Hurricanes and CO2

Nathan Schneider

CS4412 Section

Fall 2021

Professor Sharon Perry

Abstract

The correlation between human pollution and climate change has been established as fact. We have been polluting the air and water for centuries and it has only grown worse with time. One of the leading causes of global ecological downfall is carbon dioxide emissions. CO2 is a greenhouse gas produced from the consumption of fossil fuels. Without fossil fuels civilization would not be as advanced as we are today. The same thing that fuels the world is also killing the world. One of the most visible signs of climate change is an increase in extreme weather. This includes drought, flooding, extreme heat, extreme cold, and extreme storms. Different regions around the world have their own storm seasons. As CO2 emission increase, so should the quantity, and intensity of the storms. This project will identify correlations between CO2 emissions and extreme weather.

Introduction

One of the most consistent producers of CO2 is the United States. A prominent example of extreme weather in and around the United States is the Atlantic Hurricane season. The Atlantic Hurricane season officially lasts from June to November, but storms can still occur outside of this time frame. In my research I set out to find specifically the impact of CO2 from the United States on the Atlantic Hurricane season. The motivation for this project is to gain a better understanding of historical climate change. Understanding the past will allow for a better understanding of the future.

Background

The National Oceanic and Atmospheric Administration (NOAA) has compiled all of their data on hurricanes into a database called the Hurricane Database (HURDAT). The database currently consists of data ranging from 1851 to 2010. The older data is substantially less reliable than the recent data. Because of this, my research focuses on data starting in 1950. Not all of the storms made landfall, but I am looking for the existence of the storms in general. The important aspect of the database is the type of storm. The storms range from tropical depressions through category 5 hurricanes. This part of the database is what will inform my mining and experiments.

The HURDAT data will be compared against CO2 data. For this project the data relating to CO2 emissions comes from the Oxford Martin School’s “Our World in Data” project. The research institute has combined all vast amounts of CO2 emissions data into one database. Much like the HURDAT database, the older data is less reliable than the more current data. Much of the older data is derived from experiments done in labs rather than directly from samples taken from the air. To correspond with the time range for data from HURDAT, I have gathered data from 1950 to 2010. The important data categories are the total CO2 emissions from the country and the percentage of the global CO2 emissions. This will be shown against the data from HURDAT

Methodology

Because both databases are so large, I needed to extract the specific data that I was looking for. The Oxford database has over 24,00 rows and the HURDAT database has over 40,000 rows. Much of this data is not relevant to the project and can be ignored. The Oxford database is maintained in a nice csv file on GitHub. I wrote a python file that indexed into the database to find the data for the United States. Then it copied the relevant information into a separate and more manageable file for the experiments.

Unlike the Oxford database, the HURDAT database is stored in a plain text file. For HURDAT I wrote a python file to extract the relevant information. The code uses regular expressions and the Natural Language Tool Kit (nltk) for parsing. First the code identifies the header for each storm with the format “Storm [a-zA-Z]+ \*is number \*\d+ of the year \d+”. An example of the header for a storm is “Storm ABLE is number 1 of the year 1950”. The regular expression will recognize “ABLE” as “[a-zA-Z]+”, “1” as “\d+”, and the year “1950” as “\d+”. After identifying the start of a storm’s information, the code iterates through each line of the storm’s data. The lines of data represent recordings of data during the life span of the storm. Each recording will include the time stamp, location, direction, speed, wind speed, atmospheric pressure, and the type of storm. All storms start off as depressions and grow in strength. Thus, the relevant information to extract for each storm is its name, the year it occurred, and its peak intensity. Meta data is gathered from this to calculate how many storms occurred in each year and the cumulative intensity of the years. I created a scale to represent a storm’s intensity. Storms that reached a maximum level of extratropical storm, tropical depression, or tropical storm were given the value 1. Storms that evolved into hurricanes were given the value of their highest category + 1 (e.g. Category 2 would be 2 + 1). For each year the intensity of each storm is accumulated to give a total score for that year. The meta data was copied into a separate file.

Experiments

After using python files to preprocess my data, I used python in Jupyter Notebook for the experiments. The two data sets that were imported represent the relevant information from HURDAT and the meta data from Oxford. Using the software Pandas, I merged the two data sets together. The format for the code used to merge the data sets is “merged\_data = pd.merge(hurdat, co2, on='Year').dropna()”. The merged data set contains all the data from the preprocessed data with the year as the key. The categories for the merged data set are year, score, score scaled, quantity, quantity scaled, CO2, and global percentage. Score comes from the HURDAT meta data and represents the total intensity of all storms for a given year. Quantity also comes from the HURDAT meta data and represents the total number of storms for a given year. CO2 comes from the Oxford data set and represents the amount of CO2 produced by the United States in metric tons. Global percentage comes from Oxford data set and represents the portion of all global CO2 emissions that the United States produced. For the purposes of superimposing multiple graphs, I scaled up the score and quantity values so that all data would appear in the same range for y on the graphs. See figure 1 in Appendix B. I used a combination of NumPy (np) and Seaborn (sns) for generating my graphs. Seaborn is substantially more powerful than NumPy. It handles values that are usually too large for NumPy’s exp function to calculate. Additionally, I could generate graphs using only one or two lines.

From the merged data sets I initially generated multiple scatter plots to visual the data sets separately. The first scatter plot represented the trend over time between the score (i.e. total intensity) for a given year and the quantity for a given year. As stated above, the score and quantity values were scaled using a coefficient to allow for both data categories to be displayed in the same graph. The lines of best fit for quantity and score both represent very wide positive parabolas. See figure 2 in Appendix B.

Next a scatter plot was made to show the correlation between the score and quantity. The x axis represents the quantity, and the y axis represents the score. The closer to the origin of the graph the points are, the closer together they are. This shows that when a quantity is lower, then the score will also be lower. As the quantity increases the range of scores varies more widely. This makes sense because a higher number of storms would allow for more storms that were on the higher end of the intensity scale which results in a higher score. The line of best fit for this graph is a third-degree polynomial that is nearly a line at 45º. This shows that we can safely assume that in general both the quantity and score have a direct causation. See figure 3 in Appendix B.

After viewing the relationships between score and quantity, I generated scatter plots to show the isolated score and quantity over time. Both graphs had similar parabolas as stated above. At this point in the experiments, I realized that the period from 1950 to roughly 1975 shows a gradual negative trend in the scores and quantities for each year. Then from 1975 onwards the trend increases at a faster rate than the downward trend. The data is correct, and the trend line is accurate. It does not make sense for there to have been a decades long decrease in both the quantity and scores for a given year. This will be addressed later in the report. See figures 4 and 5 in Appendix B.

The next scatter plot shows the trend of CO2 emissions over time. The x axis represents a given year, and the y axis represents the amount of CO2 emitted. The graph shows that CO2 emissions have risen at a steady pace since 1950. The graph also reveals a spike in CO2 emissions from 1970 to 1975 peaking at 1973 and another spike from 1975 to 1982 peaking at 1978. These spikes average out, but still affect the line of best fit for the data. The line of best fit accurately represents the data points as a second-degree polynomial. See figure 6 in Appendix B.

After analyzing the HURDAT and Oxford datasets individually, I moved on to analyzing the correlations between the two. First, I made a graph showing the graphs for CO2 for a given year and Score Scaled for a given year superimposed. This was showing previously the analyzed graph shown in figure 6 with the addition of the scaled data for the score for a given year. As stated above the CO2 trend has been steadily rising since 1950. The trend in scores for a given year continues to show a downward trend from 1950 to 1975 before climbing again from 1975 onwards. See figure 7 in Appendix B.

I made another graph similar to the previous graph but, showing the graph for Quantity Scaled for a given year superimposed on the CO2 graph. This graph shows that from roughly 1970 the number of storms per year is increasing more rapidly. This is reassuring because it fits with my hypothesis, but it is also unsettling because it means that future years will become more dangerous. See figure 8 in Appendix B.

After generating the superimposed graphs, I generated graphs showing the correlation between CO2 and the quantitative data from HURDAT. The first of these two graphs has CO2 as the x axis and quantity as the y axis. The trend line is a second-degree polynomial that is confusing. In the range of 2500 metric tons of CO2 (the lowest) and roughly 3600 metric tons of CO2 the quantity of storms will be less frequent as the CO2 emissions grow. This follows the conclusions extrapolated from previous graphs. Over time the amount of CO2 emissions has grown each year. During the years with decreasing CO2 emissions (i.e. further in the past) the quantity of storms was decreasing from 1950 to 1975. During the years with increasing CO2 emissions (i.e. further towards the present) the quantity of storms was increasing from 1975 onwards. This fits in with previous graphs, but it does not fit with my hypothesis. See figure 9 in Appendix B.

The second graph showing the correlation between CO2 and the quantitative data from HURDAT has CO2 as the x axis and the score as the y axis. The line of best fit for the data points resembles the line from the previous graph but more exaggerated. This line is also a second-degree polynomial. In the range of 2500 metric tons of CO2 to 3600 metric tons of CO2 the score decreases as the CO2 emissions grow. This mirrors the previous graph, but it does not make logical sense. From 3600 metric tons onwards, the score grows as expected. See figure 10 in Appendix B.

Results and Discussion

When starting this project, I expected to find that historically an increase in CO2 emissions would directly result in an increase in the number of storms for a given year and the total intensity of the storms for a given year. I was partially correct and entirely confused. HURDAT data from 1975 onwards showed the increases that I was expecting. For comparison, the 2020 Atlantic Hurricane season had 31 storms and a score of 64. This is in line with the general trend from 1975 onwards. The part of the data that confuses me is what happens from 1950 to 1975. The score and quantity both decrease during this time period. I have come up with multiple different theories about why this happens. The older data is the less accurate it becomes. It would follow that there would be fewer storms recorded and lower intensities recorded. However, this is clearly not the case. Also 1950 is not so long ago that I would be worried about substantial discrepancies in data. I trust the scientists at NOAA accurately compiled the data accurately and correctly. Upon realizing that nearly half of my data did not fit my hypothesis, I could have changed the time period to remove the data, but I decided to leave it. There is a decent possibility that I did all of my data mining and analysis correctly, and that the historical data is correct. If my report is correct, then I will leave it to someone who is more knowledgeable in meteorology to figure out why the data is the way that it is. While I am disappointed that my hypothesis was not completely correct, I am happy (sarcastic) to learn that climate change and extreme weather will continue to worsen in the coming years.

Appendix A

Figure 1: Preprocessing of Oxford data in python

import json

fileName = "owid-co2-data.json"

f = open(fileName,)

data = json.load(f)

co2 = []

for designation in data:

location = designation

year = ""

localCo2 = 0

globalPercent = 0

if location == "United States":

for a in data[location]:

for b in data[location][a]:

for c in b:

if c == "year" and b.get(c) > 1949 and b.get(c) < 2011:

year = int(b.get(c))

localCo2 = b.get("co2")

globalPercent = b.get("share\_global\_co2")

newEntry = []

newEntry.append(year)

newEntry.append(localCo2)

newEntry.append(globalPercent)

newEntry.append(localCo2 \* 1)

co2.append(newEntry)

strO = str(co2)

newO = ""

prevC = ""

for c in strO:

# prevC = c

if c == "," and prevC == "]":

newO += "\n"

elif c == "]":

prevC = c

# newO += ","

elif c == "[":

prevC = c

# continue

elif c == ",":

newO += " "

elif c == " ":

preC = c

else:

newO += c

prevC = c

prevC = c

f = open("co2\_output.txt", "w")

f.write(str(newO))

f.close()

Figure 2: Preprocessing of HURDAT data in python

import nltk.tokenize

import re

stormLevels = {"Extratropical Storm": 1, "Tropical Depression": 1, "Tropical Storm":1,"Hurricane - Category 1":2 ,"Hurricane - Category 2": 3,"Major Hurricane - Category 3":4, "Major Hurricane - Category 4":5, "Major Hurricane - Category 5":6, }

re\_StormTypes = "Tropical Storm|Hurricane - Category 1|Hurricane - Category 2|Major Hurricane - Category 3|Major Hurricane - Category 4|Major Hurricane - Category 5|Topical Storm|Tropical Depression|Extratropical Storm"

re\_Ops = "(\++)|(-)|(=)|(\\*)|(/)|(%)|(--)|(<=)|(>=)|(~=)"

re\_Storm\_Name = "Storm [a-zA-Z]+ \*is number \*\d+ of the year \d+"

re\_astrixs = "\\*+"

re\_tags = "Month Day Hour Lat\. Long\. Dir\. ----Speed----- -----Wind------ Pressure ------------Type-----------"

def strip\_Spaces(hFile):

retFile = []

for l in hFile:

if (l.strip() != ''):

retFile.append(l.strip())

return retFile

def strip\_Comm(hurdat\_File):

hurdat\_Sans\_Multi\_Line\_Comm =re.sub("/\\*[^\*]\*\\*+(?:[^/\*][^\*]\*\\*+)\*/","", hurdat\_File)

hurdat\_Sans\_Snlg\_Line\_Comm = re.sub("//.\*","", hurdat\_Sans\_Multi\_Line\_Comm)

hurdat\_Sans\_All\_Comm = hurdat\_Sans\_Snlg\_Line\_Comm

return hurdat\_Sans\_All\_Comm

def printArray(arr):

for line in arr:

print(str(line))

def highestLevelStorm(levels):

highestStr = ""

highestInt = 0

for stormIndex in levels:

storm = stormIndex[0]

if stormLevels.get(storm) > highestInt:

highestStr = storm

highestInt = stormLevels.get(storm)

retVal = [highestStr, highestInt]

return retVal

fileName = "HURDAT.txt"

fileObject = open(fileName, 'r')

HURDAT\_file = fileObject.read()

hurdat\_strip\_com = strip\_Comm(HURDAT\_file)

hurdat\_split = hurdat\_strip\_com.split('\n')

hurdat\_Sans\_White\_Space = strip\_Spaces(hurdat\_split)

hurdat\_Joined = '\n'.join([str(element) for element in hurdat\_Sans\_White\_Space])

hurdat\_lines = hurdat\_Joined.split('\n')

hurdat\_list = []

for line in hurdat\_lines:

hurdat\_list.append(line)

allStorms = []

newStorm = []

maxStormLevel = []

for line in hurdat\_lines:

# print(line)

if(re.findall(re\_Storm\_Name,line)):

if len(newStorm) != 0:

maxL = highestLevelStorm(maxStormLevel)

newStorm.append(maxL[0])

newStorm.append(maxL[1])

allStorms.append(newStorm)

newStorm = []

maxStormLevel = []

stormInfo = re.findall(re\_Storm\_Name,line)

tokens = nltk.wordpunct\_tokenize(line)

newStorm.append(tokens[1])

newStorm.append(tokens[4])

newStorm.append(tokens[8])

elif(re.findall(re\_astrixs, line)):

x = 0

elif (re.findall(re\_tags, line)):

x = 0

elif (re.findall(re\_StormTypes, line)):

stormLevel = re.findall(re\_StormTypes, line)

maxStormLevel.append(stormLevel)

maxL = highestLevelStorm(maxStormLevel)

newStorm.append(maxL[0])

newStorm.append(maxL[1])

allStorms.append(newStorm)

newStorm = []

maxStormLevel = []

yearInt = {}

yearCount = {}

for i in range(61):

yearCount[str(i + 1950)] = 0

for i in range(61):

yearInt[str(i + 1950)] = 0

for storm in allStorms:

yearInt[str(storm[2])] += storm[4]

yearCount[str(storm[2])] += 1

str00 = str(yearInt)

new00 = ""

for i in range(61):

new00 += str(i+1950) + " " + str(yearInt[str(i + 1950)]) + " " + str(yearInt[str(i + 1950)]\*150) + " " + str(yearCount[str(i + 1950)]) + " " + str(yearCount[str(i + 1950)]\*200) +"\n"

print(str(new00))

print(str())

strO = str(allStorms)

newO = ""

prevC = ""

for c in strO:

if c == "," and prevC == "]":

newO += "\n"

elif c == "]":

prevC = c

elif c == "[":

prevC = c

elif c == ",":

newO += " "

elif c == " ":

prevC = c

elif c == "'":

prevC = c

else:

newO += c

prevC = c

prevC = c

f = open("HURDAT\_output.txt", "w")

f.write(str(newO))

f.close()

f = open("HURDAT\_meta.txt","w")

f.write(str(new00))

f.close()

Note: I reused some of my code from the lexical analyzer from the Concepts course

Figure 3: Jupyter Notebook Python for reading in HURDAT data and displaying head

hurdatS = pd.read\_csv("HURDAT\_scaled.csv")

hurdatS = pd.DataFrame(hurdatS)

hurdatS.head()

Figure 4: Jupyter Notebook Python for reading in CO2 data and displaying head

co2 = pd.read\_csv("co2\_output.csv")

co2 = pd.DataFrame(co2)

co2.head()

Figure 5: Jupyter Notebook Python for merging HURDAT and CO2 data and displaying head

merged\_data\_s = pd.merge(hurdatS, co2, on='Year').dropna()

merged\_data\_s.head()

Figure 6: Jupyter Notebook Python for superimposing score and quantity by year

ax = sns.regplot(x="Year", y="Score\_Scaled", data=merged\_data\_s,order=4, ci=None)

ax2 = sns.regplot(x="Year", y="Quantity\_Scaled", data=merged\_data\_s,order=4, ci=None)

Figure 7: Jupyter Notebook Python for score per quantity

ax = sns.regplot(x="Quantity", y="Score", data=merged\_data\_s,order=3, ci=None)

Figure 8: Jupyter Notebook Python for score per year

ax = sns.regplot(x="Year", y="Score", data=merged\_data\_s,order=2, ci=None)

Figure 9: Jupyter Notebook Python for quantity per year

ax = sns.regplot(x="Year", y="Quantity", data=merged\_data\_s,order=2, ci=None)

Figure 10: Jupyter Notebook Python for CO2 per year

ax = sns.regplot(x="Year", y="co2", data=merged\_data,order=2, ci=None)

Figure 11: Jupyter Notebook Python for superimposing CO2 per year and Score Scaled per year

ax = sns.regplot(x="Year", y="co2", data=merged\_data\_s,order=4, ci=None)

ax2 = sns.regplot(x="Year", y="Score\_Scaled", data=merged\_data\_s,order=4, ci=None)

Figure 12: Jupyter Notebook Python for superimposing CO2 per year and quantity scaled per year

ax = sns.regplot(x="Year", y="co2", data=merged\_data\_s,order=4, ci=None)

ax2 = sns.regplot(x="Year", y="Quantity\_Scaled", data=merged\_data\_s,order=2, ci=None)

Figure 13: Jupyter Notebook Python for quantity per CO2 metric ton

ax = sns.regplot(x="co2", y="Quantity", data=merged\_data\_s,order=2, ci=None)

Figure 14: Jupyter Notebook Python for score per CO2 metric ton

ax = sns.regplot(x="co2", y="Score", data=merged\_data\_s,order=2, ci=None)

Appendix B

Figure 1: Head of merged data

Table

Description automatically generated

Note: See figures 3 through 5 in Appendix A for corresponding code.

Figure 2: Superimposed score and quantity by year using scaled y values

Chart, scatter chart

Description automatically generated

Note: Quantity is orange and score is blue. See figure 6 in Appendix A for corresponding code.

Figure 3: Score per quantity graph showing data points and trend line

Chart, scatter chart

Description automatically generated

Note: See figure 7 of Appendix A for corresponding code.

Figure 4: Data points and trend line for score by year

Chart, scatter chart

Description automatically generated

Note: See figure 8 in Appendix A for corresponding code

Figure 5: Quantity per year graph showing data points and trend line

Chart, scatter chart

Description automatically generated

Note: See figure 9 in Appendix A for corresponding code.

Figure 6: CO2 per year graph showing data points and trend line

Chart, scatter chart

Description automatically generated

Note: See figure 10 in Appendix A for corresponding code.

Figure 7: Superimposed score scaled per year and CO2 per year with data points and trend line

Chart, scatter chart

Description automatically generated

Note: CO2 per year is in blue and score scaled per year is in orange. The y axis values for CO2 are not displayed. See figure 11 in Appendix A for corresponding code.

Figure 8: Superimposed quantity scaled per year and CO2 per year graph showing data points and trend line

Chart, scatter chart

Description automatically generated

Note: CO2 per year is shown in blue and quantity scaled per year is shown in orange. The y axis values for CO2 are not shown. See figure 12 in Appendix A for corresponding code

Figure 9: Quantity per CO2 metric ton graph showing data points and trend line

Chart, scatter chart

Description automatically generated

Note: See figure 13 in Appendix A for corresponding code.

Figure 10: Score per CO2 metric ton graph shown with data points and trend line

Chart, scatter chart

Description automatically generated

Note: See figure 14 in Appendix A for corresponding code.

Sources

Hannah Ritchie and Max Roser (2020) - "CO₂ and Greenhouse Gas Emissions". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/co2-and-other-greenhouse-gas- emissions' [Online Resource]

National Oceanic and Atmospheric Administration. “Hurricane Database.” *HURDAT Easy Read*, <https://www.aoml.noaa.gov/hrd/hurdat/easyread-> 2011.html