Temporal Trends in Earth's temperature: A surface level analysis through time series regression and graphical representations

Robby Hooker[1]

Abstract

This paper reports on the analysis of time series data of Earth's temperature from 1750 to 2015. The goal of this paper was to identify high level climate trends at a global level, as well as changes throughout countries and major cities. The research consists of statistical analysis, visual analyses, and the implementation of an auto-regressive integrated moving average analysis model. The primary data used for analyses was separated into three sets, average temperature globally, by country, and by major city. Due to uncertainty in older measurements, the data was often trimmed to more recent years for reliability purposes. The analyses reveals trends in the global climate, specifically in the past 50 years, that indicate an acceleration in Earth's temperature rise in contrast to the preceding half century. This acceleration is cause for concern for people conscious of the environment. It has already affected the behavior patterns of species that have evolved over millions of years to survive in their natural habitats, and many believe it will be the downfall of our own species. Further research on the subject might consist of utilizing added climate metrics, implementing an auto-regressive distributed lag model, and enhancing visualizations to invigorate a wider audience.

Keywords: Global Climate Change, Time Series Analysis, Python Visualizations, Econemetrics

1. Introduction

The Earth's temperature has become one of the most widely discussed topics in modern society. The effort to reduce the increase in global temperature often seems futile, but activists hope that further increasing awareness and the amount inescapable data will convince the masses that change is imperative. This paper builds upon previous econometric analyses of the global climate, of which there are many.

Inspiration for this research has been drawn from the work of brilliant climate researchers, who have published extensive works on the factors contributing to, and the trends of the climate change discussed in this paper. Hopefully, this paper serves as an accessible and digestible resource for individuals curious about the data behind climate change. It is designed to act as a gateway, providing a comprehensive yet easily understandable overview of the subject matter. It attempts also to serve as a supplement, and ideally guides people into further reading more extensive climate change research.

The data analyses conducted in this paper were carried out to identify trends in global temperature that signify an accelerated increase in our planets temperature. Simultaneously, the data visualizations in this paper are meant to increase readability, in an effort to broaden scope and reach. Data science and analysis was carried out in python, using libraries such as pandas, matplotlib, pmdarima, folium, and others.

Following this introduction, this paper provides a literature review, discussing previous research on the topic. Next, a description of the data, including source, intricacies, supplemental data, and structure. Following this the paper will go into statistical analyses of the important features across the core data sets. Finally the results of the implemented regression analysis will be discussed, and analyzed. Upon reading these interwoven sections, the reader will have a better high level understanding of the often polarizing subject matter.

2. Literature Review

This literature review aims to provide background information on the topic of global temperature analyses. The topic is widely studied by researchers across disciplines. The papers selected to be reviewed in this section all have in common the fact that they advanced the study of climate change, support their arguments with statistics, and inspired the writing of this paper. This review will detail two papers sequentially, and outline the commonality of their results at the end.

The first paper at hand is David I. Stern and Robert K. Kaufman's 1999 Publishing: *Econometric analysis of global climate change*. This paper uses rigorous statistical methods to develop an understanding of the causes of global temperature increase, mainly focusing on stochastic trends. The pair conduct a Granger causality test to determine whether or not northern hemisphere temperatures are useful in predicting southern hemisphere temperatures. The paper details the underpinnings of such a test, especially the room for omitted variable bias. Upon conducting the test with 'simple models' with no conditioning variables, the result is a rejection of the null hypothesis, meaning there is south to north causality. However, as they detail in the paper, when they include greenhouse gasses in the model, the result is a failed rejection of the null hypothesis, which is a great example of omitted variable bias and how it can skew the results of research. This is admirable as their research was clearly deliberate and carried out by two people who

have a profound understanding of econometrics. Next they test for stochastic trends, which they explain can serve as evidence for causal relation between time series, more so than linear deterministic trends. Upon the carrying out of multiple types of stochastic tests, they explain that results differ due to the different nature of the processes, however the absence of certain integration throughout all tests is helpful to their understanding of the causality. This is an important lesson in statistics, as sometimes it is the non presence of certain statistics that helps paint the picture.

Moving on to the second paper to be discussed in this review, Camille Parmesan and Gary Yohe published their paper, *A globally coherent fingerprint of climate change impacts across natural systems* in 2003. Their research focuses on the impact that global climate change has on the species that inhabit our planet, which was a key inspiration for my interest in the broader topic of global warming. The observation that species which have magnificently evolved to survive in their respective domains, have been forced to change their behavior due to the temperature change is a melancholy subject. The authors of this paper meta-analyze over 1,700 species and show biological trends that match those of climate change predictions. Their means of analysis were through species range changes and phonological shifts. Their range analysis shows that on average the range limits of species have moved 6.1 km per decade northward. They also explain that 434 species were changing ranges over the time periods of 17-1,000 years, with a median of 66 years. They add that of these 434, 80% have shifted with climate change predictions. Species are moving to where the earth was previously cooler, and arctic species are contracting their range as the outer bounds of their previous range's heat up. Additionally their phenological meta-analysis provides similar results. They explain that data from 172 species shows a mean shift towards earlier spring timing of 2.3 days per decade. This result is a simple, yet convincing evidence that the planet is warming, as it shows that our warm seasons are becoming longer, and vice versa. The paper also provides confidence intervals and p-values that show their findings are significant. Similar to the previously discussed paper, the authors are deliberate in their statistical analysis in order to its prove distinction.

From our discussion of just two papers in the field, it is clear that significant thought and resources have gone into the research of climate change. It is also evident that there is a significant and ongoing shift in our planets climate, captured well by these two statistical studies in temperatures/greenhouse gases, and natural systems behavior. I find it inspiring that these authors dedicate themselves to such a noble cause, especially considering they are well aware of how hard it is to create change in today's society.

3. Data Description

The data used in the analysis for this paper is separated into three sets, all sourced from www.berkeleyearth.org. The most important of these sets perhaps is the Global Temperatures set, as it would be used for the time series regression analysis. The data consists of monthly measurements, and the uncertainty of each measurement from 1750 to 2015. The data also has other features such as max and min temperature, but these would be dropped prior to analysis. It is important to note that all the data used in this analysis would be manipulated and transformed multiple times, depending on the plot or model which it needed to service. For example, this monthly temperature data was transformed into a yearly average for certain plots and for the ARIMA model. A snippet of this important table is shown in figure 1 below.

Year	LandAverageTemperature	LandAverageTemperatureUncertainty
2011	9.516000	0.082000
2012	9.507333	0.083417
2013	9.606500	0.097667
2014	9.570667	0.090167
2015	9.831000	0.092167

Figure 1:

The second most important data set used in the research for this paper contained monthly measures of temperature for a list of 100 major cities. The data ranged from 1849 to 2013, and had features of average temperature, measurement uncertainty, city, country, and latitude and longitude coordinates. With a large data set such as this one, it is important to ensure cleanliness. To do so, NaN (not a number) values and outliers were removed. In figure 2 below you can see the simple python calculations to remove outliers. Another part of the cleaning process for this set involves the latitude and longitude columns. These columns were original formatted as a general object, with the number coordinate followed by a cardinal direction. This format was undesirable for multiple types of analysis so it would have to be changed. Figure 3 below shows the python function used to transfer to a more usable form, which involves removing the cardinal direction, and instead changing the sign to negative if the coordinate was south or west. This function would be applied to all latitude and longitude values in the major city dataset.

```
# remove tempurature outliers
temp_Q1 = df['AverageTemperature'].quantile(.25)
temp_Q3 = df['AverageTemperature'].quantile(.75)

temp_IQR = temp_Q3 - temp_Q1

df5 = df4[df4['AverageTemperature'].between(temp_Q1 - 1.5 * temp_IQR, temp_Q3 + 1.5 * temp_IQR)]
```

Figure 2:

```
def convert_lat_lon(value):
    numeric_value = float(value[:-1])
    direction = -1 if value.endswith('S') or value.endswith('W') else 1
    return numeric_value * direction
```

Figure 3:

This city data was meant to be used to analyze trends across the globe, which it ended up being very useful for. To do so the data would be grouped into yearly averages by city, while

maintaining the latitude and longitude of each city. In python, this means using the 'groupby' method to and taking the mean of each group.

The final dataset using in the research was monthly temperatures by country. Similar to the city data, this data would prove useful in visualizing trends and the fluctuation of said trends across the globe.

In each set the date column is given as a general object. This is not ideal for the methods of analysis used in this research, so the column was transformed into a pandas date-time object. This allows the column to be parsed for year and month individually, and most python statistics packages read date-time objects. Most significant to this paper, the auto arima function from pmdarima, which would be used to perform autoregression, uses date-time objects.

Part of the data cleaning process for all three sets included a look into the uncertainty of measurements column, as using reliable measurements is important when planning to draw conclusions. Upon plotting the average temperature uncertainty by year for the dataset, it was evident that the data should be trimmed down to a period of time where it would be more reliable. Figure 4 below shows the plot used.

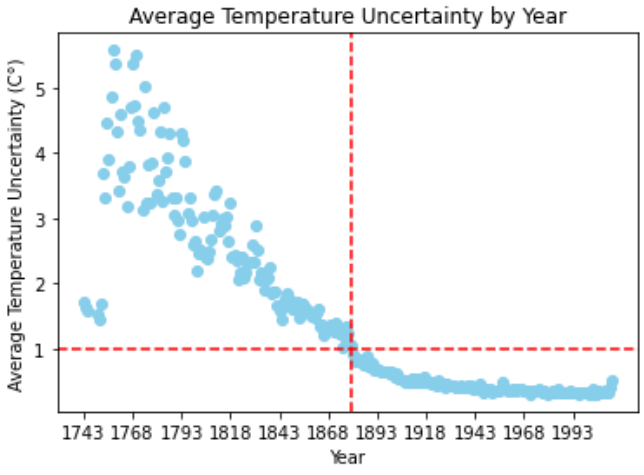

Figure 4:

From this plot you can see that around the year 1880 the uncertainty of measurements dips below 1° Celsius, which was deemed to be a good spot to trim the data from. It is important to note that when trimming data like this, it is crucial to find a balance between not removing enough bad data that may make your analysis unreliable, and removing too much data, which could make your analysis non representative of the data.

Going forward, it would be useful to have data that includes more climate metrics than the ones available in these datasets, CO2 emissions for example. Additional features could allow for different types of analysis, such as multiple linear regression. That being said, it is important that this data is from a reliable source in www.berkeleyearth.org

4. Exploratory Data Analysis

Once the data was cleaned and sorted, it could be used to understand patterns and trends in the data! These insights are largely the purpose of this paper, as they will hopefully portray climate change data to readers in a digestible form. As discussed in the data description section, the data at hand lacks a large amount of variables, however the volume of the data, along with manipulation, allow numerous possibilities for analysis. Prior to observing the analysis, it is important to fully understand the variables at hand. The two variables that appear across each dataset are temperature and temperature uncertainty. These will be used in multiple ways to make up for the lack of other features in the data. It is important to note that that temperature variables are measured in Celsius, and any statistical analysis on these variables will be in terms of Celsius.

Firstly, to better understand the temperature, and how poor the measurements were in the early years of this data, we can plot the two metrics together. Using the modified dataframe which averages the yearly temperatures and uncertainty, we can produce a time series plot that tells the story of uncertainty over time. This plot (Figure 5) is also a preview of what is to come in regards to trends in temperature over time.

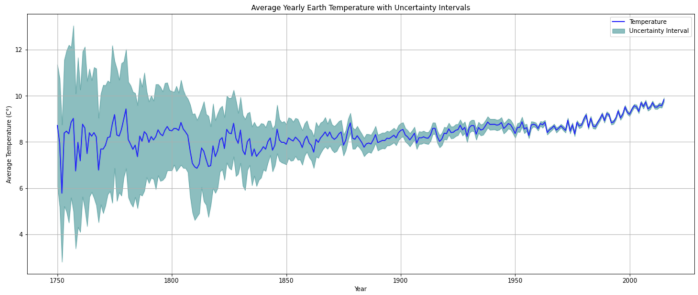

Figure 5:

In the graph of figure 5, the dark blue line represents the average temperature measurement by year in the global data, where the teal color fill shows the range of uncertainty for each point. It is clear from this visualization why we decided to trim off records of earlier measurements, as they are clearly unreliable. It also appears that in roughly the last 50 years, the rate a which temperature is rising is increasing.

Next, in preparation for time series regression, we will explore time series statistics that will explain some of the data. Specifically we perform a Dickey-Fuller test (one of the tests used in *Econometric analysis of global climate change*), and check auto correlations of 1, 3, and 5 years. Below is the equation for the Dickey-Fuller test

$$\Delta y_t = \delta y_{t-1} + u_t \tag{1}$$

Fortunately, we do not have to code this formula into python, as there is an existing function in the stats models package that does this for us. Upon running the function with our yearly average data, we should have statistics that tell us whether or not the data has stationarity.

Metric	Value
Test Statistic	-0.191872
*p-value	0.939484
No. of lags used	15
Number of observations used	250
critical value (1%)	-3.456781
critical value (5%)	-2.873172
critical value (10%)	-2.572969

Table 1: Dickey Fuller Results

From this table we conclude that that the data is non-stationary, as the test statistic is less than smaller than the critical values. This was to be expected based off the graphical representation of the data, and is accounted for in the regression analysis.

Next we the auto-correlations of the data will provide us with a measure of the relationship between the current value of temperature and its past values. Table 2 shows the auto-correlation for 1, 3, and 5 year lags.

Number of lags	Autocorrelation
1	0.72507
3	0.64145
5	0.58700

Table 2: Autocorrelations for Different Lags

We see that as the number of lags increases, the current value becomes less correlated with the lagged values, which is typical in this type of analysis, and will be something to keep in mind during regression.

Next we'll use seasonal decomposition to further identify trends in the data. First simply plugging the yearly average data into a python seasonal decomposition function shows us the trend of the data. This is displayed in figure 6.

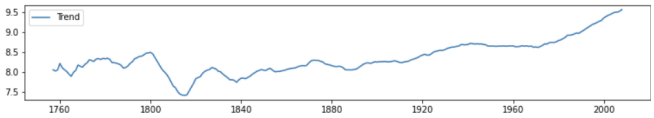

Figure 6:

This chart further solidifies our recognition the rate at which temperature is increasing is accelerating.

Next, using the same seasonal decomposition function, we can analyze trends within a year of data, by inputting monthly temperature data. These results show that global temperature fluctuates throughout the year, and can be seen in figure 7.

Moving on, we'd like to identify differences in temperature change across the globe, rather than as a whole. Using the major city and country data, we can map temperatures differences to a map of Earth to help build some intuition behind this.

Both maps will be created using geospatial data and the pandas folium package. The data displayed is a measurement of the difference between a six year average from 2008-2013 and

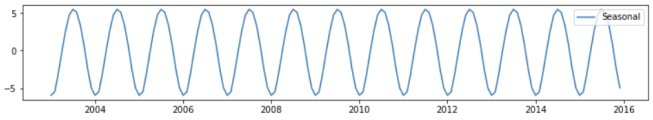

Figure 7:

1900-1905 for each city and country. This measure tell which cities and countries have experienced higher or lower temperature change. Figures 8 and 9 display the city and country map respectively.

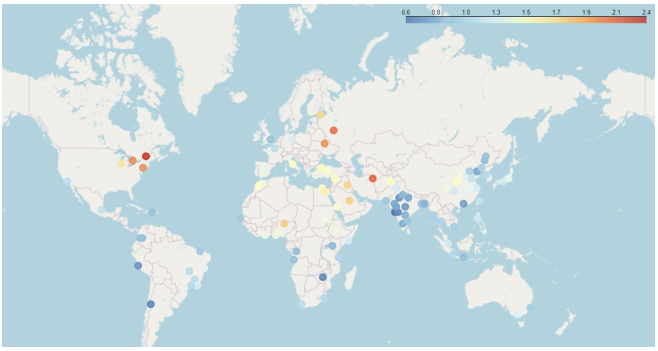

Figure 8:

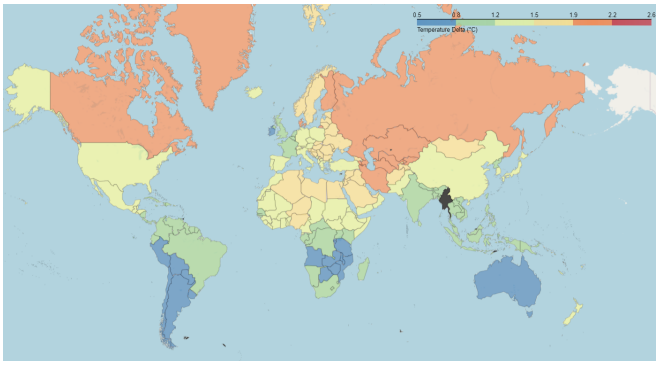

Figure 9:

The maps are a great indicator of how temperature change varies across the globe, and they also add an element of beauty to the research.

Now, with a better understanding of the data, we can perform regression analysis on both the global and more granular datasets.

5. Regression Analysis

Regression analysis of our temperature data will allow us to try to predict future values, quantify the trends identified in our data, and make a final conclusion on the acceleration of climate change.

Firstly, to analyze trends throughout some of the major cities we can plot each city's average temperature by year, and find

a simple linear regression equation to fit the data. The coefficient on the year variable will indicate the magnitude at which temperature is changing in a given city. The charts in figure 10 below provide an example of this from New York City and Santiago. We see that based off the coefficients of the regression lines, New York City's temperature is increasing at a faster rate than Santiago's.

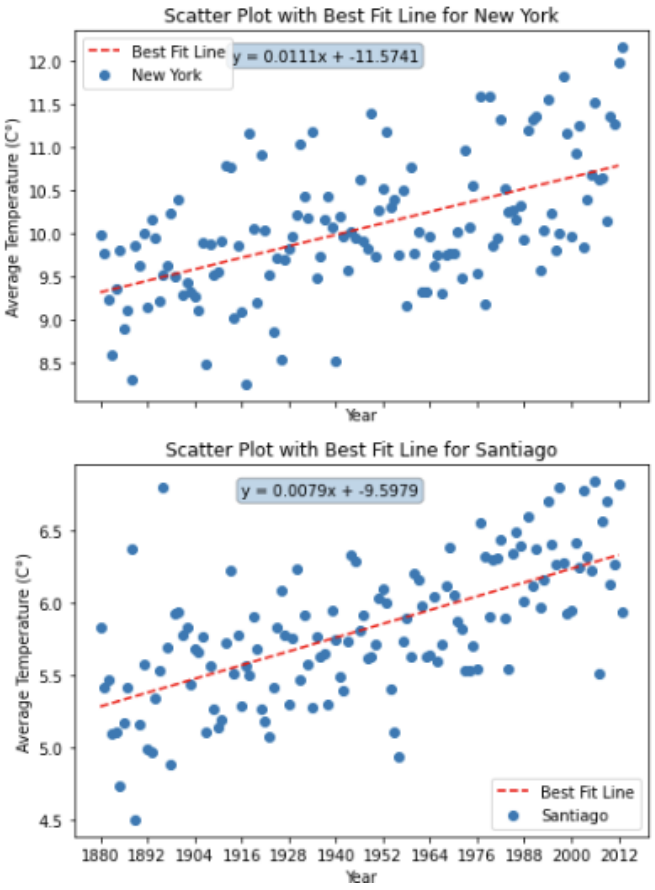

Figure 10:

Further graphs like these can be found in the GitHub repository for this project. The graphs are perhaps most useful when used with the city heat map provided in the exploratory analysis section of this paper. Doing so you can see that graphs with greater coefficients correspond to cities that have higher temperature deltas in the map.

Next we will create an autoregressive moving average model using pmdarima's auto arima function. This type of model is ideal for our non-stationary data, as it will transform it to stationary, and then implement autoregressive methods. First, the data we will be using is the global data by yearly average. Due to the uncertainty measures observed earlier in this paper, we have trimmed to data to the range 1900 - 2015 for reliability purposes.

Once the data is in this form, it needs to be split into separate train and test split datasets, where the train dataset is used to create the regression, and the test is compared against predictions from the regression to test the models accuracy. Figure 11 below shows the initial ARIMA model's forecast, as well as the train test split used to create the model.

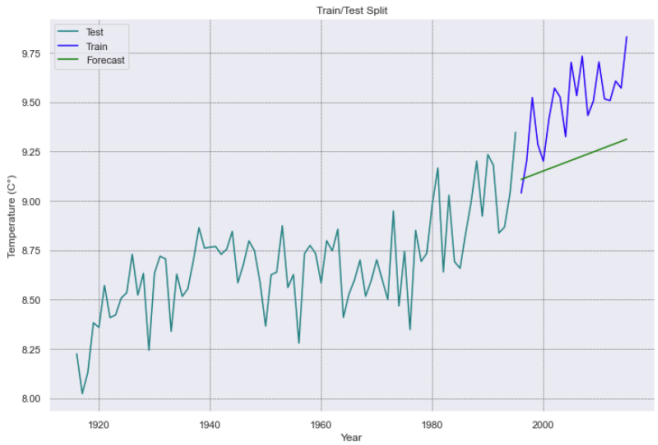

Figure 11:

As you can see, the model is an awful predictor of future values or global temperature. The root mean squared error came in at .3167, which is huge in terms of Celsius measurement. Upon looking at the train test split and the trend of our forecast, it seems the train portion of the model is a poor range to use for training, as it does not represent the more recent changes in global climate.

In an effort to make the model a better predictor, we can move the train and test split up to the portion of the data where we start to see similar trends to those of current values. In figure 12, the new train test split is shown, where it begins in 1970. Also plotted is the new model's predictions in blue, and the old model's predictions in red.

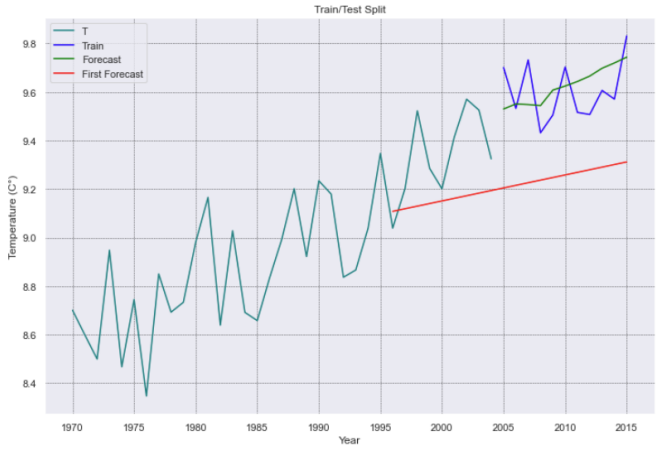

Figure 12:

Although trimming the data, reduces our volume of predictions, based on the graph, it is clearly a better predictor of future temperature values. This is also evidenced by the new, lower, root mean squared error of .1251. Table 3 below contains statistics describing the newly generated model.

We can see that the coefficients on values at t-1 and t-2 are significant, as their P-Values are less than .05, however the co-

Parameter	**Coefficient**	**Standard Error**	**Z-Score**	**P-Value**
Intercept	0.0652	0.041	1.604	0.109
*ar.L*1	-0.8058	0.201	-4.001	0.000
*ar.L*2	-0.5363	0.227	-2.359	0.018
*ar.L*3	-0.2939	0.180	-1.631	0.103
σ^2	0.0425	0.014	2.963	0.003
Additional Information				
Log Likelihood	5.011			
AIC	-0.023			
BIC	7.609			
HQIC	2.580			
Model Diagnostics				
Ljung-Box (L1) (Q)	0.01			
Jarque-Bera (JB)	1.23			
Prob(Q)	0.94			
Prob(JB)	0.54			
Heteroskedasticity (H)	0.54			
Prob(H) (two-sided)	0.32			

Table 3: SARIMAX Model Results

efficient on values at t-3 are not significant. A quick interpretation of the summary statistics provides insight to the performance and reliability of the model. The high Log Likelihood shows that the model explains the observed data well, and the low AIC suggests the same. The low P-Value for Ljung-Box (L1)(Q) suggests that there is autocorrelation in the data, which was also observed earlier in this paper.

6. Conclusion

To wrap up this paper we can conclude that all data used in this analysis of global temperatures points to an accelerated increase in the Earth's temperature. It also helps us develop intuition on how the temperature change varies across the globe. The paper builds of econometric style research, and reports similar findings to the widely accepted narrative related to this subject. The data used to perform this analysis was simple, but very versatile, and being sourced from www.berkeleyearth.org, it can be considered reliable. Upon statistical analysis of the variables in the data we see graphs and maps the indicate global temperature increase is accelerating. The same conclusion can be drawn from the regression analysis carried out for this paper, specifically showing that temperatures in roughly the last 50 years are increasing more rapidly than the preceding half-century. Hopefully after reading through the sections of this paper, you have gained an understanding of the trends in Earth's temperature. Furthermore, hopefully more research like this acts as a catalyst for change, so we as a species can prevent further damage to out planets natural systems.

References

David I. Stern, Robert K. Kaufmann, Econometric analysis of global climate change, Environmental Modelling Software, Volume 14, Issue 6, 1999, Pages 597-605, ISSN 1364-8152, https://doi.org/10.1016/S1364-8152(98)00094-2. (https://www.sciencedirect.com/science/article/pii/S1364815298000942)

Parmesan, C., Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. Integrative Biology, Patterson Laboratories 141, University of Texas, Austin, Texas 78712, USA. John E. Andrus Professor of Economics, Wesleyan University, 238 Public Affairs Center, Middletown, Connecticut 06459, USA.

Pierre, S. (2022). A Guide to Time Series Analysis in Python. https://builtin.com/data-science/time-series-python

Link to project github repo: Github Repository