CS-C3250 - Data Science Project

Climate Change Analysis with Reaktor

Jeheon Kim, Alice Boix, Mathilda Smith, Dylan Nguyen, Phi Dang

Reaktor Partner: Janne Sinkkonen

Instructor: Alena Shchevyeva

Lecturer: Jorma Laaksonen

Table of Contents

Abstract

The goal of this project was to produce a website that portrays climate change data smoothly to interested viewers. By using data points from multiple different sources, the site undoubtedly shares the most accurate and up-to-date information we could find. In order to accurately portray the effects climate change has had on the planet, we chose to illustrate data of CO2 emissions rising (along with other gases), the ice caps melting, which results in sea levels rising, the rising of temperatures, as well as included calculations that show how volcanic activity, El Niño, and other factors, are not the primary cause for the rise of CO2 in the atmosphere - but rather it is the human activity that has been the driving force.

Producing incorrect or fake data can significantly hinder the general population's understanding of what is going on around them, which makes the importance of fast, reliable data that is readable and understandable to the general viewer a crucial part of this project. This is why the site was created with clear visuals along with explicit explanations: to ensure viewers understand the issue and why it is so significant, as well as what can be done to relieve the situation as much as possible.

The front page of the website showcases the severity of the situation. The other tabs on the website go into detail for each of the following categories: sea levels rising, greenhouse gases, temperature levels, as well as arctic amplification. Each tab explains the negative effects the human population and its excessive release of CO2 has had on each of the four aspects, for example: human activity has caused sea levels to rise all over the world. Going through each tab, it is clear how they are all connected. CO2 levels go up, which causes temperatures to rise, which causes ice caps to melt, which in turn causes the sea levels to rise. They are all connected to one another, which is why it is important for the world population to understand the dire need for change if positive results are expected. As said on the front page of the website: "We can fix it."

Hopefully, websites such as ours or other similar ones can simply but methodically show the countries with the largest effects on the Earth and encourage them to make a change.

Motivation

Advancements in the Earth Observation (EO) technology, such as satellite systems, have made it possible for researchers to see the bigger picture of climate change by providing various types of information, such as physical, chemical, and biological systems of the planet, about earth and its climate on a global scale.

According to research, climate change has already brought multiple observable impacts to our environment. Glaciers have shrunk and a number of animal and plant species are in danger of extinction due to climate change. Such impacts can fundamentally transform whole ecosystems and the intricate webs of life. Furthermore, it has a significant effect on our livelihoods, health, and future.

There is no more time to wait. Although we cannot stop climate change overnight, we still can slow down the pace of it. And for this, we first must understand how the climate is changing and why it is happening.

Thus, in this research, we examine some representative scientific evidence of climate changing, and attempt to model and extrapolate the global mean temperature using various prediction models.

Data Description

Temperature Analysis

A temperature anomaly is the difference from a baseline temperature, which is typically computed by averaging 30 years of temperature data. In our analysis, the base period is set from 1951 to 1980.

A positive anomaly indicates the observed temperature was warmer than the baseline temperature, while a negative anomaly indicates the observed temperature was cooler than the baseline temperature. (NOAA, 2020)

- **Data**

Two datasets we used are provided by the global component of Climate at a Glance (GCAG) and the GISS Surface Temperature (GISTEMP). Both datasets contain the yearly temperature anomaly data from 1880 to 2016.

- **Method & Result**

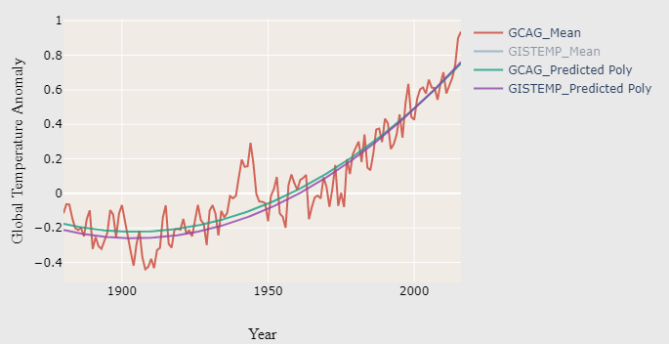


Figure 1. Change in Global Average Temperature Anomaly

For the time series visualization of temperature anomaly, we used the Python Plotly library. To sufficiently capture the fluctuation dynamics of long time-series data, we utilized the polynomial regression for which we used sklearn's PolynomialFeatures to transform original features into their higher degree (2-degree) terms and LinearRegression to fit the converted features.

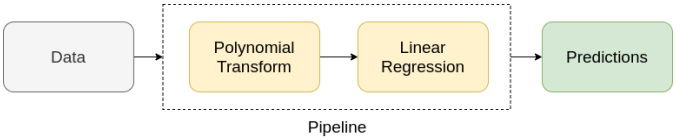


Figure 2. Polynomial Regression Workflow

This method satisfactorily smoothed out yearly fluctuation in the data and showed better overall temperature trend. The outcome, figure 1, illustrates long-term global mean warming trend, started in the mid 1900's and continues on through the present day. Furthermore, the almost identical results of two datasets, GCAG and GISTEMP, gave us a greater confidence in our conclusion.

Greenhouse Gases

We decided to make our main focus on carbon dioxide, although we drew up graphs and considered including other gases in our website as well. Here is a visual of the top contributors to climate change due to CO2 emissions:

➢ Data

The data points presented here come from a dataset the United Nations Climate Change Greenhouse Gas Inventory Data. This dataset consists of near 10 different gases, from most of the countries all around the world. Although the data on our site shows data points from 1990s onward, the dataset consisted of data points dating all the way back to the early 1700s, but because we wanted reliable and consistent results, we chose to use the data from the 1990s onwards, which also ensured there were data points from every single year for all the countries we looked at.

➢ Method & Result

We wanted to use only the most relevant data, so we decided to create comparisons according to such. We compared the CO2 levels of the top contributors in the world, filtering out all other gases and countries with less effect on the planet, then we compared the Nordics amongst one another, since their numbers are relatively similar, and lastly, we compared CO2 vs temperature as well as sun spots vs CO2, to show the amount of damage that human-created CO2 has done.

Here is a visual of the top contributors to climate change due to CO2 emissions:

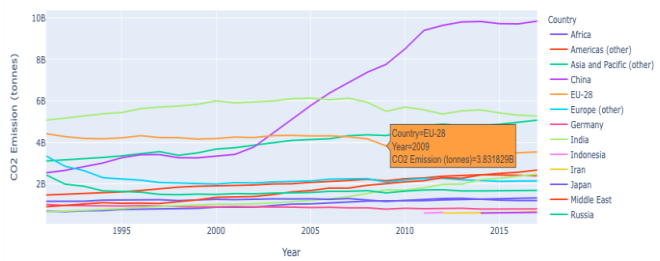


Figure 3. CO2 emissions of the top contributors of the world

Because we live in Finland and we wanted to show data specifically showing Finland, we decided to compare Finland to the other Nordic countries. Unfortunately, due to time constraint, it was unable be included in the website, but we can show the results anyway:

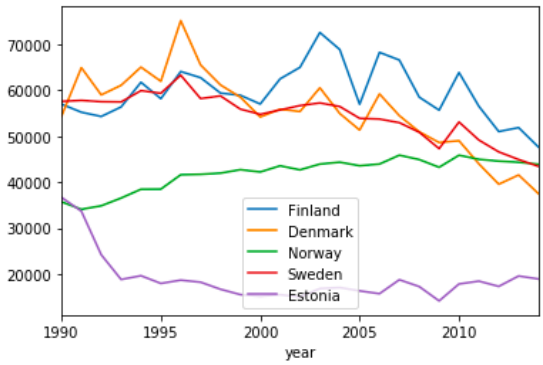


Figure 4. Nordic countries CO2 levels comparison

Here it can be seen that all countries besides have been in a slow decline for the last 10 years, but overall (other than Estonia) have stayed roughly the same. The Nordics lead the world in many aspects, with climate change reduction being one of them. The data presented here came from the same dataset as that of the above.

Because of the limited time frame, we had to decide on what we found most relevant and most important to the project. We decided to share on the site the greenhouse gases of the top countries (above) as well as the correlation between temperature vs CO2, and sunspots vs. CO2. Temperature and CO2 show a strong correlation, as can be seen here:

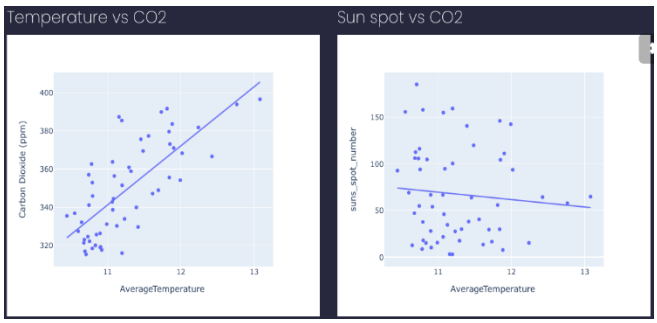


Figure 5. Correlation between CO2 vs temperature, and CO2 vs sunspots

Sharing these graphs show the negative effects that human activity has had on the planet, since we know that the rise in CO2 levels is a result of human activity.

Sea Ice Extent

➢ Data

The dataset we used comes from the National Snow & Ice Data Center website and it has been collected by NASA. This dataset reports the sea ice extent with day-by-day measure from 1978 until 2019. The data has been collected for the north hemisphere and for the south hemisphere. There were some missing values but as we

wanted to study a global trend over decades, we did not need to replace them.

➢ **Method & Result**

Using Plotly we were able to visualize on the same plot the North hemisphere and the South hemisphere sea ice extent fluctuations over years. There are significant seasonal variations. In both hemispheres, sea-ice extent fluctuates sinusoidally. The North hemisphere reaches a local maximum every winter and a local minimum every summer. It is the perfect opposite for the southern hemisphere.

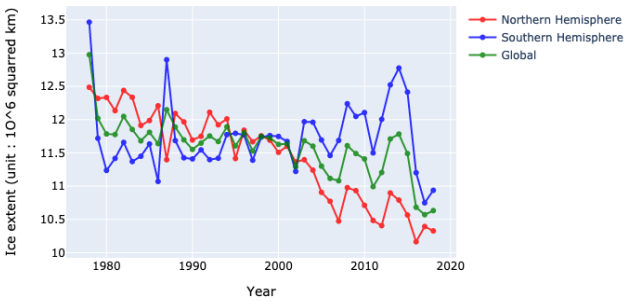


Figure 6. Evolution of the yearly average ice extent

To get rid of yearly fluctuations, we studied the average sea-ice extent yearly average. Both North and South sea-ice extent is decreasing. This was evidenced by data visualization, linear regression, and polynomial regression. However, we noticed that the North Sea-ice extent is decreasing way faster than the South sea-ice extent. The climatic mechanisms governing the two poles are different indeed, the north pole is a sea whereas the south pole is a continent. Therefore, we can identify some differences in the way ice is melting in the two hemispheres. However, the global trend is the melting of the ice, validating the thesis of global warming.

Solar Activity

➢ **Data**

The dataset we used was collected by Solargis, a company specialized in solar power investment. The captors were in southern Spain, they gathered pieces of information about global horizontal irradiation, direct normal irradiation, diffuse horizontal irradiation, global tilted irradiation. We wanted to measure solar activity, so we focused on global horizontal irradiation. The dataset goes from 1994 to 2020 and the measures are taken once a month.

➢ **Method & Result**

After a first visualization, we can see that the solar irradiance received by the captors described sinusoidal yearly fluctuations. The yearly average value is quite stable. To model properly these data, we first determine the period T of the sinusoid. Then we identify the trendline of the maximum and minimum peaks and finally we can predict the future values using a sinusoidal prediction. The amplitude of the solar irradiance received seems to be decreasing slowly. Regarding our analysis there is no evidence that the sun has an increasing activity responsible for global warming.

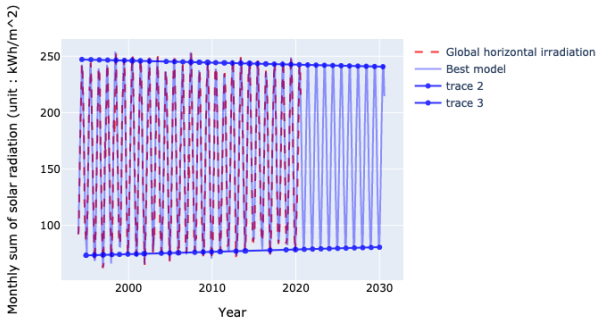


Figure 7. Historic global horizontal irradiation & predictions

Volcanic Activity

➢ **Data**

The dataset we used has been gathered by the Smithsonian Institution's Global Volcanism Program. The data includes information on all volcanic eruptions over several millennia. There is also a description of the type of eruption. We focused on eruptions including Tephra emissions (Tephra are gases and particles that alter the sun penetration into the atmosphere).

➢ **Method & Result**

The amount of Tephra released is measured by the Volcano Explosivity Index (VEI) in a logarithmic way. After visualizing the amount of Tephra released since 1900, we noticed that the eruptions take place in a perfectly random way and the amount of Tephra released is not large enough to have an impact comparable to the effects of greenhouse gases.

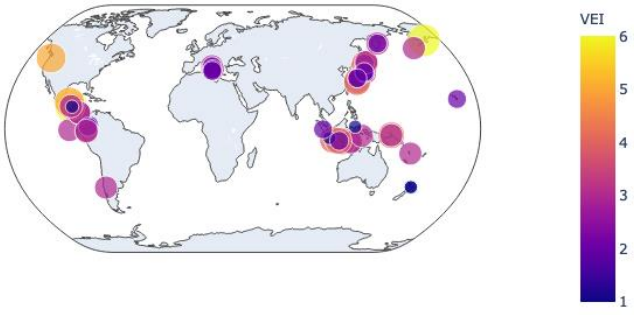


Figure 8. Choropleth map of volcanic activity

Sea Level Rise

Sea level is an essential climate change indicator. As the temperature rises, so does sea level. The spike in sea level is primarily caused by two reasons associated with global warming: the addition of water from melting

glaciers and the extension of seawater when the temperature rises inside. Our project makes two analyses about the sea level: sea level anomalies (variations) and sea level rise impact for coastal developing countries.

➢ **Data**

1. An excel file with 6 sheets with 6 corresponding factors: Land, Population, GDP, Agriculture, Urban Extent, and Wetland affected by Sea Level Rise (SLR), from The World Bank (download here)
2. A text file contains Global Mean Sea Level (GMSL) variations compared to the 20-year collinear mean reference from 1996 - 2016, from NASA (download here).

The dataset represents GMSL variation in several conditions (column description in the data file). However, we only use two columns named GMSL2 and smGMSL3:

- GMSL2: GMSL with Global Isostatic Adjustment (GIA) applied
- smGMSL3: A smoothed GMSL2 with annual and semi-annual signals removed

➢ **Method & Result**

- *Sea level variations*

Extrapolation: At first, we use such simple models as Linear Regression and Polynomial Regression to quickly visualize the general trend. We can see that both lines from the models represent the upward trend, although Polynomial Regression tends to go in a better direction. Then, we try to apply the Prophet model in the data, and this seems to result in a better extrapolation.

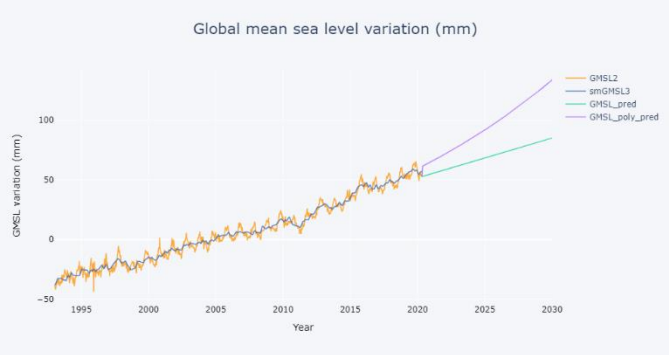


Figure 9. Global mean sea level variation

Accuracy: We split the data into two parts before and after 2015 to make a train and a test set. Here is the result of 3 models.

	RMSE	MAE	R2
LR	11.4	10.32	−1.42
PR	5.86	4.52	0.35
Prophet	5.55	4.77	0.42

Figure 10. Result of global mean sea level predictions

Based on the prediction models, this analysis concludes that the global sea height increases. In 10 years from now, the variation might even double.

- *Sea level rise impact*

We made multiple visualizations from this dataset to draw some comparisons, rankings, and conclusions.

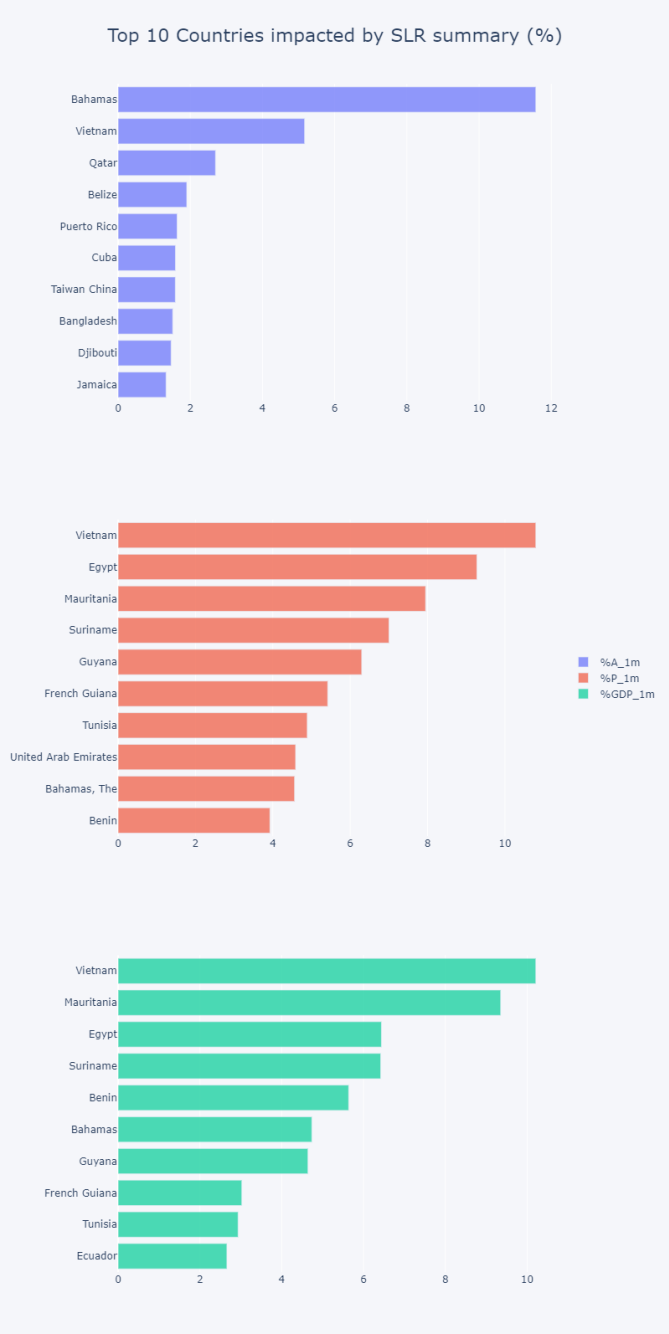


Figure 11. Top 10 countries impacted by SLR summary (%)

Figure 12. Sea Level Rise 1m & 5m impact summary (%)

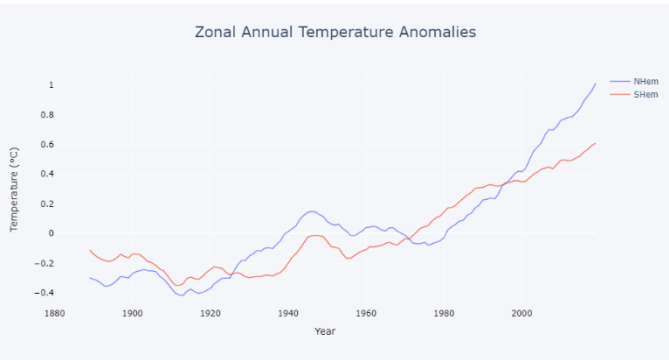


Figure 13. Temperature anomalies in two hemispheres (°C)

Label interpretation (Figure 11 & 12): The percentage of impact on a factor when sea level rises by a certain level. For example, %L_1m is the percentage of lost land when the sea level rises by 1m. We have L = L and, P = Population, and GDP.

Figure 12 depicts the comparisons of the sea level rise impact on the continental scale. We can observe that the most severe impact lies in East Asia when the increases are 1m and 5m. On the other hand, South Asia seems to have the least negative impacts when most of its index is the lowest.

Figure 11 shows the impact on the country scale. Our results are extremely skewed, with severe impacts limited to a relatively small number of countries. For these countries (e.g., Vietnam, A.R. of Egypt, The Bahamas); however, the consequences of SLR are potentially catastrophic.

Arctic Amplification

The global surface temperature has warmed up around 0.6°C over the past 30 years, but not uniformly worldwide. In the Arctic, the temperature rose almost twice as quickly as in the equator. This is a phenomenon known as Arctic amplification.

The dataset is the zonal annual means of Combined Land-Surface Air and Sea-Surface Water Temperature Anomalies from NASA GISS (download here)

The time series fluctuates quite significantly. To emphasize the trend, we use the moving average methods. Using a rolling mean of 10 years, each data point becomes the mean of the last ten years until then.

The mean sea level is directly linked to the average temperature. Indeed, due to the thermal expansion of water and ice melting, the higher the temperature, the higher the sea level. Figure 13 shows a clear trend of an increasing sea level in both hemispheres over the past decades. This increase is even more significant in the northern hemisphere.

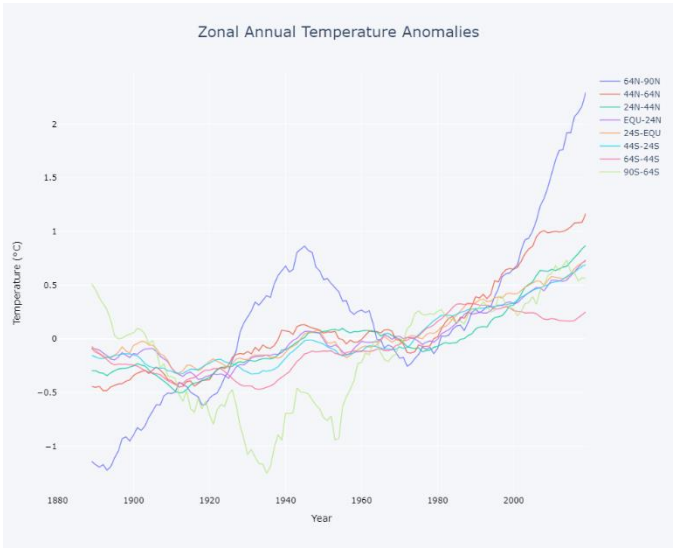


Figure 14. Temperature anomalies by latitudes (°C)

Next, Figure 14 shows the different temperature anomalies depending on the latitude. Earth has been split into eight geographical areas from the north pole to the south pole. The global trend shows an increase in the mean sea level in all these areas. However, the fluctuations are not all the same. The two poles are susceptible and observe a higher variation. The north pole is the area with the highest increase since this area is only composed of water. The melting of sea ice and a large

amount of water in this area make it sensitive to thermal expansion resulting in a high increase in the mean sea level. Because it is a continent, the south pole does not observe such a huge increase. However, this area also describes significant variations, which is also a marker of climate change.

GDP vs CO2

First, we want to explore the relationship between the GDP and the CO2 emissions of a country. This would be a good starting point to visualize the difference between poor and rich nations in producing emissions. We calculated the average Pearson correlation between two factors of each country. The result was about 0.65 for both GDP and GDP per capita data, so the correlation is not clear for every country. After that, we take the top 15 countries that release the most CO2 in 2018. Surprisingly, these countries have very high correlations (see Figure 15).

In our website, we narrowed down further the countries to the top three world's largest CO2 emitters: China, United States, and India; all of which showed relatively stronger correlation than others.

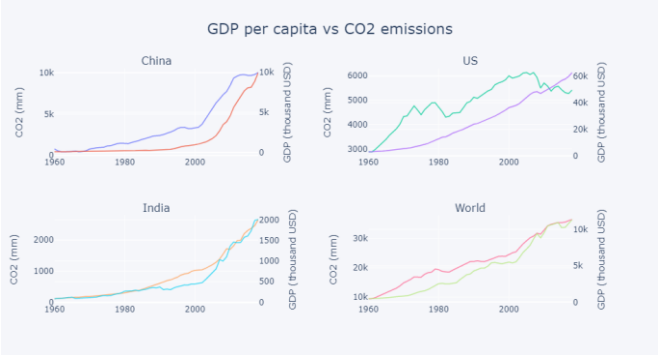


Figure 15. GDP & CO2 emissions correlation of the top three CO2 emitters in the world

From our analysis, all the top three CO2 emitters are found to have strong correlation between their increasing amount of CO2 emission and GDP growth, started since the mid-1990s.

The correlation was especially stronger for China and India which indicates that their rapid and explosive economic growth over the past decades is heavily based on the energy-intensive industry, such as pulp, paper, refining, iron, and steel, which requires an extensive burning of fossil fuels (Coal, oil, and natural gas) to generate energy which is later converted to electricity and heat for production.

On the contrary, the countries that have been developed for much longer (Check starting GDP value on the y-axis of US), such as United States and many Western European countries, were also helpful in observing the long-term effects but their correlation in time series was not as clear as China and India to discern particular relationship. It is also due to the fact that these developed countries already had changed their overall industry into environmentally friendly production in an attempt to reduce the amount of CO2 emissions.

Methodology

In this section, we discuss the main methodologies that are utilized in the paper in order to model the average global temperature. Various prediction models, both linear and non-linear were implemented for comparison purposes. However, most of the naive models in question were not able to capture the temporal relationship between temperature and its past values, as well as the lagged correlation between temperature as the dependent variables, against independent variables. Therefore, in this section, we discuss the algorithms that were found to perform the best on the task of predicting the temperature. The model performance, result and findings are examined later in the results and discussion section.

Prophet

Prophet is a powerful and user-friendly tool from Facebook. Its advent has solved two main issues in extrapolating a time series in practice: The inflexibility of automatic models in parameter tuning; and the lack of professional skills for the analysts to produce high-quality time series prediction.

Theoretically, Prophet represents the data as being made up of several components, following the equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

g(t): The trend function that models non-periodic changes exploiting two trend models. By default, the piecewise linear regression model with changepoints is used for the forecasts. When the problems exhibit saturating growth, we can switch to a nonlinear model with logistic growth and set maximum and minimum achievable points.

s(t): The periodic changes (yearly, monthly, daily seasonality) based a partial Fourier sum. We can disable certain kinds of seasonality and change the frequency changes.

h(t): The effect of holidays and special events. We can add manually a set of holidays and events or use built-in data from the library. The prior scale could be reduced to dampen the holiday effect.

The approach takes inspiration from generalized additive models (GAMs), where forecasting is considered a curve-fitting exercise.

With Prophet, our team has created better predictions with acceptable errors. The same API with sci-kit learn models (fit, predict, plot methods) enable less complication in implementation and code comprehending.

Besides, Prophet's measurements do not require preprocessing missing values or outliers. There was also wide space for tunings, such as additional regressors or

adjusting prior scales, which significantly improved our temperature extrapolation accuracy. We have also tried utilizing the "holidays and special events" feature to improve the temperature predictions by using volcanic activity data.

To evaluate and compare model performances, we need some validation. Usually, this could be done manually by, for example, splitting the data with the ratio 80-20 into training and testing sets. Fortunately, Prophet offers a built-in cross-validation method, which splits the data into several segments, performs training in one segment, and validates in other segments.

Last but not least, we conducted some parameter tuning using an exhaustive grid search.

XGBoost

The eXtreme Gradient Boosting (XGBoost) is an application of gradient boosted decision tree algorithm, designed to solve regression and classification predictive modeling problems with high model performance and computational speed.

Boosing is an ensemble learner, where the final model is created based on a collection of individual models. Models are built sequentially by minimizing the errors from previous models while boosting the influence of high-performing models. And Gradient Boosting is one type of boosting where the Gradient Descent Algorithm is used to minimize errors. Lastly, XGBoost is built on the principles of Gradient Boosting but pushes the extreme of computation limits of machines to find the best tree model (XGBoost, 2020). Its combination of software and hardware optimization techniques can quickly yield superior results in a short time by using significantly less computing resources.

There are several advantageous characteristics of XGBoost. First, it uses a regularization method to prevent over-fitting, ensuring that the final model we end up with is generalized. Second, it can deal with large sparse data by internally handling missing values. In brief, it can automatically choose the best imputation value for a dataset based on the reduction in training loss. Third, it supports k-fold cross-validation, which enables more accurate estimates with efficient use of available data. Lastly, it takes advantage of parallel computing, which facilitates a more efficient and scalable tree construction.

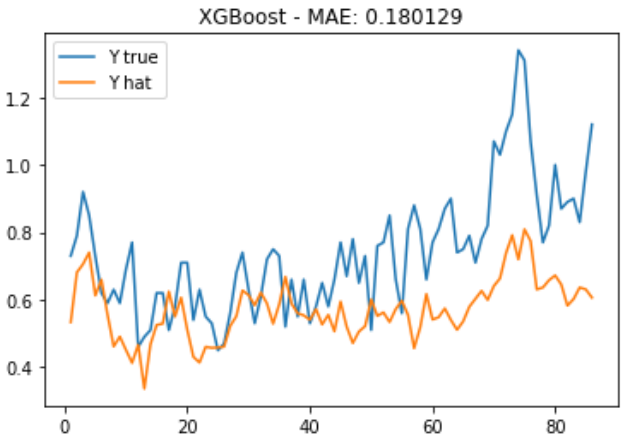


Figure 16. XGBoost fitting result (Orange line)

With XGBoost, our group could create the second-best fitting curve to our data, behind the Prophet's, with fairly good Pearson correlation value of 0.638647 and significantly low values of both RMSE (0.049148) and MAE (0.180129), all of which indicate a good fit.

Correlation

Modeling of the dependent variable is a difficult problem to solve, because of the complexity involved in separating the cause-effect relationship of the dependent variable, which is the global mean temperature, and independent variables, which are discussed later in this section, in an extremely complicated system powered by various different factors.

In our studies, we look at the relationship between the global mean temperature, with 6 following time series:

- Atlantic Multidecadal Oscillation (AMO): This data series represents the mean SST of North Atlantic, i.e., within the latitude 0 °–70 °N, detrended to remove the influence of global warming.
- Greenhouse gases (CO2): The long-time yearly time series of the concentration of co2.
- North Atlantic Oscillation (NAO): An index calculated from the measurements of air pressure at two locations: Ponta Delgada, Azores, and Stykkisholmur/Reykjavik in Iceland.
- Sunspots Number (SSN): The number of sunspots.
- El Niño/Southern Oscillation (ENSO): Temperature fluctuations expressed by the average SST anomaly of the region 20 °N–20 °S minus 90 °N–20 °N and 20 °S–90 °S, relative to the base period 1950–1979. This has 3 separate time series.
- Volcanic Explosivity Index (VEI): An index marking major volcanic explosions.

The data frame contains 8 columns for the 6 independent time series (NINO has 3 different time series), 1 column for the dependent variable, and 1 column for the timestamp. Due to the difference in availability of data, the time frame is sliced to the period of 1st of January 1984 to 1st of February 2017.

The correlation matrix of the data can be seen below:

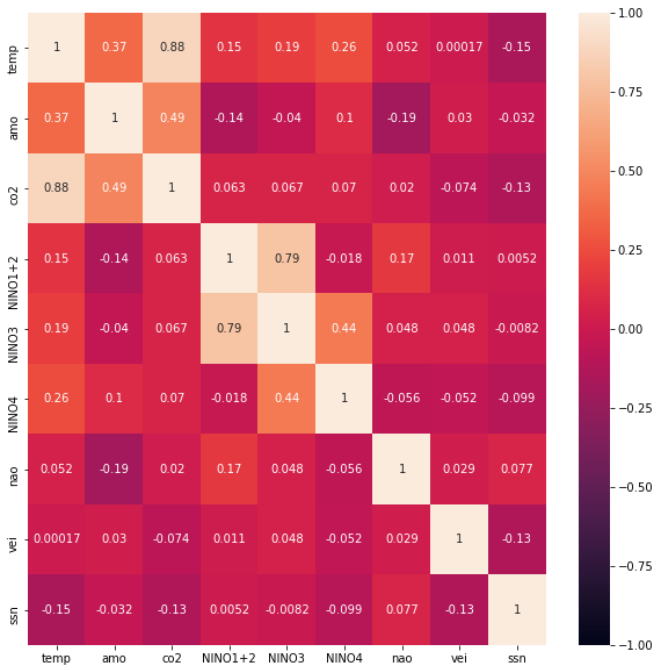


Figure 17. Heatmap matrix of dataset features

The figure shows the strongest correlation for the global temperature, which is CO2, which is 0.88. The other time series correlation with temperature is non-significant. Furthermore, we also see a strong correlation between NINO1+2 and NINO3, suggesting multicollinearity in the ENSO's 3 different time series.

The correlation is also explored thanks to XGBoost regressor, and the feature importance score can be seen from the figures below:

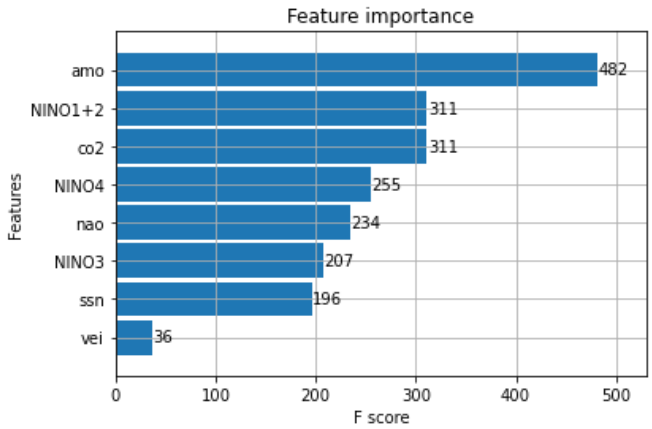


Figure 18. Feature Importance with type 'weight'

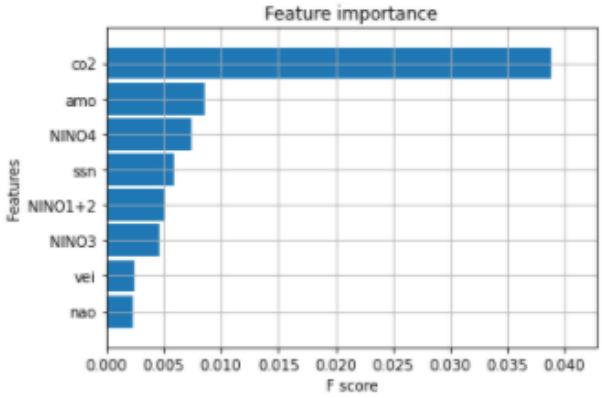


Figure 19. Feature Importance with type 'gain'

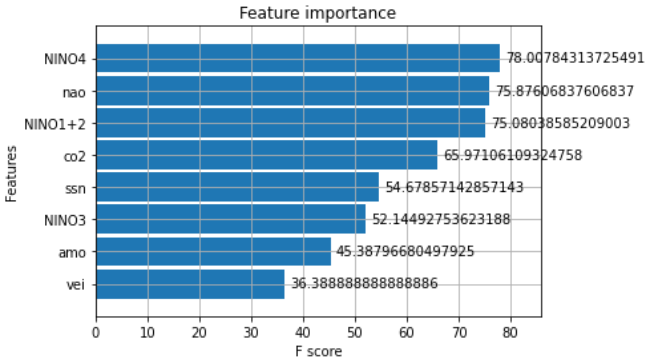


Figure 20. Feature Importance with type 'cover

XGBoost provides different indicators for the calculation of importance. We used the default type 'gain', which implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. A higher value implies more importance for generating a prediction.

Each gain, specifically from two leaves, can be calculated as:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

which can be decomposed as 1) The score of the new left leaf, 2) The score of the new right leaf, 3) The score on the original leaf, 4) Regularization on the additional leaf (XGBoost, 2020).

The second type is called 'cover', which implies the relative quantity of observations related to this feature.

The last type is 'weight', which is the percentage representing the relative number of times a particular feature occurs in the trees of the model.

Noted that, depending on the data, the feature importance orderings can be varied for different types of importance. Our research generated all different features as the most predictive feature of each type. With the metric 'gain', CO2 was found to be the dominant feature with a very high gain while all others of similarly low importance. And with the metric both 'weight' and 'cover', CO2, and the combination of NINO1+2 and NINO4 are found to have relatively high predictive power.

Another thing to notice is that AMO and NAO, a climate cycle that affects the sea surface temperature (SST), are found to have a strong predictive power for each 'weight' and 'cover' respectively, but not together. It is probably due to the inverse relationship between AMO and NAO decadal tendencies. When the Atlantic is cold (AMO negative), the AO and NAO tend more often to the positive state, when the Atlantic is warm, on the other hand, the NAO/AO tend to be more often negative (Easterbrook, 2011).

Model Performance

- Dataset

The data has 398 rows and 10 columns in total, which includes variables such as 'ds', 'y', 'amo', 'co2', 'NINO1+2', 'NINO3', 'NINO4', 'nao', 'vei', 'ssn'. The naming convention are explained in the previous Correlation part. Due to the limited availability and poor quality of the data, the period of time in consideration was sliced down to 1984 - 2017. The train-test split was such that the train dataset was from January 1984 to December 2009, and the test set was in January 2010 to February 2017. The number of rows on the training set and test set are 312 and 86 rows, whose percentage on the total amount of rows are 78% and 22%.

Volcanic Explosivity Index dataset (VEI) required some additional preprocessing since the dataset consists of volcanic as well as non-volcanic natural events that occur from 4360 BC to 2014. Therefore, in order to fit the scheme of monthly frequency with other time series, the data of non-volcanic events were filtered out and VEI was aggregated by month.

- Evaluation

In evaluation phase, three main fitness definition were used to assess the performance of each model, which includes the following:

- Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

- Pearson Correlation (PC)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Correlation Coefficient as shown in the paper

$$f_{COR}(p) = |1 - \rho(\tau, p)|$$

	Model	MAE	Pearson Correlation	Cor
Non-linear model	K Nearest Neighbors	0.14534	Nan	Nan
	Decision Tree	0.20627	0.31699	0.683003
	Extra Tree	0.17034	0.30856	0.691438
	Support Vector Regressor	0.29224	0.63606	0.363935
Ensemble models (number of trees: 100)	Adaboost	0.16533	0.50133	0.498665
	Bagging Regressor	0.16730	0.66230	0.337692
	Random Forest Regressor	0.17077	0.62655	0.373448
	Extra Trees Regressor	0.164556	0.658584	0.341416
	Gradient Boosting	0.150280	0.533882	0.466118
	XGBoost	0.180129	0.638647	0.361353
Prophet	Univariate Prophet	0.123903	0.587539	0.412461
	Multivariate Prophet with 8 extra regressors, no hyper parameters tuning	**0.094520**	**0.791021**	0.208979

Prophet performed best on in the list of prediction methods that we looked at, with the lowest MAE and best PC score overall. This was due to the fact that Prophet was able to capture the trend, seasonality, as well as incorporating the lagged effects of the extra regressors into predicting the future values.

From here, with the prophet model in mind, we performed a permutation test of the set of 8 variables, to determine the sets of variables that produce the best predictive powers with respect to the temperature dependent variable. In total, there are 2^8 distinct sets of different combinations of 8 variables, including {{}, {amo}, {amo, co2}, {amo, nino1+2}, ...}. The performance of Prophet is reported below:

prior_amo	prior_co2	prior_nao	prior_nino1	prior_nino3	prior_nino4	prior_ssn	prior_vei	MAE	corr
False	True	True	True	False	True	False	False	0.090354	0.790683
False	True	False	True	False	True	False	False	0.090443	0.788172
False	True	True	False	True	True	False	False	0.090634	0.794093
False	True	True	False	True	True	False	True	0.090769	0.788302
False	True	True	True	True	True	False	False	0.090859	0.783543
False	True	True	True	False	True	False	True	0.090909	0.781511
False	True	False	False	True	True	False	False	0.090944	0.793609
False	True	False	True	False	True	False	True	0.090968	0.782692
False	True	False	True	True	True	False	False	0.091153	0.783691
False	True	False	False	True	True	False	True	0.091301	0.788306

Figure 21. Top 10 models' performance

These are performance of the Prophet models which are trained with on the set of variables. Each of the "prior_" columns determine whether the time series are considered as an extra regressor in the model. From there, we can obtain a better understanding on which independent variables or set of variables provides the most predictive power for the temperature.

In the top 10 models, we can see that CO2 and NINO4 appeared 10 out of 10 times, indicating an individual effect or a pair effect on the prediction of the temperature. Atlantic Multidecadal Oscillation (AMO) and Sunspot number (SSN) and AMO do not appear anytime in the top 10 models. And the other time series appear 4-6 times. From here, we addressed the low power of the other time series by assigning them a low prior scale.

The hyper parameters of the final prophet model can be found in the appendix, which obtained the MAE of 0.076708 and Pearson Correlation of 0.848478. The temperature prediction of this model is presented below:

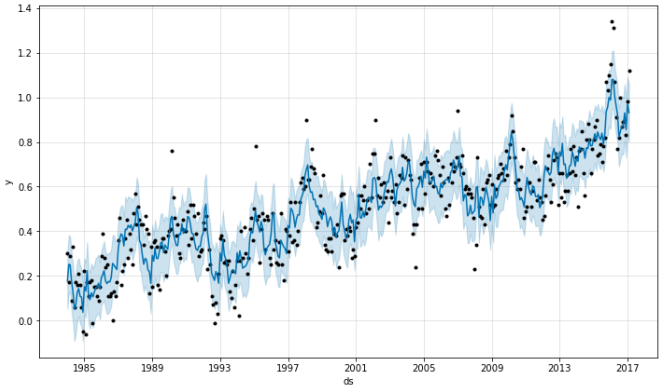


Figure 22. Model Fitting of the past data 1985 to 2017

The model was able to capture the past data quite well, with only a few outliers data not being captured. We were not able to make the prediction for after 2017, due to unavailability of CO2 data from 2017.

Cross-Validation

As aforementioned, Prophet provided an automatic cross-validation tool to evaluate the forecasts. However, the API is quite different from that of sci-kit learn. For our temperature extrapolation prediction, we have specified:

- Horizon = 365 days = 1 year: The forecast horizon's length will be out-of-sample forecasted and validated with true values. By default, Initial is set to three times, and Period is set to half the Horizon.
- Initial = 1460 days = 4 years: The initial training set length.
- Period = 180 days = ½ year: The spacing between each starting point of validation set.

Here the cross-validation evaluates the forecasts on a 365-day horizon, making predictions every 180 days after starting with the initial 730-day training set. For our time series 1984 - 2017, there are totally about (2017 - 1984 - 4) / (½) = 58 validations.

After that, we can use the library's performance_metrics utility to calculate the errors correlated to the horizon length. This way answers the question of how far the prediction can still behave well. The result is illustrated in Figure 23: the blue line is the MAE, where the mean is taken over a rolling window of absolute percent error for each validation (the dots). We can see that the mean absolute error (MAE) fluctuates around 0.1, no matter how long the horizon is.

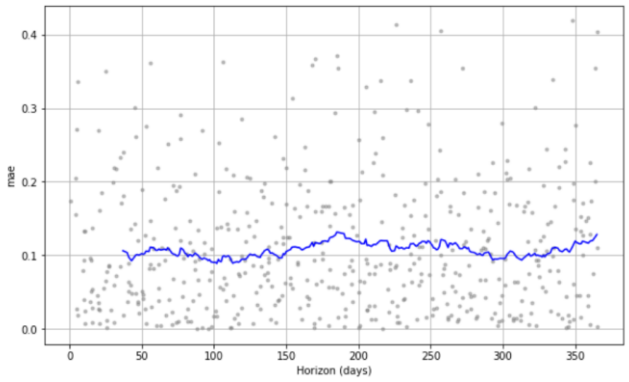


Figure 23. Performance_metric result (MAE) correlated to the horizon length

Hyperparameter parameter

There are several parameters in Prophet that can be tuned to increase the model's performance, such as changepoint_prior_scale and seasonality_prior_scale. Our project conducted an exhaustive grid search to find the best changepoint_prior_scale and seasonality_prior_scale using the built-in cross_validation to evaluate. Note that by default, changepoint_prior_scale = 0.05 and seasonality_prior_scale = 10.

	changepoint_prior_scale	seasonality_prior_scale	rmse	mae
0	0.001	0.01	0.145416	0.111594
1	0.001	0.10	0.150175	0.114740
2	0.001	1.00	0.159324	0.119823
3	0.001	10.00	0.160817	0.119961
4	0.010	0.01	0.130029	0.099586
5	0.010	0.10	0.133446	0.101880
6	0.010	1.00	0.140319	0.104705
7	0.010	10.00	0.143898	0.105498
8	0.100	0.01	0.132380	0.101908
9	0.100	0.10	0.137838	0.105403
10	0.100	1.00	0.148612	0.111207
11	0.100	10.00	0.153169	0.113130
12	0.500	0.01	0.145343	0.112280
13	0.500	0.10	0.164662	0.119340
14	0.500	1.00	0.182655	0.125094
15	0.500	10.00	0.188476	0.127068

Figure 24. Temperature extrapolation Prophet grid search results

The grid search indicated that changepoint_prior_scale = 0.01 and seasonality_prior_scale = 0.01 are the best parameters (see Figure 24). Although the performance did not significantly improve, hyperparameter tuning is always worth giving a try.

Ethical Issues

One of the largest challenges that affects the entire globe right now is climate change. This is due to many reasons, one of the largest contributors being greenhouse gases. Once greenhouse gases are emitted into the atmosphere, they can have large climate effects anywhere on the planet, regardless of where they came from. (IPCC 2007) Although all countries would collectively like to limit global emissions, especially to reduce the risk of environmental disasters, each country individually chooses to continue to pollute, which could be considered a sort of 'tragedy of the commons' situation (e.g., Soroos 1997, Helm 2008, but see Gardiner 2011a). Despite this collective continuance of destroying the environment, the countries that contribute the least to the problem at hand, tend to suffer the most, at least short term. This includes lots of third world countries that financially cannot rebuild themselves after ecological disasters. This imbalance casts a remarkable shadow over both practical, and theoretical efforts to secure any sort of global cooperation. In order to minimise the harmfulness of climate change, developed countries - those that contribute a larger part of the problem and can financially afford to do so - must take responsibility, and take action.

Another challenge involved is the lifespan of many of the greenhouse gases that have been released into the atmosphere already. The most prominent-carbon dioxide-exists in the atmosphere for a long time - an estimated 300 to 1000 years - which means negative impacts on the Earth for centuries. This means that making changes immediately is absolutely necessary.

Because the timeline of climate change is still relatively short, having become relevant only in the last two decades, it is very easy to kick the can down the road and pretend it is an issue for a different day. In only 70 years, humans have managed to destroy the planet completely, with damages lasting possibly thousands of years, as stated earlier. From the perspective of future generations, it would be best if emissions were substantially reduced as soon as possible, in order to minimise future damages to the climate. Unfortunately, this would be costly for the current generation, with the benefits being mostly for the future generations, thus, encouraging to kick the can down the road (as mentioned previously).

We mentioned how climate change hits hardest to those who emit the least (developing countries), but it is also important to mention the most innocent in this process: the animals. Despite having nothing to do with the release of large amounts of emissions, they are being affected the most; with an approximate one million species going extinct due to climate change (Danise, 2019). We have an obligation to protect coral reefs in order to preserve biodiversity, as well as unique ecosystems and the animals living in such ecosystems.

Conclusion

Although the conclusions drawn may seem daunting, in order to move forward and resolve the issues at hand, one must understand the reality of the situation. Seeing that some countries have CO2 levels that have actually *decreased* in the last 10 years could allow for inspiration and share their knowledge on how it can be done. For example, many Nordic countries have this concept of 'energy waste,' where the contents separated into this pile, are burned, and used for energy.

Considering we live in Finland, we thought it would be interesting to see how Finland compares to the rest of the world. Finland's CO2 emissions peaked in 2003 but have been on the decline since. Finland has promised to be carbon neutral by the year 2035, with the EU following in their footsteps and promising carbon neutrality by 2050. If countries with much higher levels of CO2, which can be seen in the results section, were to take inspiration from the EU or the Nordics, the planet could still be saved. Some scientists speculate that if nothing is done in the next decade or so, the changes are irreversible. Using the Facebook Prophet method on our website, we predicted more negative results, which would unfortunately mean that the future of the planet is bleak. However, the Prophet method is purely based on past data points, since it is near impossible to predict the future without including other factors such as the deals that the EU is promising to implement.

Future Idea

Although we are happy with the results, we were able to produce, with more time and better knowledge prior to the course, there are a few things we would have done differently, or can recommend for future prospects.

For one example, we recommend that in the future, it could be better to create a site that updates continuously, rather than using pre-existing pre-downloaded data (.csv's) such as we did.

Moving forward, in order to make a prediction from the Prophet model, recent data need to be collected for the extra regressors, such as CO2, AMO, VEI, etc. Without this extra regressors data, Prophet cannot make a prediction. This can be solved by imputing the missing time series value. However, the result would only be as reliable as the predicted values for the extra regressors.

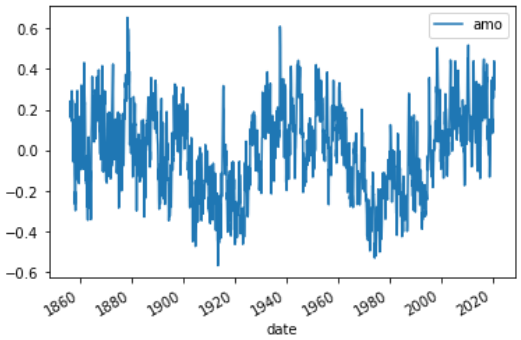
Appendix

1. Summary of Statistics of datasets

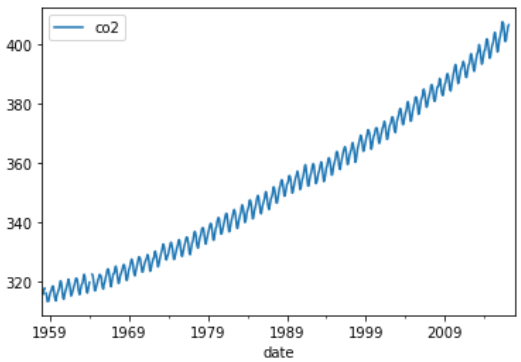
Name	#Rows	#Cols	Time frame	Frequency
Greenhouse gases	8406	4	1990-2014	Yearly
Temperature Anomaly	1680	2	1880-2019	Yearly
Sea-ice extent	26353	1	1978-2019	Every other day
Solar activity	320	1	1994-2020	Monthly
Volcanic activity (VEI)	658	36	4360 BC-2020	When volcanic activity occurs
GDP / GDP per capita	264	63	1960 - 2018	Yearly
Zonal Temperature Anomaly	140	14	1880 - 2019	Yearly
Sunspot Number	3262	1	1749-2020	Monthly
Atlantic Multidecadal Oscillation (AMO)	1980	1	1856 - 2020	Monthly
North Atlantic Oscillation (NAO)	852	1	1950-2020	Monthly
El Niño/Southern Oscillation (ENSO)	466	3	1982-2020	Monthly

2. Distribution of independent variables for prediction models

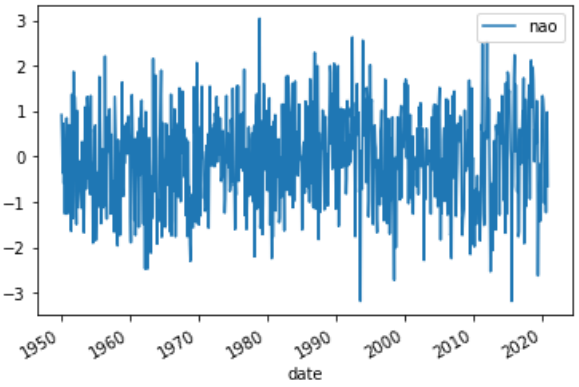
- AMO time series



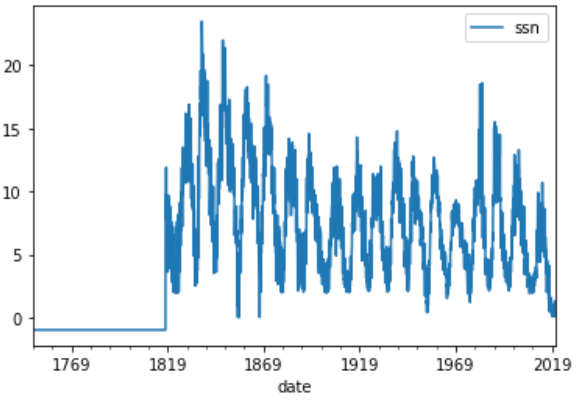
- CO2 Time Series



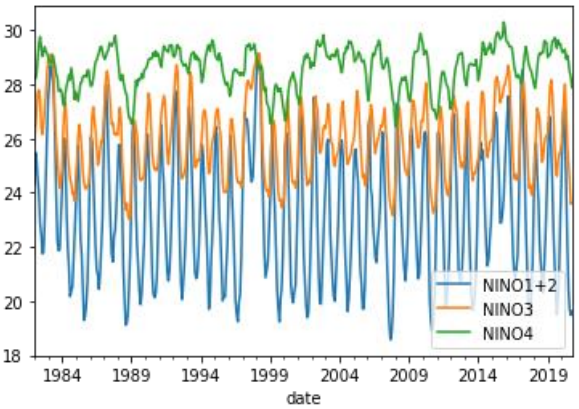
- NAO time series



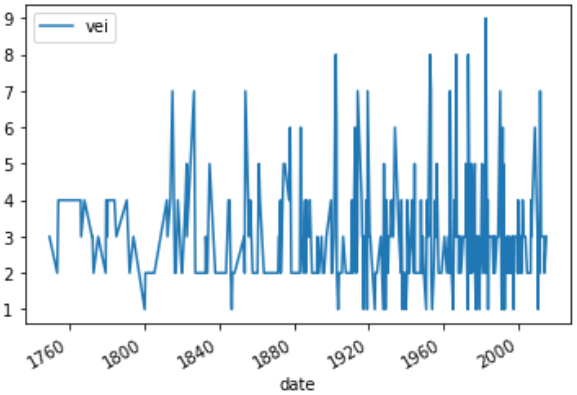
- SSN time series



- NINO time series



- VEI time series



3. Independent variables permutation

- Top 10 models' performance

prior_amo	prior_co2	prior_nao	prior_nino1	prior_nino3	prior_nino4	prior_ssn	prior_vei	MAE	corr
False	True	True	True	False	True	False	False	0.090354	0.790683
False	True	False	True	False	True	False	False	0.090443	0.788172
False	True	True	False	True	True	False	False	0.090634	0.794093
False	True	True	False	True	True	False	True	0.090769	0.788302
False	True	True	True	True	True	False	False	0.090859	0.783543
False	True	True	True	False	True	False	True	0.090909	0.781511
False	True	False	False	True	True	False	False	0.090944	0.793609
False	True	False	True	False	True	False	True	0.090968	0.782692
False	True	False	True	True	True	False	False	0.091153	0.783691
False	True	False	False	True	True	False	True	0.091301	0.788306

4. Hyperparameters of final Prophet model

```
m = Prophet(n_changepoints=100,
changepoint_prior_scale=0.5)
m.add_regressor("amo", prior_scale=0.01)
m.add_regressor("co2", prior_scale=0.5)
m.add_regressor("NINO1+2", prior_scale=0.1)
m.add_regressor("NINO3",prior_scale=0.1)
m.add_regressor("NINO4",prior_scale=0.5)
m.add_regressor("nao", prior_scale=0.1)
m.add_regressor("vei",prior_scale=0.1)
m.add_regressor("ssn", prior_scale=0.01)
```

Reference

National Oceanic and Atmospheric Administration (NOAA). 2020. Anomalies vs. Temperature. Available at: https://www.ncdc.noaa.gov/monitoring-references/dyk/anomalies-vs-temperature#:~:text=A%20temperature%20anomaly%20is%20the,average%2C%20or%20baseline%2C%20temperature.&text=A%20positive%20anomaly%20indicates%20the,was%20cooler%20than%20the%20baseline

XBGoost. 2020. Introduction to Boosted Trees. Available at: https://xgboost.readthedocs.io/en/latest/tutorials/model.html

Easterbrook, D. 2011. Evidence-Based Climate Science: Data Opposing CO2 Emissions as the Primary Source of Global Warming. ScienceDirect. Available at: https://www.sciencedirect.com/book/9780128045886/evidence-based-climate-science

Danise, C. 2019. 1 million species under threat of extinction because of humans, biodiversity report finds. Available at: https://www.nbcnews.com/mach/science/1-million-species-under-threat-extinction-because-humans-report-finds-ncna1002046