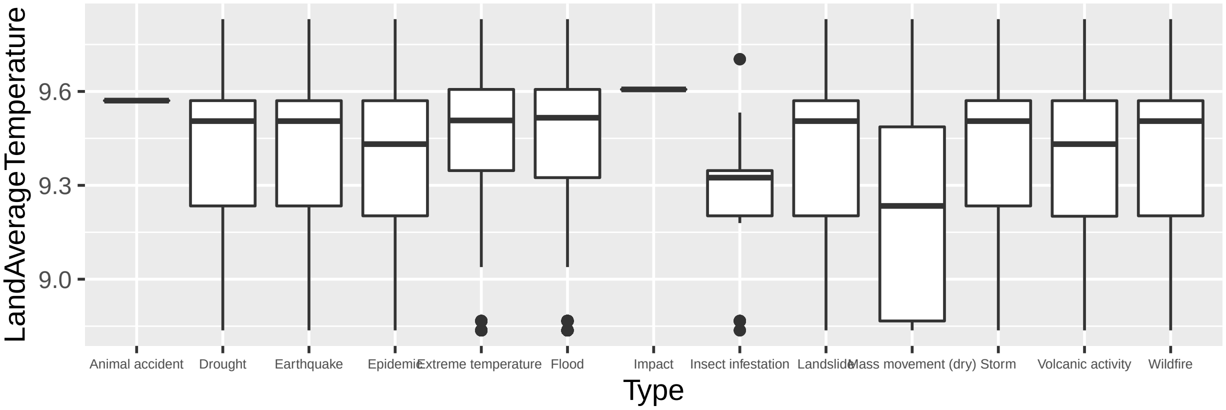
As wee can see from the plots above, CO_2 emissions and land and land-and-ocean average temperatures have rised since 1990. The overall trend is the same for methane, but with a noticeable dip between 1990 and around 1996.

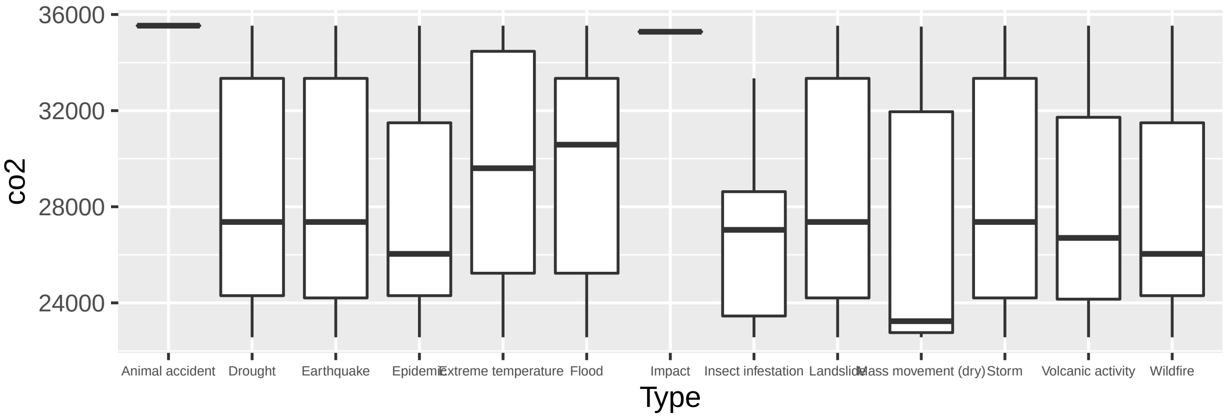
The data that we have available for **all** of these variables is on a shorter time period than for temperature alone (smaller sample size), so we must be sure to handle the data carefully and make sure no assumptions are violated in our analyses.



The two plots above (plotly not visible in PDF) are a representation of average land temperature and CO_2 emissions data points for each natural disaster type. There isn't a very clear way to categorize natural disasters by the temperature and CO_2 emissions that we can observe by simply looking at these plots. Therefore we will require more analytical methods to determine a way to separate them.

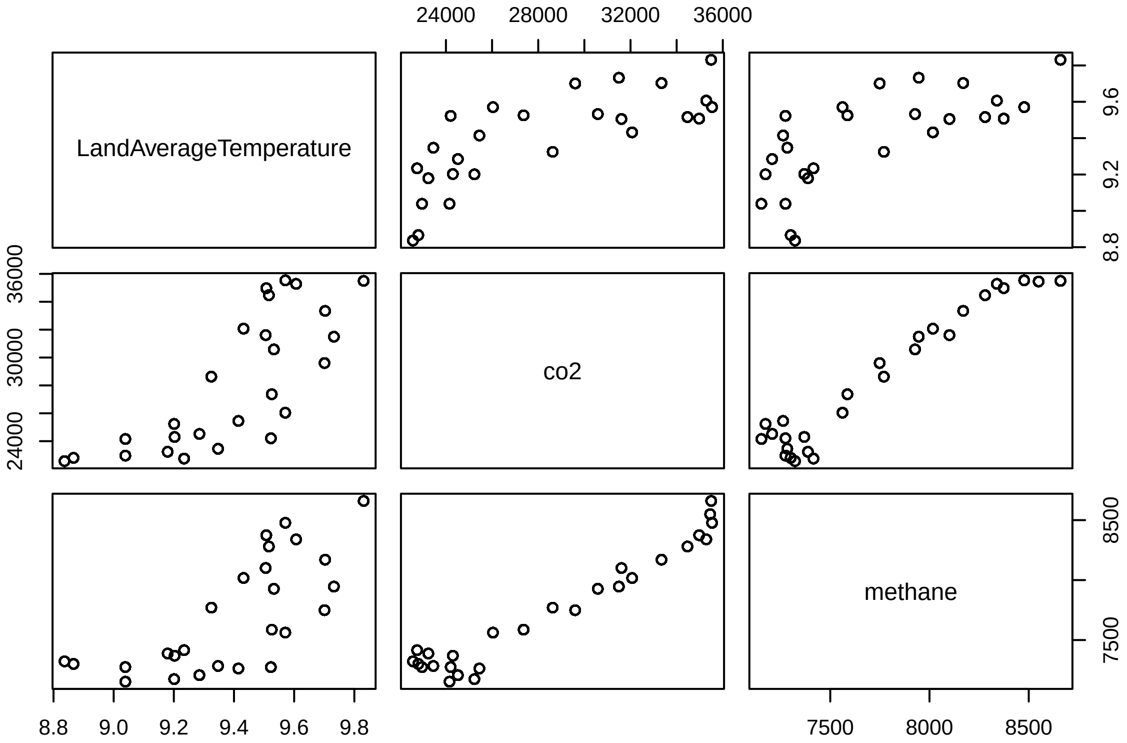
One thing to notice from the color-coded plot of temperature is that it appears that, though usually directly proportional to emissions levels, there are some higher temperature points that occur near the middle of the emissions spectrum. Therefore, emissions cannot be the only driver of increases in global temperature.



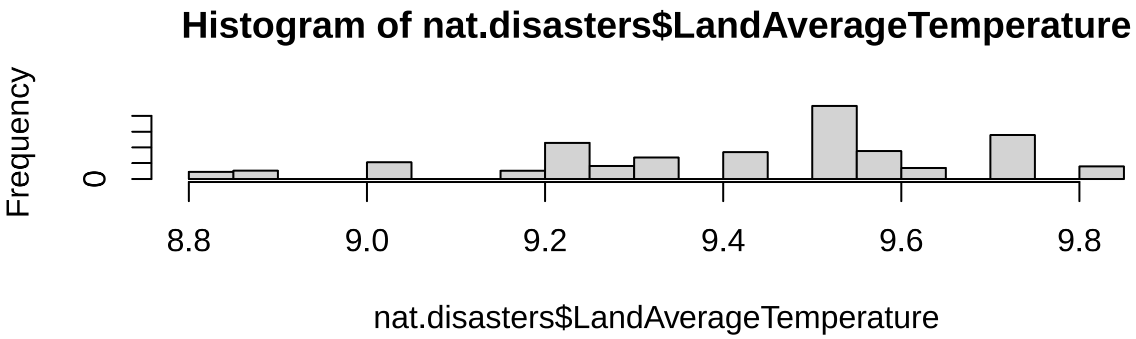


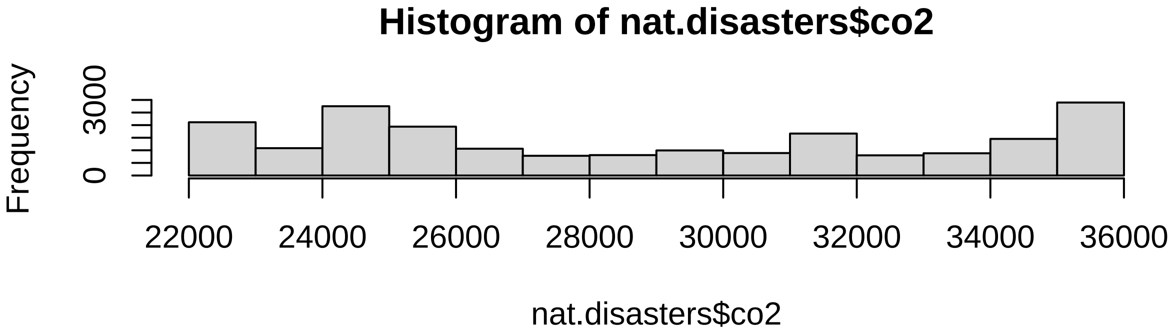
Here we separate out the factors of CO_2 emissions and average land temperature for each disaster type into two boxplots. For the first plot, it appears most disaster types have roughly similar median temperatures, though there are some deviations. We also observe a few outliers for the disaster types “Extreme temperature” (somewhat obviously), “Flood”, and “Insect infestation”.

For the second plot, the median CO_2 levels do seem a bit more spread out than for temperature. It is important to mention, however, that while differences observed between medians among boxplots may indicate a *noteable* difference in the true mean values of the measured variable, they do not necessarily indicate a statistically *significant* difference.



We create plots of all three predictor variables against one another in order to determine if there is multicollinearity. Multicollinearity appears to exist for CO_2 and methane. Because of this, one of the variables should be removed from our analysis. We choose to remove methane.





We plot histograms of the remaining two predictors, land average temperature and CO_2. Neither appears to be normally distributed, so we should not use a method of classification which requires normality of predictors such as linear discriminant analysis (LDA). We choose to use quadratic discriminant analysis (QDA), which will work with non-normal predictors.

The QDA method involves finding the class, k, which maximizes the following function:

$$\delta_k(x) = -\frac{1}{2}log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + log\pi_k \tag{2}$$

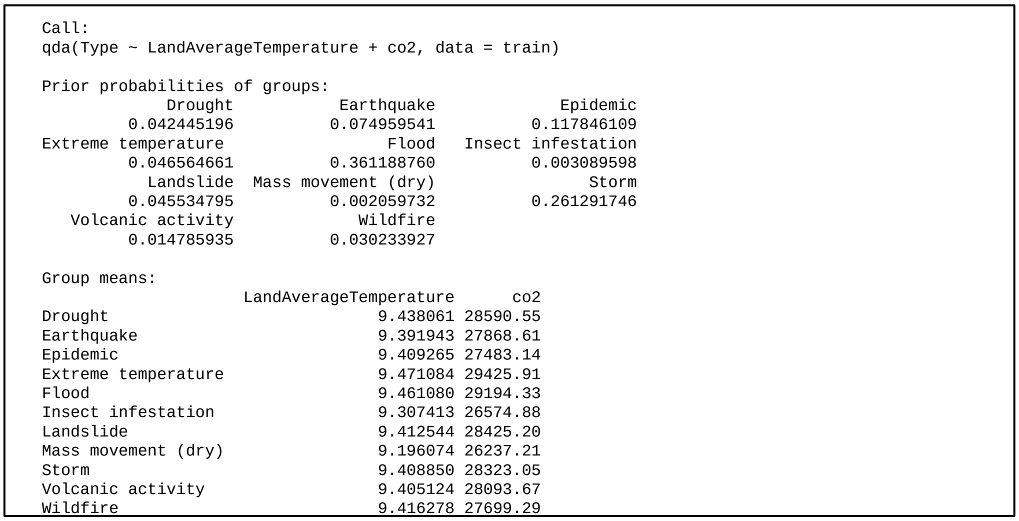
Where

Σ_k is the variance-covariance matrix for class k

μ_k is the mean of class k

π_k is the prior probability of class k[6]

Before beginning our classification of disaster type by average land temperature and CO_2 emission level, we filter “Animal accident” and “Impact” out of our dataset because they have a very small number of observations. We then randomly split all of our roughly 20,000 observations into a training set with 70% of the observations and a test set with the remaining 30% of observations. We train the training set on a QDA model and get the following results:

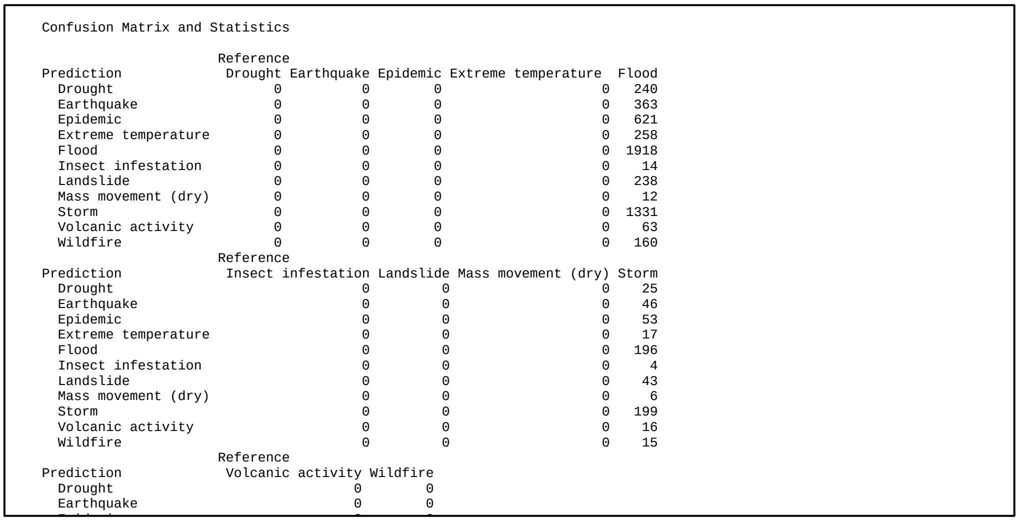


```
Call:
qda(Type ~ LandAverageTemperature + co2, data = train)

Prior probabilities of groups:
             Drought           Earthquake             Epidemic 
         0.042445196          0.074959541          0.117846109 
Extreme temperature                 Flood   Insect infestation 
         0.046564661          0.361188760          0.003089598 
           Landslide  Mass movement (dry)                Storm 
         0.045534795          0.002059732          0.261291746 
   Volcanic activity             Wildfire 
         0.014785935          0.030233927 

Group means:
                     LandAverageTemperature      co2
Drought                            9.438061 28590.55
Earthquake                         9.391943 27868.61
Epidemic                           9.409265 27483.14
Extreme temperature                9.471084 29425.91
Flood                              9.461080 29194.33
Insect infestation                 9.307413 26574.88
Landslide                          9.412544 28425.20
Mass movement (dry)                9.196074 26237.21
Storm                              9.408850 28323.05
Volcanic activity                  9.405124 28093.67
Wildfire                           9.416278 27699.29
```

We then use our QDA model to predict the data from our test set and check our predictions against the actual observations. We obtain the following confusion matrix:



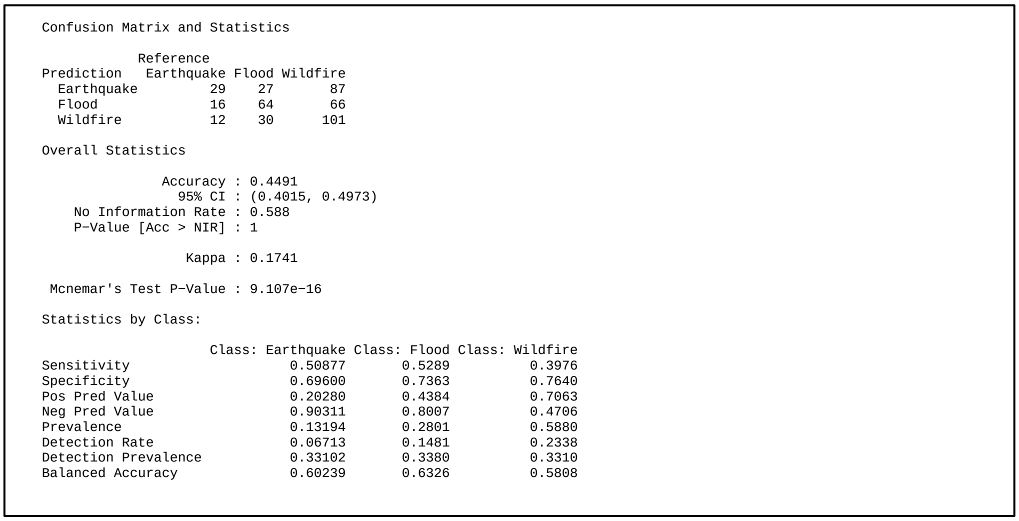
```
Confusion Matrix and Statistics

                      Reference
Prediction             Drought Earthquake Epidemic Extreme temperature  Flood
  Drought                    0          0        0                   0    240
  Earthquake                 0          0        0                   0    363
  Epidemic                   0          0        0                   0    621
  Extreme temperature        0          0        0                   0    258
  Flood                      0          0        0                   0   1918
  Insect infestation         0          0        0                   0     14
  Landslide                  0          0        0                   0    238
  Mass movement (dry)        0          0        0                   0     12
  Storm                      0          0        0                   0   1331
  Volcanic activity          0          0        0                   0     63
  Wildfire                   0          0        0                   0    160
                      Reference
Prediction             Insect infestation Landslide Mass movement (dry) Storm
  Drought                               0         0                   0    25
  Earthquake                            0         0                   0    46
  Epidemic                              0         0                   0    53
  Extreme temperature                   0         0                   0    17
  Flood                                 0         0                   0   196
  Insect infestation                    0         0                   0     4
  Landslide                             0         0                   0    43
  Mass movement (dry)                   0         0                   0     6
  Storm                                 0         0                   0   199
  Volcanic activity                     0         0                   0    16
  Wildfire                              0         0                   0    15
                      Reference
Prediction             Volcanic activity Wildfire
  Drought                              0        0
  Earthquake                           0        0
```

We can see from the confusion matrix that the model predicted every test point to of type "Flood" or "Storm", which led to a prediction accuracy of only 36.26%. There could be several reasons for this behavior including the number of observations from each class being widely different or the mean values of the two predictors being too similar among the classes.

For this reason, we chose to replicate the test with a subset of the original observations. We took 500 randomized observations from the disaster types "Flood", "Wildfire", and "Earthquake". These were chosen because we hypothesized that they were events that would occur in different climatological settings (i.e. different mean temperature and CO_2 levels) and they had enough observations to obtain a reasonably-sized and equal sample from each.

Repeating the test with this subset yielded the following results:



```
Confusion Matrix and Statistics

            Reference
Prediction   Earthquake Flood Wildfire
  Earthquake         29    27       87
  Flood              16    64       66
  Wildfire           12    30      101

Overall Statistics

               Accuracy : 0.4491
                 95% CI : (0.4015, 0.4973)
    No Information Rate : 0.588
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1741

 Mcnemar's Test P-Value : 9.107e-16

Statistics by Class:

                     Class: Earthquake Class: Flood Class: Wildfire
Sensitivity                    0.50877       0.5289          0.3976
Specificity                    0.69600       0.7363          0.7640
Pos Pred Value                 0.20280       0.4384          0.7063
Neg Pred Value                 0.90311       0.8007          0.4706
Prevalence                     0.13194       0.2801          0.5880
Detection Rate                 0.06713       0.1481          0.2338
Detection Prevalence           0.33102       0.3380          0.3310
Balanced Accuracy              0.60239       0.6326          0.5808
```

We can see that we now get predictions in all three categories and the prediction accuracy has increased to 44.91%. This is still not an optimal prediction accuracy, so other solutions to the classification problem should be attempted.

Conclusion

In our analysis of the three continuous factors of number of disasters, year, and land average temperature we found all three to be positively related. Since the 1900s, global disasters have increased exponentially while temperature increased mostly linearly except between around 1945 and 1975. Perhaps this time period should be studied in more depth to understand what factors may have led temperatures to stop increasing. It is important to consider outliers and verify that data is complete for all of these factors. We found the spline method to be the best way to generate a model of the S-shaped regression curve of number of disasters over temperature.

We attempted three ways of modeling the time series of average land temperature: rough plus trend, first difference, and auto-generated seasonal. We got the best results with regard to AIC criterion from the rough plus trend model and the best looking forecast from the first difference model. It could be useful to re-attempt a seasonal model using a monthly level of granularity instead of yearly to capture actual variations due to the seasons (i.e. Fall, Winter, Spring, Summer).

Our categorization of disaster type based on land average temperature, CO_2 emissions, and methane emissions proved difficult as there were problems of multicollinearity, non-normality of data, similarity of mean levels of predictors for each disaster type, and different numbers of each disaster type recorded. It would be helpful to look at more predictors that are less correlated with one another and for which the means for each disaster type differ more. These could include factors like geographical factors, wind speed, atmospheric measures, and sunlight.

The discoveries made in this report lead us to believe that more research must be done to discover which of these factors and other factors may be responsible for *causing* the increase in natural disasters, rather than just being correlated. It is our hope that the information in this report may provide a baseline for others to answer this question as well as figure out potential solutions to the problem.

References

1. https://www.kaggle.com/brsdincer/all-natural-disasters-19002021-eosdis - Natural disasters dataset
2. https://www.kaggle.com/amelinvladislav/map-of-temperatures-and-analysis-of-global-warming/data?select=GlobalTemperatures.csv - Temperature dataset
3. https://ourworldindata.org/co2-emissions (https://github.com/owid/co2-data) - Greenhouse gas emissions dataset
4. https://www.usnews.com/news/best-countries/slideshows/here-are-10-of-the-deadliest-natural-disasters-in-2021?slide=11
5. https://en.wikipedia.org/wiki/Logistic_function
6. https://online.stat.psu.edu/stat508/lesson/9/9.2/9.2.8
7. https://link.springer.com/article/10.1007/s10113-013-0499-2 - EURO-CORDEX projections study
8. https://www.liebertpub.com/doi/pdf/10.1089/big.2014.0026 - Using big data for climate change
9. https://stackoverflow.com/questions/49719660/r-markdown-to-pdf-printing-console-output - Used for console printing function

Code Appendix

```r
# Load libraries and data
library(datetime)
library(dplyr)
library(ggplot2)
library(sigmoid)
library(cowplot)
library(plotly)
library(forecast)
library(MASS)

disasters1.data <- read.csv('./data/DISASTERS/1900_2021_DISASTERS.xlsx - emdat data.csv')
disasters2.data <- read.csv('./data/DISASTERS/1970-2021_DISASTERS.xlsx - emdat data.csv')
temp.data <- read.csv('./data/GlobalTemperatures.csv')
emissions.data <- read.csv('data/owid-co2-data.txt', header = TRUE)

# Combine data from two disasters datasets
common <- match(colnames(disasters1.data), colnames(disasters2.data))
disasters2.data <- disasters2.data[,common]
disasters.data <- rbind(disasters1.data, disasters2.data)
```

```r
# Histogram of all disasters per year
hist(disasters.data$Year, xlab = "Year", main = "Total Global Disasters per Year")

# Get data on CA fires
fire.data <- disasters.data[disasters.data$Disaster.Type == "Wildfire",]
row.condition <- Reduce(union, list(grep("cali", fire.data$Location, ignore.case = TRUE),
                        grep("los angeles", fire.data$Location, ignore.case = TRUE),
                        grep("Oakland", fire.data$Location),
                        grep("San Diego", fire.data$Location)))
cali.fires <- fire.data[row.condition,]

# Histogram of CA fires per year
hist(cali.fires$Year, xlab = "Year", main = "California Fires per Year")

# Get temperature data in desired format
temp.data.mod <- temp.data
temp.data.mod$dt <- as.date(temp.data.mod$dt, format = "%Y-%m-%d")
temp.data.mod <- temp.data.mod[temp.data.mod$dt > as.date("1899", format = "%Y"),]
temp.data.mod$dt <- as.numeric(format(temp.data.mod$dt, format = "%Y"))
temp.data.mod <- temp.data.mod %>% group_by(dt) %>% summarize(LandAverageTemperature = mean(LandAverageT
                                                              LandAverageTemperatureUncertainty = mean(L
                                                              LandAndOceanAverageTemperature = mean(Land
                                                              LandAndOceanAverageTemperatureUncertainty

# Plot of average land temperature over time
ggplot(data = temp.data.mod, mapping = aes(x = dt, y = LandAverageTemperature)) +
  ylim(min(temp.data.mod$LandAverageTemperature), max(temp.data.mod$LandAverageTemperature)) +
  labs(title = "Land Average Temperature by Year", x = "Year", y = "Temperature") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_point() + geom_smooth()

# Get number of disasters per year up to 2016
disasters.data.mod <- disasters.data
disasters.data.mod <- disasters.data.mod[disasters.data.mod$Year < 2016,]
disasters.data.mod <- disasters.data.mod %>% group_by(Year) %>% summarize(n = n())

# Join temperature and disaster datasets
temp.and.disasters <- dplyr::inner_join(temp.data.mod, disasters.data.mod, by = c("dt" = "Year"))

# Plot of number of disasters against average land temperature
ggplot(data = temp.and.disasters, mapping = aes(x = LandAverageTemperature, y = n)) +
  ylim(min(temp.and.disasters$n), max(temp.and.disasters$n)) +
  labs(title = "Number of Disasters for each Temperature", x = "Temperature", y = "Disasters") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_point() + geom_smooth()

# Linear model with response number of disasters and predictor average land temperature using sigmoid
num.disasters.v.temp <- lm(n ~ sigmoid(LandAverageTemperature - 50, "logistic"), data = temp.and.disaste
p1 <- ggplot(data = temp.and.disasters) +
  labs(title = "Land Average Temperature by Year (Fitted)", x = "Temperature", y = "Disasters") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_point(mapping = aes(x = LandAverageTemperature, y = n)) +
  geom_line(mapping = aes(x = LandAverageTemperature, y = num.disasters.v.temp$fitted.values), size = 1
# Fit using splines
spline.model <- smooth.spline(x = temp.and.disasters$LandAverageTemperature, y = temp.and.disasters$n, n
p2 <- ggplot(data = temp.and.disasters) +
  labs(x = "Temperature", y = "Disasters") +
  geom_point(mapping = aes(x = LandAverageTemperature, y = n)) +
  geom_line(mapping = aes(x = spline.model$x, y = spline.model$y), size = 1, color = 'blue')
plot_grid(p1, p2, ncol = 1)

## Time series projections ##

# Fit a trend model of land temperatures against time
temp.trend <- smooth.spline(x = temp.and.disasters$dt, y = temp.and.disasters$LandAverageTemperature, nk

# Plot ACF and PACF for raw data, first difference of data, and data with trend removed
orig.par <- par()
par(mfrow = c(3,2), mar = c(3,4,3,2))
acf(temp.and.disasters$LandAverageTemperature, main = "ACF (raw data)")
pacf(temp.and.disasters$LandAverageTemperature, main = "PACF (raw data)")

acf((temp.and.disasters$LandAverageTemperature - temp.trend$y), main = "ACF (rough part)")
pacf((temp.and.disasters$LandAverageTemperature - temp.trend$y), main = "PACF (rough part)")

acf(diff(temp.and.disasters$LandAverageTemperature, lag = 1), main = "ACF (first difference)")
pacf(diff(temp.and.disasters$LandAverageTemperature, lag = 1), main = "PACF (first difference)")
par(orig.par)

# Generate rough, first difference, and seasonal models
temp.ts.rough <- arima((temp.and.disasters$LandAverageTemperature - temp.trend$y), order = c(0,0,3))
temp.ts.diff <- arima(temp.and.disasters$LandAverageTemperature, order = c(3,1,0))
temp.ts.seas <- auto.arima(temp.and.disasters$LandAverageTemperature, seasonal = TRUE)

# Generate forecasts for the next 10 years from each model
temp.fc.rough <- forecast(temp.ts.rough, h = 10)
temp.fc.diff <- forecast(temp.ts.diff, h = 10)
temp.fc.seas <- forecast(temp.ts.seas, h = 10)
temp.fc.rpt <- temp.fc.rough$mean + predict(temp.trend$fit, 2016:2025)$y

# Plot the fitted models and forecasts
par(mfrow = c(3,2), mar = c(3,4,3,2))
plot.ts(fitted(temp.ts.rough) + temp.trend$y, main = "Rough Plus Trend (Fitted)")
plot.ts(temp.fc.rpt, main = "Rough Plus Trend (Forecasts)")
plot.ts(fitted(temp.ts.diff), main = "First Difference (Fitted)")
plot.ts(temp.fc.diff$mean, main = "First Difference (Forecasts)")
plot.ts(fitted(temp.ts.seas), main = "Seasonal Model (Fitted)")
plot.ts(temp.fc.seas$mean, main = "Seasonal Model (Forecasts)")
par(orig.par)

# Plot forecast and historical data (last 10 years) from assumed best models
combined.rough.ts <- ts(c(fitted(temp.ts.rough)[107:116], temp.fc.rough$mean))
p1 <- ggplot() +
  labs(title = "Rough Plus Trend (Historical and Forecast)", x = "Year", y = "Temperature") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_line(mapping = aes(x = 2007:2026, y = combined.rough.ts)) +
  geom_vline(xintercept = 2017, color = 'red')

combined.diff.ts <- ts(c(fitted(temp.ts.diff)[107:116], temp.fc.diff$mean))
p2 <- ggplot() +
  labs(title = "First Difference (Historical and Forecast)", x = "Year", y = "Temperature") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_line(mapping = aes(x = 2007:2026, y = combined.diff.ts)) +
  geom_vline(xintercept = 2017, color = 'red')

plot_grid(p1, p2, ncol = 1)

## Predictors of types of natural disasters ##

# Modify emissions data to only include global data and select needed variables
emissions.data.mod <- emissions.data %>% dplyr::select(country, year, co2, methane) %>% filter(country =

# Plots of emissions/temperatures over time
p1 <- ggplot(data = emissions.data.mod[!is.na(emissions.data.mod$methane),]) + labs(y = "Emissions (mil
  geom_smooth(mapping = aes(x = year, y = co2))
p2 <- ggplot(data = emissions.data.mod[!is.na(emissions.data.mod$methane),]) + labs(y = "Emissions (mil
  geom_smooth(mapping = aes(x = year, y = methane))
p3 <- ggplot(data = temp.data.mod) +
  geom_smooth(mapping = aes(x = dt, y = LandAverageTemperature))
p4 <- ggplot(data = temp.data.mod) +
  geom_smooth(mapping = aes(x = dt, y = LandAndOceanAverageTemperature))
plot_grid(p1, p2, p3, p4, ncol = 2)

# TODO: filter out only Climatological subgroup

# Combine natural disasters, temperature, and emissions datasets for years in which there is
# data for all three of them (1990-2016)
nat.disasters <- data.frame(Year = disasters.data$Year, Type = disasters.data$Disaster.Type) %>% filter
nat.disasters <- nat.disasters %>% left_join(temp.data.mod, by = c("Year" = "dt"), copy = TRUE) %>% left

plot_ly(data = nat.disasters, x = nat.disasters$Type, y = nat.disasters$co2, z = nat.disasters$LandAvera
ggplot(data = nat.disasters) +
  labs(title = "Temperature and CO2 Emissions for Disaster Type", x = "Disaster Type", y = "Million Tonn
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(size = 9)) +
  geom_point(mapping = aes(x = Type, y = co2, colour = LandAverageTemperature), size = 5) +
  scale_color_gradient(low = "yellow", high = "red")

# Boxplots of temperatures and CO2 levels at which disasters occur
p1 <- ggplot(data = nat.disasters) +
  geom_boxplot(mapping = aes(x = Type, y = LandAverageTemperature)) +
  theme(axis.text.x = element_text(size = 5))
p2 <- ggplot(data = nat.disasters) +
  geom_boxplot(mapping = aes(x = Type, y = co2)) +
  theme(axis.text.x = element_text(size = 5))
plot_grid(p1, p2, ncol = 1)

# ANOVA on temperature, CO2, and methane
aov.temp <- aov(LandAverageTemperature ~ Type, data = nat.disasters, na.action = na.exclude)
tukey.temp <- TukeyHSD(aov.temp)
```

```r
aov.co2 <- aov(co2 ~ Type, data = nat.disasters)
aov.methane <- aov(methane ~ Type, data = nat.disasters)

# Classification of disaster type based on temperature and CO2
pairs(nat.disasters[,c("LandAverageTemperature", "co2", "methane")])
hist(nat.disasters$LandAverageTemperature)
hist(nat.disasters$co2)

# CO2 and methane appear to be multicollinear and LandAverageTemperature and CO2 are both NOT normally
nat.disasters.mod <- nat.disasters %>% filter(!(Type %in% c("Animal accident", "Impact"))) %>% dplyr::se

set.seed(123)
train.ind <- sample(1:nrow(nat.disasters.mod), size = 0.7 * nrow(nat.disasters.mod), replace = FALSE)
train <- nat.disasters.mod[train.ind,]
test <- nat.disasters.mod[-train.ind,]
temp.co2.qda <- lda(Type ~ LandAverageTemperature + co2, data = train)

# Obtained from https://stackoverflow.com/questions/49719660/r-markdown-to-pdf-printing-console-output
print_output <- function(output, cex = 0.7) {
  tmp <- capture.output(output)
  plot.new()
  text(0, 1, paste(tmp, collapse='\n'), adj = c(0,1), family = 'mono', cex = cex)
  box()
}

print_output(temp.co2.qda)

pred.qda <- predict(temp.co2.qda, test)
conf.qda <- confusionMatrix(as.factor(test$Type), pred.qda$class)

print_output(conf.qda, cex = 0.4)

set.seed(456)
nat.disasters.mod <- nat.disasters %>% filter((Type %in% c("Flood", "Wildfire", "Earthquake"))) %>% dply

train.ind <- sample(1:nrow(nat.disasters.mod), size = 0.7 * nrow(nat.disasters.mod), replace = FALSE)
train <- nat.disasters.mod[train.ind,]
test <- nat.disasters.mod[-train.ind,]
temp.co2.qda <- qda(Type ~ LandAverageTemperature + co2, data = train)

pred.qda <- predict(temp.co2.qda, test)
conf.qda <- confusionMatrix(as.factor(test$Type), pred.qda$class)

print_output(conf.qda, cex = 0.4)
```