A Statistical Model of CO2 Emissions and Sea Ice Extent

ENG - 270

BÄCHLER Alicia, PRATS-GENRE Chloé

Fall Semester 2024

Contents

1 Presentation of the Project

Since the end of the 1970s, satellites have monitored sea ice overall decline, which is largely caused by global warming and increasing Earth temperature, which in turn is due to increasing CO_2 global emissions due to human activities. In our initial project we wanted to model a reduced form of an energy-balanced model, in order to describe the positive feedback loop between temperature rising, increased melting of sea ice, decrease in albedo coefficient and, as a result, rise again in temperature. However, while expanding on this idea, we were faced with major obstacles. Indeed, this approach required us to make numerous assumptions about the interactions between temperature, albedo and ice melt processes. These assumptions, while simplifying the model, introduced considerable uncertainty and limited the reliability of our results.

Moreover, as we were exploring our data, we were struck to discover differing relationships between greenhouse gases emissions and sea ice surface in the Northern and Southern Hemisphere. Hence we sought to understand why they were differing. This led us to question our initial simple linear regression estimation model. This avenue of research led us to explore how to code various statistical tests to assess the significance of our models and add new explanatory variables to better model the Southern Hemisphere relationship between sea ice and CO_2 emissions.

Ultimately, we also developed a prediction model for the Northern Hemisphere fed with three different IPCC emission scenarios in order to predict sea ice extent depending on a pessimistic, optimistic and reasonable scenario.

2 Motivation

We settled on this project because it helped us get a deeper understanding of different statistical models and run a concrete exercise on how to model and predict sea ice extent. In general, this project is interesting to highlight links between CO_2 and ice melting mirroring climate change. It provides a "low cost", publicly accessible insight, as well as demonstrates a tangible, not too complex, effect of emissions on the Earth's ecosystem, which could inspire some straight forward understanding and action given its more relatable style.

3 Data Description

3.1 Data Sources

We are using several annual datasets from various sources. The dataset from the National Snow and Ice Data Center[1] offers information on sea ice extent [million km^2] in both the northern and southern hemisphere since 1979 to the present. Sea Ice Extent refers to the total area of the ocean where sea ice is present and has at least 15% ice concentration. The Emissions Database for Global Atmospheric Research (EDGAR)[2] gives us global CO_2 emissions [Gg] by country since 1970 to the present. The last two data sets come from the National Oceanic and Atmospheric Administration (NOAA). One grants us access to precipitation data[3] [mm] over the same time span we covet. The last one provides data on the average land-ocean temperature[4] [°C] on our selected latitude (60-90°S).

3.2 Assumptions

The ice extent data is categorised by the Northern and Southern Hemispheres. Since the majority of the ice is situated at each pole, we assumed the totality of the hemisphere's ice to be limited to its respective poles. Hence, hereafter, we are going to use the terms "pole" and "hemisphere" interchangeably.

This assumption is mainly utilised for the calculation at the south pole, where we take land-ocean temperature data within the polar latitudes (60-90°S).

3.3 Preliminary Observations

We first plotted CO_2 emissions and Sea Ice Extent against time since 1979 in order to visually confirm a likely relationship between both variables.

On Figure 1, we can see an expected decrease in sea ice extent as global CO_2 emissions increase. This result is consistent with the fact that CO_2 emissions contribute to the increase in greenhouse gases, which in turn leads to higher global temperatures, and warmer temperatures lead to melting of sea ice. For the Northern Hemisphere, a strong negative correlation is observed between CO_2 emissions and sea ice extent.

The Southern Hemisphere, on the other hand, shows a less clear linear correlation. It's interesting to note that sea ice extent initially increases slightly until about 2015, after which a tipping point becomes apparent on the graph. That may indicate a change in the underlying dynamics.

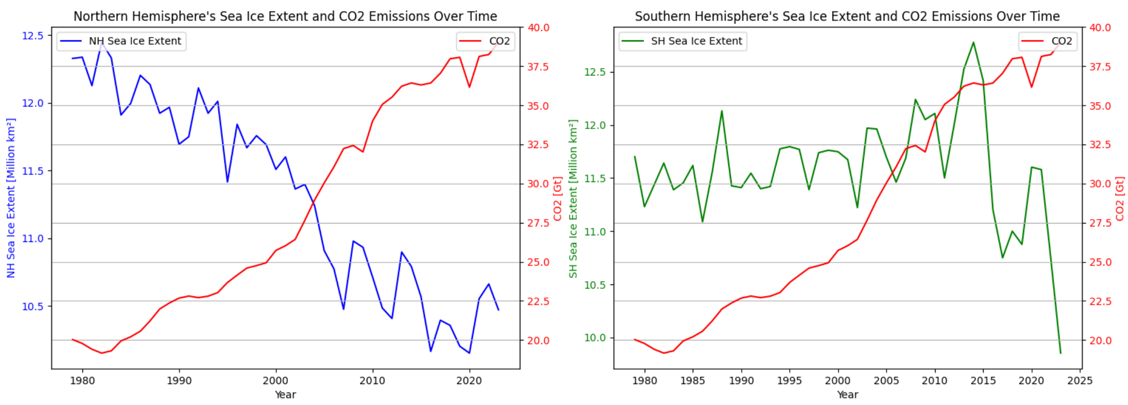

Figure 1: Average annual sea ice extent and CO_2 emissions over time in the Northern and Southern Hemisphere[5]

4 Model Description

4.1 Regression Models

For the Northern Hemisphere, we used a simple linear regression to model the relationship between sea ice extent (y) and the explanatory variable (x = CO_2 emissions)

The regression curve can be described by the following equation:

$$\hat{y} = mx + b \tag{1}$$

where b is the intercept, m is the slope, and $\hat{y}$ is the predicted value of y.

For the Southern Hemisphere, we initially applied a simple linear regression model, as we did for the Northern Hemisphere. However, this approach failed to capture the complexity of the relationships between variables (see Appendix A.1, Figure5). We then explored a quadratic regression model, aiming to account for potential non-linear interactions. Unfortunately, this also did not produce satisfactory results (see Appendix A.1, Figure 6). Finally, we developed a multiple linear regression model. After a trial and error process using various possible explanatory variables ("El nino" occurrences, winds, etc.), we settled on a model incorporating three explanatory variables: $x_1 = CO_2$ emissions, $x_2 =$ temperature of the sea, $x_3 =$ precipitation.

This model is expressed as :

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \tag{2}$$

Where b_0 is the intercept and b_1, b_2, b_3 are the regression coefficient.

4.2 Errors Metrics

To assess the model's performance and its ability to explain observed data, two statistical measure were computed :

- The Root Mean Square Error, RMSE: It calculates the average discrepancy between the actual values (y_i) and the predicted values ($\hat{y}_i$) of a statistical model.

  RMSE quantifies how dispersed the residuals (differences between predicted and observed values) are, revealing how tightly the observed data clusters around the predicted values. It can take any value from 0 to infinite, with the same unit as the dependent variable. The smaller the RMSE, the better the model's predictions are at approximating the actual observed values. It indicates that the residuals are smaller on average, meaning the model is more accurate.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

  Where n is the total number of observations.

- The coefficient of determination, R^2: It measures the proportion of variance in the dependent variable (y) that is explained by the independent variables (x_i) in the model. So, the proportion that remains, $(1 - R^2)$ is the variance that is not predicted by the model.

  In simpler terms, R^2 measures how well the model matches the data. The better a model is at making predictions, the closer its R^2 will be to 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

  Where $\bar{y}$ is the mean of y.

4.3 Statistical Tests

In addition, we used several different statistical tests to assess the accuracy of the model and the impact of its predictive parameters. Specifically, these tests evaluate the null hypothesis. The null hypothesis, H_0, assumes that the variables included in the model have no significant effect on the dependent variable (sea ice extent). Rejection of this null hypothesis through low p-values indicates that the predictors contribute meaningfully to the model.

- The F-test : A significance test for the entire regression model.

  It indicates whether our linear regression provides a better fit to our data than a model that would contain no independent variables. The null hypothesis, H_0 is that all regression coefficients are equal to zero. In other words, none of the explanatory variables has any effect on the dependent variable.

$$F = \frac{R^2}{k} \times \frac{n-k-1}{1-R^2} \tag{5}$$

Where : k is the number of independent variables (x_i)

- The t-test: A significance test for regression coefficients.

  It is a test to be applied to each estimated coefficients in the regression. It tests whether they are significantly different from zero, which means they are indeed significant. The null hypothesis, H_0 states that $\hat{b} = 0$.

$$t = \frac{\hat{b}}{\mathrm{SE}(\hat{b})} \tag{6}$$

  Where : $\hat{b}$ is the estimated regression coefficient and $\mathrm{SE}(\hat{b})$ is the standard error of the estimated coefficient.

5 Results

For greater clarity, a summary table presenting the different regression models, their respective R^2, RMSE values, and statistical test is included in the Appendix A.2 (Table 1).

5.1 Results of the Regressions

The graphs below (Figure 2) illustrate the linear regression between sea ice extent and carbon emissions. On the left, sea ice extent is plotted against global CO_2 emissions, while on the right, it is plotted over time starting at 1979.

The linear regression model for the Northern Pole shows an excellent fit to the observed data, as evidenced by the R^2 of 0,9265. This illustrates that 92,65 % of the variability in sea ice extent is explained by variation in global CO_2 emissions. The very low RMSE (0,01920 million km²) further supports the model's accuracy. In addition, the F-test of 541.98 with p-value of 1.11×10^{-16}, strongly rejects the null hypothesis that the model coefficients are zero. The very low p-value indicates that the overall regression model is highly significant.

However, while the R^2 is very high, it may reflect a spurious relation between sea ice and CO_2 emission. However, since we want to use this model as a prediction model, we kept it unchanged.

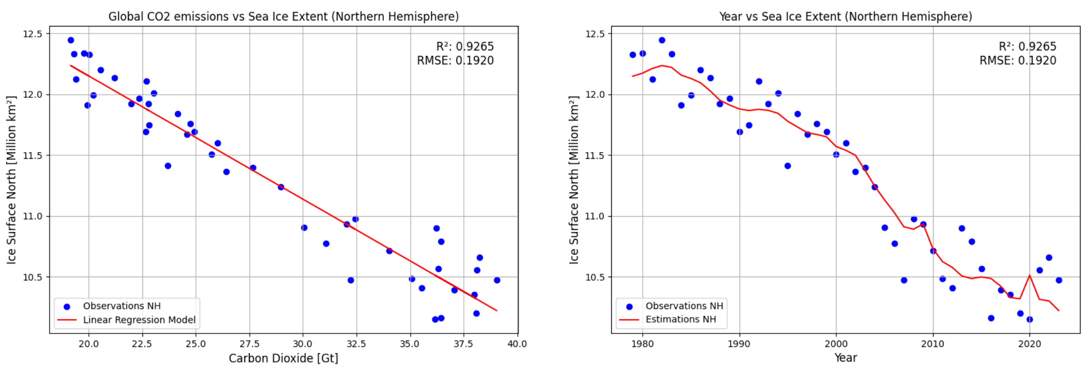

Figure 2: Linear Regression for North Pole [5]

Figure 3 illustrates the multiple regression between sea ice extent and CO_2 emissions, sea temperature and precipitation in the Southern Hemisphere, plotted against time for better visualisation.

Although the low value of R^2 (0.37%) suggests that the linear model poorly explains the variance in sea ice extent for the Southern Hemisphere; the t-test indicate that the three predictors (x_1, x_2 and x_3) are all statistically significant ($p < 0.05$), justifying their selection in the model.
Furthermore, the F statistic of 6.8349 with its low p-value (7.6922×10^{-4}) suggests that the overall regression model is statistically significant, though less than the Northern Hemisphere model.

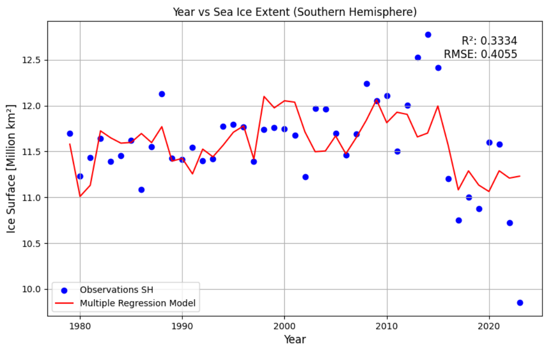

Figure 3: Multiple regression for South Pole [5]

The divergence between the two Hemisphere could be explain by the Southern Hemiphere's distinct climate dynamics. Having analysed the results for the Southern Hemisphere, we can now look at the potential reasons for the more complex relationship observed at the South Pole.

First, the Southern Ocean, which encircles Antarctica, modifies sea ice dynamics in a special way. Sea ice extent may be influenced by oceanic currents such as the Antarctic Circumpolar Current (the strongest and most continuous ocean current on Earth) differently than in the Arctic, which would make the connection with CO_2 emissions less obvious. Second, sea ice extent fluctuates drastically in the Southern Hemisphere during specific seasons, which could mask long-term patterns in response to greenhouse gas emissions.

Finally, unlike the Arctic, where ice melt and warming trends are closely linked, Antarctic ice trends may be influenced by wider environmental variability. So much so that, in some places, cooling may be caused by stratospheric ozone depletion.

5.2 Predictions

The Intergovernmental Panel on Climate Change (IPCC) [6] provides several scenarios for future CO_2 emissions. The Shared Socioeconomic Pathways (SSPs) are a collection of five scenarios that present different possible futures for global society throughout the 21st century. The IPCC established these scenarios to investigate the potential effects of socioeconomic factors on greenhouse gas emissions and attempts to mollify climate change.

We chose three scenarios to test our model and predict sea ice extent: a pessimistic, an optimistic and a reasonable scenario. Each assumes varying levels of greenhouse gas emissions, which were fed into our linear model to predict corresponding changes in sea ice extent (Figure 4, left panel). As these selected scenarios only contains CO_2 emissions, they can only be applied to the North Hemisphere model.

- Optimistic SSP1 (Sustainability): It represents a sustainable trajectory with fast CO_2 reductions and global cooperation. It is an ambitious scenario to reach the 1.5 $^\circ$ C target of the Paris Agreement.
- Intermediate SSP3 (Regional Rivalry): It represents a fragmented world with high challenges to both mitigation and adaptation.

- Pessimistic SSP5 (Fossil-fuelled Development): It represents a world of rapid economic growth based on fossil fuels, with high challenges to mitigation but low challenges to adaptation. It reflects a future with high emissions, minimal cooperation, and significant warming.

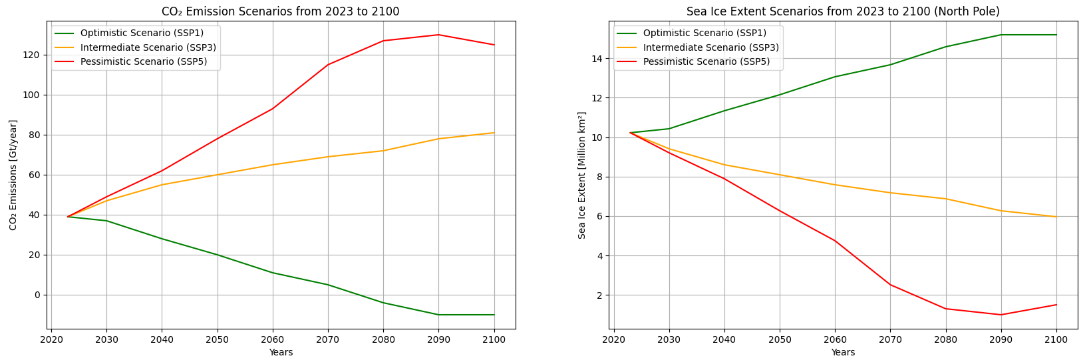

Figure 4: Predictions curve

The right panel of Figure 4 shows that sea ice extent will continue to diminish as CO_2 emissions continue to increase. However, an optimistic scenario of rapidly decreasing emissions shows that sea ice could start expanding again quite quickly, providing an helpful feedback loop to the global climate.

6 Conclusion

This project has shown us that a linear regression is very adequate to easily show the correlation of the CO_2 emissions and the sea ice though we cannot establish a distinct causality between both variables. Due to the Southern hemisphere's intricate data, we found the multiple regression variable to be the best approximation we managed, with the addition of precipitation and sea temperature data. This model confirms that the ice melting is not merely caused by CO_2 emissions but also many other factors. It is also pleasant to note that under an optimistic emissions scenario, our results suggest the potential for stabilisation and even recovery of sea ice extent. This underlines the importance of a continued global efforts towards reducing the CO_2 emissions.

In order to further improve our results, we could upgrade our estimations by taking into account additional relevant data such as other greenhouse gases like methane and nitrous oxide. Other important influences could be wind, oceanic currents, cloud coverage and salinity levels. It would also be interesting to further try different non-linear regressions and introducing lags (such as the effect of past CO_2 emissions). The idea of processing the data on a monthly or seasonal basis is interesting to improve precision; nevertheless it would in turn need a more processing power given the big time span.

This project could be expanded in numerous varying ways, reflecting the vast complexity of climate system. This small yet valuable project highlights not just the challenges but also the exciting potential for ongoing exploration in understanding and mitigating the impacts of climate change.

Credit

We took inspiration from the work of Noah Alviz (2020), who uses linear and polynomial regression models to predict the extent of Arctic sea ice. However, our approach modifies and improves on this idea by incorporating new explanatory variables, such as temperature and precipitation, as well as applying multiple regression models to better capture the complex dynamics of the Southern Hemisphere[7].

We also worked with the reference course "Probabilité et Statistique, Spring 2024" by Linda Mhalla, a Professor of Probability and Statistics at EPFL.

References

[1] F. Fetterer et al. *Sea Ice Index (G02135, Version 3)*. Sea_Ice_index_Monthly_Data_by_Year_G0213 Boulder, Colorado, USA, 2017. URL: `https://doi.org/10.7265/N5K072F8`.

[2] EDGAR (Emissions Database for Global Atmospheric Research) Community. *IEA-EDGAR CO2 (v3), a component of the EDGAR Community GHG Database version EDGAR_2024_GHG*. Based on data from IEA (2023) Greenhouse Gas Emissions from Energy. European Commission, Joint Research Centre (JRC) and International Energy Agency (IEA), 2024. URL: `https://edgar.jrc.ec.europa.eu/report_2024`.

[3] NOAA National Centers for Environmental Information. *Climate at a Glance: Global Time Series*. Dec. 2024. URL: `https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series`.

[4] B. Huang et al. *NOAA Global Surface Temperature Dataset (NOAAGlobalTemp), Version 6.0.0*. aravg.ann.ocean.90S.60S.v6.0.0.202410. 2024. URL: `https://doi.org/10.25921/rzxg-p717`.

[5] Credit : Sea Ice Index, National Snow and Ice Data Center.

[6] IPCC. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by Core Writing Team, H. Lee, and J. Romero. Geneva, Switzerland: IPCC, 2023, pp. 35–115. DOI: `10.59327/IPCC/AR6-9789291691647.001`.

[7] Noah Alviz. *Using Linear and Polynomial Regression to Predict Arctic Sea Ice Extent*. Accessed: [16 november 2024]. 2020. URL: `https://noahalviz.medium.com/using-linear-and-polynomial-regression-to-predict-arctic-sea-ice-extent-98863c87555c`.

Appendix A.1

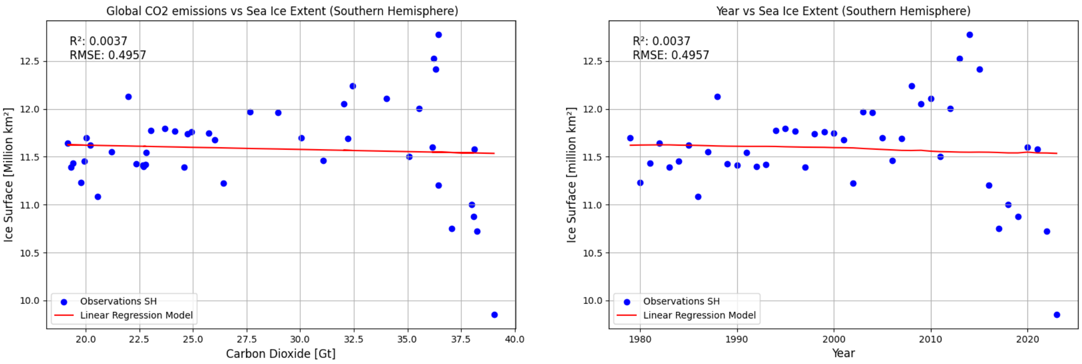

Figure 5: Unsatisfactory Linear Regression for South Pole [5]

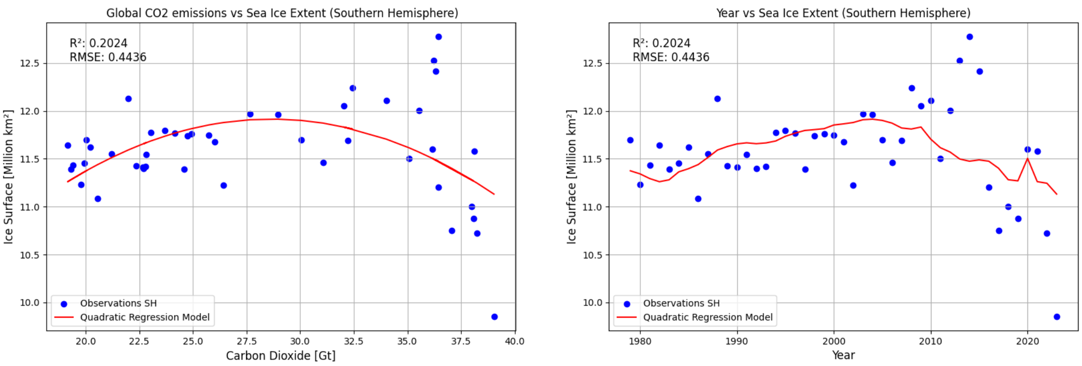

Figure 6: Unsatisfactory Quadratic Regression for South Pole [5]

Appendix A.2

Variable	Coefficient	t-value	p-value	R^2
Northern Hemisphere				
m_N (CO2)	-0.101450	-23.2806	0.0000e+00	0.9265
b_N	14.180093	112.3895	0.0000e+00	
F-statistic (NH)		541.9819	1.1102e-16	
Southern Hemisphere				
b_0	3.523724	1.0550	2.9762e-01	
b_1 (CO2)	-0.033887	-2.6288	1.2005e-02	
b_2 (Sea Temperature)	-3.928913	-3.9547	2.9649e-04	
b_3 (Precipitation)	0.008648	2.5354	1.5141e-02	
F-statistic (SH)		6.8349	7.6922e-04	0.3334
Regression Models				
North Pole	$y = -0.101450 \cdot x + 14.180093$		**RMSE**: 0.191951	$\mathbf{R^2}$: 0.9265
South Pole (Lin)	$y = -0.004470 \cdot x + 11.709924$		**RMSE**: 0.495735	$\mathbf{R^2}$: 0.0037
South Pole (Quad)	$y = -0.007275 \cdot x^2 + 0.416823 \cdot x + 5.943379$		**RMSE**: 0.443555	$\mathbf{R^2}$: 0.2024
South Pole (Mult)	$y = 3.523724 - 0.033887 \cdot x - 3.928913 \cdot \text{temp} + 0.008648 \cdot \text{precip}$		**RMSE**: 0.405475	$\mathbf{R^2}$: 0.3334

Table 1: Results of Statistical Tests