Statistička analiza podataka o klimatskim promjenama

Sorić, Ante

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

BACHELOR THESIS No. 1570

STATISTICAL ANALYSIS OF CLIMATE CHANGE DATA

Ante Sorić

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

BACHELOR THESIS No. 1570

STATISTICAL ANALYSIS OF CLIMATE CHANGE DATA

Ante Sorić

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Zagreb, 04 March 2024

BACHELOR THESIS ASSIGNMENT No. 1570

Student:	**Ante Sorić (0036539765)**
Study:	Electrical Engineering and Information Technology and Computing
Module:	Computing
Mentor:	assoc. prof. Marina Bagić Babac

Title: **Statistical analysis of climate change data**

Description:

This thesis explores diverse data from the internet on climate change with the aim of identifying relevant variables and building regression models to analyze correlations and forecast changes in climatic conditions. Through descriptive statistical analysis, the basic characteristics of the data need to be explored to identify key variables that could impact climate change. The objective is to determine statistically significant relationships between different variables and their potential influence on climatic patterns, as well as to develop regression models to assess the impact of various factors on climatic conditions, providing a foundation for understanding the dynamics of climate change and its potential consequences.

Submission date: 14 June 2024

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1570

Pristupnik:	**Ante Sorić (0036539765)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentorica:	izv. prof. dr. sc. Marina Bagić Babac

Zadatak: **Statistička analiza podataka o klimatskim promjenama**

Opis zadatka:

Ovaj završni rad istražuje raznolike podatke s interneta o klimatskim promjenama s ciljem identifikacije relevantnih varijabli i izgradnje regresijskih modela za analizu povezanosti i prognoziranje promjena u klimatskim uvjetima. Kroz deskriptivnu statističku analizu treba istražiti osnovne karakteristike podataka kako bi se identificirale ključne varijable koje bi mogle utjecati na klimatske promjene. Cilj je utvrditi statistički značajne veze između različitih varijabli te njihov potencijalni utjecaj na klimatske obrasce te izraditi regresijske modele za procjenu utjecaja različitih faktora na klimatske uvjete koji pružaju temelj za razumijevanje dinamike klimatskih promjena i njihovih potencijalnih posljedica.

Rok za predaju rada: 14. lipnja 2024.

Acknowledgements

I am grateful to my mentor, prof. dr. sc. Marina Bagić Babac, for her valuable advice, help and patience in writing this paper.

Content

Introduction

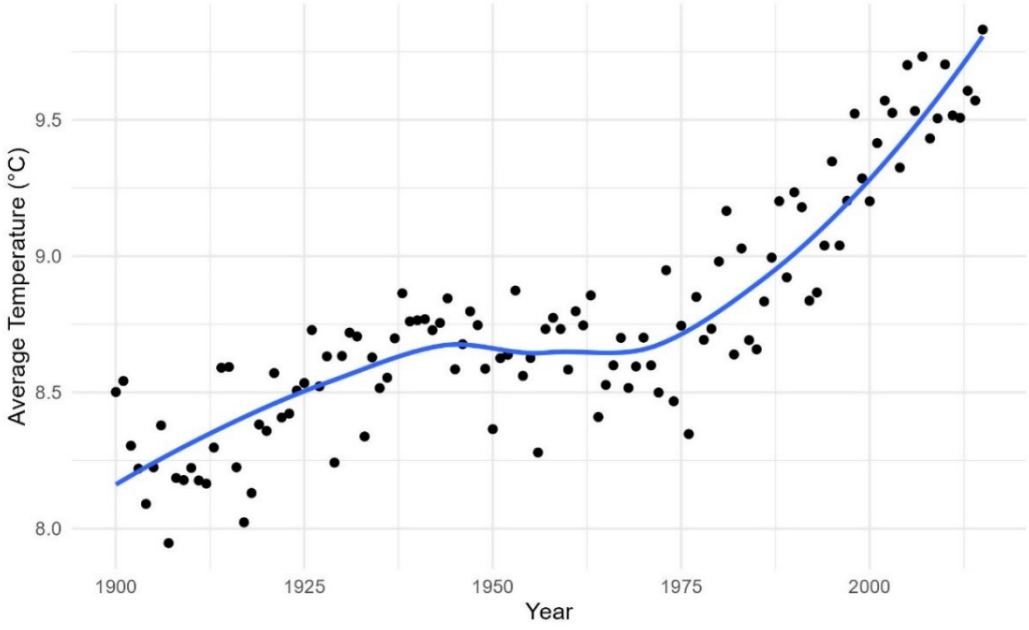
Climate change is a fundamental problem, and its consequences can be felt worldwide. We often think it's going to affect us badly in the future, but it is already an ongoing process, and it's only getting worse. It is currently arguably the most pressing issue of our time, affecting every corner of our planet Earth. Lack of rain in places where it is needed can cause a shortage of specific foods all around the world. Extreme summer heat can lead to water shortage, especially in countries with less natural water (water sources, rivers etc.).

As we can see in (NOAA, Climate Change Impacts), climate change's impacts vary across sectors. Droughts disrupt agriculture, floods threaten infrastructure, and extreme weather poses health risks. Vulnerable populations are disproportionately affected due to social inequities. Investing in clean energy fosters economic growth while curbing health hazards.

The global average temperature has been rising consistently, especially in comparison to the period between 1961 and 1990. In recent decades, there has been a significant and noticeable increase in global temperatures, approximately by 0.7°C compared to the temperatures recorded during 1961–1990. Moreover, if we trace back to 1850, temperatures were even lower, by about 0.4°C, compared to the 1961–1990 period. Combining these changes, we see an overall temperature rise of around 1.1°C. The average global temperature has increased by approximately 1.1 to 1.2°C from 1850 to 2019.

Temperature change is observed with respect to a baseline climatology, corresponding to 1951–1980. The following Figure shows a change in average temperature for each year since 1900 and every country. Every black dot represents a unique value for each country. It is easily seen that the values are rising, as we can see that the average before 1920 is around 8.25, while post-2010 averages are even higher than 9.5. We can see that the increase is around or even higher than the already mentioned critical 1.5 degrees Celsius.

Figure 1 – Average temperature change



Source: figure by the author

The research is focused on multiple questions.

First, a time series analysis of temperature data is done, and possible time patterns are investigated. Also, statistical methods are used to establish correlations between multiple factors, including temperature rise, ocean level rise, severe weather reports (Pacific and Atlantic hurricane reports), global mean sea level etc.

1. Related work

(Naiqian, et al., 2019) aims to model and forecast global land-surface air temperatures to understand the ongoing trend of global warming. The researchers utilized monthly mean temperature data from 1880 to 2018, sourced from NASA's Goddard Institute for Space Studies (GISS). They first processed the data to ensure it was stationary by applying differencing techniques. After dealing with outliers, they tested several ARIMA models, ultimately selecting the ARIMA(2,1,1) model as the most accurate for their data. This model indicated that the overall temperature trend shows a significant increase, especially after 1980. The study found that the global temperature has been steadily rising, particularly in recent decades, and predicts this warming trend will continue. The forecast for 2019 and 2020 also suggested a continued increase in global temperatures. The researchers concluded that global warming is a persistent and accelerating problem, necessitating effective measures to address it. The study emphasizes the importance of short-term temperature forecasting as a tool for understanding and responding to climate change.

(Pielke Jr, et al., 2005) examines the complex relationship between global warming and hurricane activity. The researchers aimed to clarify whether global warming influences the frequency and intensity of hurricanes and their associated impacts. The study reviewed existing peer-reviewed literature and historical data on hurricanes, particularly focusing on event risk (the occurrence and intensity of hurricanes) and outcome risk (the impact of hurricanes on society). The study found no clear evidence linking the increase in hurricane frequency or intensity directly to global warming. While some studies suggest a potential connection, the findings are not definitive, and the data often shows considerable variability rather than a clear trend. The researchers highlighted that the increase in the number of storms and their intensity in recent decades, particularly since 1995, could largely be attributed to natural multidecadal variability rather than climate change. They also noted that global modelling studies on the future impact of global warming on hurricanes yield contradictory results, making it difficult to draw firm conclusions. The study concludes that claims of a direct connection between global warming and hurricane impacts are premature, given the current state of scientific understanding. The researchers argue that more robust evidence and consistent modelling results are needed before any definitive link can be established.

(Chen, et al., 2023.) utilized a linear regression to explore the relationship between atmospheric CO_2 levels and global temperature changes. Despite the complexity of climate dynamics, linear regression is shown to be a robust and effective tool for certain aspects of climate modelling. One of the key strengths of linear regression highlighted in the paper is its ability to provide a clear and quantifiable relationship between variables. In this case, the study found a strong correlation between CO_2 levels and global temperatures, with a correlation coefficient of 0.96. This high correlation demonstrates that linear regression can capture significant trends in the data, making it a valuable tool for predicting how increases in CO_2 might affect global temperatures. In the study, linear regression helped to establish a baseline understanding of the relationship between CO_2 and temperature, which was then used to guide the development of more sophisticated models like ARIMA and LSTM. The paper demonstrates that even with more advanced models available, linear regression still plays a vital role in the overall analysis, providing reliable and interpretable predictions. While the paper acknowledges the limitations of linear regression, especially when dealing with non-linear or complex interactions over long periods, it also emphasizes that linear regression is "good enough" for certain applications. It effectively captures the main trend in the relationship between CO_2 and temperature, which is essential for understanding the broader implications of rising greenhouse gas levels.

(Bhagat, et al., 2023) forecasted future global sea levels using machine learning techniques. The researchers employed linear regression and gradient descent to build predictive models. Linear regression was used to establish a relationship between time and sea level, allowing for predictions based on this linear relationship. Gradient descent, an optimization algorithm, was utilized to minimize errors in the predictive model. The data used in this study included global mean sea level (GMSL) records from 1880 to 2013. The study's findings indicate that global sea levels will continue to rise significantly in the coming years, with an estimated increase of approximately 20 centimetres from 2014 to 2050. This prediction is supported by a notable rise in sea levels starting around the year 2000, which is consistent with trends related to global warming and the melting of ice sheets. The accuracy of the models was evaluated using the R^2 score, with the linear regression model achieving the highest accuracy at 94.6%. In conclusion, the study predicts a continued rise in global sea levels due to ongoing climatic changes, posing potential risks to coastal communities and ecosystems.

(Wang & Wu, 2021) examined changes in polar sea ice over 30 years using data from the National Snow and Ice Data Center (NSIDC). The study shows a general decrease in sea-ice concentration in the Arctic, with significant regional variability in the Antarctic. The total sea ice extent decreased by about 2.07% per decade, with multiyear ice declining 12.31% per decade, while first- and second-year ice slightly increased by 2.13% per decade. Linear regression was used to analyse trends, revealing a significant overall decline in sea ice. The Arctic showed a consistent decrease across most areas, while the Antarctic exhibited a mix of increasing and decreasing trends, with regions like the Weddell Sea seeing increases and the Amundsen and Ross seas experiencing decreases. This contrast highlights different responses to climate change in the Arctic and Antarctic, with the Arctic showing more uniform ice loss and the Antarctic displaying varied regional patterns.

(Magi, 2008) examines the impact of increasing atmospheric CO_2 on ocean acidification and evaluates the effectiveness of CO_2 Ocean sequestration as a mitigation strategy. The study highlights that as atmospheric CO_2 levels rise, a significant portion of this CO_2 is absorbed by the ocean, leading to increased acidification of the ocean's surface layers. The CO_2 dissolves in seawater, forming carbonic acid, which dissociates into bicarbonate and hydrogen ions, thereby lowering the pH of the water. This ongoing acidification poses a serious threat to marine life, particularly organisms that rely on calcium carbonate for their shells and skeletons, as lower pH levels reduce the availability of carbonate ions needed for calcification. The simulations show that without mitigation, the continued increase in atmospheric CO_2 would lead to significant acidification of the ocean's surface, with potentially devastating effects on marine ecosystems.

2. Methodology

In this paper, various datasets from Kaggle and other resources on the topic of climate change indicators are used.

(Kaggle, Climate change – earth surface temperature data) is used for tracking Earth's surface temperature data.

The dataset contains global temperature records, including measurements of average, maximum, and minimum monthly temperatures for both land and combined land-ocean surfaces. The dataset spans from 1750 to recent years, totalling 3192 entries with nine columns. In the next Table, a random dataset sample is provided.

Table 1 – random dataset sample (earth surface temperature)

Date	Land Average Temperature	Land Average Temperature Uncertainty	Land Max Temperature	Land Max Temperature Uncertainty	Land Min Temperature	Land Min Temperature Uncertainty
1988-10-01	9.979	0.062	15.719	0.194	4.363	0.151
1763-10-01	5.535	2.961	NaN	NaN	NaN	NaN
1898-07-01	14.138	0.367	20.145	0.495	8.039	0.306
1936-03-01	5.031	0.194	11.032	0.197	-1.154	0.218
1893-12-01	3.155	0.399	9.030	0.429	-2.790	0.321

Source: table by the author

The research uses data starting in 1750 for average land temperature and 1850 for maximum and minimum land temperatures, as well as global ocean and land temperatures. The columns include *LandAverageTemperature*, which represents the global average land temperature in Celsius, and *LandAverageTemperatureUncertainty*, indicating the 95% confidence interval around the average.

There are missing values in many columns, particularly those related to maximum and minimum temperatures and combined land-ocean temperatures, so the analysis focuses on average temperature readings. The dataset spans from January 1750 onwards, with

temperature values recorded in Celsius and uncertainties provided to indicate measurement reliability.

The column statistics are as follows: *LandAverageTemperature* has a mean of 8.48°C, a standard deviation of 4.35, and a range from -6.36°C to 22.66°C; *LandMaxTemperature* has a mean of 14.35°C, a standard deviation of 4.31, and a range from 5.90°C to 21.32°C; *LandMinTemperature* has a mean of 2.74°C, a standard deviation of 4.16, and a range from -5.41°C to 9.72°C; and *LandAndOceanAverageTemperature* has a mean of 15.21°C, a standard deviation of 1.27, and a range from 12.48°C to 17.61°C.

(Kaggle, Hurricane database) is used for hurricane data in the Atlantic and Pacific Oceans. It consists of two datasets, containing Atlantic and Pacific data, which are combined.

The combined dataset contains hurricane records from both the Atlantic and Pacific regions, spanning from 1851 to recent years, totalling 75242 entries with eleven columns. In the next Table, a random dataset sample is provided.

Table 2 – random dataset sample (hurricane database)

ID	Name	Date	Time	Status	Latitude	Longitude	Maximum Wind (knots)	Minimum Pressure (mb)	Ocean
EP041983	DALILIA	1983-07-11	12:00	TD	19.7N	125.5W	25	-999	Pacific
EP081993	GREG	1993-08-28	12:00	TD	21.3N	140.6W	25	1010	Pacific
AL021982	UNNAMED	1982-06-18	18:00	SS	31.4N	80.3W	60	992	Atlantic
AL081871	UNNAMED	1871-10-10	18:00	HU	24.4N	64.1W	70	-999	Atlantic
AL011945	UNNAMED	1945-06-23	0	TS	25.9N	86.6W	50	-999	Atlantic

Source: table by the author

The research uses columns including the date of tropical storm occurrence, which start from 1750 but focuses more on later years due to the availability of more detailed information. The status column indicates the storm's classification, such as *Tropical Wave*, *Tropical Depression*, *Tropical Storm*, and *Hurricane*, with the research primarily focusing on hurricanes. The *Maximum Wind* column records the highest wind speed, and the *Minimum Pressure* column captures the lowest air pressure recorded during the storms. Initial observations indicate some missing values, particularly in the *Minimum Pressure* and *Maximum Wind* columns. The data spans from June 25, 1851, to November 29, 2015. The wind speed is provided in knots and converted to kilometres per hour (km/h) for easier interpretation, while air pressure is measured in millibars (mb). Column statistics reveal that the mean maximum wind speed is 93.95 km/h with a standard deviation of 49.56, ranging from 18.52 to 342.62 km/h. The mean minimum pressure is 993.40 mb with a standard deviation of 18.75, ranging from 872.0 to 1024.0 mb.

(Our world in data, CO_2 emissions) is used for global air pollution data. The dataset contains CO_2 emissions records from various countries and regions, spanning from 1750 to recent years, totalling multiple entries with four columns. In the next Table, a random dataset sample is provided.

Table 3 – random dataset sample (CO_2 emissions)

Entity	Code	Year	Annual CO_2 Emissions
World	OWID_WRL	1910	3,034,090,000
Algeria	DZA	1955	4,070,131
Argentina	ARG	2000	136,498,734
Australia	AUS	1970	157,460,918
Austria	AUT	1980	56,333,437

Source: table by the author

The research utilizes columns that include the entity name (country or region), the corresponding country or region code, the year of the CO_2 emissions record, and the total annual CO_2 emissions measured in tonnes. There are missing values in some columns,

particularly in the *Annual CO_2 Emissions* column. The data spans from 1750 to recent years, providing a long-term view of CO_2 emissions over time. The *Annual CO_2 Emissions* values show a mean of 391,272,200 tonnes, with a standard deviation of 1,855,825,000 tonnes, and a range from 0 to 37,149,790,000 tonnes. The research primarily focuses on global CO_2 emissions data.

(Kaggle, sea level change dataset) is used for tracking global mean sea level (GMSL) changes over time. The dataset contains sea level records spanning from 1880 to recent years, totalling 1608 entries with three columns. In the following Table, a random dataset sample is provided.

Table 4 – random dataset sample (sea level change)

Time	GMSL	GMSL Uncertainty
1880-03-15	-164.3	24.2
2000-07-15	52.4	3.4
1955-11-15	-9.2	4.6
1990-05-15	24.8	4.3
2015-01-15	70.6	3.2

Source: table by the author

The research dataset includes columns for the date of sea level measurement, global mean sea level (GMSL) in millimetres, and the uncertainty in the GMSL measurement. Dataset is complete with no missing values. The data spans from 1880 to recent years, providing a long-term view of sea level changes. The *GMSL* column records global mean sea level measurements, with a mean of -66.08 mm, a standard deviation of 62.89 mm, and a range from -184.5 to 82.4 mm. The *GMSL Uncertainty* column captures the measurement uncertainty, with a mean of 11.30 mm, a standard deviation of 5.28 mm, and a range from 6.2 to 24.2 mm.

(Kaggle, Daily Sea ice extent) is used for tracking global ice sea extent on both hemispheres. The research dataset includes columns for the year, month, and day of the sea ice measurement, as well as the sea ice extent recorded in million square kilometres. It also includes a column to indicate missing data, the source of the data, and the hemisphere (north or south) where the measurement was taken. The dataset spans from 1978 to recent years, containing a total of 26,354 entries across these seven columns. A random sample of this dataset is provided in Table 5 for further analysis.

Table 5 – random dataset sample (daily sea ice extent)

Year	Month	Day	Extent	Missing	Hemisphere
1978	10	26	10.231	0	north
2000	07	15	9.456	0	north
1995	12	10	12.345	0	north
1985	04	25	14.231	0	north
2015	01	03	13.678	0	north

Source: table by the author

Dataset is complete with no missing values. The data spans from 1978 to recent years, with sea ice extent measurements recorded in the *Extent* column, expressed in million square kilometres. The column statistics for sea ice extent show a mean of 11.49 million square kilometres, a standard deviation of 4.61 million square kilometres, and a range from 2.08 to 20.20 million square kilometres. For further research, the dataset is transformed to calculate yearly mean sea ice extent values, separated by hemisphere, to facilitate analysis. A sample of this transformed dataset is presented in the following Table.

Table 6 – random dataset sample (transformed dataset)

Year	Hemisphere	Mean Extent (million sq. km)
1978	north	10.231
1978	south	12.345
1979	north	11.001

1979	south	13.678
1980	north	12.345
1980	south	14.231

Source: table by the author

For mean sea ice extent research, dataset with average temperatures by hemispheres was useful (NCEI, Global time series). The dataset is divided into two sections: the Northern Hemisphere and the Southern Hemisphere, each detailing yearly temperature anomalies relative to a baseline. The dataset spans 174 years, containing 348 records with three columns. In the following Table, a random dataset sample is provided.

Table 7 – random dataset sample (average temperature anomaly on each hemisphere)

year	anomaly	hemisphere
1909	-0.50	north
1929	-0.11	north
1940	-0.04	south
1998	0.63	north
2001	0.74	north

Source: table by the author

The research dataset includes columns for the year, temperature anomaly in degrees Celsius, and the corresponding hemisphere (*north* or *south*). The data spans from 1851 to 2024, providing insights into long-term temperature trends. For the Northern Hemisphere, the average anomaly is 0.13°C with a standard deviation of 0.84°C, and values range from -0.83°C to 2.13°C. For the Southern Hemisphere, the average anomaly is 0.05°C with a standard deviation of 0.55°C, ranging from -0.71°C to 1.18°C.

(EEA, Decline in ocean pH) is used for tracking ocean acidification. The dataset consists of columns including the year of the pH measurement, the specific date of the measurement, the pH value, and the uncertainty in the pH measurement. The dataset spans from 1985 to 2022, containing a total of 39 entries. A random sample of this dataset is provided in Table 8.

Table 8 – random dataset sample (ocean pH)

Year	Date	pH	Uncertainty
2010	1.7.2010	8.071523	0.013109959
1996	1.7.1996	8.094436	0.01376242
2004	1.7.2004	8.081276	0.013095527
1989	1.7.1989	8.103392	0.015003047
2018	1.7.2018	8.055486	0.013702162

Source: table by the author

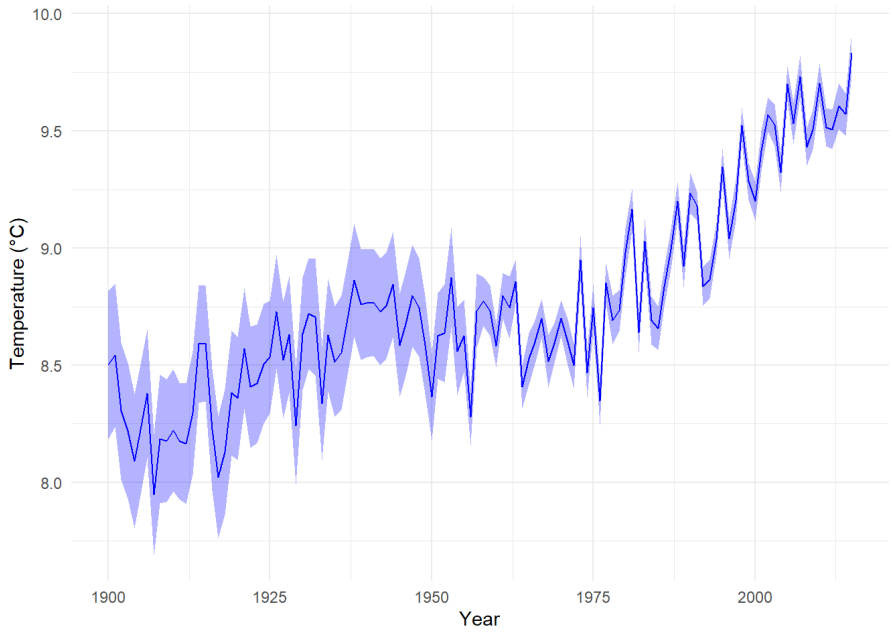
Initial observations indicate that the dataset is complete with no missing values across all columns. The data spans from 1985 to 2022. The *pH* column contains ocean pH measurements, while the *Uncertainty* column provides the associated measurement uncertainty. The column statistics for pH show a mean of 8.0803, a standard deviation of 0.0218, and a range from 8.047 to 8.110.

3. Time series analysis of temperature data

Effective forecasting begins with a clear definition of the forecasting problem, followed by the selection of relevant variables and appropriate methods. This process requires a deep understanding of the data and context to ensure accurate and useful predictions. Developing a reliable forecast involves defining the problem, gathering data, exploring, and visualizing the data, choosing and applying a forecasting method, and evaluating the model's performance.

The first thing to do is to plot the data. Visualization helps us understand patterns even before analysis is done. Changes over time, relationships between variables, outliers etc. can be spotted in the next Figure. It represents time series data as a plot, with included uncertainties. It should be noted that this plot only represents yearly mean values. A clear upward trend can be seen.

Figure 2 – Yearly land average temperature with uncertainty over time

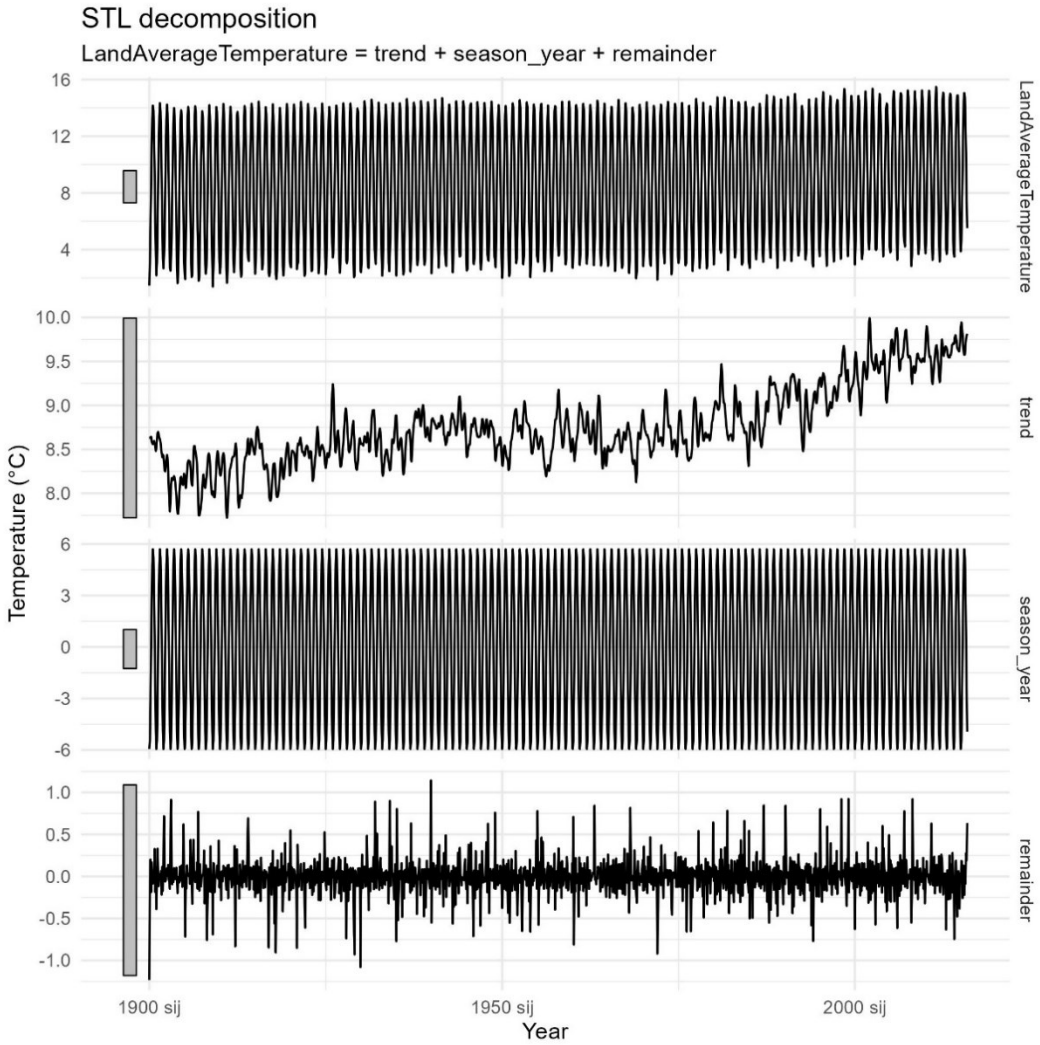


Source: figure by the author

Time series decomposition is essential for breaking down a series into its underlying components – trend, seasonal, and residual. This helps in understanding the patterns and behaviours in the data, making it easier to analyse and forecast accurately. In this paper, STL decomposition is used. STL (Seasonal-Trend decomposition using LOESS) is a powerful method for decomposing time series data into three components: trend, seasonal, and remainder. This method is highly flexible and can handle any type of seasonality, making it

robust to outliers and useful for a wide range of time series applications. STL uses LOESS (locally estimated scatterplot smoothing) to iteratively estimate the trend and seasonal components, allowing for clear separation and a better understanding of the underlying patterns in the data. The Figure below illustrates our STL decomposition, where 7 years were used for the trend-cycle window parameter, which represents the number of consecutive observations to be used when estimating the trend-cycle.

Figure 3 – STL decomposition of time series



Source: figure by the author

There is a clear upward trend in temperatures over the past century, suggesting a long-term increase in global average temperatures. There is a strong seasonal pattern in the data, with temperatures fluctuating in a regular, annual cycle. This is because of the nature of the temperature changes throughout the months. The residuals (random noise) do not show any

obvious pattern, indicating that the model has effectively captured the trend and seasonal components. The remainder component suggests that while there are short-term irregularities, the major patterns are well-captured by the trend and seasonality.

ARIMA models are used in this research paper. The Autoregressive Integrated Moving Average (ARIMA) model is a widely used statistical method for time series forecasting. It combines three key components: autoregression (AR), moving average (MA) and differencing. The autoregression part models the relationship between an observation and several lagged observations, the differencing part makes the time series stationary by subtracting observations from previous time steps, and the moving average part models the relationship between an observation and a residual error from a moving average model applied to lagged observations. ARIMA is particularly useful for understanding and predicting future points in a time series by considering past values and the noise in the data. ARIMA models are denoted as ARIMA(p, d, q), where p represents the number of lag observations in the autoregressive component, d signifies the degree of differencing needed to make the series stationary, and q stands for the size of the moving average window.

A Seasonal ARIMA(SARIMA) model incorporates both non-seasonal and seasonal components, allowing it to effectively model data with regular seasonal fluctuations. The identification process involves using ACF and PACF plots to determine the appropriate orders for autoregressive and moving average terms, as well as the degree of differencing needed to achieve stationarity.

An ARIMA model with seasonal components is written as ARIMA(p,d,q)$(P,D,Q)_m$ Here:

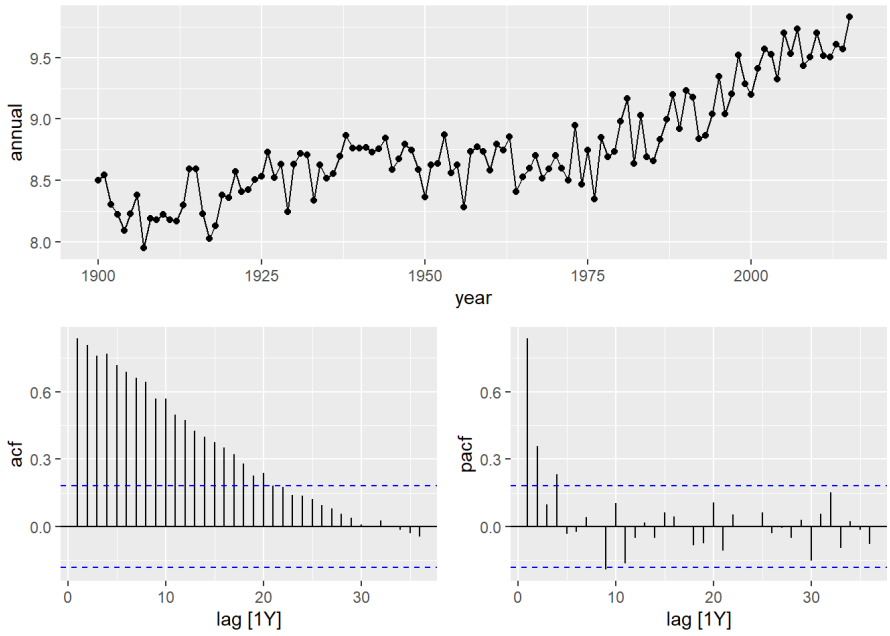
- (p,d,q) represent the non-seasonal parts of the model: p – non-seasonal autoregressive order, d – non-seasonal differencing order, q – non-seasonal moving average order.
- $(P, D, Q)_m$ represent the seasonal parts of the model: P – seasonal autoregressive order, D – seasonal differencing order, Q – seasonal moving average order, m – seasonal period (e.g., 12 for monthly data)

In evaluating model's performance, AIC and BIC metrics are used. AIC stands for Akaike Information Criterion. It's a measure used in statistics to compare different models and determine which one best explains a given dataset. The AIC considers the goodness of fit of the model and penalizes it for complexity (i.e., the number of parameters used). In general, a lower AIC value indicates a better model, as it strikes a balance between fit and simplicity. It's

commonly used in model selection processes to help choose among various competing models. The Bayesian Information Criterion (BIC) is more useful in selecting a correct model while the AIC is more appropriate in finding the best model for predicting future observations. Models with lower BIC are generally preferred

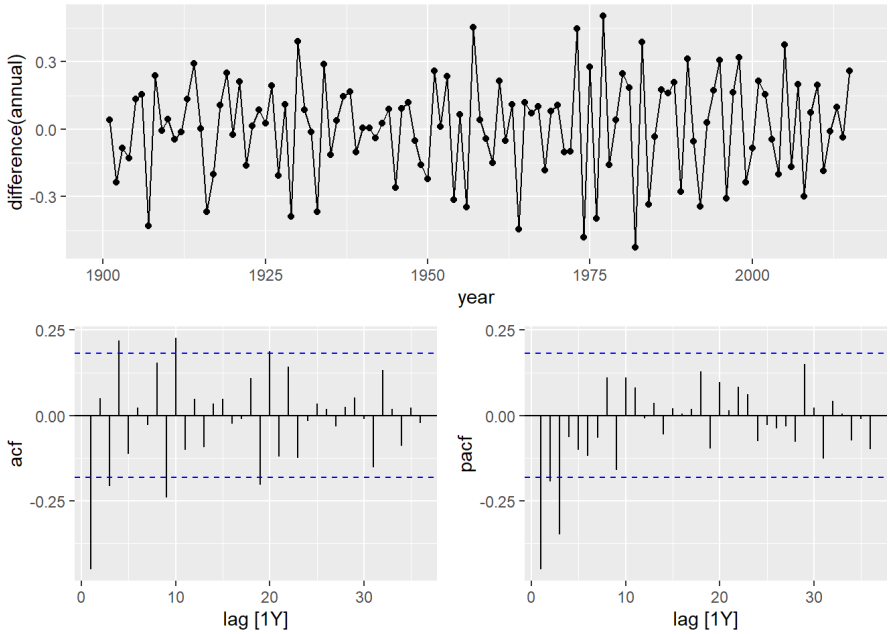
If we focus only on yearly averages, ACF and PACF graphs will look as can be seen in the following Figure. The first set of plots in the figure displays the annual mean temperatures from the dataset, along with the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The time series plot shows a clear upward trend indicating that the series is non-stationary. This is further confirmed by the ACF plot, which shows a slow decay, and the PACF plot, which has significant spikes. The second set of plots (Figure 5) shows the differenced annual mean temperatures, which helps in removing the trend and making the series stationary. The time series plot of the differenced data fluctuates around a constant mean, suggesting that the series is now stationary. The ACF and PACF plots of the differenced data, as can be seen in the second figure, show rapid decay, confirming that the differencing has successfully removed the trend, and the series is suitable for ARIMA modelling.

Figure 4 – ACF and PACF of yearly temperature time series



Source: figure by the author

Figure 5 – ACF and PACF of yearly temperature differenced series



Source: figure by the author

ARIMA model is used for this time series data, as there is no seasonality. To determine the best ARIMA model for this differenced time series, analysis of the ACF and PACF plots is needed. The ACF plot shows significant spike at lag 1, indicating a potential moving average component. The PACF plot shows significant spikes at lag 1 and lag 3, indicating a potential autoregressive component. Based on the plots, a few ARIMA models might be suitable: ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(3,1,0), ARIMA(1,1,1), ARIMA(3,1,1)

The models are tested, and results are presented in following Table.

Table 9 – ARIMA models (annual temperatures data)

Model	AIC	BIC	RMSE
ARIMA(0,1,1)	-61.4	-53.17	0.179
ARIMA(1,1,0)	-42.14	-33.9	0.195
ARIMA(3,1,0)	-58.42	-44.7	0.178
ARIMA(1,1,1)	-59.61	-48.63	0.179
ARIMA(3,1,1)	-59.76	-43.3	0.176

Source: table by the author

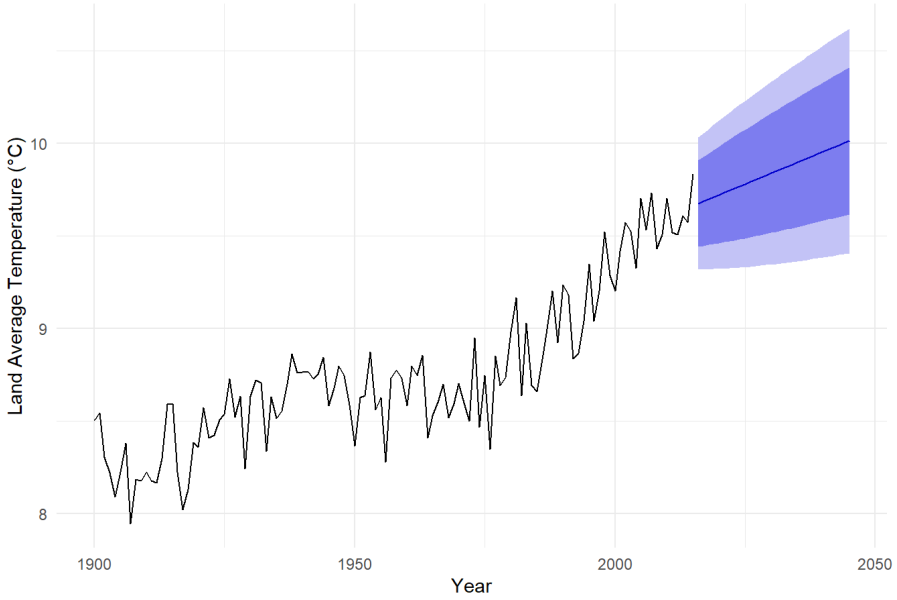
Since ARIMA(0,1,1) has lowest AIC and BIC it is the model used in forecasting. Model equation is following:

$$y_t = y_{t-1} + 0.0117 - 0.7433\varepsilon_{t-1} + \varepsilon_t$$

The equation describes how the value at time t depends on the previous value y_{t-1}, a small upward trend of 0.0117, and the influence of the previous period's error. The MA term of -0.7433 means that last period's error reduces the current value. Additionally, random noise is added to account for unpredictable fluctuations.

The results of forecasting are presented in the following Figure. Blue lines represent 80% and 95% confidence intervals.

Figure 6 – ACF and PACF of differenced data

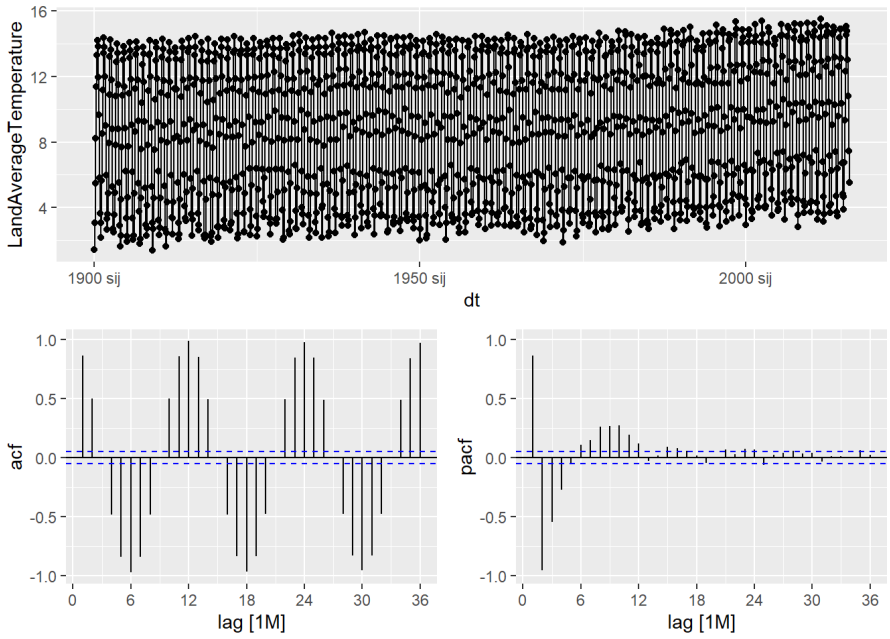


Source: figure by the author

As expected, the forecast shown in blue indicates that land average temperatures are projected to continue rising in the future. The central blue line represents the most likely trend based on current data, while the shaded areas around it represent uncertainty intervals. These intervals suggest a range of possible temperature outcomes depending on various factors. Overall, the forecast highlights a strong likelihood of continued warming through 2050.

Now the focus switches to monthly data. Before fitting models, the following Figure featuring ACF and PACF measurements helps us understand our data better. The time series plot shows a clear periodic pattern with a regular cyclic behaviour. This suggests a seasonal component, which is a strong indicator of non-stationarity. The amplitude of the cycles is constant, but the presence of a periodic pattern itself indicates non-stationarity.

Figure 7 – ACF and PACF (monthly temperature averages)



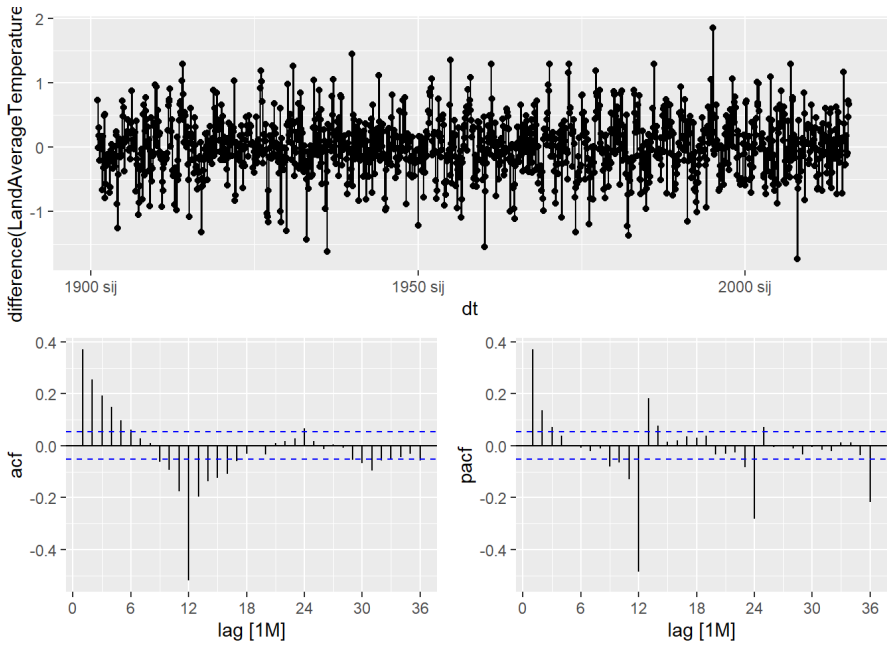
Source: figure by the author

The ACF plot shows significant spikes at regular intervals, corresponding to the periodic cycles observed in the time series plot. This pattern is indicative of seasonality and suggests that the series is non-stationary because of the repeating cycles. The PACF plot also shows significant spikes at regular intervals, reinforcing the evidence of a seasonal component. This periodic pattern in the PACF plot suggests that the time series has strong autocorrelation at seasonal lags. To make the series stationary, seasonal differencing is applied. After differencing, we get the following Figure.

The ACF plot shows that autocorrelations drop off relatively quickly, with most values falling within the confidence bounds after a few lags. This indicates that there is no significant long-term autocorrelation, which is consistent with stationarity. The PACF plot shows significant spikes at the initial lags but then quickly drops off, suggesting that the series has minimal

autocorrelation beyond the initial lags. Based on these observations, the time series is stationary after seasonal differencing.

Figure 8 – ACF and PACF (differenced monthly temperature averages data)



Source: figure by the author

Based on the previous figure, the conclusion is that:

- The ACF plot shows significant spikes at lag 1 and multiples of 12 (12, 24, etc.), indicating seasonality.
- The PACF plot shows significant spikes at lag 1 and lag 12 (and its multiplies), suggesting an autoregressive component.

Based on everything previously stated the model could be SARIMA(1,0,1) $(1,1,1)_{12}$. Model has AIC of 617.4 and RMSE of 0.2979. The equation of the model is following:

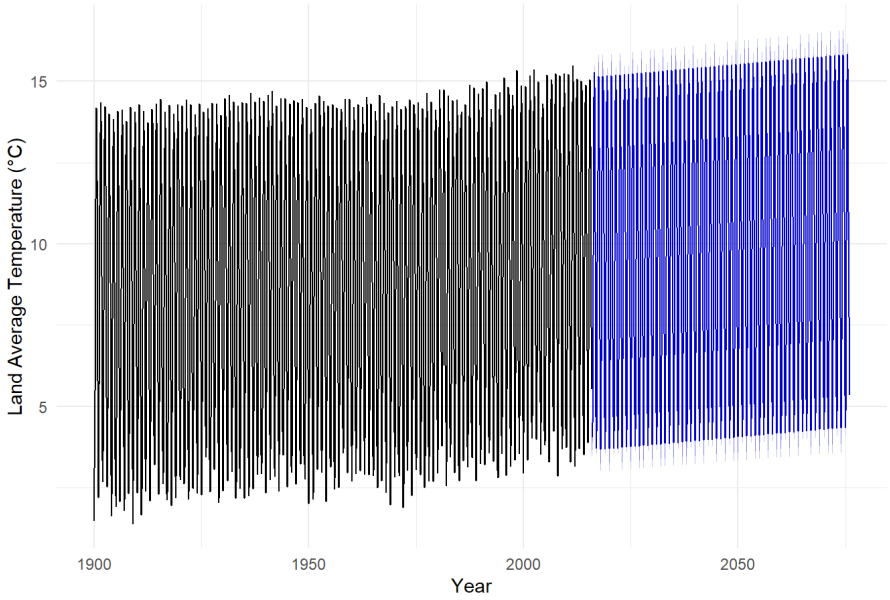
$$y_t = 0.8934y_{t-1} - 0.595\varepsilon_{t-1} - 0.0530y_{t-12} - 0.9018\varepsilon_{t-12} + \varepsilon_t$$

The current value y_t is influenced by both the previous month's value y_{t-1} and the error from the previous month ε_{t-1} (non-seasonal AR(1) and MA(1)). Additionally, model incorporates a seasonal AR(1) and MA(1) component, which means that the value 12 months ago (y_{t-12}) and the error from 12 months ago (ε_{t-12}) also affect the current value.

The following Figure represents the result of forecasting with this model, with included 80 and 95% confidence intervals. It is easily seen that monthly values are predicted to increase

each year, suggesting that global temperatures will continue to increase, based on available data.

Figure 9 – SARIMA model forecast



Source: figure by the author

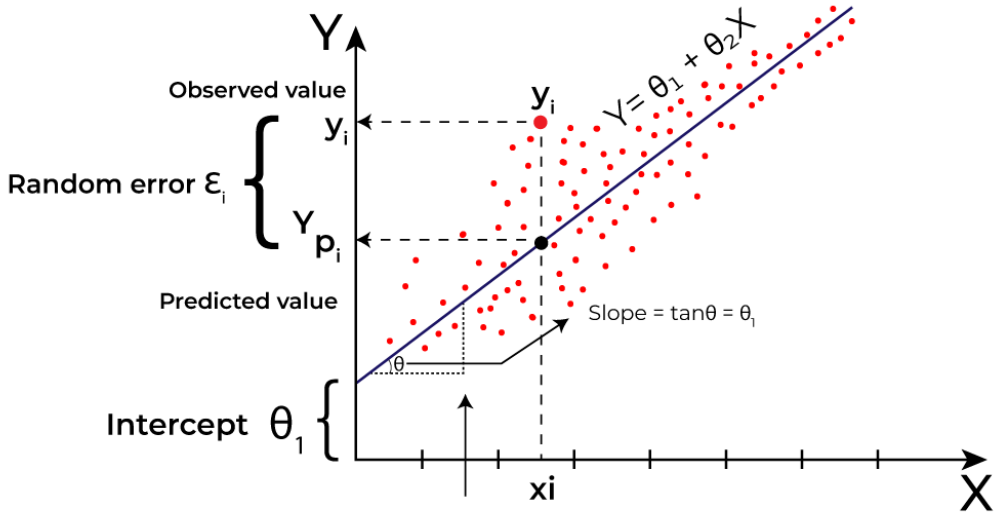
The next paragraphs will focus on regression analysis, considering temperature rise and all conclusions that were made using time series analysis.

4. Regression analysis

Linear regression is a statistical method employed to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It aims to estimate the coefficients of the linear equation, representing the slopes and intercepts, that best describe the relationship between the variables. The following Figure represents a simple linear regression model, showing the relationship between an

independent variable X and a dependent variable Y. The blue line represents the regression line defined by the equation $Y = \theta_1 + \theta_2$, where θ_1 is the intercept and θ_2 is the slope. The red dots are the observed data points. For a given point (x_i, y_i,) y_i is the observed value, and y_{pi} is the predicted value on the regression line. The difference between y_i and y_{pi} is the random error ε_i

Figure 10 – linear regression model



Source: GeeksForGeeks

Additionally, a random forest machine-learning algorithm is used. Random forest is an algorithm that combines the output of multiple decision trees. Multiple decision trees are trained on data and their outputs are combined to improve accuracy, improve generalisation and lower high variance – small differences in training data lead to completely different decision trees.

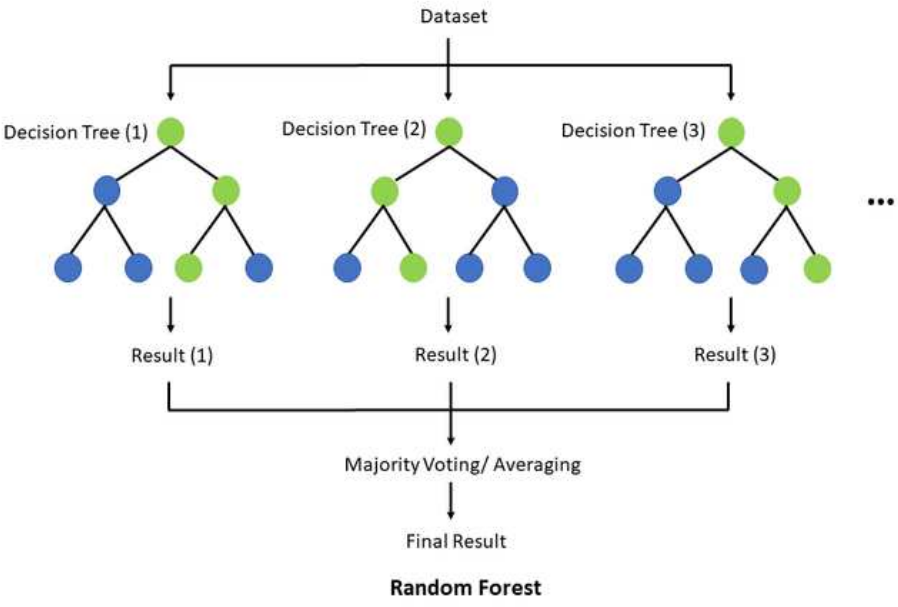
The following Figure illustrates the Random Forest algorithm. It consists of following:

1. Model Training:
    - Training Data Instance: The process begins with a training dataset.
    - Decision Trees: Multiple decision trees are generated from different subsets of the training data using random sampling. Each tree is trained independently.
2. Model Testing:

- Bagging (Voting Majority): During prediction, each decision tree provides a classification.
- Prediction Output: The final prediction is determined by majority voting, where the class with the most votes from the individual trees is selected. This method improves accuracy and robustness by averaging out the errors from individual trees.

Random Forest enhances model performance by combining predictions from multiple decision trees, reducing overfitting and improving generalization.

Figure 11 – random forest model

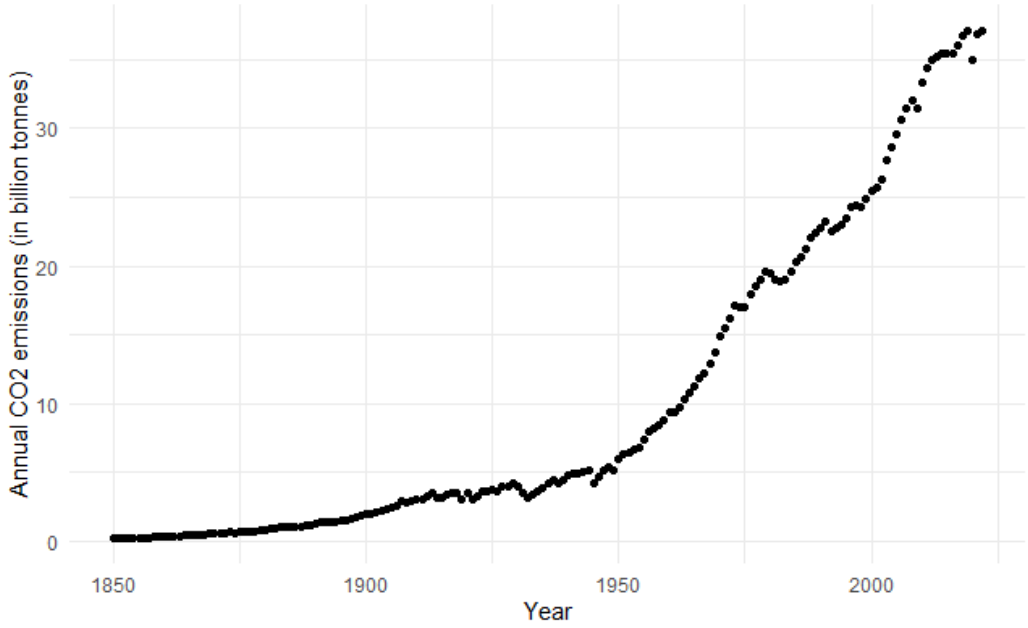


Source: Wikimedia Commons

4.1. Global air pollution (CO_2) impact on temperature changes.

A dataset with CO_2 emissions each year is used to test this possible correlation.

Figure 12 – annual CO_2 emissions

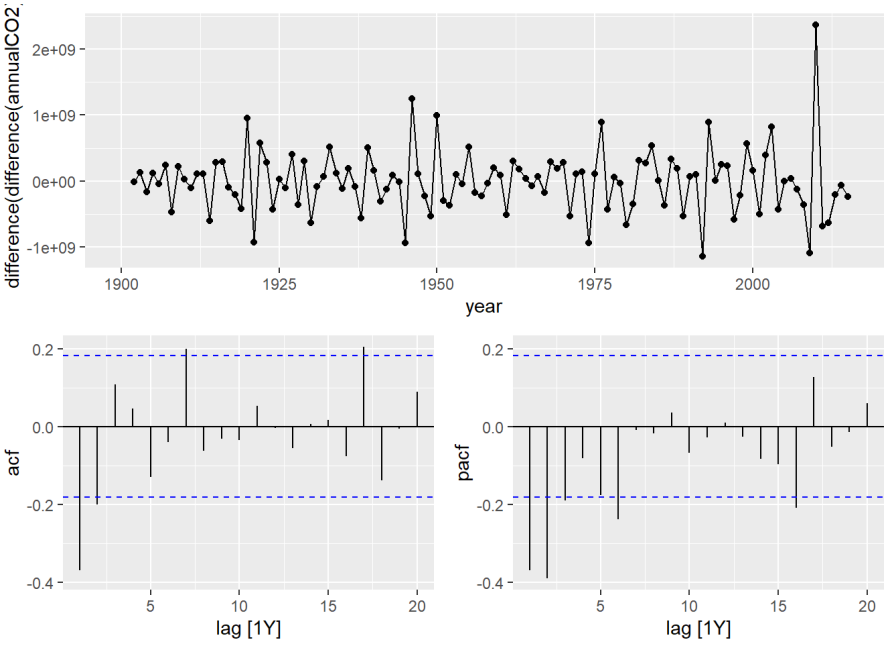


Source: figure by the author

In previous Figure, we can see annual CO_2 emissions worldwide each year since 1850. Exponential growth is easily seen, as we reach 37 megatons of CO_2 emissions in 2022. The impact of carbon dioxide emissions is nowadays one of the world's most talked-about problems and is claimed to be the first factor responsible for global warming. We tried to test if these claims are true and if there is a significant correlation between global warming and CO_2 emissions increase. Greenhouse gases include carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and water vapour (H_2O). CO_2 emissions take the first place in total gas emissions, averaging about 74.4% of total gas emissions in 2020. Methane takes the second place with 17.3%, followed by other gasses. We observed correlation by focusing solely on CO_2, as it has the most extensive historical data available.

Firstly, ARIMA model is used to make forecasts. Theoretical background is explained in Time series analysis. To make this series stationary, second differencing was needed. Next Figure illustrates ACF and PACF metrics of our data, after it was differenced two times.

Figure 13 – ARIMA models (CO_2 emissions)



Source: figure by the author

Based on this figure, few ARIMA models seem possible, based on two significant spikes on both plots. 5 ARIMA models were tested, and the results are presented in Table 10.

Table 10 – ARIMA models (CO_2 emissions)

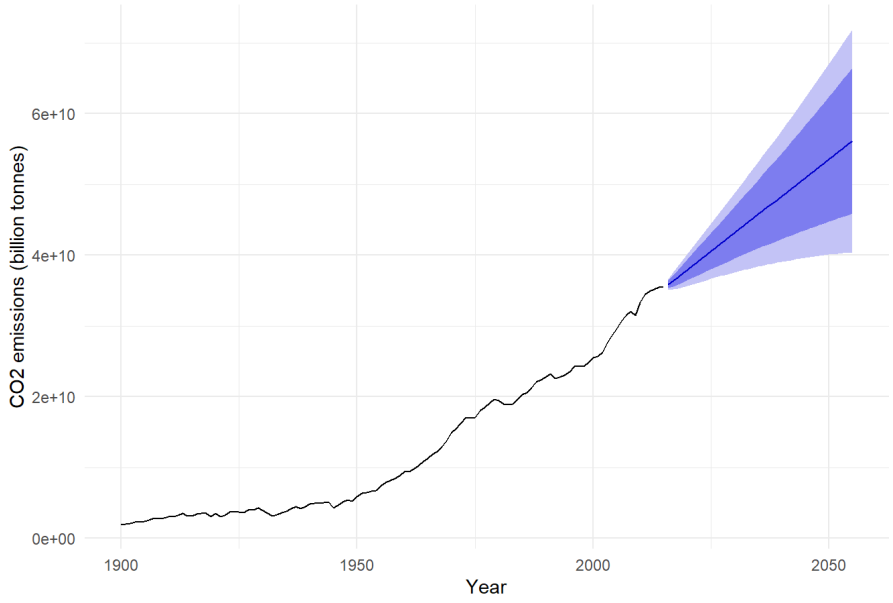
Model	AIC	BIC	RMSE (billion tonnes)
ARIMA(0,2,1)	6897.38	6903.58	0.3210
ARIMA(0,2,2)	6894.86	6904.16	0.3165
ARIMA(2,2,0)	6917.86	6927.16	0.3406
ARIMA(2,2,1)	6896.53	6908.93	0.3162
ARIMA(2,2,2)	6897.80	6913.30	0.3155

Source: table by the author

Considering this, ARIMA(0,2,2) is ultimately selected as our model, with Figure 14 illustrating forecasting done with that model, with included 80% and 95% confidence intervals in blue. As can be seen, CO_2 emissions are still forecasted to rise in the future. Model equation is following:

$$y_t = 2y_{t-1} - y_{t-2} - 0.7170\varepsilon_{t-1} - 0.1845\varepsilon_{t-2} + \varepsilon_t$$

Figure 14 – ARIMA model forecast (CO_2 emissions)



Source: figure by the author

Linear regression is now used to test if there is a correlation between CO_2 emissions and temperature rise. It should be noted that annual CO_2 emissions of the whole world were used, as we can't exclude some countries who pollute the most, such as China, since greenhouse gas emissions affect the whole planet, not just those countries. Linear regression is used on a dataset which combines 2 distinct datasets: (Kaggle, climate change) and (Kaggle, CO_2). The resulting dataset sample is represented in the following Table.

Table 11 – random dataset sample

Year	Average Temperature	Entity	Annual CO_2 emissions
1922	8.408000	World	3240313900
2012	9.507333	World	34935450000
1934	8.628333	World	3634241800
1909	8.178250	World	2890495000
1940	8.764667	World	4861346000

Source: Table by the author

Unnecessary columns are removed. Linear regression with following equation will be tested:

$$avgTemp = \alpha + \beta \times annualCO_2$$

Linear regression gives us the following output in Table 12.

Table 12 – linear regression results (CO_2)

Dependent variables				
Independent variables	Coefficient	SE	t	p
Annual CO_2 emissions	3.655×10^{-11}	2.382×10^{-12}	15.35	$< 2 \times 10^{-16}$

*Notes: R^2 = 0.7488; adjusted R^2 = 0.7456; significance codes: 0 *** 0.001 ** 0.01 **

The model equation with the given coefficients is:

$$avgTemp = 8.316 + 3.655 \times 10^{-11} \cdot annualCO_2$$

We can observe in Table that annual CO_2 emissions are highly statistically significant, with a p-value of less than 2×10^{-16}. The R^2 value is 0.7488, meaning that approximately 74.88% of the variance in annual global temperature rise is explained by the model. RMSE value is 0.566.

Random forest model is used to capture any possible non-linear correlations. Train data contains 70%, while data for testing contains 30% of whole dataset. Regression random forest model with 500 trees is used. At each decision point in the trees, only one variable is considered for splitting. The model explains about 77.57% of the variance in the target variable and has a RMSE on test data of 0.2237. In the next Table we can see a sample of model prediction on a test data.

Table 13 – random forest predictions (CO_2 emissions impact on temperatures)

Year	annualCO_2	avgTemp	predicted avgTemp	Difference
1923	3676233200	8.422167	8.568133	0.145966
1954	6789497300	8.560667	8.693920	0.133253
1976	17985243000	8.347250	8.916148	0.568898

1995	23524491000	9.347083	9.132039	0.215044
2004	28620194000	9.324583	9.574845	0.250262
2014	35466195000	9.570667	9.586688	0.016021

Source: table by the author

Lastly, linear regression and random forest are used but with lagged variables of $annualCO_2$. Specifically, lagged variables at intervals of 1, 2, 3, 5, 10, and 15 years were incorporated into different models. The training process did not involve randomly selecting data as done in previous models. Instead, the first 80% of the dataset was used for training, ensuring the sequential nature of the data was preserved for the lagged variables. The models were then tested on the most recent 20% of the dataset to evaluate their performance. The results are presented in the following Table.

Table 14 – predictions with lagged variables

	Linear regression		Random forest	
Number of lags	R^2	RMSE	R^2	RMSE
0	0.3865	0.50383	0.3141	0.4687
1	0.4218	0.47621	0.3363	0.4650
2	0.4221	0.47414	0.3492	0.4664
3	0.4243	0.47237	0.3445	0.4641
5	0.4255	0.47720	0.3071	0.4738
10	0.5330	0.43702	0.3607	0.4960
15	0.5883	0.29571	0.3752	0.4867

Source: table by the author

As can be seen in linear regression, the more lagged variables there are in the model, the better the results. With 15 lagged variables RMSE falls to around 0.296. R^2 is a bit higher than with simple linear regression with only annual CO_2 as predictor variable, likely because there was random sampling done for extracting dataset for training. After 2000, CO_2 emissions increased more than ever before, and because of that model couldn't capture everything completely accurately. For random forest model, results haven't gotten better with increasing

number of lags. In this scenario, the linearity and potential multicollinearity of the lagged variables might limit its effectiveness, explaining the smaller improvements in performance compared to linear regression.

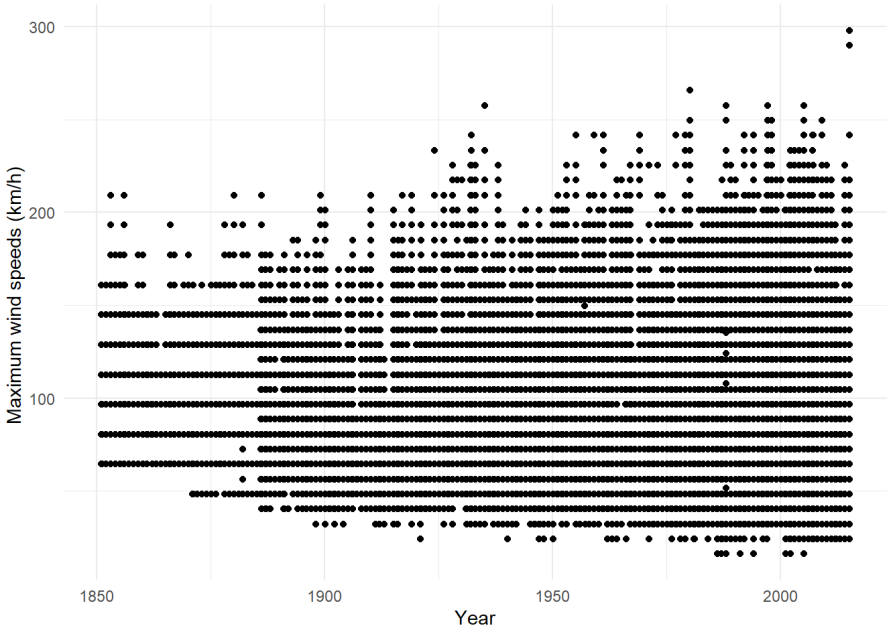
It should be noted that attributing global temperature rise solely to annual CO_2 emissions is not entirely correct, since there are other factors contributing to global warming, which are addressed later in the research, but this model gives us a clear overview of the high correlation in CO_2 emissions and global warming. As already mentioned, greenhouse gas emissions do not consist solely of CO_2, but of other gasses too, but CO_2 takes most of the emissions percentage. Also worth noting is that most of the other gasses are measured with a unit called a CO_2 equivalent. A CO_2 equivalent is a unit of measurement that is used to standardise the climate effects of various greenhouse gases.

4.2 Impact of global warming on hurricane wind strengths

For this analysis reports of hurricanes and their respective wind strength and air pressure are used, specifically hurricanes in the Atlantic and Pacific Oceans. A hurricane is a strong tropical cyclone that occurs in the Atlantic Ocean or northeastern Pacific Ocean, and in other parts of the world has other names, like typhoons in northwestern Pacific Ocean.

The Figure below illustrates the occurrences of all tropical storms in the Atlantic and Pacific Oceans after 1850 and their maximum recorded wind speeds. It should be noted that weather instruments weren't as advanced as today, so it is likely that 100 years ago not every hurricane was recorded, or at least not their wind speeds. Based on this figure only we can't deduce anything significant, but we try to find correlations later in this paragraph.

Figure 15 – highest recorded wind speeds of tropical storms

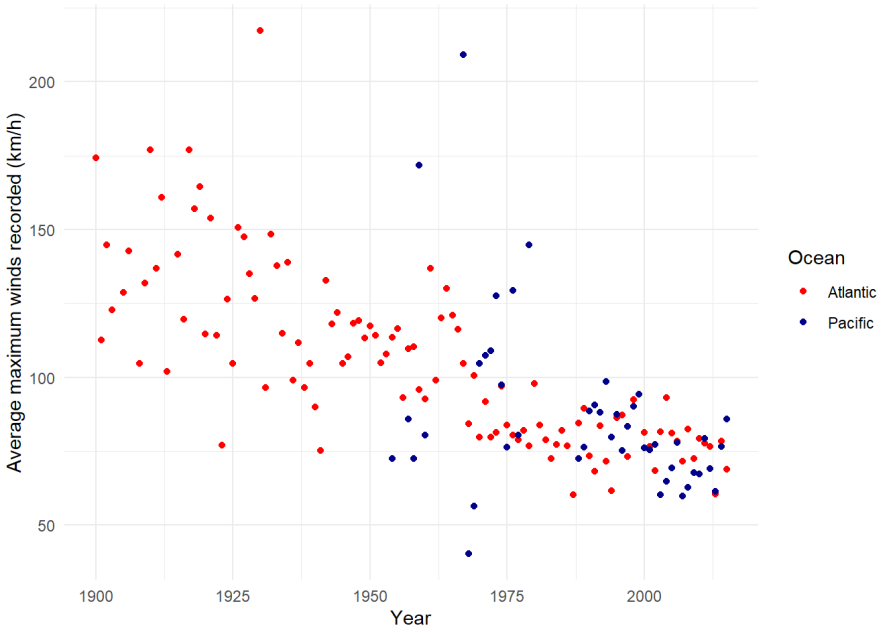


Source: figure by the author

It should be noted that these include all types of storms, not only hurricanes but their weaker forms, including tropical storms, tropical depressions, tropical waves etc. In this research, we focus only on hurricanes because they are well-documented, especially in past.

If we make a mistake and analyse every data available, we get a Figure like below. Based on this we could falsely conclude that the average maximum speed of hurricane winds is decreasing every year, and that would be wrong because in the past not every tropical storm was recorded, and they were more focused on recording only the strongest ones, while nowadays with all the technology we have it is easy to document every single storm.

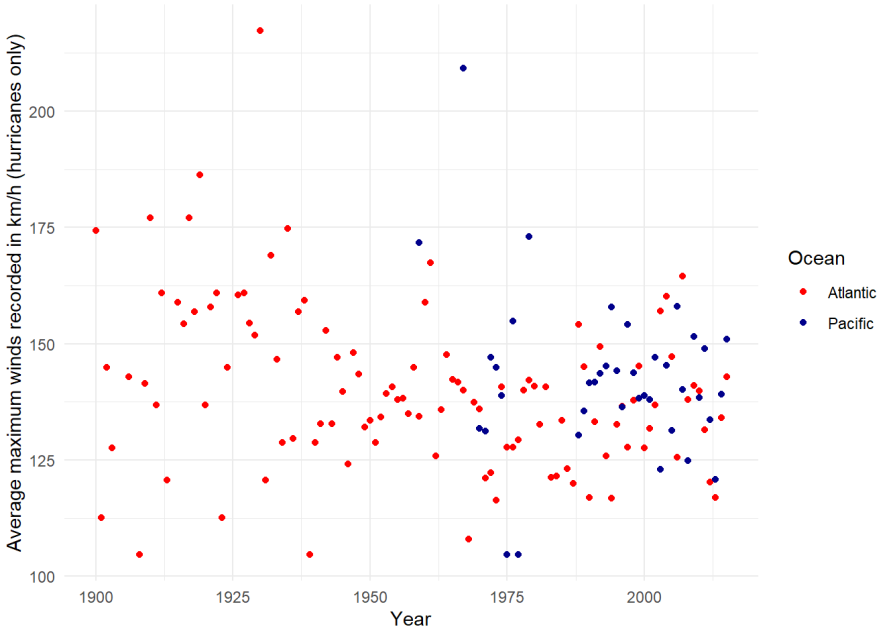
Figure 16 – average maximum wind speeds recorded



Source: figure by the author

So, if everything except hurricanes is removed, we get the Figure as below.

Figure 17 – average maximum wind speeds recorded (hurricanes only)

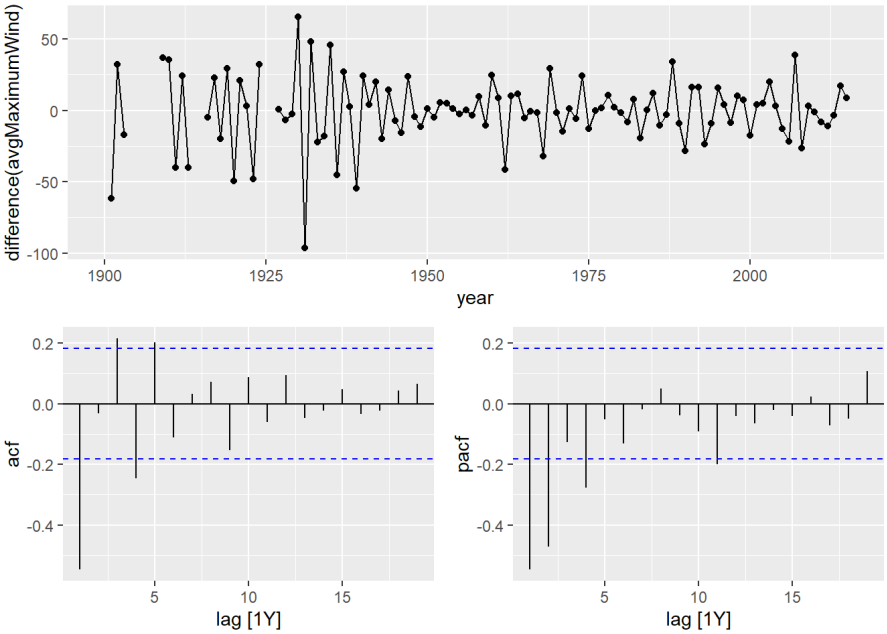


Source: figure by the author

The *X*-axis represents every year from 1900 to 2015, and the *Y*-axis represents the average maximum wind speed of every hurricane, calculated as the average maximum wind speeds of every hurricane in a specific year.

Now, ARIMA models for both Pacific and Atlantic Oceans are built. Since Atlantic Ocean has more data, 2 ARIMA models were made. For Atlantic Ocean, differencing was needed to make series stationary and ACF and PACF metrics can be seen in Figure 18.

Figure 18 – ARIMA models (hurricane wind strength – Atlantic Ocean)



Based on this few ARIMA models seem to be the best possible, one with moving average order of 2 and second of autoregressive order of 1. Out of 3 models presented in Table 15, MA(1) process seems to be the best, with lowest AIC and BIC values. Taking this into account, ARIMA(0,1,1) is used, and it is illustrated in Figure 19, with 80% and 95% confidence intervals.

Model equation is following:

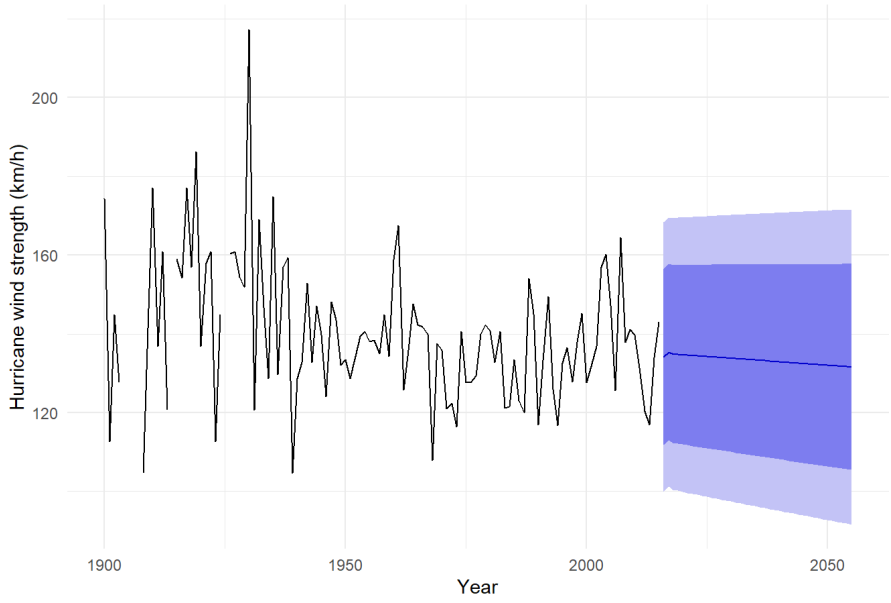
$$y_t = y_{t-1} - 0.9071\varepsilon_{t-1} + \epsilon t$$

Table 15 – ARIMA models (hurricane wind strength – Atlantic Ocean)

Model	AIC	BIC	RMSE
ARIMA(2,1,0)	957.78	968.76	17.873
ARIMA(0,1,1)	944.55	950.04	17.119
ARIMA(2,1,1)	948.78	962.51	17.025

Source: table by the author

Figure 19 – ARIMA forecast (Atlantic Ocean)



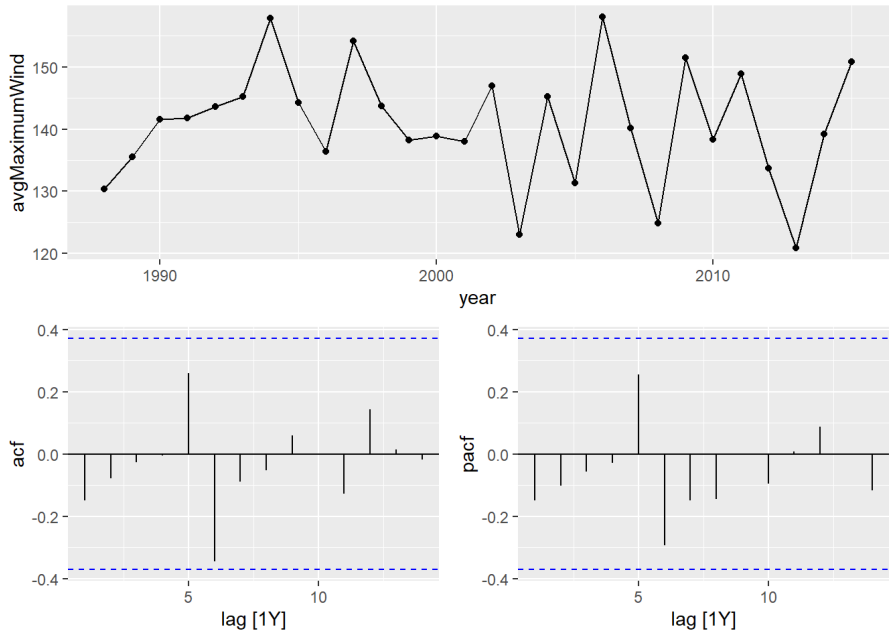
Source: table by the author

As can be seen, prediction is a slightly downwards pointed line, with added uncertainty levels.

Now the focus switches on Pacific Ocean. It's ACF and PACF plots (Figure 21) have no significant autocorrelations, therefore the time series is random. Modelling white noise is difficult, and the best we can do is model it with ARIMA(0,0,0) which has the following equation:

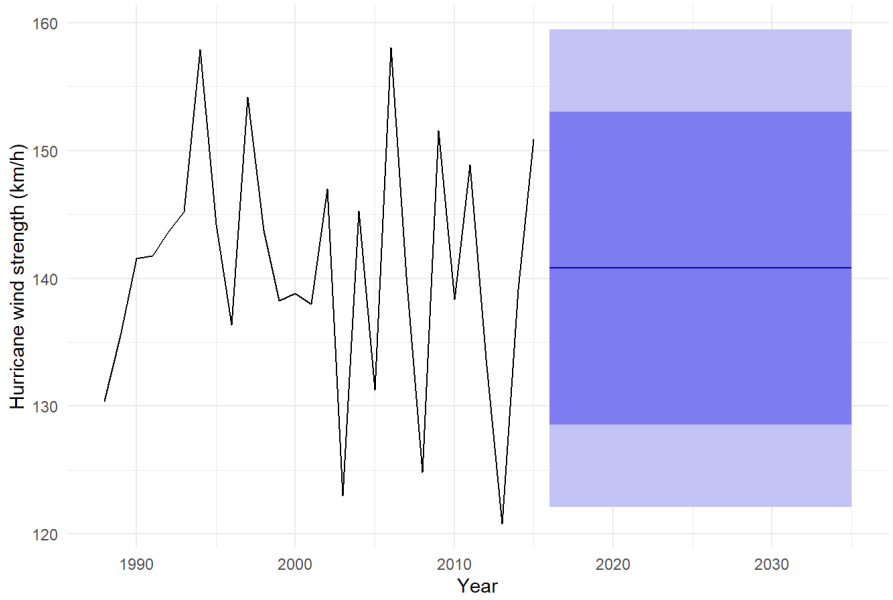
$$y_t = 140.7963 + \varepsilon_t$$

Figure 20 – ACF and PACF (hurricane wind strength – Pacific Ocean)



Source: figure by the author

Figure 21 – ARIMA forecast (Pacific Ocean)



Source: figure by the author

As can be seen in Figure 21, we got nothing of value. Part of the problem is not enough data for Pacific Ocean, while it is also possible that hurricanes haven't recorded any trend throughout the past, making forecasts difficult.

Now, regression analysis is done. A linear regression model with this equation is now built:

$$avgMaxWind = \alpha + \beta \times avgTemperature + \gamma \times avgLatitude + \delta \times avgLongitude$$

avgTemperature – average yearly temperature

avgLatitude – average latitude of recorded hurricane

avgLongitude – average longitude of recorded hurricane

Dataset sample is provided in the next Table, combining 2 datasets:

Table 16 – random dataset sample (combined temperature and hurricane dataset)

Year	average Maximum Wind	average Latitude	average Longitude	Ocean	average Temperature
1988	154.1256	20.16044	-68.64396	Atlantic	9.201583
2007	140.1470	NA	NA	Pacific	9.732167
1952	134.2532	27.83158	-71.01053	Atlantic	8.638250
1967	139.9555	26.93929	-71.21607	Atlantic	8.700083
1942	152.8877	28.33333	-96.63333	Atlantic	8.728417

Source: Table by the author

As can be seen, dataset consists of year, average maximum recorded winds of all hurricanes during that year, average latitude and longitude of recorded hurricanes, ocean where storms occurred, and average recorded temperatures. The regression results are contained in Table 16.

Table 17– linear regression model (hurricane wind strength)

Dependent variables				
Independent variables	Coefficient	SE	t	p
Temperature	-8.5084	4.9473	-1.720	0.0896
Latitude	-2.4612	0.4874	-5.050	3.01×10^{-6}
Longitude	-0.1394	0.2124	-0.656	0.539

Notes: $R^2 = 0.2984$*; adjusted* $R^2 = 0.2666$*; significance codes: 0 *** 0.001 ** 0.01 **

Source: Table by the author

Equation representing model is following:

$$avgMaxWind = 270.1728 - 8.4065 \cdot avgTemperature - 2.4453 \cdot avgLatitude - 0.1372 \cdot avgLongitude$$

The regression analysis suggests that average latitude is the most significant predictor of average maximum wind speed, with a highly significant negative relationship. Average temperature also shows a negative relationship, but its significance is weaker. Average longitude does not have a significant impact on wind speed. The model explains a moderate portion of the variability in wind speed, with the potential for further improvement.

We also try to build a random forest model with this same equation, in hope of better results. Random forest uses decision trees, which can account for non-linear relationships between variables. In our model, 500 trees were used. After constructing the model and testing its capabilities on the test dataset, we got root mean squared error on the test set of 16.293, suggesting that the model couldn't too effectively generalize to unseen data. This high mean squared error indicates that the model's predictions on the test set are, on average, far from the actual values. Results are presented in the next Table.

Table 18 – random forest prediction (hurricane data)

year	avgTemp	avgMaximumWind	avgMaximumWind prediction	Difference
1915	8.593167	158.9227	148.1163	10.8064

1921	8.571000	157.9169	143.4137	14.5032
1938	8.863667	159.3251	118.5383	40.7868
1950	8.365250	133.4706	145.7025	12.2319
1993	8.866583	125.9312	135.7248	9.7936
2003	9.525583	156.9408	132.8445	24.0963

Source: table by the author

Finally, linear regression and random forest models with lagged *avgTemp* variables are used. Results are presented in the following Table. As can be seen, R^2 values are very low for both models, indicating that hurricane data is very unpredictable, since no model can capture any significant correlations.

Table 19 – predictions with lagged variables (hurricane data)

	Linear regression		Random forest	
Number of lags	R^2	RMSE	R^2	RMSE
0	0.0054	11.4254	~ 0	12.0724
1	~ 0	11.3033	~ 0	11.7237
2	0.0008	11.4435	~ 0	11.7070
3	~ 0	11.4298	~ 0	12.3582
5	~ 0	10.7776	~ 0	12.3102
10	~ 0	11.5184	~ 0	12.4310
15	~ 0	12.1558	~ 0	12.4837

Source: table by the author

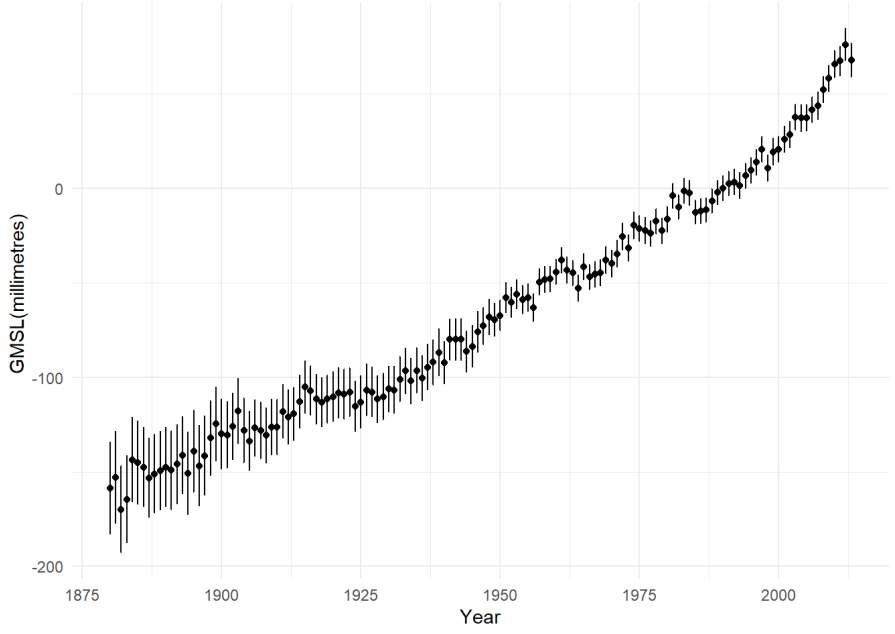
Based on all the values from previous models we can say that average temperatures don't play a very significant role in increase of hurricane strength. This could be expected based on previous graphs shown. We can confidently say that hurricane strength is currently not seriously affected by increase of global temperatures, but if this will remain like that is unknown. As we could see before, there is some slight correlation, but still not high enough.

In media, we can often hear about extreme weather happening daily, but as we can see global warming still isn't the main culprit. It should be noted that linear regression is not the best method for evaluating this hypothesis, but it is still good enough to find that significant correlation does not exist. Advanced machine learning models could be used to evaluate this, but not much difference would be found. We conclude that there is no significant correlation between the strength of hurricanes and rise of temperature. In the linear regression model, we can also see that latitude is highly statistically significant, while longitude is not. This should be expected, since most hurricanes spawn near equator, while also being spread around the entire world, so longitude isn't statistically significant.

4.3 Global warming and its impacts on ocean levels

For analysing this potential correlation, firstly we visualised changes in GMSL over time, since 1880. In Figure 22 we can observe change of GMSL, also accounting for uncertainty, since historical measurements were not as precise as today.

Figure 22 – global mean sea level change over years



Source: figure by the author

The *X*-axis represents a year, and the *Y*-axis is GMSL, with added uncertainties. It is easily seen that GMSL has risen over time. Since 1993, satellite altimeters have recorded a GMSL

rise of about 3.4 millimetres per year. The data reveals that both mass addition (from ice melt) and volume expansion (from warming) contribute to this increase. Random sample of dataset used in this model is represented in the following Table.

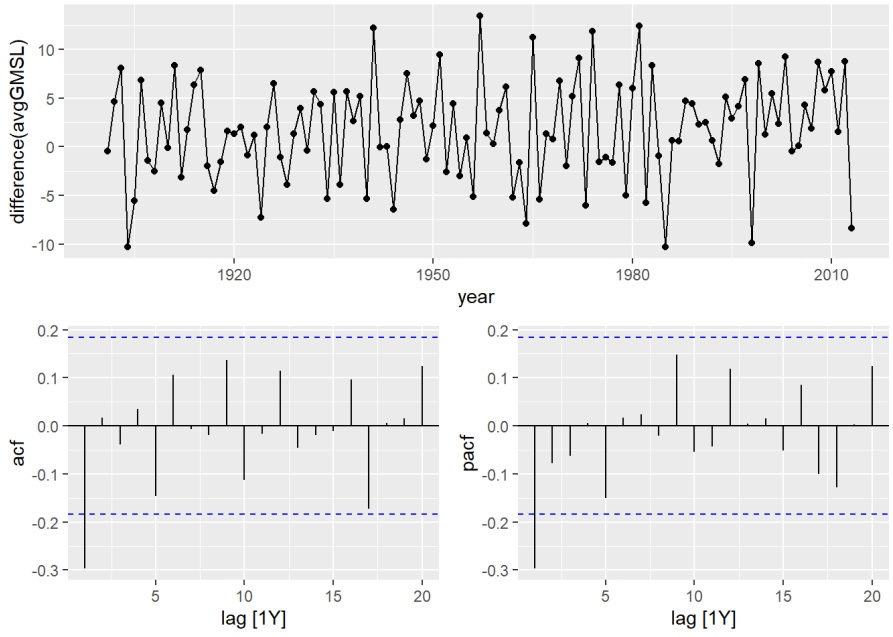
Table 20 – random dataset sample

Year	Average Temperature	Average GMSL	Average GMSL uncertainty
1903	8.220167	-117.85000	16.983333
1905	8.225167	-133.73333	15.400000
1981	9.165833	-4.07500	6.483333
1985	8.658000	-12.68333	6.266667
2004	9.324583	37.18333	6.925000

Source: Table by the author

Firstly, time series analysis is performed. First degree differencing is needed to make the series stationary. Autocorrelation function and Partial Autocorrelation Function are shown in next set of plots in Figure.

Figure 23 – ACF and PACF (GMSL)



Source: figure by the author

Based on significant spikes on both lag 1 of ACF and PACF, we could try few models.

Table 21 – ARIMA models (GMSL)

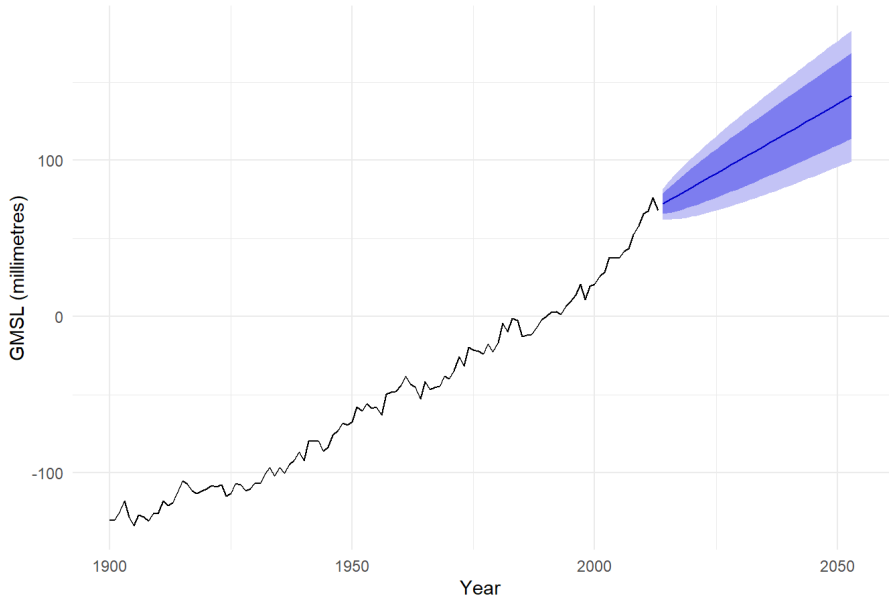
Model	AIC	BIC	RMSE
ARIMA(0,1,1)	386.1	392.48	5.141
ARIMA(1,1,0)	384.95	391.33	5.092
ARIMA(1,1,1)	386.95	395.46	5.093

Source: table by the author

Based on these values, ARIMA(1,1,0) seems most suitable. Using this model on dataset gives us the prediction in the following Figure, with confidence intervals included. Model equation is next:

$$y_t = 2.0922 - 0.3803(y_{t-1} - y_{t-2}) + \epsilon t$$

Figure 24 – ARIMA forecast (GMSL)



Source: figure by the author

Prediction is very clear, with continuous trend of GMSL increase in next 40 years.

If we now try to build a regression model with temperature change as the response variable and GMSL as a predictor variable, we get the following Table.

The temperature variable represents an average change in yearly temperature.

Table 22 – linear regression model (GMSL)

Dependent variables				
Independent variables	Coefficient	SE	t	p
avgTemp	125.95	12.99	9.693	3.51×10^{-11}

*Notes: R^2 = 0.7522; adjusted R^2 = 0.7494; significance codes: 0 *** 0.001 ** 0.01 **

Source: Table by the author

Formula representing previous model is following:

$$avgGMSL = -1147.67 + 125.95 \cdot avgTemp$$

Previous table shows us that there is significant correlation between yearly change of temperature and GMSL values, as was expected. The linear regression model shows that *avgTemp* is a statistically significant predictor of *avgGMSL*, with a high R^2 value of 0.7522, indicating that about 75.22% of the variability in *avgGMSL* is explained by *avgTemp*. The coefficients are significant, with a very low p-value. Using this model on test data gave us RMSE of 30.08.

To additionally test these claims, random forest model was used. As before, random forest model with 500 trees is used. Train data contains 80%, while data for testing contains 20% of whole dataset. The model explains about 76.41% of the variance in the target variable and has a RMSE on test data of 32.57, which seems high. Just like with linear regression. Reason for this is that GMSL wasn't too accurate throughout history, with uncertainties in datasets of up to 30 meters. To partially fix this, dataset of GMSL measurements after 1950 was used. RMSE of this new random forest model prediction is 18.45 and sample of prediction can be seen in following Table.

Table 23 – random forest model (GMSL)

year	average GMSL	predicted GMSL	difference
1955	-57.975000	-50.943283	7.031717

1962	-43.316667	-30.200661	13.116005
1987	-11.458333	-12.865827	1.407494
1991	2.508333	8.746761	6.238427
2007	43.48333	47.942215	4.458882
2008	52.200000	45.368588	6.831412

Source: table by the author

Lastly, linear regression and random forest but with lagged variables are built. Results are presented in the following Table.

Table 24 – linear regression and random forest models with lagged variables (GMSL)

	Linear regression		Random forest	
Number of lags	R^2	RMSE	R^2	RMSE
0	0.2941	42.19346	0.0995	64.10982
1	0.3816	17.37523	0.2580	65.02155
2	0.4238	16.19037	0.3629	64.17195
3	0.4720	29.68798	0.4141	64.71955
5	0.5015	43.91759	0.4040	65.58931
10	0.5211	41.25745	0.4393	67.43883
15	0.5870	32.73224	0.5378	69.11711

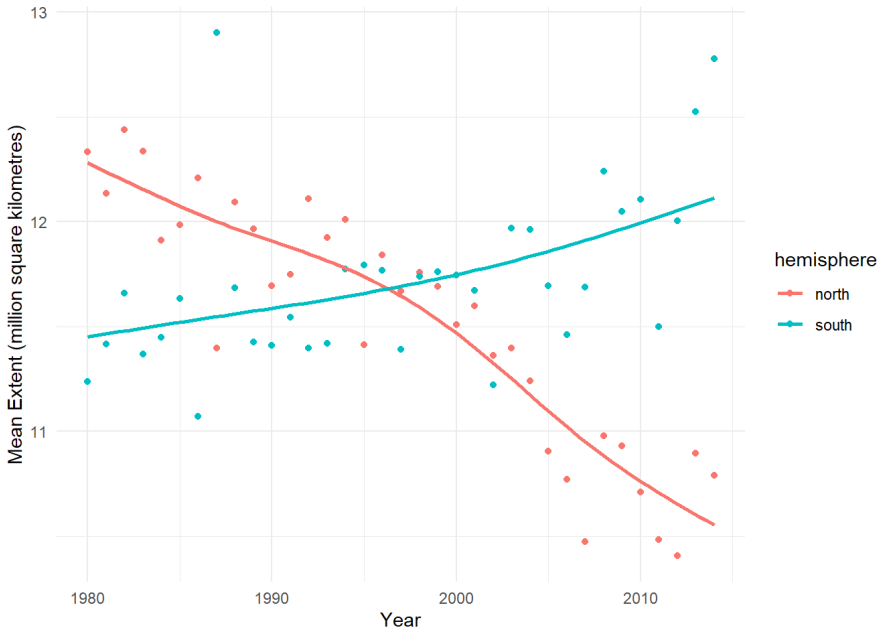
Source: table by the author

NASA's overview of global sea level change explains that global mean sea level (GMSL) is influenced by several factors, including the melting of ice sheets and glaciers, thermal expansion of seawater as it warms, and changes in land water storage. All these factors are influences by global warming, meaning that our model successfully captured the temperature rise effect on global mean sea level. Since GMSL wasn't too accurate until recent years, model couldn't capture completely accurately all the correlations, which is visible in table before. The ongoing rise in sea levels poses significant risks to coastal communities and ecosystems, necessitating continuous monitoring and research.

4.3.1 Global warming impact on sea ice extent in south and north hemispheres

Next, possible connections between rise of temperature, GMSL and sea ice extent in the world are explored. Figure 26 demonstrates interesting pattern, as can be seen that extent of sea ice in south hemisphere is rising, and in the north hemisphere dropping down. Measurements are in million square kilometres.

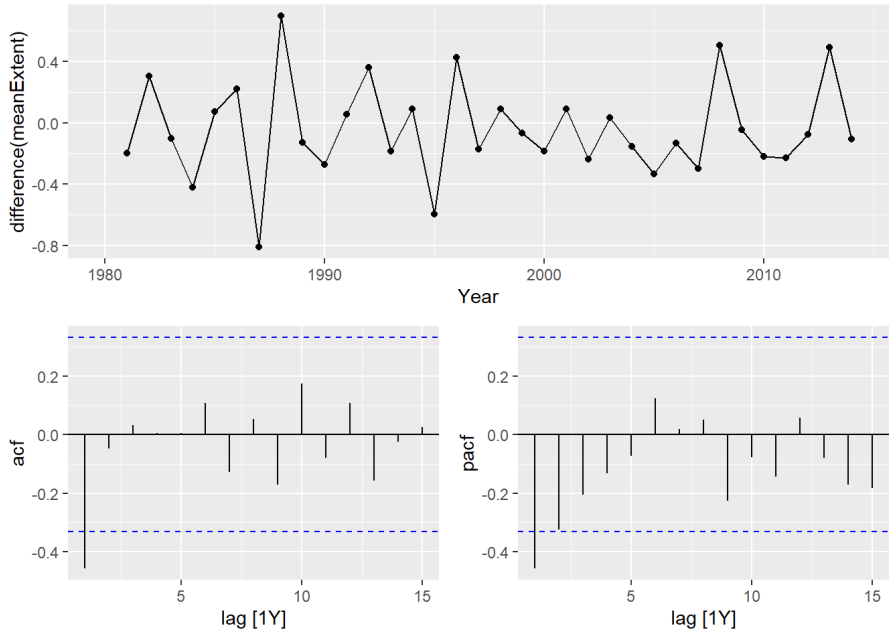
Figure 25 – sea ice mean extent by year and hemisphere



Source: figure by the author

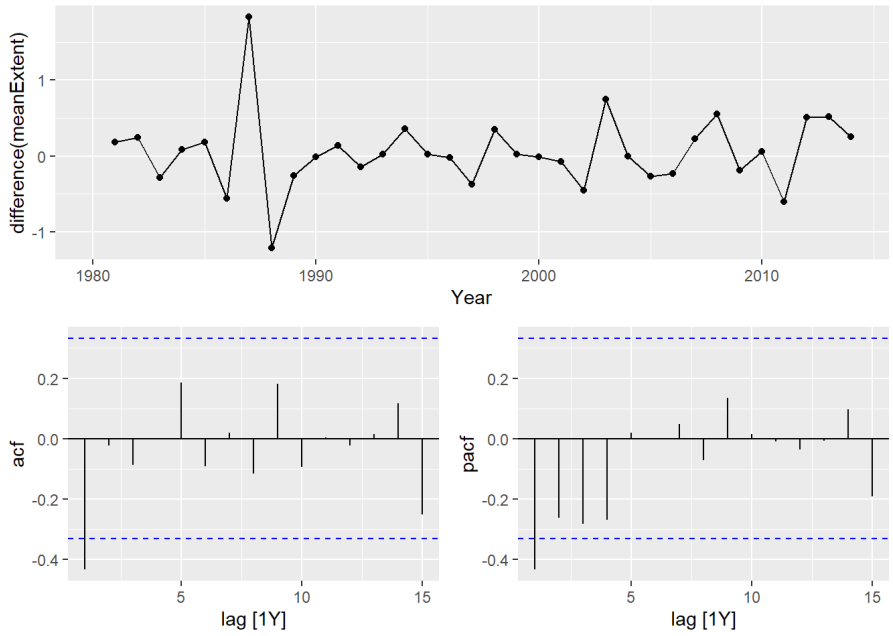
Before starting regression analysis on correlation between mean sea ice extent and temperature change, time series analysis is done, to observe possible changes in future. Since sea ice on different hemispheres seems to be acting differently, 2 ARIMA models are used, one for south and other for north hemisphere. Figure 26 shows metrics of time series where only north hemisphere data is used, while Figure 27 shows opposite.

Figure 26 – ACF and PACF (mean sea ice extent – northern hemisphere)



Source: figure by the author

Figure 27 – ACF and PACF (mean sea ice extent – southern hemisphere)



Source: figure by the author

Based on this few ARIMA models where tested, as can be seen in Table 25 and Table 26.

Table 25 – ARIMA models (northern hemisphere)

Model	AIC	BIC	RMSE
ARIMA(1,1,0)	14.84	19.42	0.2707
ARIMA(0,1,1)	8.93	13.51	0.2462
ARIMA(1,1,1)	10.91	17.01	0.2460

Source: table by the author

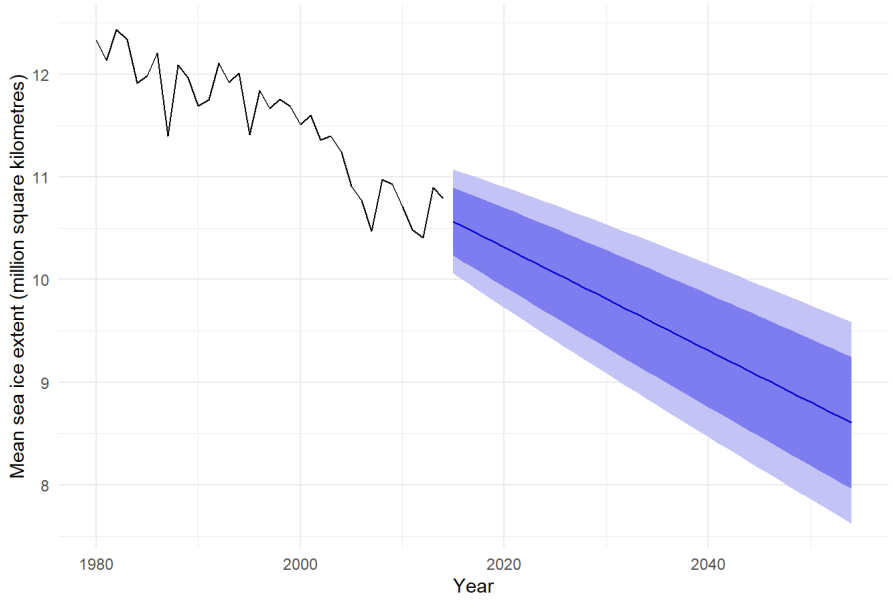
Table 26 – ARIMA models (southern hemisphere)

Model	AIC	BIC	RMSE
ARIMA(1,1,0)	46.48	51.06	0.4313
ARIMA(0,1,1)	35.73	40.31	0.3505
ARIMA(1,1,1)	37.69	43.8	0.3506

Source: table by the author

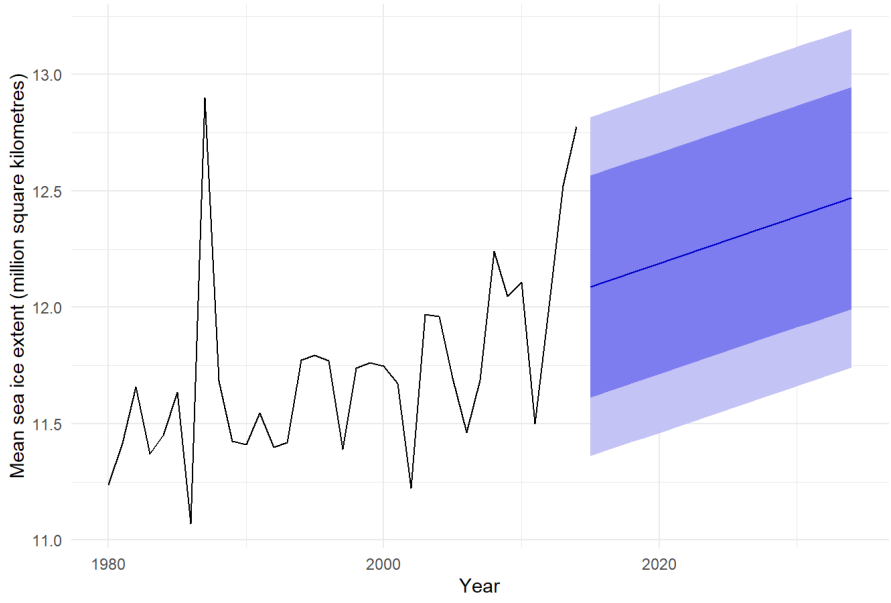
Taking these metrics into account, ARIMA(0,1,1) was used for both north and south hemisphere. Forecasts are presented in Figure 28 and Figure 29. 80% and 95% confidence intervals are included.

Figure 28 – ARIMA forecast (mean sea ice extent – northern hemisphere)



Source: figure by the author

Figure 29 – ARIMA forecast (mean sea ice extent – southern hemisphere)



Source: figure by the author

What is interesting is an increase of ice sea extent in the south hemisphere and a decrease in the north hemisphere. The probable reason for this is that The Arctic is an ocean surrounded by land masses, while Antarctica is a large continent surrounded by oceans, which makes ice expand easier to surrounding oceans.

Now, linear regression is used to test whether there is any correlation between temperature rise and mean sea ice extent changes. Linear regression on north hemisphere data is presented in Table 27, while on south hemisphere data is in Table 28.

Table 27 – linear regression model (mean sea ice extent on north hemisphere)

Dependent variables				
Independent variables	Coefficient	SE	t	p
temperature anomaly	-1.2254	0.1268	-9.664	1.42×10^{-10}

*Notes: R^2 = 0.7631; adjusted R^2 = 0.7549; significance codes: 0 *** 0.001 ** 0.01 **

Source: Table by the author

Table 28 – linear regression model (mean sea ice extent on south hemisphere)

Dependent variables				
Independent variables	Coefficient	SE	t	p
temperature anomaly	0.7095	0.2913	2.435	0.0213

*Notes: R^2 = 0.1698; adjusted R^2 = 0.1411; significance codes: 0 *** 0.001 ** 0.01 **

Source: Table by the author

The first linear regression model corresponds to the northern hemisphere, where the relationship between *anomaly* and *meanExtent* is strong, with an R^2 of 0.7631, indicating that 76.31% of the variance in *meanExtent* is explained by *anomaly*. The negative coefficient for *anomaly* (-1.2254) suggests that as the anomaly increases, the mean ice extent decreases significantly.

The second model represents the southern hemisphere, where the relationship between *anomaly* and *meanExtent* is much weaker, with an R^2 of 0.1698. The positive coefficient (0.7095) indicates a less pronounced, but still significant, increase in mean ice extent with increasing anomaly, though this relationship is less robust compared to the northern hemisphere.

Random forest analysis of this data was also done. As a response variable, the mean extent of sea ice was used, and as predictor variables, hemisphere (north or south), average global temperature for a specific year, average maximum temperature and average minimum temperature. The mean squared error of the built model is 0.1172. Data used for testing is 20% of the original data, while training data is 80%

Model prediction on a test data is visible in the following Table.

Table 29 – random dataset sample

Year	Hemisphere	Mean Extent	Average Temperature	Predictions	Difference

1980	south	0.07	11.236	12.012	0.776
1981	south	0.57	11.417	11.525	0.108
1983	north	0.07	12.336	12.166	0.170
1985	north	0.09	11.987	12.106	0.119
1986	south	0.12	11.071	11.644	0.573
1997	north	0.37	11.668	11.948	0.280
1998	south	0.39	11.738	11.660	0.078
1998	north	0.63	11.757	11.838	0.081
2009	north	1.14	10.932	10.828	0.104
2009	south	0.58	12.049	11.568	0.481

Source: Table by the author

We can see that the model predicts quite well, while not entirely correctly.

Lastly, linear regression and random forest with lagged variables is used. 15 lagged variables weren't used here because of not enough data. Results are in the following Table.

Table 30 – linear regression with lagged variables

	Northern hemisphere				Southern hemisphere			
	Linear regression		Random forest		Linear regression		Random forest	
Number of lags	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
0	0.4468	0.2897	0.1296	0.2610	0.0675	0.5921	0	0.2610
1	0.5591	0.2413	0.3916	0.3262	0.0644	0.5985	0	0.3262
2	0.5371	0.2540	0.3567	0.3219	0.1053	0.5779	0.0281	0.3219
3	0.6635	0.2326	0.5857	0.2951	0.0695	0.6091	0	0.2951
5	0.6764	0.1582	0.5703	0.2940	0	0.5713	0	0.2940
10	0.6048	0.3235	0.4398	0.2723	0.6507	0.8267	0	0.2722

Linear regression for northern hemisphere seems to be working best when 5 previous values are considered, while linear regression for southern hemisphere demonstrates worse performances, model couldn't capture any linear trend present. Random forest demonstrated good results for northern hemisphere, while the results for southern hemisphere indicate that the data is probably too unpredictable and/or not enough data points are there to test on.

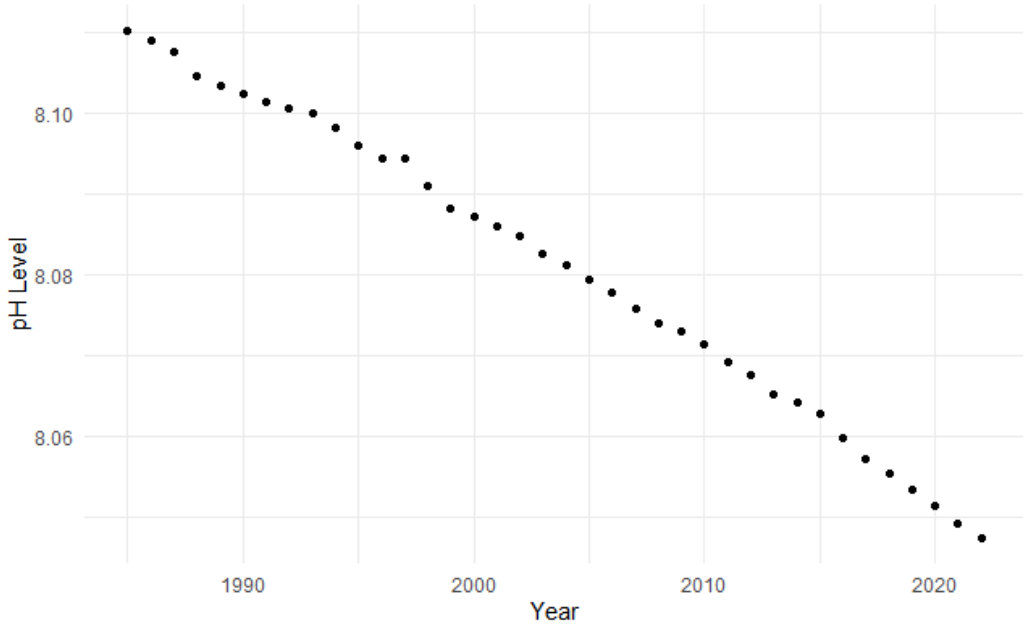
4.4 CO_2 levels and it's impacts on ocean acidification

Ocean acidification refers to the increase in ocean acidity due to the absorption of excess atmospheric CO_2, primarily from fossil fuel combustion, cement production, and land-use changes. This leads to a decrease in pH, posing a significant threat to marine life globally as the ocean and atmospheric circulation distribute the effects widely.

Ocean acidification significantly endangers marine ecosystems. The United Nations' Sustainable Development Goals (SDGs) and the Paris Agreement aim to combat this by reducing greenhouse gas emissions. Accurate monitoring of ocean acidity, facilitated by the EU's Copernicus Marine Service, is crucial for implementing effective policies to mitigate these impacts and ensure sustainable futures.

Figure 30 illustrates changes in ocean pH values over the years. We can see that values start from around 8.11, and in 2022 they are around value of 8.05. These changes may seem minimal, but they could impact marine life deeply.

Figure 30 – ocean pH values over years



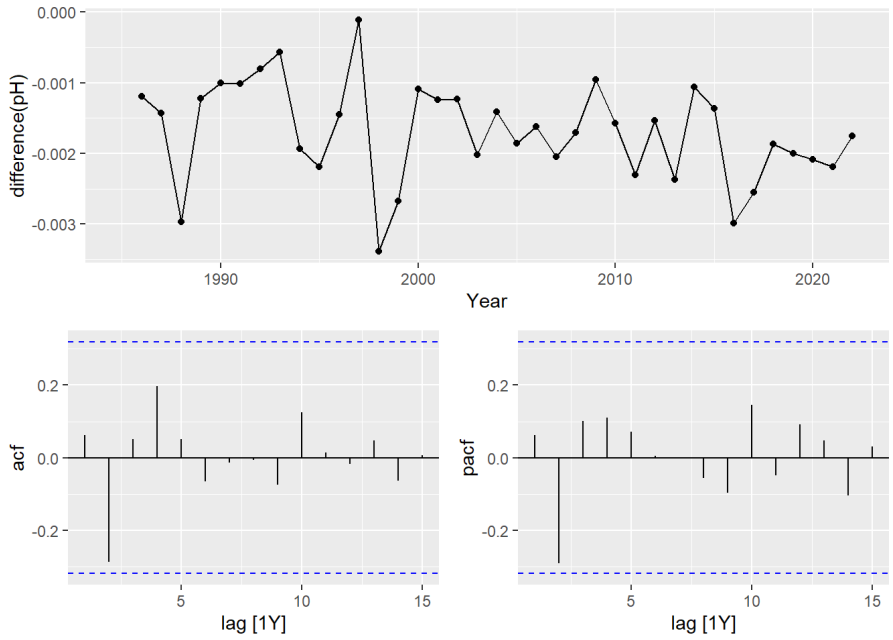
Source: figure by the author

We start with ARIMA modelling, as before. To make the series stationary first-degree differencing is needed. After differencing the series, we get following Figure. The series does not exhibit strong autocorrelation or partial autocorrelation at any lag, suggesting that ARIMA(0,1,0) is the most suitable model. Equation of the model is following:

$$y_t = y_{t-1} - 0.017 + \varepsilon_t$$

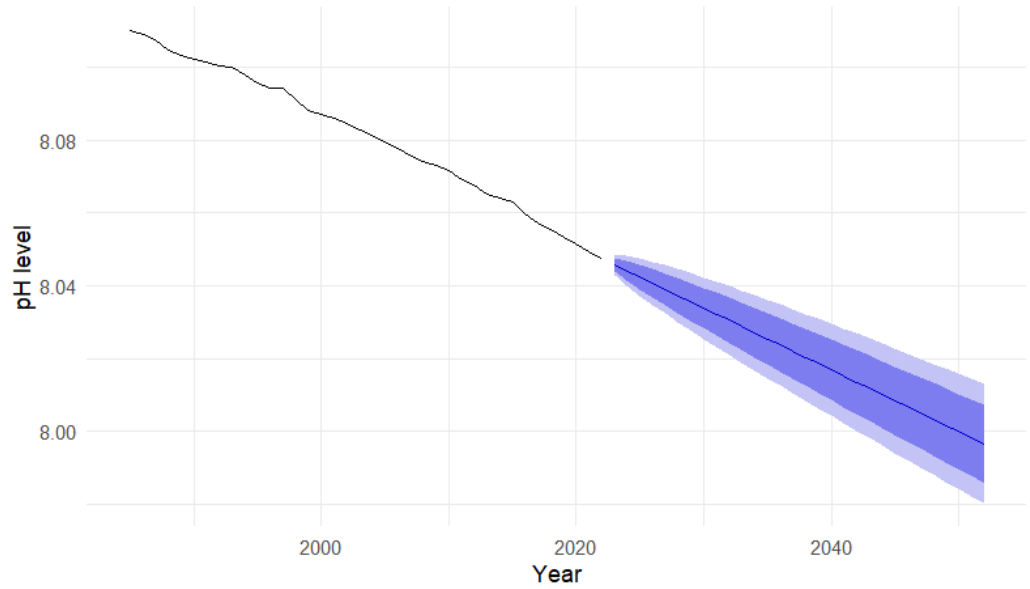
Its forecast is presented in Figure 32. Confidence intervals of 80% and 95% are also included. Model has a RMSE of 0.00148.

Figure 31 – ACF and PACF (pH levels)



Source: figure by the author

Figure 32 – ARIMA forecast (pH values)



Source: figure by the author

The next linear regression model is focused on predicting possible outcomes of increase in CO_2 emissions to decrease in ocean acidification. The model has pH values of oceans as predictor variables, and CO_2 levels around the world as response variables.

Table 31 – linear regression model (ocean acidification)

Dependent variables				
Independent variables	Coefficient	SE	t	p
Annual CO_2 emissions	3.176×10^{-12}	1.356×10^{-13}	1.356×10^{-13}	$< 2\times 10^{-16}$

*Notes: $R^2 = 0.9581$; adjusted $R^2 = 0.9563$; significance codes: 0 *** 0.001 ** 0.01 **

Source: Table by the author

The linear regression model, fitted on the training data, demonstrates a highly significant relationship between pH and annual CO_2 levels, as can be seen in Table 31. The coefficient estimate for *annual CO_2 emissions* suggests a negative association, indicating that higher CO_2 levels correspond to lower pH values. The model accounts for 95.81% of the variance in pH, with a residual standard error of 0.003859.

The mean squared error on the test set, reflecting the model's predictive accuracy, is 1.4637×10^{-5}. Resulting model formula is:

$$pH = 8.173 + (-3.176 \times 10^{-12}) \cdot annualCO2$$

Additionally, random forest model is also used. Variables used are the same as in linear regression. RMSE value is 0.0041. Model explains 92.36% of variance in the target variable. Model captures correlation between annual CO_2 change and pH change especially good. Following Table contains 5 random predictions from data used for testing purposes, which contains 30% of whole data.

Table 32 – random forest model (ocean acidification)

Year	Annual CO_2	pH	predicted pH
1997	24395952000	8.094322	8.090347
2002	26248288000	8.084703	8.089378
2003	27648650000	8.082688	8.080687

2017	36025455000	8.057357	8.059190
2019	37040103000	8.053486	8.054931

Source: table by the author

Finally, both models with lagged versions of annual CO_2 emissions are used. Results are in the next Table.

Table 32 – random forest model (ocean acidification)

	Linear regression		Random forest	
Number of lags	R^2	RMSE	R^2	RMSE
0	0.9516	0.01052	0.9192	0.01206
1	0.9698	0.01005	0.9399	0.01172
2	0.9794	0.00961	0.9513	0.01149
3	0.9847	0.00812	0.9559	0.01144
5	0.9883	0.00709	0.9542	0.01125
10	0.9965	0.00481	0.9581	0.01126

Source: table by the author

Linear regression seems to get better as there are more lagged variables and the one with 15 lagged variables has the lowest RMSE of all. All models work well, with very low RMSE. Random forest seems to work approximately the same with all models.

5. Conclusions

Throughout this research paper, various datasets were used with connections to climate change. Regression models were used on various datasets, with a goal of finding statistical connections between climate change indicators.

Time series analysis was used on temperature change on the monthly and yearly level. An ARIMA model was used, which analyses time series data by incorporating autoregressive, differencing, and moving average components to predict future temperature changes.

The findings suggest that global warming is poised to persist in the foreseeable future, with no signs of relenting.

Regression models were used on various datasets with climate change indicators, and different results were collected.

We have shown that CO_2 emissions directly impact yearly changes in temperatures. Linear regression was used, and all its analyses indicated that CO_2 concentrations significantly influence temperature variations.

The absence of modelled correlations between global warming and tropical storm wind strengths implies that the influence of global warming on tropical storms remains insufficiently captured by current modelling approaches.

Using linear regression, a significant correlation was identified between global warming and global mean sea level rise. Furthermore, contrasting trends were observed in sea ice extent, with an increase noted in the Southern Hemisphere and a decrease in the Northern Hemisphere.

Linear regression modelling revealed a slight annual decrease in ocean pH levels, indicating a direct impact of CO_2 emissions on ocean acidification.

In this study, we examined diverse datasets featuring potential climate change indicators, highlighting the widespread and significant nature of climate change as a global issue. This emphasizes the importance of collaborative global action to tackle this complex challenge and develop sustainable solutions for mitigation and adaptation. Climate change is impacting every corner of the world and will continue to get worse.

It is very important to improve the education quality on climate change and its consequences, through different educational centres, schools, universities etc., as suggested by (Patlins et al., 2020)

6. References

Change, C. (2001). Climate change. Synth. Rep.

G. Naiqian, G. Yuxin and S. Xuelian, "Global Temperature Forecast Based on ARIMA Model," *2019 4th International Conference on Communication and Information Systems (ICCIS)*, Wuhan, China, 2019, pp. 108-112, doi: 10.1109/ICCIS49662.2019.00026.

Pielke Jr, R. A., Landsea, C., Mayfield, M., Layer, J., & Pasch, R. (2005). Hurricanes and global warming. *Bulletin of the American Meteorological Society*, *86*(11), 1571-1576

P. Mangal, A. Rajesh and R. Misra, "Big Data in Climate Change Research: opportunities and Challenges," *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, London, UK, 2020, pp. 321-326, doi

M. Tavassoli and A. Kamran-Pirzaman, "Comparison of effective greenhouse gases and global warming," *2023 8th International Conference on Technology and Energy Management (ICTEM)*, Mazandaran, Babol, Iran, Islamic Republic of, 2023, pp. 1-5, doi: 10.1109/ICTEM56862.2023.10083954.

T. Chen, Z. Zhang, Z. Yi, W. Xu and K. Yang, "A Hybrid Mathematical Models for Predicting Global Climate Change," *2023 3rd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, Shenyang, China, 2023, pp. 357-367, doi: 10.1109/ACCTCS58815.2023.00052.

A. Patlins, J. Caiko, N. Kunicina, A. Zhiravetska and V. Riashchenko, "Climate Education: Challenges of Climate Change and Energy Policies," *2020 IEEE 61st International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON)*, Riga, Latvia, 2020, pp. 1-7, doi:

A. Bhagat, S. Thakre, S. Dongre and P. Maidamwar, "Prediction of Global Sea Water Level using Linear Regression and Gradient Descent," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1-5, doi: 10.1109/ICCCNT56998.2023.10306648.

X. Wang and Z. Wu, "Variability in Polar Sea Ice (1989–2018)," in *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 9, pp. 1520-1524, Sept. 2021, doi: 10.1109/LGRS.2020.3004257

M. Magi, "Prediction of Acidification in the Ocean Surface, and Benefits and Risks of the CO_2 Ocean Sequestration as a Mitigation Technology," *OCEANS 2008 - MTS/IEEE Kobe Techno-Ocean*, Kobe, Japan, 2008, pp. 1-4, doi: 10.1109/OCEANSKOBE.2008.4531104

Kaggle, Daily Sea ice extent, https://www.kaggle.com/datasets/nsidcorg/daily-sea-ice-extent-data/data

Kaggle, sea level change dataset, https://www.kaggle.com/datasets/somesh24/sea-level-change/data

Our world in data, CO_2 emissions, https://ourworldindata.org/CO_2-emissions

Kaggle, Hurricane database, https://www.kaggle.com/datasets/noaa/hurricane-database/data

Kaggle, Climate change – earth surface temperature data, https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

NOAA, Climate change impacts, https://www.noaa.gov/education/resource-collections/climate/climate-change-impacts

European Environment Agency, Decline in ocean pH measured at the Aloha station and yearly mean surface seawater pH reported on a global scale, https://www.eea.europa.eu/data-and-maps/figures/decline-in-ocean-ph-measured-4/

GeeksForGeeks, Linear regression in Machine learning, https://www.geeksforgeeks.org/ml-linear-regression/

Wikimedia Commons, Random forest, https://commons.wikimedia.org/wiki/File:Random_forest_explain.png

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. https://otexts.com/fpp3/

NCEI, Global time series, https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series/nhem/land/12/1/1850-2024

Abstract

The purpose of this statistical analysis was to identify and analyse variables that demonstrate statistically significant relationships with climate change. By determining the influential factors, we aim to gain insights into the complex dynamics driving climate variability. We found that global warming is heavily influenced by CO_2 and other greenhouse gas emissions, significantly affecting global mean sea levels and polar ice extent variations across hemispheres. While the relationship between global warming and tropical storm intensity remains inconclusive, CO_2 emissions notably drive ocean acidification. Understanding these relationships helps clarify the primary drivers of climate change.

Keywords – statistical analysis, climate change, linear regression, random forest, time series

Sažetak

Svrha ove statističke analize bila je identificirati i analizirati varijable koje pokazuju statistički značajne veze s klimatskim promjenama. Određivanjem utjecajnih čimbenika, cilj nam je dobiti uvid u složenu dinamiku koja pokreće klimatske varijabilnosti te informirati učinkovite strategije ublažavanja i prilagodbe. Otkrili smo da je globalno zagrijavanje uvelike pod utjecajem emisija CO_2 i drugih stakleničkih plinova, što značajno utječe na prosječne globalne razine mora i promjene u opsegu polarnih ledenih pokrova između hemisfera. Iako je odnos između globalnog zagrijavanja i intenziteta tropskih oluja i dalje neuvjerljiv, emisije CO_2 značajno utječu na smanjenje pH oceana. Razumijevanje ovih odnosa pomaže razjasniti glavne pokretače klimatskih promjena.

Ključne riječi – statistička analiza, klimatske promjene, linearna regresija, slučajna šuma, vremenski niz