Capstone Project 1 – Final Report:
Analyze Global Sea Ice,
CO2 levels and Forest Fires

Introduction

The chosen project aims to understand the worldwide connections between sea ice, CO2 levels and fire data. This is a topic that is included in the wider problem problem of global warming. The report is written for general audience therefore the analysis will avoid complex model that require a strong technical background.

The data is accessed from public sources and consists mainly in satellite data available for different time frames. The goal is to understand the relationships between different features and to make predictions on different dependent variables.

Overview of the Data Set

Linking the multiple sources for data-sets is challenging, because of the following reasons:

- the data consists of time series, therefore there is a trend, a seasonality and auto-correlation
- some data is missing
- the data format is different: csv, xlsx with many tabs, JSON
- time sampling is different for the different data sets; it is either days or months

Data Acquisition

The first step is Data acquisition. Details about data sources are available in the Annex but at a high level the data sources are:

- for sea ice: CSV files and xlsx file. The data is recorded daily and ranges from 1978 to 2019. The xls file has several tabs with data and needs to be converted to a flat file.
- for CO2, there is only one csv file. It is recorded monthly between 1958 and 2019.
- for fires, there are daily recordings from satellites as described in data wrangling file. The data is available between 2000 and 2019, in JSON format.

Data Wrangling

Data Wrangling consists in transforming the data into tidy Python data frames and cleaning the data. The main challenge is the size of the forest fires data (hundreds of GB).

Fire data has several issues. There is missing data especially for the first years of recordings related to issues with satellite instruments. This data was linearly interpolated by using the values before and after the missing data. In the JSON files used as inputs, there are some missing quotas on a few fields and this issue had to be fixed before importing the data in the Pandas data frame. Furthermore, there are two data sets with data acquired from two different satellites and there are different field names and different units of measurement. The data was consolidated into two data frames, one for each satellite. The columns unrelated to direct measurements were dropped.

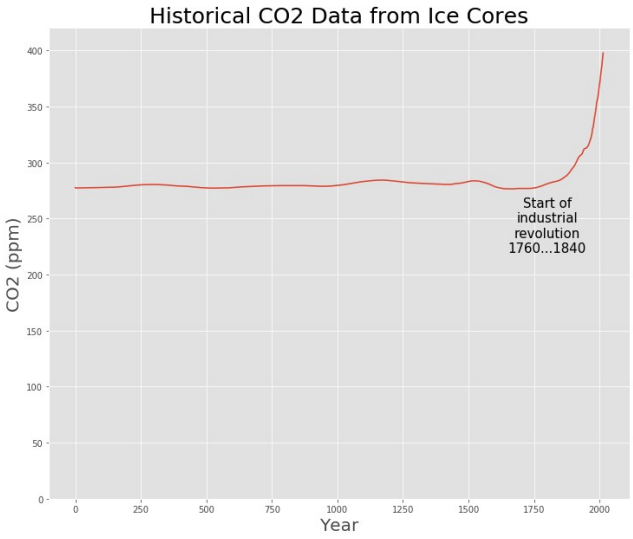
CO2 data is in a csv format and there were no issues while reading or interpreting this file.

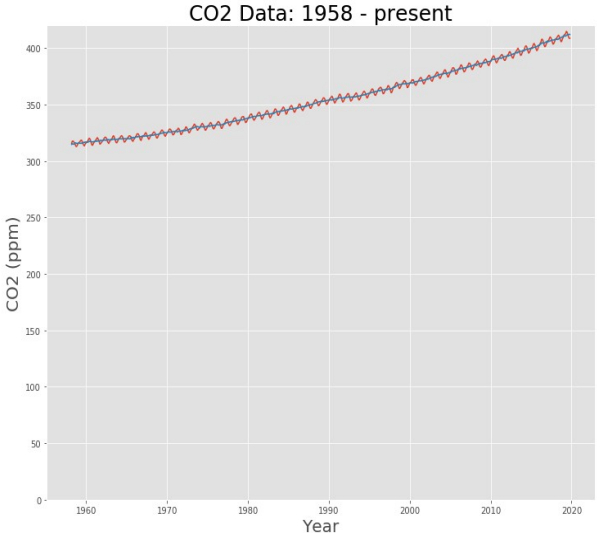
Ice data sources are Excel files with North and South ice daily data and daily data for all the Northern Seas. The data was consolidated into one Pandas DataFrame object in order to analyze the correlations between features and observations. As for fire data, there were several incomplete observations that had to be dealt with.

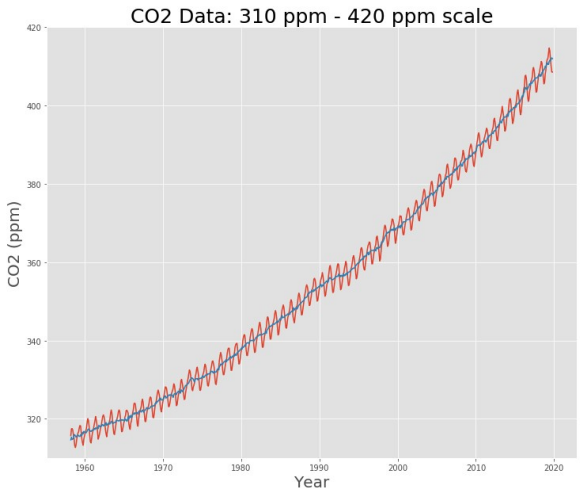
Exploratory Data Analysis

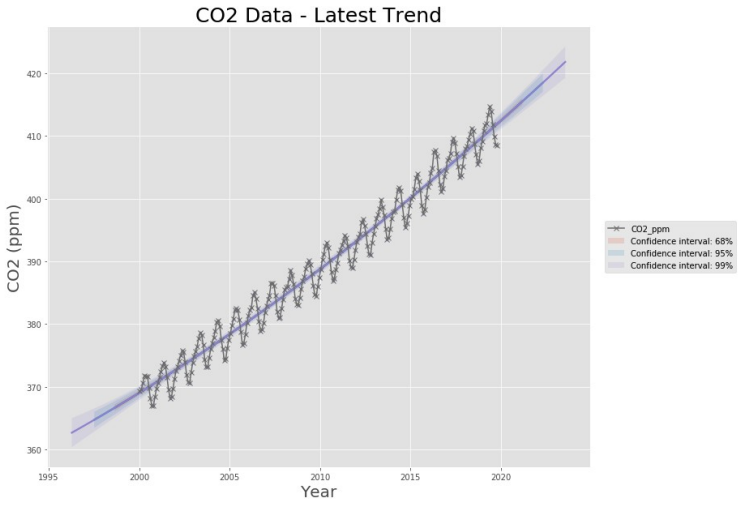
This stage is a preliminary analysis of the clean data sets for a better understanding of the features and the data trends and correlations.

CO2 data is easier to analyze. As shown below, hystorical CO2 concentration measured from ice cores was relatively constant until the industrial revolution. Exploratory Data Analysis (EDA) will focus on the recent period (1958 – present) where the data has higher confidence.



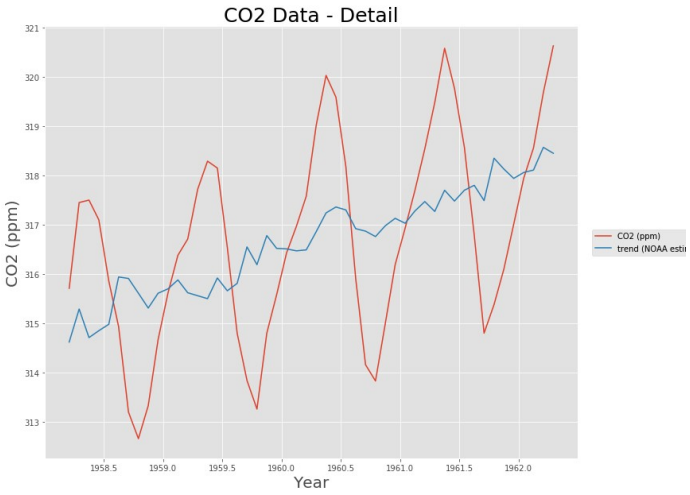


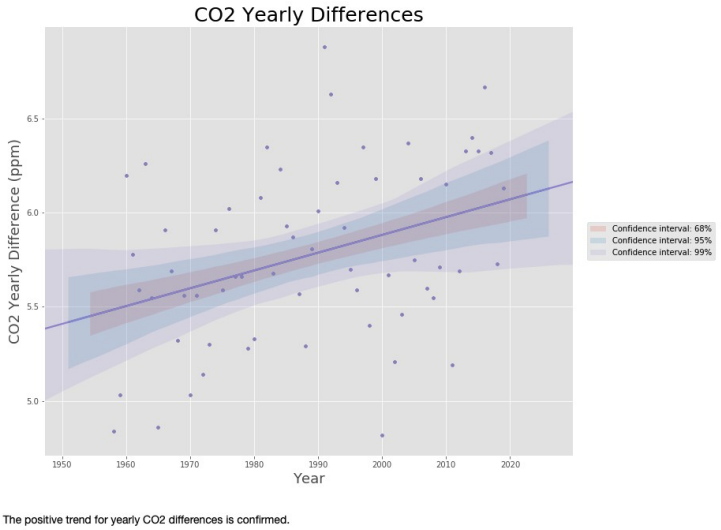




As shown above, on a historical scale CO2 concentration becomes almost a vertical line.

There is an upward trend and a seasonal variation. The seasonal variation is caused by the different land mass covered with forests for the Northern and Southern Hemisphere. For a statistical analysis, I need to examine the seasonal component (annual min and max together with the overall trend).

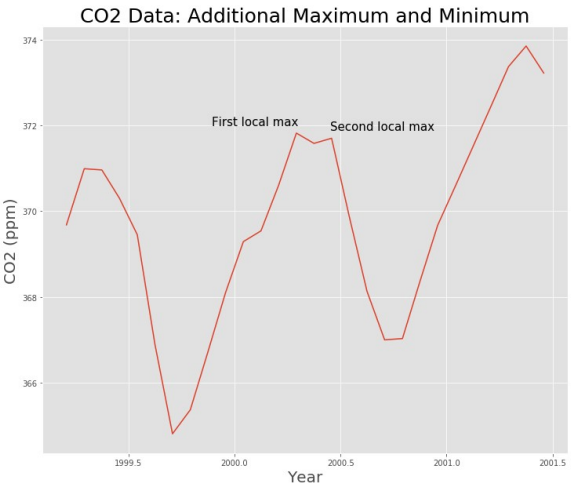


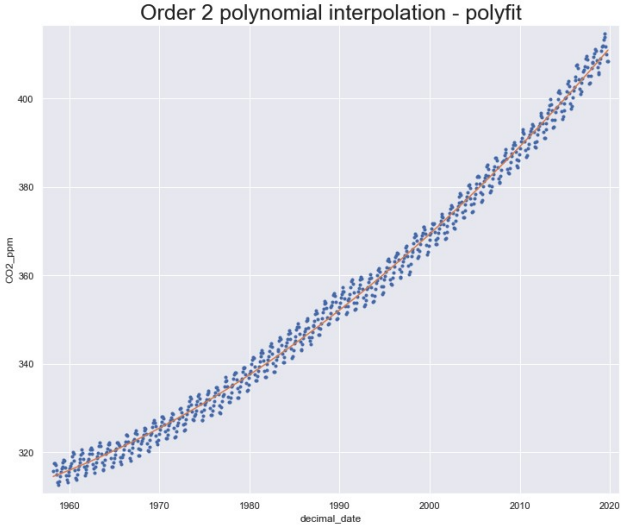


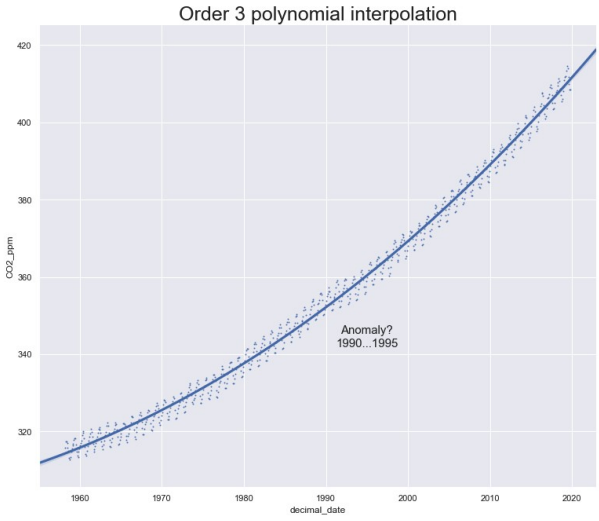
The positive trend for yearly CO2 differences is confirmed.

Data sampling and measurement errors may create false local maximums and minimums as shown in the diagram. These false extremes must be ignored. Using the correct minimums and maximums, it looks like the yearly difference between these local extremes is increasing.

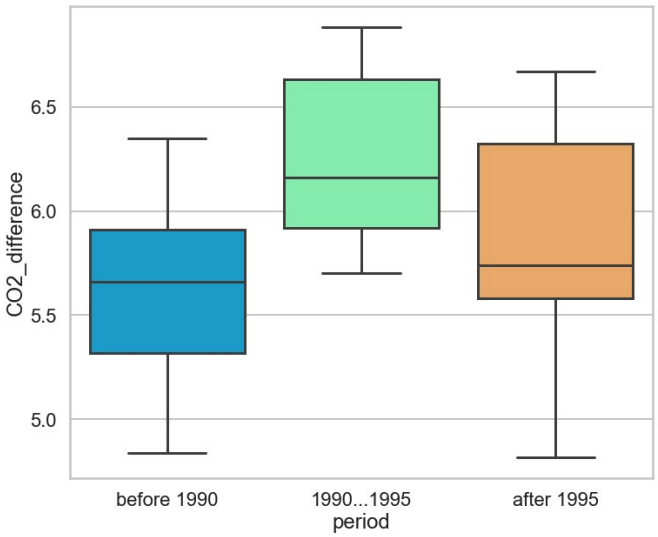
There is an upward trend with a positive second derivative so a polynomial fit should have at least order 2.

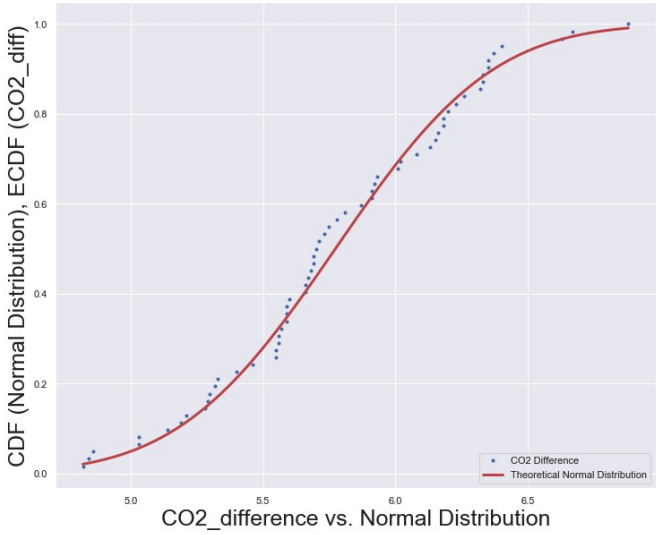




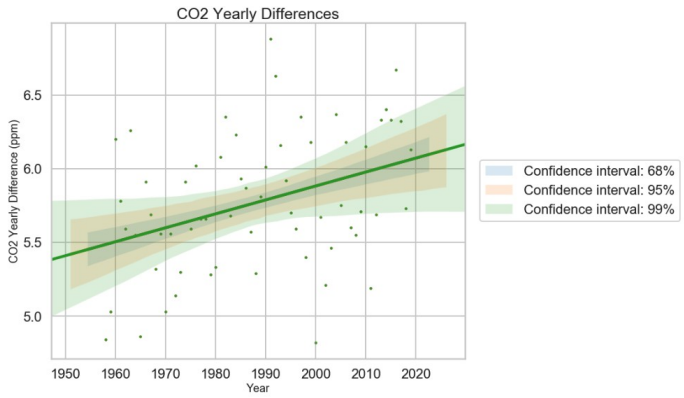


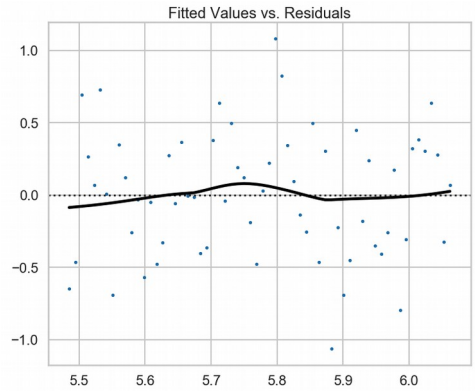
There is no visible difference between order 2 and order 3 interpolation and going to higher orders may end up in overfitting. There is a curious anomaly around 1990…1995 CO2 data, it looks like the CO2 emissions did not increase as fast as before. This anomaly must be investigated.
From the diagram above, the CO2 ECDF follows more or less the theoretical CDF for a normal distribution. However, here is a curious anomaly around 1990…1995 CO2 data, it looks like the CO2 emissions did not increase as fast as before. The box plots show a clear difference for the





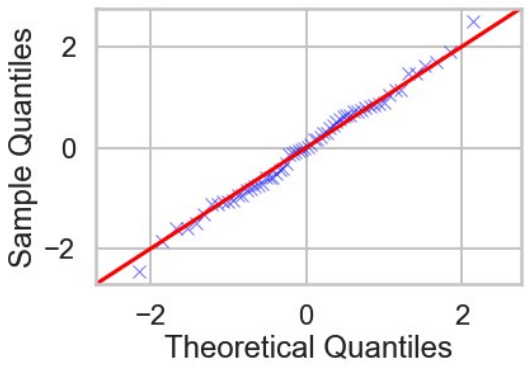
1990…1995 period. It looks like the data is not symmetrical, the position for the median is close to the 3rd quartile before 1990 and close to the first quartile after 1995. I am showing again the yearly CO2 difference in order to compare it with the fitted against residuals diagram. The residuals look good, with a slow increase between 5.5 and 6.3 and a slow decrease around 5.8. In the diagram at the , the 5.8 value is exactly in the period 1990…1995 so the 1990…1995 anomaly is real.

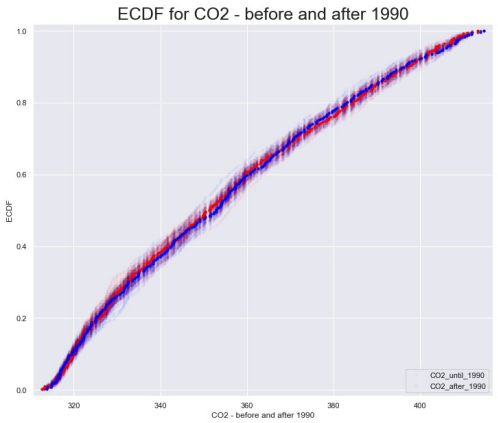




As shown above, the theoretical quantiles align well with the sample quantiles so the CO2 yearly difference distribution is normal.

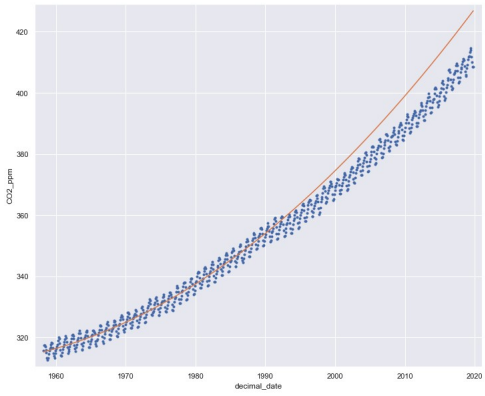
I am defining H0 by stating that the standard deviations of the two distributions before 1990 and after 1995 are identical. From the discussion above they might be different so H0 might be rejected but only by using statistical models. Using poly fit order 2 with the data until 1990 and extrapolating until 2020 shows that CO2 increase slowed down after 1990.

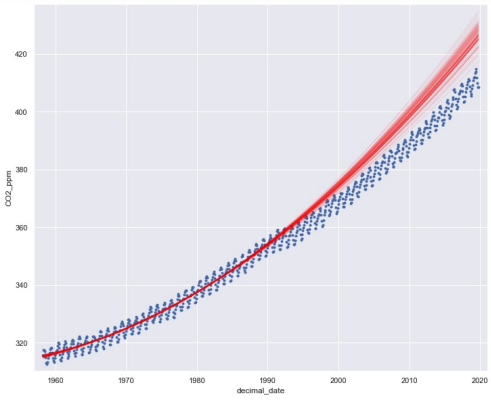




To the left I am showing the order 2 polynomial interpolation for CO2 data until 1990 and extrapolation for the rest of the data. While the fit until 1990 looks good, due to the 1990...1995 anomaly the curve starts to diverge more and more from the observed data.

A better way to explore the variance inherent in the data is to use bootstrap replicates. I am showing 1000 bootstrap replicates calculated only with the data until 1990:

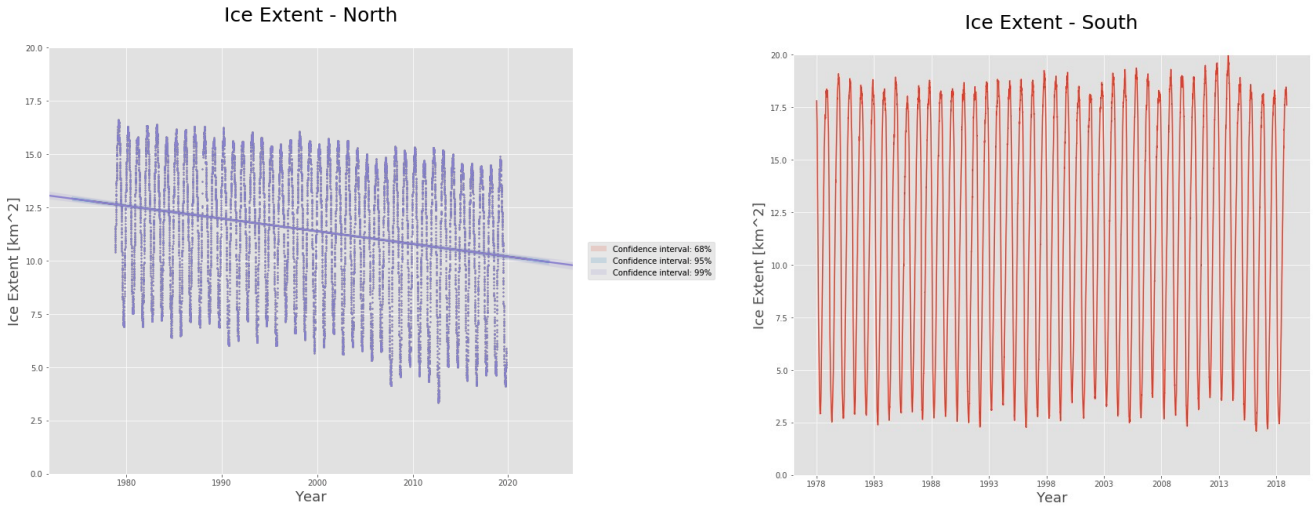




Even by choosing the lowest curve from the 1000 bootstrap replicates, the predicted CO2 value for 2020 (417.39 ppm) is higher than the maximum CO2 value measured so far (405.12 to 414.66 ppm in 2018 and 2019).

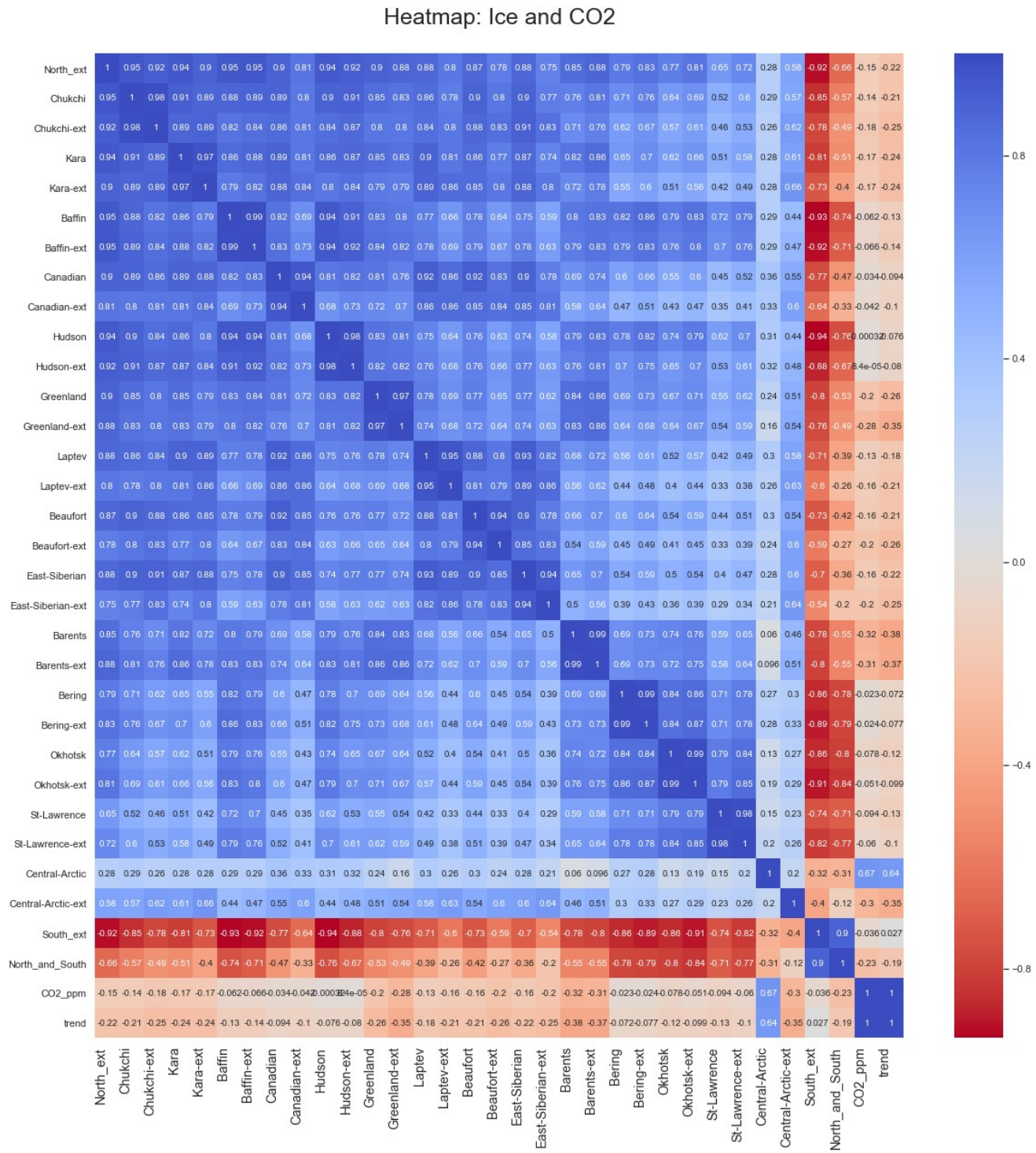
This should make us believe that H0 can be rejected. However, for rejecting H0, ECDF must be calculated for the data before 1990 and after 1995.

However, there is no obvious difference between the two ECDF curves and the bootstrap replicas created from the full dataset. Furthermore, p-value is calculated as 1.0. Therefore, H0 cannot be rejected if the statistic test consists of the standard deviation. There might be other tests that would give different results. Actually, the trend did not change, the 1990…1995 period is just a delay in the long-term CO2 trend.



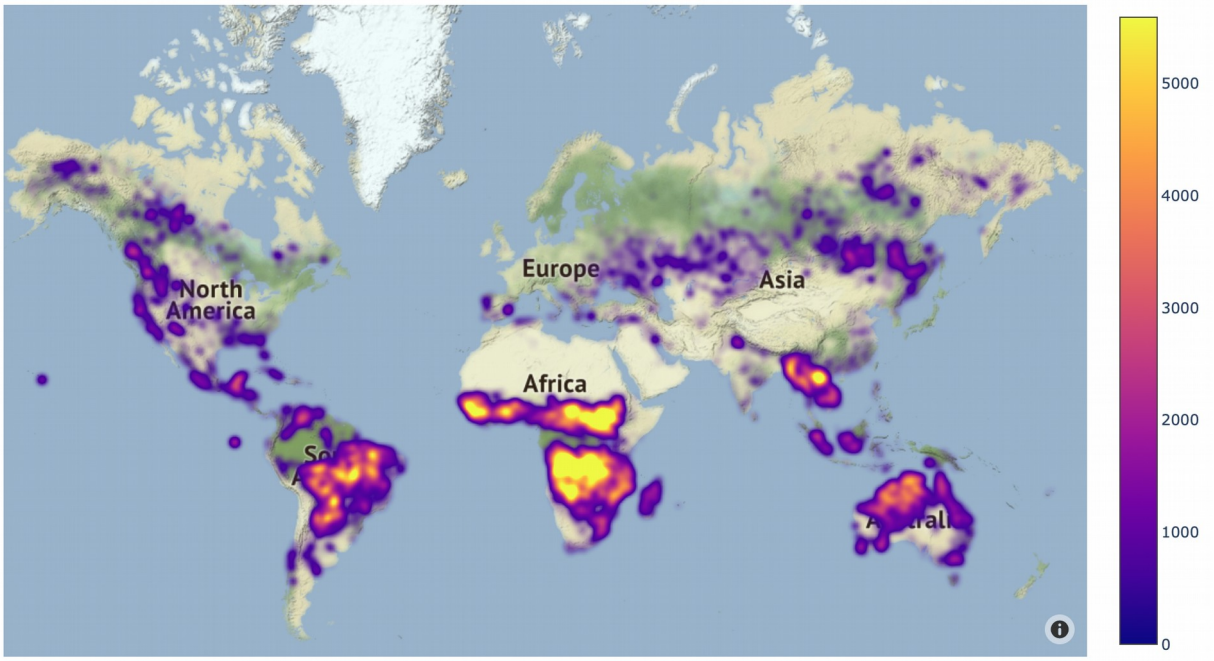
By adding all the Northern Seas to the North and South ice extent, I created a heat map ordered by Pearson correlation. As a metric I used the absolute value of L1. All the ice extents from the Northern Seas are positively correlated. Chukchi Sea is also highly correlated to the North extent, because it is just North of the Bering strait and there is a lot of ice movement around that area. At the other end, St. Lawrence and Central Arctic are less correlated because there are either no currents around (Central Arctic) or a very slow ice and water movement (St. Lawrence). As expected, North and South ice extents have the highest (negative) correlations.

The added feature “North and South” just adds up the data for North and for South. There is a strong correlation of 0.9 between South ice and North and South. This means that 90% of the variance of the total ice is explained by the variance of the ice on the South. This is to be expected, taking into account the huge ice sheet of Antarctica.

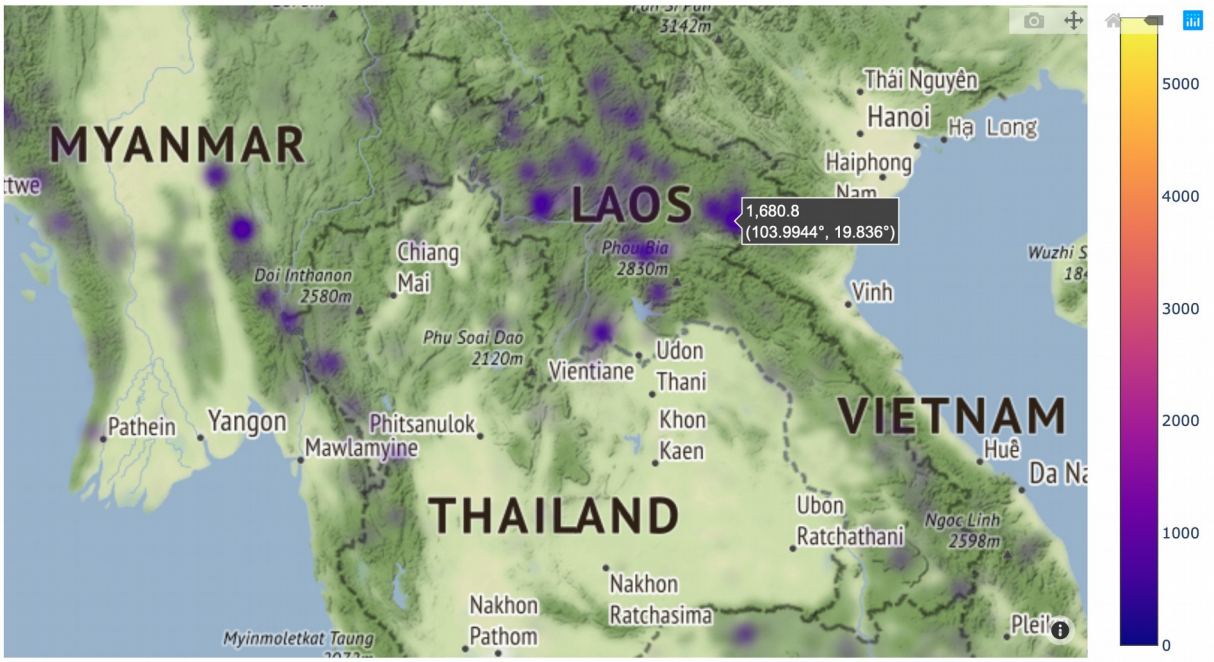


Since the Northern sea ice decreases during summer when the CO2 levels have a seasonal decrease as well, the correlation values are positive but with lower values than the correlation between different seas. A strange exception is Central Arctic.

The last type of data is related to forest fires. A quick look on 20 years of forest fires data is shown below. It is interesting to see that the forest fires in Africa were much more intense than the fires in the Amazonian forest or the ones in Australia. The data does not include the Australian forest fires from December 2019 and January 2020. Data acquisition ended in October 2019. Plotly library allows showing on the map any features for observations that have a longitude and latitude. A world map for forest fires created from the processed data is shown below:

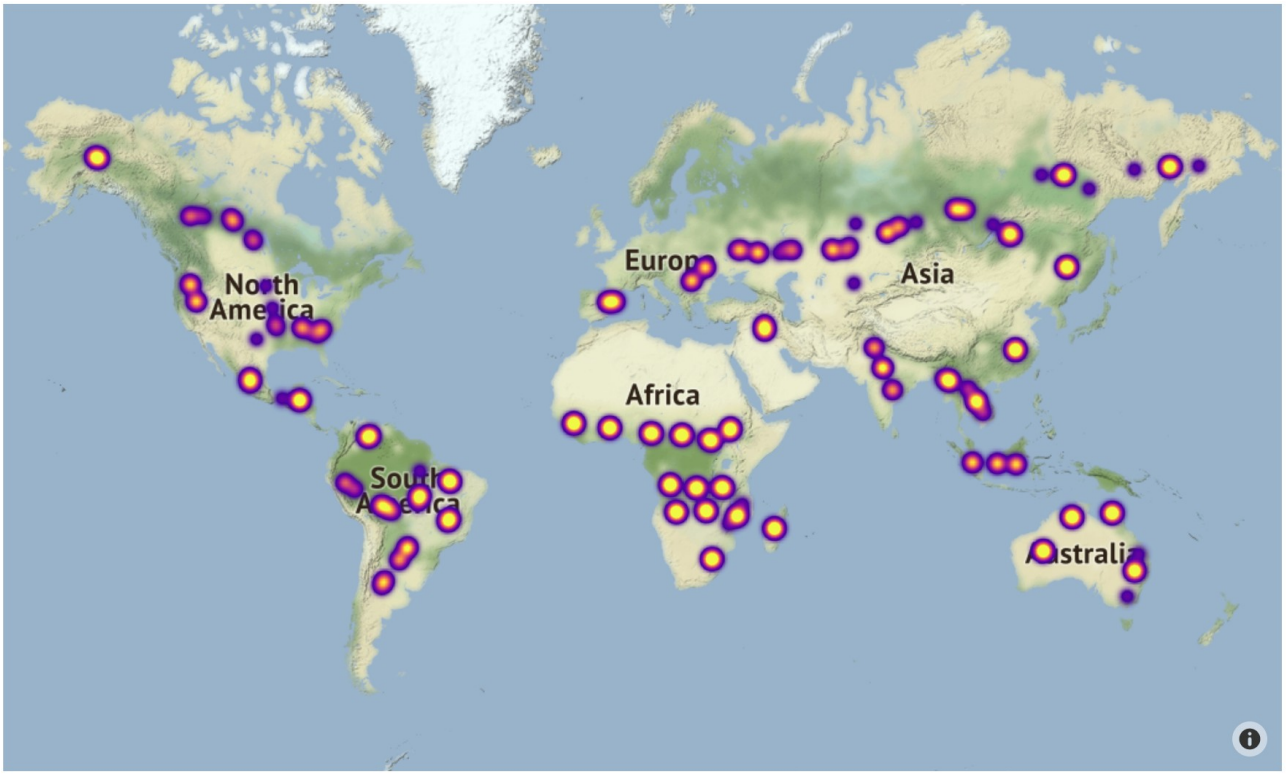


Zooming in can provide useful insights regarding the data, before statistical analysis. For example, this is a detail for South-East Asia:

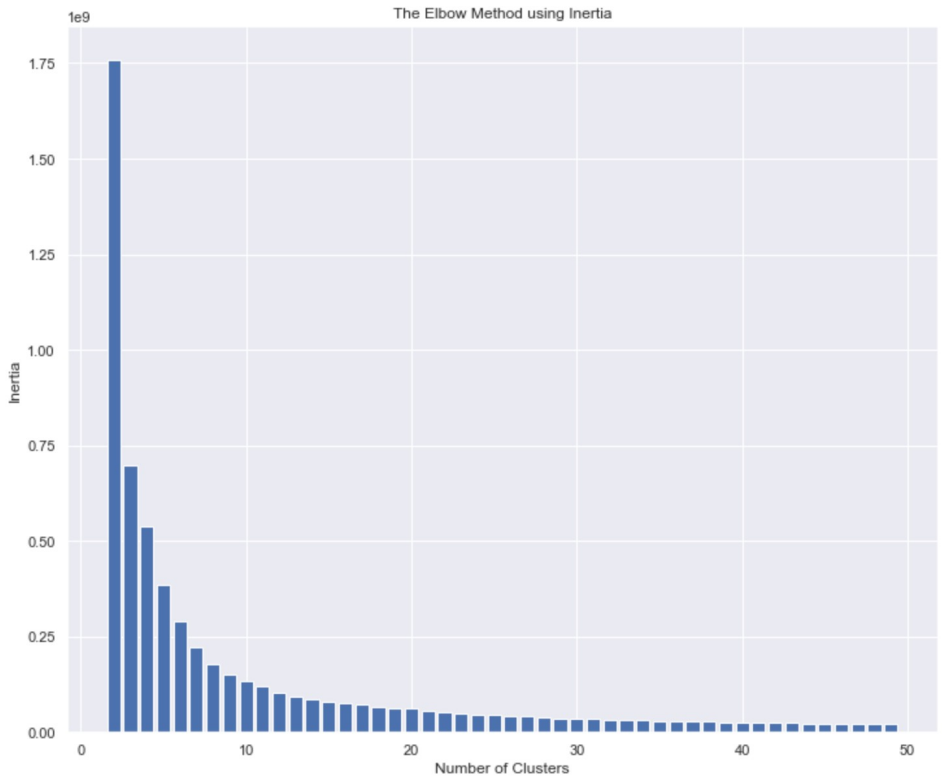


The interactive plot shows that the highest radiative power (1680.8 W/m2) is located in a small area in the North-East of Laos. Future analysis will focus on trends of these forest fires and correlations to CO2 increase and ice loss worldwide.

I am using K-means clustering for grouping the fores fires in clusters and calculate the cluster radiative power (frp). For 50 clusters the centroids are shown on the map:



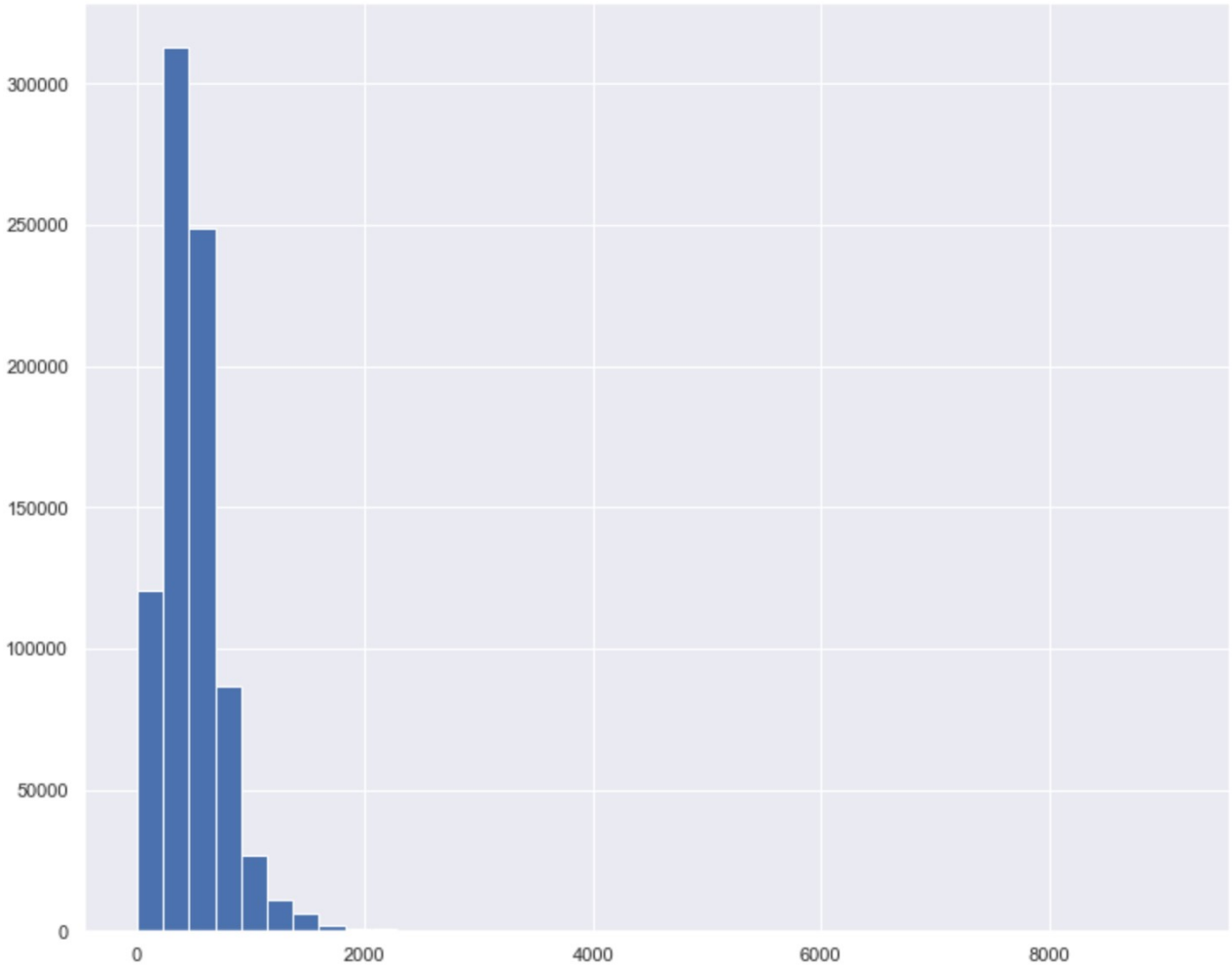
The size and color of the centroids are proportional to frp per cluster. This is a good match with the overall forest fires (previous page). The most important parameter for K-Means clustering is the number of clusters. This number must be chosen for minimizing the chosen loss function (in this case, the inertia parameter). I am showing below the inertia for different numbers of clusters:



Other possible options for the metric might be the median instead of the mean and discarding the fires that are far away from any determined centroids and are not often in the time series. This would eliminate many fires in remote islands that increase artificially the inertia.

It is difficult to chose the “best” number of clusters because there is no “elbow” where the loss functions starts to become flat. Such a point might be k=10 but for a more detailed analyze I will chose k=43.

For the 43 clusters, the histogram for the distances to the cluster center is shown below:



The mean is 471 km and the median is 424 km.

I am now using the clusters for creating a wide-format data frame with clusters as rows, years as columns and the content of the cells representing the frp per cluster for the chosen year.

Unfortunately, this processed data does not show any trend; there is a lot of noise with total radiative power varying randomly, more that 5 times from one year to another. With current data, a statistical analyze cannot give any clue on the forest fires forecasts even at global level.

Conclusion

The purpose of this project was to understand the relationships between CO2, ice and forest fires on a global scale, without using very complex models. It seams that there is a correlation between ice and CO2 but the project does not infer any causality between these. For the forest fires, the situation is more complicated. There is a lot of noise in the data. There may be missing data, wrong readings due to sensors (increased non-linearity towards the edges of the satellite sensors).

Furthermore, there were lots of features that were not taken into account due to lack of resources and reliable data:

- wind patterns
- humidity in the atmosphere
- temperature
- deforestation that has significant values during the last decades
- direct human intervention like fires for clearing the land for agriculture
- etc.

However, the project shows some ideas for data cleaning and data visualizations for the general audience. It gives some hints on the scale of the changes especially for CO2 and ice cover for the Northern Hemisphere. More features and better models could bring a better picture about the relationships between CO2, sea ice and forest fires.

Annex - References for Datasets

Sea Ice Data (daily):

Fetterer, F., K. Knowles, W. N. Meier, M. Savoie, and A. K. Windnagel 2017, updated daily. Sea Ice Index, Version 3. [subset Sea_Ice_Index_Regional_Daily_Data_G02135_v3.0]. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. doi: https://doi.org/10.7265/N5K072F8. [Date Accessed].
As specified in the User Guide for the data, “On monthly extent images, ice ends and water begins where the concentration estimates of grid cells for that month drop below 15 percent.”
Data downloaded from:
ftp://sidads.colorado.edu/DATASETS/NOAA/G02135/north/daily/data/
N_seaice_extent_daily_v3.0.csv
ftp://sidads.colorado.edu/DATASETS/NOAA/G02135/north/daily/data/
S_seaice_extent_daily_v3.0.csv
ftp://sidads.colorado.edu/DATASETS/NOAA/G02135/seaice_analysis/
Sea_Ice_Index_Regional_Daily_Data_G02135_v3.0.xlsx

CO2 Data (monthly):

C. D. Keeling, S. C. Piper, R. B. Bacastow, M. Wahlen, T. P. Whorf, M. Heimann, and H. A. Meijer, Exchanges of atmospheric CO2 and 13CO2 with the terrestrial biosphere and oceans from 1978 to 2000. I. Global aspects, SIO Reference Series, No. 01-06, Scripps Institution of Oceanography, San Diego, 88 pages, 2001.
Data downloaded from: scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/monthly/monthly_in_situ_co2_mlo.csv

Forest Fires (daily):

NRT VIIRS 375 m Active Fire product VNP14IMGT. Available on-line [https://earthdata.nasa.gov/firms]. doi: 10.5067/FIRMS/VIIRS/VNP14IMGT.NRT.001.
MODIS Collection 6 NRT Hotspot / Active Fire Detections MCD14DL. Available on-line [https://earthdata.nasa.gov/firms]. doi: 10.5067/FIRMS/MODIS/MCD14DL.NRT.006
Data downloaded from: https://firms.modaps.eosdis.nasa.gov/download/