American Journal of Electronics & Communication

Vandana Jagtap et. al., American Journal of Electronics & Communication, Vol. III (2), 1-7

# **Climate Change**: *Preliminary analysis and prediction using global temperature data.*

Vandana Jagtap[1] , Yash Rathore[2] , Sarthak Oke[3] , Parth Gawande[4] , Pooshan Singh[5]

Department of Computer Science and Engineering, MIT-WPU, Pune, India

***Abstract*—** Little has to be said when the global populus is experiencing climate change firsthand. But relations have to be established between organic data and the visible change if we must understand it better. Except for strengthening the arguments in favor of working against climate change, it also ensures we use the right tools to do so. Using the UN's Global Temperature change data. Which we first clean and restructure it to our ease. The aim is to understand in depth, the patterns and relationships in temperature rise over time. Next logical thing to do is co-relate it with rising CO-2 emissions, loss of forest cover, Air Quality index as factors. Linear and Polynomial Regression is used to train our data and get temperature rise predictions till 2059. The forecast is executed utilising 2 and 5 degree polynomial regression.

***Keywords—temperature change, Regression, Standard Deviation, Pandas, NumPy, sklearn***

## I. INTRODUCTION

FAOSTAT's Temperature Change dominion promulgates yearly statistics of mean surface temperature change by country. The period of 1961-2019 has been covered in this distribution. We have monthly, seasonal and annual anomalies in change of temperature i.e. changes that occur against climatological baseline, ranging from 1951-1980. standard deviation of the baseline method in temperature change is also given. Datapoints have been set in accordance with publicly available GISTEMP data. NASA's Goddard Institute for Space Studies administers the Global Surface Temperature Change data

## **II. L**ITERATURE REVIEW

| Sr No. | Author(s) | Year of Publication | Title | Techniques used | Research Gap |
|---|---|---|---|---|---|
| 1 | Eric Wolff FRS, (UK lead), University of Cambridge Inez Fung (NAS, US lead), University of California, Berkeley<br><br>Brian Hoskins FRS, Grantham Institute for Climate Change<br><br>John F.B. Mitchell FRS, UK Met Office | 2020 | Climate Change: Evidence and causes | RFC<br><br>SVM | Absent Standard Deviation results |
| 2 | Mukhtar Ahmed | 2020 | Introduction to Modern Climat | RFC<br><br>SVM<br><br>Regression | Noisy global average plots |

| | | | | | |
|---|---|---|---|---|---|
| | | | e Change. Andrew E. Dessler: Cambridge University | | |
| 3 | Brain O'Neill<br><br>Timothy R. Carter<br><br>Kasper Kok<br><br>Paula A. Anderson | 2020 | Achievements and needs for the climate change scenario framework | SSP-RCP Matrix | Integrated<br><br>scenarios give rise to indiscernible readings |

## III. PROPOSED METHODOLOGY

Best way to understand patterns in such datasets is in tranches. Thus our charted course includes namely:

***A.*** Import required libraries

***B.*** Load data and Preliminary Analysis

***C.*** Cleaning

**D.** Antarctica as a Case Study

**E.** Test-Train Split (Linear and Polynomial Regressions)

**G**. Results and Discussion

***A.*** We'll primarily make use of the libraries below to attain our task:

- *Pandas*
- *Matplotlib*
- *SNS Seaborn*
- *NumPy*
- *sklearn*
- *Pipeline*

```
#import libraries
import pandas as pd
import datetime
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import warnings

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import Pipeline
from sklearn import metrics
```

*Figure 1.1*

***B.*** Cleaning the dataset is an essential step as irregularities and false entries contribute to an inaccurately trained model.

| Country | Months Code | Months | Element Code | Element | Unit | Y1961 | Y1962 | Y1963 | ... | Y2010 | Y2011 | Y2012 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Indonesia | 7012 | December | 7271 | Temperature change | °C | 0.191 | -0.066 | 0.221 | ... | 0.063 | 0.416 | 0.570 |
| Austria | 7005 | May | 7271 | Temperature change | °C | -1.180 | -1.586 | -0.047 | ... | 0.531 | 2.037 | 2.207 |
| Serbia and Montenegro | 7001 | January | 7271 | Temperature change | °C | NaN | NaN | NaN | ... | NaN | NaN | NaN |

*Figure 1.2*

A heatmap (figure1.3) is generated to highlight the present **NaN/null** values so they can be removed.

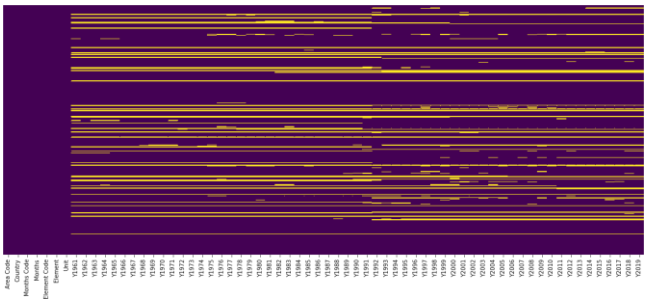


*Figure 1.3*

Dropping the rows containing these, we move ahead. Additionally, we also delete any month wise groupings with other unnecessary columns. (area, element & month codes, etc.). In order to work with simpler manipulations all the years are moved to a single column which barely affects memory usage although

we have more rows now. DF 'temperature_change' is created to house these points.

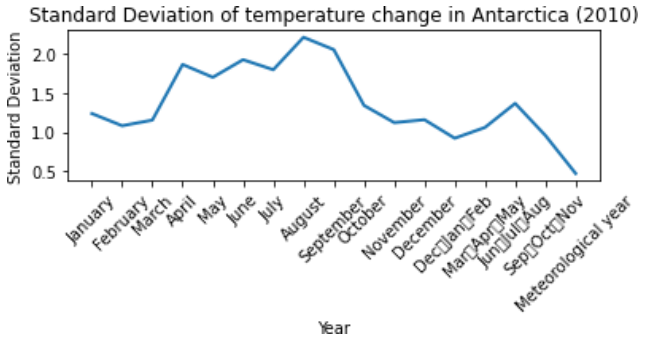
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2006 entries, 0 to 2005
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Country             2006 non-null   object
 1   Months              2006 non-null   object
 2   Element             2006 non-null   object
 3   Year                2006 non-null   object
 4   temperature_change  2006 non-null   float64
dtypes: float64(1), object(4)
memory usage: 78.5+ KB
```

*Figure 1.4*

***C.*** The country wise grouped data allows us to further analyse climate change patterns in individual regions/countries. Learnings from these can be later applied to gain better insight into inter-regional relationships.

*Figure 1.5*

***D.*** An SNS line plot shows average monthly temperature change for 4 different years. We can further analyse the overall impact by considering monthly Standard Deviation so we can deduce possible minute changes over a larger spread.




- Summer 2010 in the Southern Hemisphere began on Wednesday, 22 December. Least deviation occurs in the following months. The winter months show a larger spread where July harboured highest. Standard Deviation is calculated by looking through the temperature of every month in Antarctica for each year. If made over a larger spread, we see the deviation to turn positive.

- To further extract incrementing patterns, we use a scatter plot to have their means help us understand the magnitude of rise in mercury.

(All temperature values in °C)

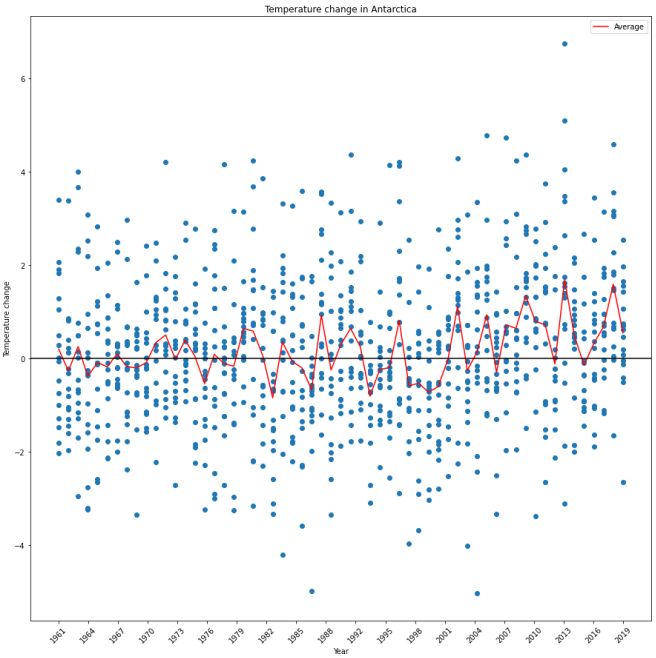



Figure 1.6

A clearer relationship between the mean temperature and its changes can be seen through histograms. Antarctica shows higher numbers than the world baseline, as average temperature change also increases every year. Insights gathered from the case study are used to jot down parameters on which the analysis focuses.Namely the density of global temperature rise vs avg ideal values, Mean temperature change around the world and the overall gains taken place. Then we move onto training our model using the dataset and get predictions for the imminent rise using Linear and Polynomial Regression. Regression is one of the method to assess the required component.

The linear regression defines $X\text{trend}(i)$ by using two parameters, one is the intercept, $\beta 0$, and the other is slope, $\beta 1$.

The model is “on the process level” given by

$$X(i)= \beta_0 + \beta_1 \times T(i)+S(i)\times X_{\text{noise}}(i)$$

***E.*** When we move onto every other country in the dataset, we need to avoid skewing calculations.

Thus we drop any country groupings present. Another column consists of the year and the corresponding change in temperature is added. If needed later, we make a new dataset containing just the regions. We calculate the average global change in temperature and move it under a new dataframe. Similar operations are performed for each country. A scatter plot can be prepared with these data frames showing temperature change for all the countries over years and the world average.

| | Country | Months | Element | Year | temperature |
|---|---|---|---|---|---|
| 0 | Afghanistan | January | Temperature change | 1961 | 0.777 |
| 1 | Afghanistan | January | Standard Deviation | 1961 | 1.950 |
| 2 | Afghanistan | February | Temperature change | 1961 | -1.743 |
| 3 | Afghanistan | February | Standard Deviation | 1961 | 2.597 |
| 4 | Afghanistan | March | Temperature change | 1961 | 0.516 |

(The word “data” is plural, not singular.)

Distribution data is now ready to be plotted. The histogram below shows a similar density distribution plot as Antarctica.

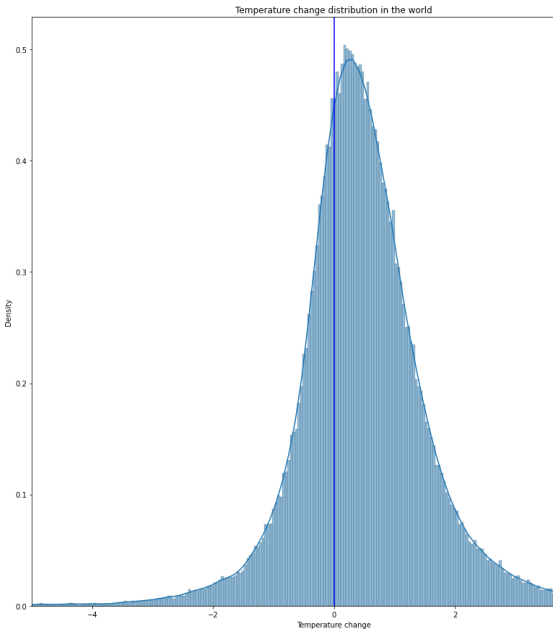



Figure 2.1

We calculate the average global change in temperature and move it under a new dataframe. Similar operations are performed for each country. A scatter plot can be prepared with these data frames showing temperature change for all the countries over the years and the global average. The global average can also be cast on top of the temperatures of each country.

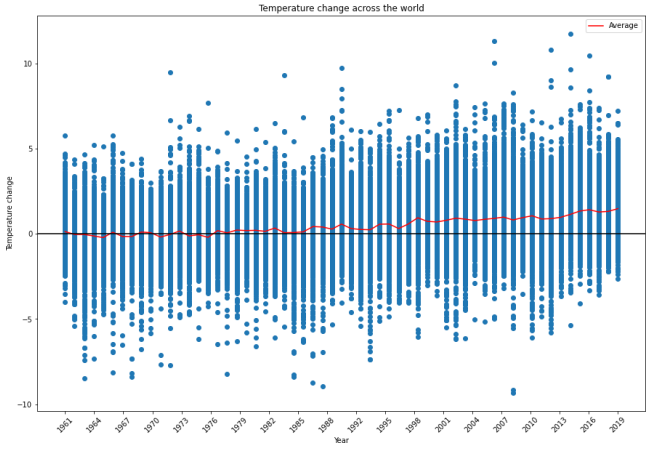



Figure 2.2

a. Test-Train split

We split the data into traditional tests and train datasets to verify our model's predictability.Linear Regression’s residuals against time are portrayed. The patterns suggest residuals being structured, these trends being:

long-term and

Relatively shorter terms such as positive year-round residuals throughout 1960. This makes the linear model really unsuitable for a GISTEMP time series. Root Mean Squared Error (RMSE) is a measure of how far from the regression

line our predicted data points lie. A high concentration of these around the baseline indicates a good fit. We can now use the entire dataset to train the model.

We can now use the entire dataset to train the model.

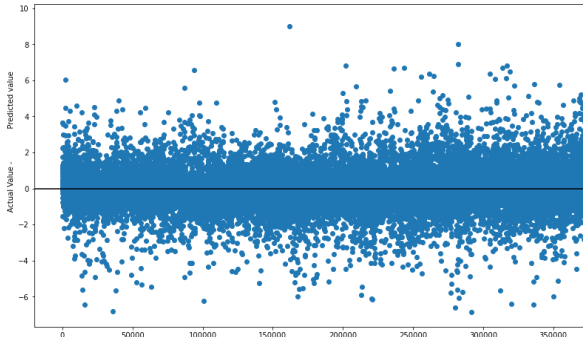



Figure 2.3

The ordinary least-squares (OLS) estimation minimizes the sum of squares of differences between data and linear fit. Predicted to actual value comparison density plot shows most data points nearing

We use the same logic to test the trained data under Polynomial Regression. The obtained values are visualised in the grouped graphs showing both the compiled temperature changed data, linear as well as degreed regressions.

Predicted actual value comparison density plot shows most data points nearing the baseline. We use the same logic to test the trained data under Polynomial Regression. The obtained values are visualised in the grouped graphs showing both the compiled temperature changed data, linear as well as degreed regressions.

## IV. RESULTS AND DISCUSSION

```
RMSE: 0.9721743424324878
Text(0, 0.5, '                    Actual Value -                    Predicted
value')
```

Figure 2.4

Root Mean Squared Error (RMSE) (Figure 2.4) is a measure of how far from the regression line our predicted data points lie. A high concentration of these around the baseline indicates a good fit.

We can now use the entire dataset to train the model.

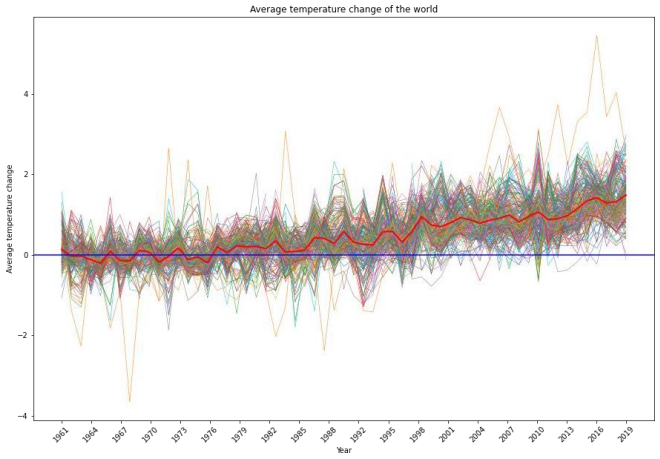



Figure 2.6

Prediction points are randomly generated and sorted by values as strings in order to get the test data in the primary dataframe, giving us the final results. The RMSE score reads 0.9721.

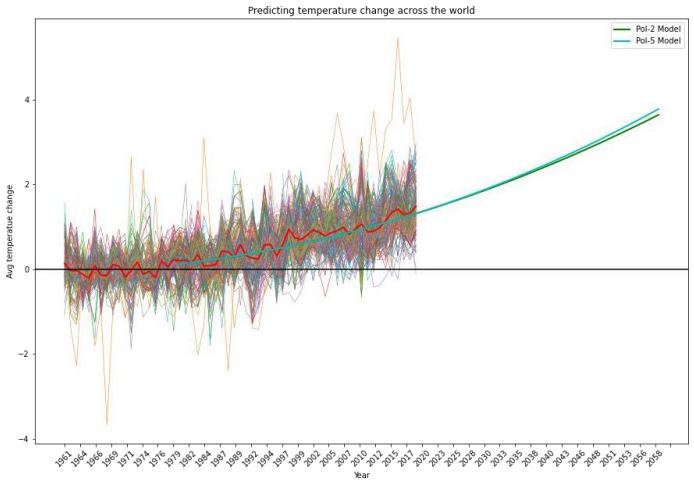



Figure 2.7

The learnings we get from preliminary results are utilised to devise detailed plans for further data wrangling, analysis, and visualisations.

We use these results to try and first find the global average temperature change so we can visually second the increasing trends earlier noticed.

Linear and Polynomial Regression models were trained with the dataset to realise a prediction of temperature change in the foreseeable future. Final inference indicated the global average above 4°C, unsurprisingly higher than the desired 1.5°C mark.

The thresholds we seem to be breaking with ease remain to inch us closer to the imminent point of no return. COVID-19 brought a transient phase of better air and water qualities, which was caught sight of globally. The immediate effects of lockdown on air quality can be seen through reduced levels of the 5 criterion 'very unhealthy' pollutants (namely, $PM_{10}$ and $PM_{2.5}$, $SO_2$, $NO_2$, and CO) depending on the

particulate size. These 5 constitute to formulate an Air Quality Index for any region. Indian Institute of Tropical Meteorology, Pune holds hourly AQI by aggregating values of every region into a sole index.

A 24 hour mean can be thus calculated and viewed. Each pollutant's index is correlated to a unique sub-index thus resulting in an association between the concentration of pollutants and their impact. Thus,

$$Ii = f(Xi), i = 1,2,\dots,n$$

A function for our corresponding sub-indices would read to be:

$$I = F(I, I2, \dots, In)$$

To garner a relation between a pollutant's concentration and subsequent upshot in ecology:

$$I = \alpha X + \beta \ (\alpha = line's\ slope)$$

When segmented with each sub-index (Ii) for a concentration ($C_p$) we can calculate these points to aggregate them later.

$$I_i = [\{(I_{HI} - I_{LO}) / (B_{HI} - B_{LO})\} * (C_p - B_{LO})] + I_{LO}$$

$B_{HI}$ occurs when the breakpoint concentration exceeds known value and $B_{LO}$ when it's lower. $I_{HI}$ and $I_{LO}$ are AQI's equivalent to $B_{HI}$ and $B_{LO}$. If we substitute this equation in a root -sum- power form we have an aggregated index, giving us a daily reading of our sub indices**.**

$$I = Aggregate\ Index = [\sum I^p]^{(1/p)} \ \dots (p>1)$$

and the RM Square form gives us:

$$I = [1/k(I_1^2 + I_2^2 + \dots + I_k^2)]^{0.5}$$

24 hour average plots for the NCR sector have us see the prominent drops in pollutant concentrations and realise that organised reductions can be consciously achieved if efforts are made in the right directions.

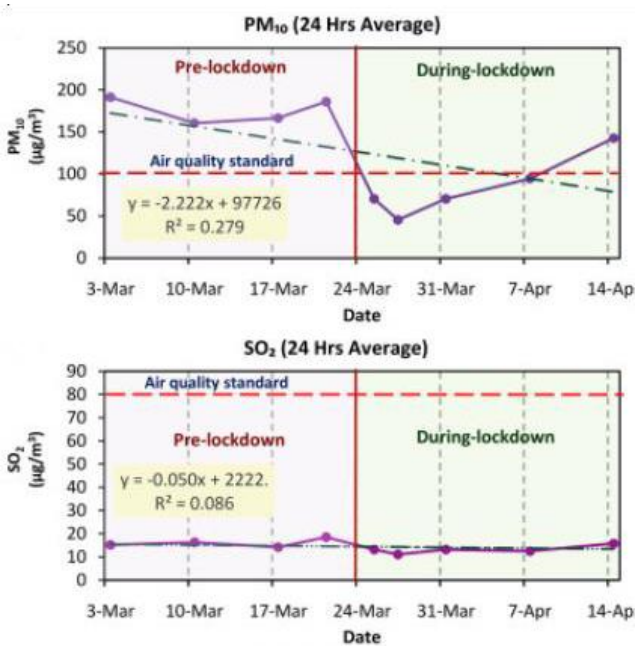



Figure 2.8

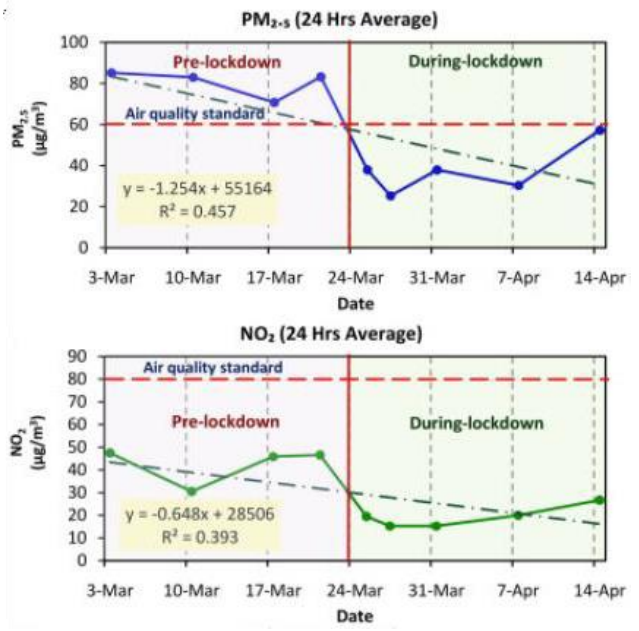



Figure 2.9

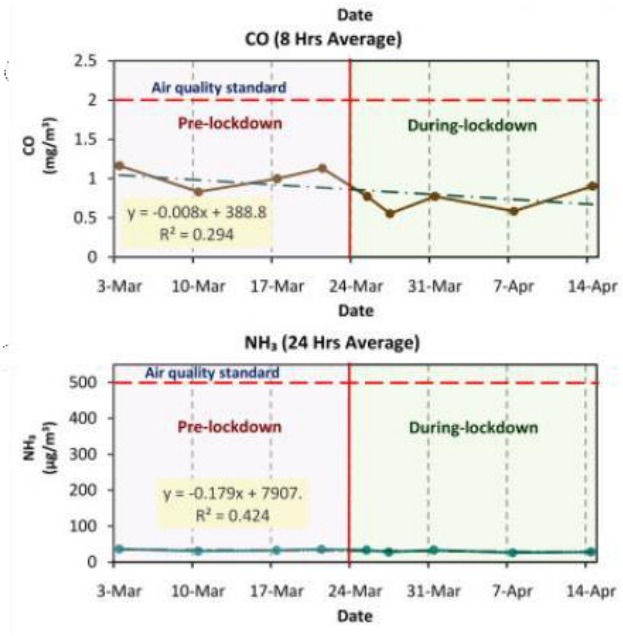



Figure 2.10

## VI. CONCLUSION

The tested dataset can also be realised using a median function representing an individual grouping of temperature ranges. It gives us an idea of how far along the current standards can be considered accurate, and when we may need to revise our study. One thing this study and climate change both make transparent is that its speed only seems to accelerate by the day.

The models used in order to make predictions from the dataset comply with many published authors and the idea of using median instead of mean gives us much more complex insights with comparatively lower memory usage.

## VII. REFERENCES

[1] Rich Seymour, *Understanding the Global Warming Discussion: Climate Change as a Context for Developing Standards-Based Research*

[2] Houghton, J T; Jenkins, G J; Ephraums, J J, Pg: 54-59

[3] National Academies of Sciences, Engineering, and Medicine (NASEM), 2019: *Negative Emissions Technologies and Reliable Sequestration: A Research Agenda* [https://www.nap.edu/catalog/25259]

[4] Intergovernmental Panel on Climate Change (IPCC), 2019: *Special Report on the Ocean and Cryosphere in a Changing Climate* [https://www.ipcc.ch/srocc

[5] Royal Society, 2018: *Greenhouse gas removal* [https://raeng.org.uk/greenhousegasremoval]

[6] U.S. Global Change Research Program (USGCRP), 2018: *Fourth National Climate Assessment Volume II: Impacts, Risks, and Adaptation in the United States* [https://nca2018.globalchange.gov]

[7] Online Links:

- https://data.ucar.edu/
- https://climatedataguide.ucar.edu
- https://iridl.ldeo.columbia.edu
- https://ess-dive.lbl.gov/
- https://www.ncdc.noaa.gov/
- https://www.esrl.noaa.gov//gmd/ccgg/trends/
- http://scrippsco2.ucsd.edu
- http://hahana.soest.hawaii.edu/hot/