Investigation on Ocean Acidification

Table of Contents

0 Introduction *3*

0.1 What is ocean acidification? 3

0.2 Why is this worth investigating? 3

1 Visualise, Summarise & Communicate the Data *4*

1.1 Table Summarising Key Metrics and Statistics 4

1.2 Why has the data been summarised in this way? 4

1.3 Main Trends and Patterns 5

2 Model and Evaluate Linear Behaviours in the Data *6*

2.1 Linear Regression 6

2.2 Scatter Plots 7

2.3 Forecasting concentration of CO_2 in the atmosphere 8

3 Model and Evaluate Seasonal Patterns in the Data *9*

3.1 Real-world/Engineering examples and analogies of k_1, k_2, k_3 and k_4 9

3.2 Formula for (RMS) of a sine wave of amplitude A over one period 9

3.3 Initial Guesses for k_2, k_3 and k_4 10

3.4 Finding Optimum Approximations 11

4 Reflection of Learning *12*

4.1 What I have learned 12

4.2 What does the data tell us? 12

4.3 Implications of results for the environment, society and economy 12

5 Appendix *13*

5.1 References 13

5.2 MATLAB Code 14

0 Introduction

The Great Barrier Reef (GBR) is a World Heritage Site and a habitat supporting a variety of marine life. Located near Queensland, Australia and comprising of over 2,900 individual reefs, the GBR is unfortunately now under threat as it has been estimated that it has lost half of its corals in the last 30 years [1]. The primary reason behind this is said to be ocean acidification and rising ocean temperatures.

0.1 What is ocean acidification?

Ocean acidification occurs when the ocean absorbs excessive carbon dioxide in the atmosphere. At nominal carbon dioxide levels, this isn't a concerning issue. However, carbon dioxide levels have now risen by 30% since the Industrial Revolution [2] which causes the ocean to uptake millions of tonnes of carbon dioxide every year [3]. This causes a chemical reaction, like the one below, to take place which lowers the ocean pH levels – making it more acidic. This acidity then makes it difficult for marine calcifying organisms, such as coral and some plankton, to form shells and skeletons – making them vulnerable to dissolution [4].

$$CO_2 + H_2O \rightarrow (H^+) + (HCO_3{}^-)$$

0.2 Why is this worth investigating?

Figure 1.
Map showing data collection site

Coming to why this is a problem, certain harmful algae species that are found in the GBR produce more toxins and bloom faster in acidified waters [5]. These toxins can damage marine life and in the long term, can have a negative effect on food chains and thus a disrupted food supply to humans [6]. A solution to this issue would be to reduce the amount of carbon dioxide and greenhouse gases released into the atmosphere. This could be by reducing energy consumption, preventing deforestation, using carbon-neutral products/solutions, etc. To better understand the correlation between the date (d), days from start of data collection (d_), atmospheric CO_2 content (y), water CO_2 content (x), ocean surface temperature (T) and atmospheric pressure (P), this report will analyse and use data from the Australian National Mooring Network (ANMN which has been collected over a decade from 24th March 2011 to 24th March 2021 at the coordinates: 23°27'29.9"S 151°55'35.8"E.

1 Visualise, Summarise & Communicate the Data

1.1 Table Summarising Key Metrics and Statistics

Table 1. Statistics and Key Metrics for atmospheric pressure, ocean surface temperature, water CO_2 content and atmospheric CO_2 content

Statistic/Key Metric	**P (kPa)**	**T (°C)**	**x (ppm)**	**y (ppm)**
Total readings	3666	3684	3642	3682
Maximum value	102.6528	29.1896	513.7365	423.1927
Minimum value	100.1097	20.0061	264.0551	381.7259
Mean value	101.3858	24.4364	389.7687	398.1049
Median value	101.4046	24.4868	385.8369	397.4781
Range	2.5431	9.1835	249.6814	41.4668
Standard Deviation	0.4716	2.1659	39.5905	7.1143
Lower Quartile	101.0729	22.4517	356.4129	391.7062
Upper Quartile	101.7086	26.4218	418.2741	404.4075
Interquartile Range	0.6357	3.9701	61.8612	12.7013
Variance	0.2224	4.6910	1.5674e+03	50.6135

1.2 Why has the data been summarised in this way?

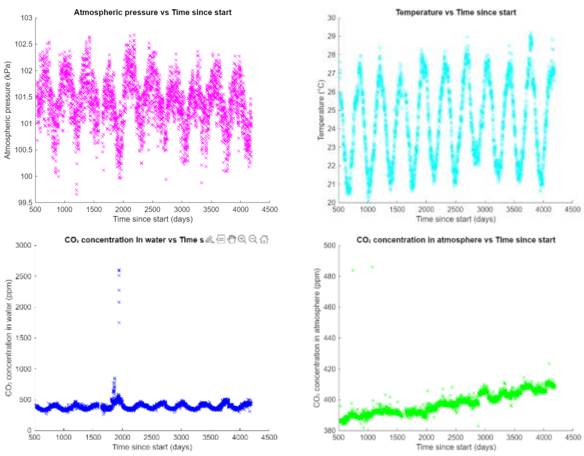


Figure 2.
Graphs showing all values from 09/10/2009 for P, T, x and y (including outliers)

When I first imported the data into MATLAB, I found that the data had a few repeated readings and had actually been collected from 9th October 2009. So, I first removed the repeated data and data from 09/10/2009 to 23/03/2011 so that the new data started from 24/03/2011, as stated in the brief. However, when I plotted test graphs for P, T, x and y against d_, I found sharp rises and drops in readings due to outliers. These could have been caused by faulty equipment, extreme weather (e.g. storms), improper calibration, etc. Thus, it wouldn't be fair to include these values for calculating key metrics, so I removed them. In the process, I found that only P, x and y had outliers. After making 4 separate tables for the variables, I then calculated the total, maximum, minimum, mean, median, range, standard deviation, LQ, UQ, IQR and variance for P, T, x and y. The mean and median aid in understanding the average for each variable, standard deviation and variance help us understand how "spread-out" the readings are to the mean. And other metrics such as UQ, LQ, IQR, max and min values also contribute to show how readings vary compared to one another and help explain how the data is distributed (i.e. closer to/further away from the mean).

1.3 Main Trends and Patterns

Table 2. Yearly summary (mean and standard deviation) for ocean surface temperature, water CO_2 content and atmospheric CO_2 content

Year	Total T values	T - Mean value	T – Standard deviation	Total x values	x – Mean value	X – standard deviation	Total y values	y – Mean value	y – standard deviation
2011	277	23.1464	1.8424	277	359.6940	27.1071	276	388.4677	1.9562
2012	357	24.0074	2.1813	357	371.9338	32.3701	356	390.3384	1.8688
2013	358	24.2458	2.0081	358	380.8300	29.5615	358	392.2324	1.3804
2014	305	23.7813	1.9853	285	397.0922	48.6863	305	391.4947	1.1311
2015	361	24.5768	2.1526	340	388.6580	47.9621	361	394.2245	2.7768
2016	363	24.6064	2.1536	363	395.7999	29.1742	363	397.8526	1.7697
2017	349	24.7691	1.9479	349	394.8294	36.9974	349	400.4795	3.4573
2018	361	24.4398	2.1048	361	386.2726	37.1169	361	402.3987	2.3643
2019	361	24.3887	2.0844	361	392.2214	35.6681	361	405.2720	2.6377
2020	426	24.7727	2.3256	425	403.8086	42.2622	426	407.9107	2.3240
2021	166	26.9838	0.4280	166	429.7822	22.5252	166	409.5040	0.8757

To help better understand trends and patterns in the data, I decided to do a yearly summary for T, x and y. Initially, I thought of adding P but after looking at Table 1, I found that the standard deviation for atmospheric pressure was relatively low, and as the mean is roughly 101kPa, it's representative of true atmospheric pressure which is constant. So instead, I calculate the mean and standard deviations for oceanic surface temperature, water carbon dioxide content and atmospheric carbon dioxide content. For oceanic surface temperature, the results show that over time, the temperature increased, dropped in 2014, 2018 and 2019 and continues to increase as the highest recorded average temperature was 26.9838 degrees Celsius which is from 2021. For water carbon dioxide content, the data seems to follow a similar pattern as the water CO_2 content rises over time but drops in 2015, 2017, 2018 and peaks in 2021 with a record 429.7822 ppm. Lastly, atmospheric CO_2 levels also increase over time, dropping slightly in 2014 and show that there was an average of 409.5040 ppm of CO_2 in the atmosphere in 2021. These drops are possibly caused by the El Nino and La Nina events that took place [8]. Water temperatures dropped slightly in 2014, and this caused more CO_2 to be dissolved, raising the x values slightly and once the water was hot enough, CO_2 dissolved less as cold waters tend to dissolve CO_2 better than warm waters [9]. From this data, we can see that water CO_2 concentrations have increased the fastest compared to other variables as, for the same time frame of 10 years, it has a range of 249.6814, standard deviation of 39.5905 and variance of 1.5674e+03 which is the largest amongst P, T, x and y. This shows that the data for x varies greatly from it's mean and that is due to the large yearly increases in x. This emphasises how rapidly ocean acidification takes place and how important it is that we tackle the issue of oceanic acidification. Furthermore, we can see that temperature and atmospheric CO_2 content is almost proportional. This is because CO_2 is a greenhouse gas and as levels of CO_2 increases, more heat is trapped in the atmosphere – causing temperatures to increase. In a similar way, water CO_2 levels can be correlated with atmospheric CO_2 levels because the abundance of CO_2 in the atmosphere causes oceans to absorb more CO_2 than previously which leads to increased water CO_2 levels, contributing to ocean acidification.

2 Model and Evaluate Linear Behaviours in the Data

2.1 Linear Regression

Linear regression is a way of describing a relationship between an independent variable (what is being changed) and a dependent variable (what is being measured). A basic linear regression follows the below format where a is the gradient of x (i.e. how f(x) changes with respect to x) and b is the offset (f(x) when x = 0):

$$f(x) = a \cdot x + b$$

In our context, the variables being measured are atmospheric pressure (P), ocean surface temperature (T), water CO_2 content (x) and atmospheric CO_2 content (y) while the variable being changed is the time or more specifically the number of days (t/d_). So, using the model from above, we can modify it so that it can be applied to P, T, x and y:

$$P(t) = a_P \cdot t + b_P$$
$$T(t) = a_T \cdot t + b_T$$
$$x(t) = a_x \cdot t + b_x$$
$$y(t) = a_y \cdot t + b_y$$

When applied to the existing data, a_P, a_T, a_x and a_y correspond to the gradient (how much either atmospheric pressure, ocean surface temperature, water CO_2 content or atmospheric CO_2 content increases/decreases per day). A positive gradient means an increasing trend while a negative slope means a decreasing trend. Similarly, b_P, b_T, b_x and b_y are the offsets (initial values for atmospheric pressure, ocean surface temperature, water CO_2 levels and atmospheric CO_2 levels). Using the fitlm function, I calculated the coefficient of determination, R^2, to be as follows:

- Atmospheric pressure (kPa) $\rightarrow P(t) = (-9.9496 \cdot 10^{-5}) \cdot t + (101.63)$
  - $a_P = -9.9496 \cdot 10^{-5}\ kPa\ per\ day$
  - $b_P = 101.63\ kPa$
  - $R^2 = 0.0531$ so, 5.31% of model's variance is accurate, atmospheric pressure doesn't increase linearly over time and is influenced by other factors or has a more complex pattern
- Ocean surface temperature (°C) $\rightarrow T(t) = (0.00038675) \cdot t + (23.492)$
  - $a_T = 0.00038675\ ℃\ per\ day$
  - $b_T = 23.492\ ℃$
  - $R^2 = 0.038$ so, 3.8% of model's variance is accurate, ocean surface temperature doesn't increase linearly over time and is influenced by other factors or has a more complex pattern
- Water CO_2 content (ppm) $\rightarrow x(t) = (0.010103) \cdot t + (365.03)$
  - $a_x = 0.010103\ ppm\ per\ day$
  - $b_x = 365.03\ ppm$
  - $R^2 = 0.0783$ so, 7.83% of model's variance is accurate, water CO_2 content doesn't increase linearly over time and is influenced by other factors or has a more complex pattern
- Atmospheric CO_2 content (ppm) $\rightarrow y(t) = (0.0061743) \cdot t + (383.02)$
  - $a_y = 0.0061743\ ppm\ per\ day$
  - $b_y = 383.02\ ppm$
  - $R^2 = 0.898$ so, 89.8% of model's variance is accurate, atmospheric CO_2 content increases steadily and almost linearly with time.

As we can see, by looking at the R^2 values for P, T, x and y, the linear model is most appropriate for atmospheric CO_2 content compared to other variables as its R^2 value is very close to 1. This means that atmospheric CO_2 content changes consistently and linearly over time. However, P, T and x have R^2 values closer to 0 so these variables show a very weak linear relationship with time. This could mean that they require a more complex model and/or are affected by other factors.

2.2 Scatter Plots

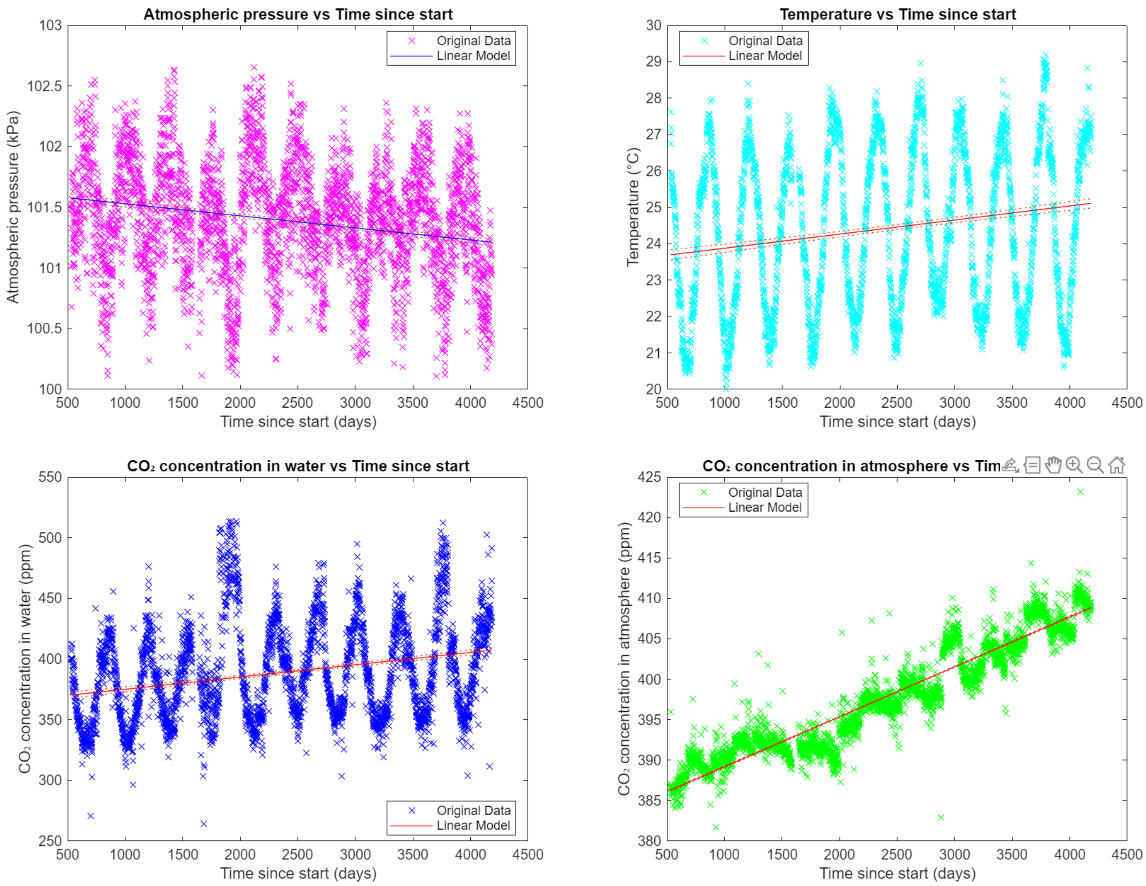


Figure 3.
Graphs showing data and fitted linear models for P, T, x and y

2.3 Forecasting concentration of CO_2 in the atmosphere

To forecast CO_2 content in the atmosphere, I extrapolated the existing linear regression. As my data starts from 24/03/2011, I decided that for my predictions, I will be using the same dates so 24/03/2030, 24/03/2050 and 24/03/2100. I did this by converting the dates to days from first measurement (d_) and then using the predict function to find future y values. I then added these values to the existing graph:

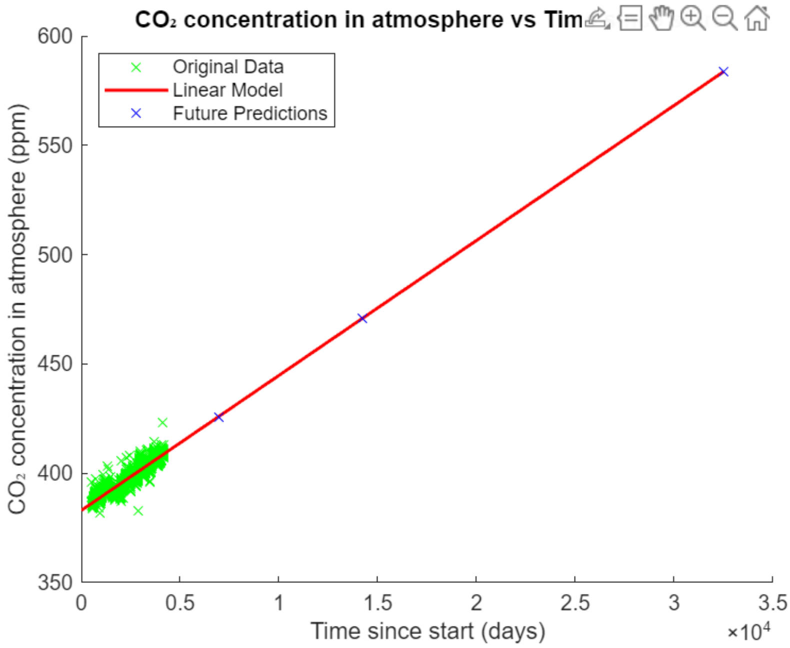


Figure 4.
Graphs showing data, fitted linear models and future predictions for y

The blue crosses indicate the points that have been plotted due to the extrapolation of the linear model. From left to right, these points are (6940 days,425.8683 ppm), (14245 days,470.9717 ppm) and (32507 days,583.7270 ppm) where 6940, 14245 and 32507 are 24/03/2030, 24/03/2050 and 24/03/2100 converted to days since 09/10/2009, respectively. I believe that these values for y are only true if the existing trend of CO_2 concentration in the atmosphere continues at the same rate. However, many countries have made statements that they will be going carbon-neutral or reducing their carbon emissions from 2050 onwards [10] which could suggest that the red line may hopefully start to dip near the second blue cross. Therefore, a more refined model may be needed to accurately predict futuristic values. Furthermore, I think the oscillating trend will continue for all 4 variables as a primary cause for this is due to the seasonal changes that occur across the year. Also, as we know that colder waters absorb more CO_2 than warmer ones, it can be said that as ocean surface temperatures increases linearly, the CO_2 content in water will start to decrease. And finally, there is still the possibility of El Nino and La Nina events occurring, where temperature and CO_2 levels vary, which could change the existing trend meaning that the future predictions would have to change.

3 Model and Evaluate Seasonal Patterns in the Data

3.1 Real-world/Engineering examples and analogies of k_1, k_2, k_3 and k_4

First, let's look at equation 1:

$$T(t) = k_1 \sin(2\pi k_2 t + k_3) + k_4 \quad (1)$$

Upon inspection, I found that k_4 is the vertical translation of the function Similarly, k_3 seems to be a horizontal translation of the function or a phase shift which may have been used to map the sine wave onto the T(t) function (or vice versa). k_2 gives the frequency of the function/number of cycles per day and therefore the period of each cycle would be given by $\frac{1}{k_2}$. Lastly, k_1 is the amplitude of the function and is the maximum/minimum temperature from the mean at which the function oscillates. Coming to real-world analogies, k_1 would be the highest/lowest recorded temperature which shows how extreme the temperature can get (the amplitude of seasonal variations). k_2 is the frequency and refers to the number of times per day that the temperature reaches a maximum, decreases to the average, continues to decrease to a minimum and finally rises back up to the average. This can be altered to model temperature changes in seasons from summer to winter, for example. k_3 is the horizontal translation that helps us calibrate and align the amplitude(s) of T to its appropriate t values (e.g. peak temperatures in summer). An example of this is phase shift in AC power supplies. Lastly, k_4 refers to the vertical translation of the function and is needed to change the average temperature that the function oscillates around.

3.2 Formula for (RMS) of a sine wave of amplitude A over one period

$$RMS = \sqrt{\frac{1}{\tau}\int_0^{\tau} [A\sin(\omega t + \phi)]^2 \, dt}$$

First, let's ignore the square root sign:

$$\frac{1}{\tau}\int_0^{\tau} [A\sin(\omega t + \phi)]^2 \, dt$$

Next, we will expand the brackets by squaring the expression inside the integral and take A out of the integral as it is a constant:

$$= \frac{1}{\tau}\int_0^{\tau} A^2 \sin^2(\omega t + \phi) \, dt$$

$$= \frac{A^2}{\tau}\int_0^{\tau} \sin^2(\omega t + \phi) \, dt$$

Using the identity: $\sin^2(x) = \frac{1-\cos(2x)}{2}$ we can redo the integral and take out the ½ to give:

$$= \frac{A^2}{2\tau} \int_0^\tau 1 - \cos(2\omega t + 2\phi)\, dt$$

$$= \frac{A^2}{2\tau} \int_0^\tau 1\, dt - \frac{A^2}{2\tau} \int_0^\tau \cos(2\omega t + 2\phi)\, dt$$

If we think of integrating as finding the area under the curve, we can see that over one time period, the area under a cosine curve will be 0 as the area above the axis is equal to the area under the axis. As a result, the definite integral of the cosine function over its time period will be 0.

$$= \frac{A^2}{2\tau}(\tau - 0)$$

$$= \frac{A^2}{2}$$

And bringing back the square root sign, this gives:

$$\therefore RMS = \sqrt{\frac{A^2}{2}} = \frac{A}{\sqrt{2}}$$

To use this for my temperature data, I first rearranged the above equation for A to give:

$$A = RMS \cdot \sqrt{2}$$

I then calculated the RMS value of the temperature from 24th March 2011 to 24th March 2021 by first finding the difference between each temperature value and the mean and then using the rms function which gave me an RMS of 2.1656°C. Therefore, to find the amplitude of seasonal variations, k_1, I multiplied this by $\sqrt{2}$ to give 3.0626:

$$k_1 = A = RMS \cdot \sqrt{2} = 2.1656 \cdot \sqrt{2} = 3.0626$$

3.3 Initial Guesses for k_2, k_3 and k_4

Since k_2 is the frequency and temperature oscillated in a yearly cycle, the time period of the cycle would be one year or 365.25 days. Therefore, the frequency, k_2 would be:

$$k_2 = \frac{1}{365} \approx 2.739 \cdot 10^{-3}$$

Moving on to k_4, since this would be the reading at which the graph oscillates, I decided to use the mean of the temperature data by using the mean function. This gave me a mean value of 24.4364. Lastly, to find k_3, I used trial and error. First, I plotted a graph with my known variables: k_1, k_2 and k_4. I then started off with $k_3 = 0$. As the graph appeared to be starting off a bit to the left, I knew that my horizontal translation would have to be negative. So, then I tested -pi and the graph looked like it shifted too far to the right. I then tried -pi/2 and as it looked a little too far to the right again, I tried -pi/4. This looked acceptable but just to check, I also tried -pi/8 but found that the graph had not shifted enough to the left which showed that -pi/4 (-0.785) was optimal.

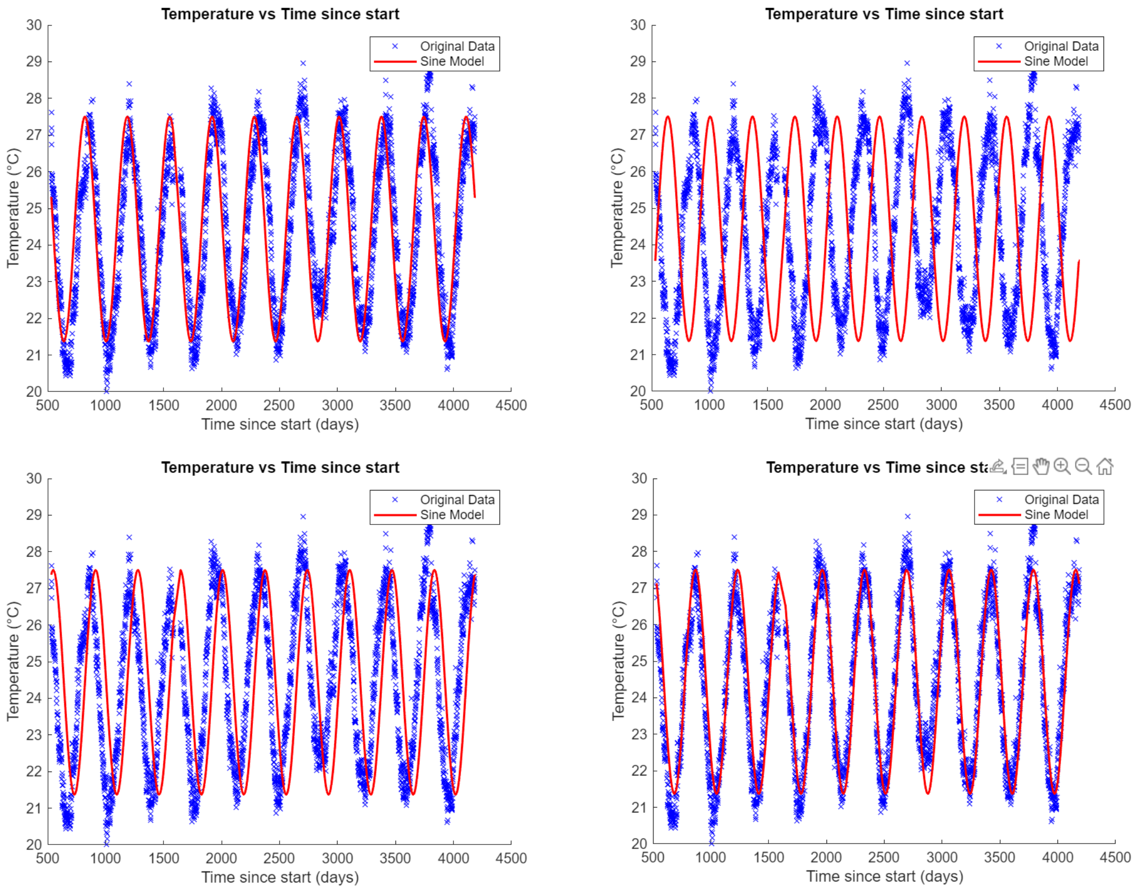


Figure 5.
Graphs showing different values of k_3 (top left – k_3 = 0, top right – k_3 = -pi, bottom left – k_3 = -pi/2, bottom right – k_3 = -pi/4)

3.4 Finding Optimum Approximations

After using the provided code and inputting my values for initial parameters, lower and upper bounds, I found that the optimum values for k_1, k_2, k_3 and k_4 were 2.9685, 0.0027, -0.5302 and 24.4225 respectively. Comparing these to my initial guesses, I feel that the largest error came from the k_3 value. This is because I didn't use a wide range of values to test how much each graph shifts by, so it was quite inaccurate. However, I feel that the remainder of my values are quite accurate.

Variable Name	Initial	Lower Bound	Upper Bound	Optimum
k_1	3.0626	1.0626	5.0626	2.9685
k_2	1/365 (0.0027)	1/370	1/350	0.0027
k_3	-pi/4 (-0.785)	-pi (-3.142)	0	-0.5302
k_4	24.4364	21.4363	27.4364	24.4225

Table 3. List of all k values I used to find the optimum

4 Reflection of Learning

4.1 What I have learned

Being someone who is completely new to coding and also MATLAB, this coursework helped me understand both the importance of knowing how to use MATLAB but also perseverance as many things don't often work the first time around and you have to analyse the questions from different perspectives. In Section 1, I learnt about how to clean data by removing outliers and plotting basic graphs. A major takeaway from that was indexing as it proved to be very useful. For Section 2, I learnt how to make clear, descriptive graphs and ways to customise them by changing the colours of the lines and markers, for example. For Section 3, I enjoyed going back to the mathematical side of data analysis as I understood how to map different functions onto given data so that the model can then be used to extrapolate accurately.

4.2 What does the data tell us?

After reviewing the results of the investigation, I believe that the concentration of carbon dioxide in both the atmosphere and water in the GBR is increasing at an alarming rate. We have seen that carbon dioxide in water in particular has seen the largest increase since 24th March 2011 and it is vital that we take actions to prevent further damage to the marine environment. The model from Section 3 has shown that the ocean surface temperature of the GBR is periodic as it peaks and drops at certain times in the year. However, looking at the yearly statistics for T, x and y, it is clear that this periodicity still has an increasing trend. Furthermore, looking at the extrapolation for years 2030, 2050 and 2100, we can see that CO_2 levels in the atmosphere are set to almost double. This ties in with CO_2 levels in the water as an increased abundancy in CO_2 causes more CO_2 to be absorbed by the water and lastly, since CO_2 is a greenhouse gas, it traps heat – causing even higher temperatures.

4.3 Implications of results for the environment, society and economy

Of all the factors being considered, the environment is arguably to get damaged the most. Marine life, for example, are already being challenged by the less alkali waters as some organisms with shells and skeletons face being dissolved [4], a result of a ruined habitat. Furthermore, with fewer microorganisms in the GBR, the food chain will be disrupted as nearby algae and corals won't receive sufficient nutrients causing microorganisms to look for food elsewhere or possible even becoming extinct. As for society, certain toxins released by plants and corals in acidic waters can be harmful for humans too – possibly resulting in allergic reactions, infections or other complications. Also, being a heritage site, the GBR deserves to be conserved and failing to preserve it can lead to problems even to do with the economy. Many fish thrive off algae and plants in the GBR so if these corals are damaged, more fish will go to other locations for food – possibly travelling further away from the coast. This makes fishing much more difficult as boats will have to travel further for fewer fish – affect livelihoods of many fishermen and thus the economy as it may be some people's only source of income. To me, analysing sustainability problems is all about offering possible solutions rather than fixating on past data. This is because action needs to be taken as soon as possible to avoid devastating effects to the environment. Finally, I feel that engineering students should be able to look at these issues from different points of view as some trends may be difficult to spot but using a basic understanding of topics such as rates of change can go a long way to help understand complex patterns.

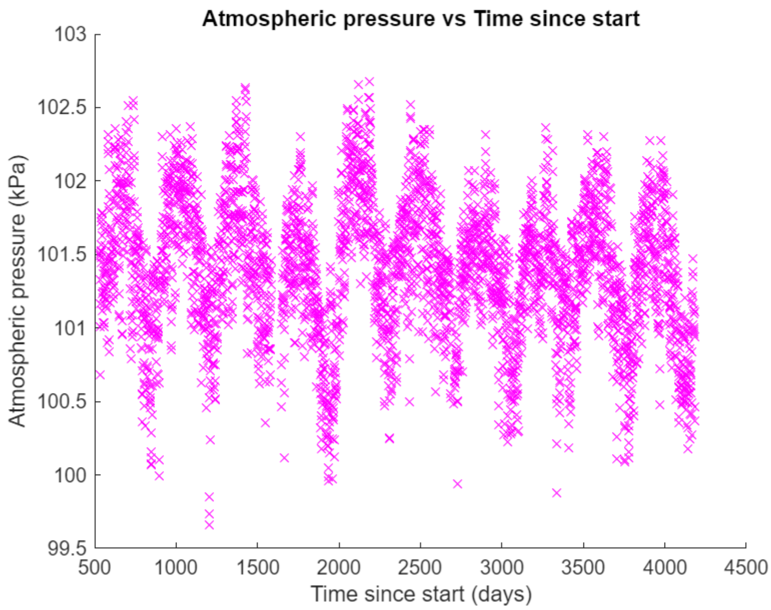
5 Appendix

5.1 References

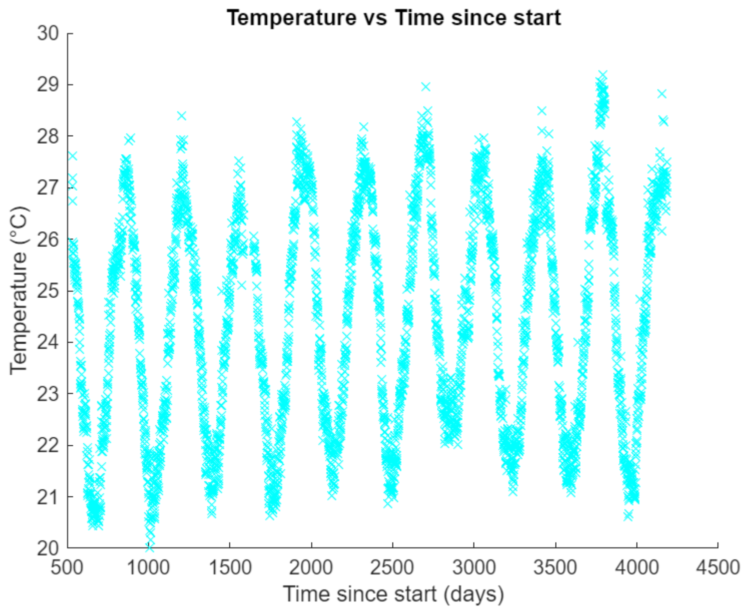
[1] CNN, B.A.W. (2020). *The Great Barrier Reef has lost half its corals within 3 decades*. [online] CNN. Available at: https://edition.cnn.com/travel/article/great-barrier-reef-coral-loss-intl-scli-climate-scn/index.html.
[2] University of Plymouth (2021). *What is ocean acidification? Find out how research at Plymouth is tackling this global carbon dioxide problem*. [online] University of Plymouth. Available at: https://www.plymouth.ac.uk/research/ocean-acidification.
[3] University of Plymouth (2021). *What is ocean acidification? Find out how research at Plymouth is tackling this global carbon dioxide problem*. [online] University of Plymouth. Available at: https://www.plymouth.ac.uk/research/ocean-acidification.
[4] Coast Adapt (2017). *Ocean acidification and its effects*. [online] CoastAdapt. Available at: https://coastadapt.com.au/ocean-acidification-and-its-effects.
[5] Fisheries, N. (2020). *Understanding Ocean Acidification | NOAA Fisheries*. [online] NOAA. Available at: https://www.fisheries.noaa.gov/insight/understanding-ocean-acidification#why-is-ocean-acidification-a-problem?.
[6] Fisheries, N. (2020). *Understanding Ocean Acidification | NOAA Fisheries*. [online] NOAA. Available at: https://www.fisheries.noaa.gov/insight/understanding-ocean-acidification#why-is-ocean-acidification-a-problem?.
[7] Mathworks.com. (2024). *Detect Outliers in Vector*. [online] Available at: https://uk.mathworks.com/help/matlab/ref/isoutlier.html#bvlllts-method [Accessed 1 Nov. 2024].
[8] Wikipedia Contributors (2023). *2014–2016 El Niño event*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/2014%E2%80%932016_El_Ni%C3%B1o_event.
[9] Isaacs-Thomas, B. (2022). *When it comes to sucking up carbon emissions, 'the ocean has been forgiving.' That might not last*. [online] PBS NewsHour. Available at: https://www.pbs.org/newshour/science/the-ocean-helps-absorb-our-carbon-emissions-we-may-be-pushing-it-too-far.
[10] Energy & Climate Intelligence Unit (2023). *Net Zero Tracker*. [online] Energy & Climate Intelligence Unit. Available at: https://eciu.net/netzerotracker.

5.2 MATLAB Code

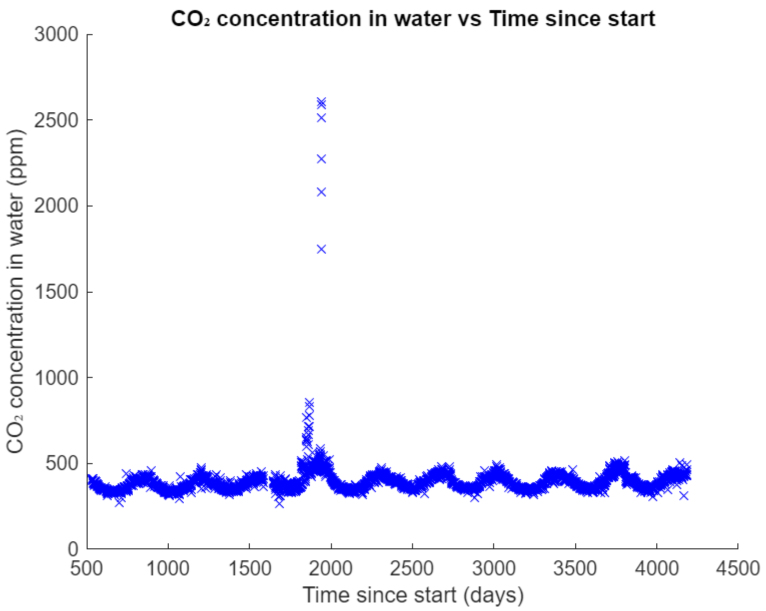
```
initialTable = table(d,d_,P,T,x,y);
% An initial table containing all 4210 values for date, ocean surface
temperature,
% atmospheric pressure, seawater carbon dioxide content and air carbon
dioxide content


uniqueTable = unique(initialTable, "rows");
% Each entire row in initialTable is treated as a single entity and
% duplicate rows are removed. This new table is uniqueTable.


uniqueTable(1:526,:) = [];
dclipped=d(527:end);
d_clipped=d_(527:end);
Pclipped=P(527:end);
Tclipped=T(527:end);
xclipped=x(527:end);
yclipped=y(527:end);


Pvalues = table(dclipped,d_clipped,Pclipped);
Tvalues = table(dclipped,d_clipped,Tclipped);
xvalues = table(dclipped,d_clipped,xclipped);
yvalues = table(dclipped,d_clipped,yclipped);
scatter(d_clipped,Pclipped,"mx"); % Plots with outliers (days)
xlabel('Time since start (days)');
ylabel('Atmospheric pressure (kPa)');
title('Atmospheric pressure vs Time since start');
```



```
scatter(d_clipped,Tclipped,"cx");
xlabel('Time since start (days)');
ylabel('Temperature (°C)');
title('Temperature vs Time since start');
```

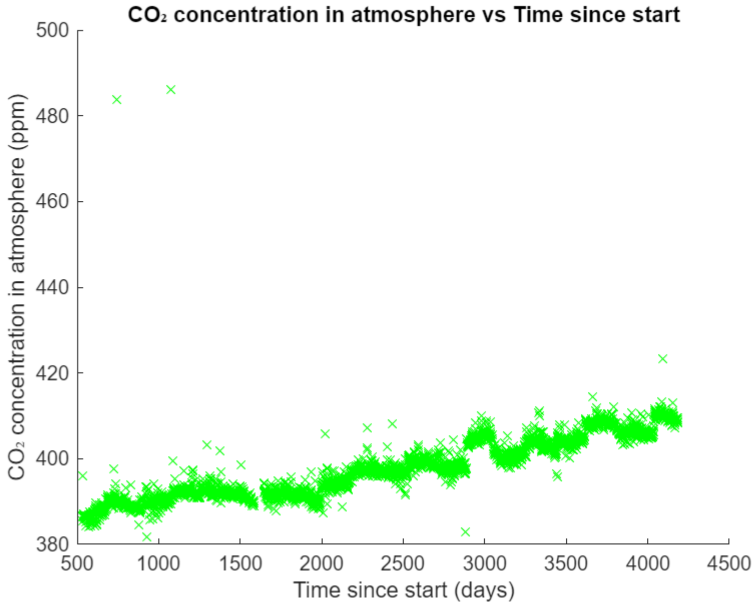



```
scatter(d_clipped,xclipped,"bx");
xlabel('Time since start (days)');
ylabel('CO₂ concentration in water (ppm)');
title('CO₂ concentration in water vs Time since start');
```



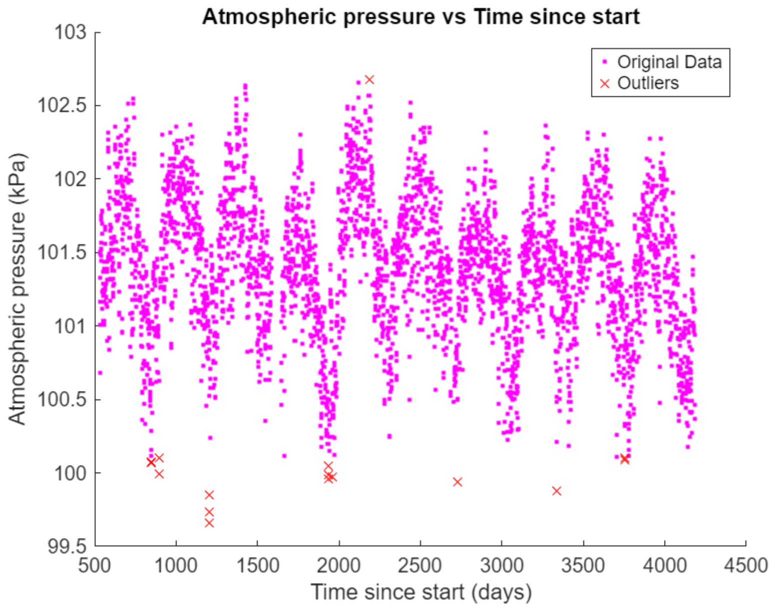

```
scatter(d_clipped,yclipped,"gx");
```

```
xlabel('Time since start (days)');
ylabel('CO₂ concentration in atmosphere (ppm)');
title('CO₂ concentration in atmosphere vs Time since start');
```



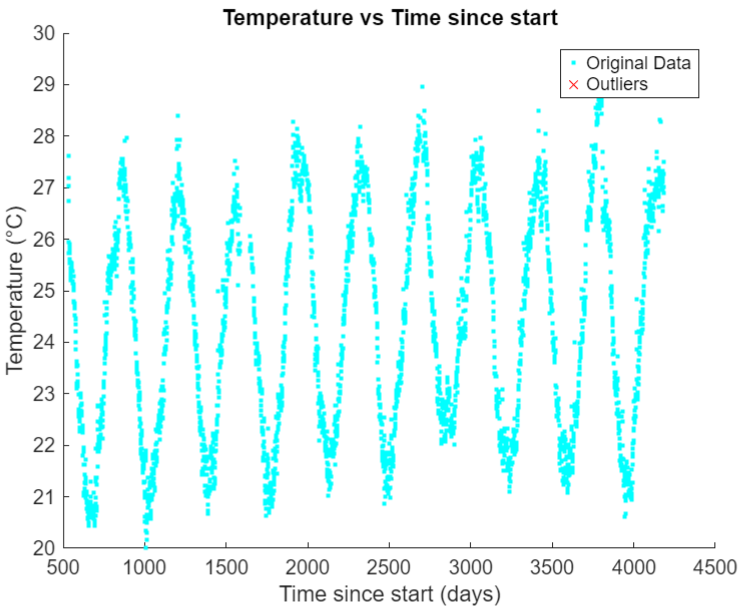

```
detection_P = isoutlier(Pclipped,"quartiles");
% Checks for outliers in atmospheric pressure
detection_T = isoutlier(Tclipped,"quartiles");
% Checks for outliers in ocean surface temperature
detection_x = isoutlier(xclipped,"quartiles");
% Checks for outliers in water carbon dioxide content
detection_y = isoutlier(yclipped,"quartiles");
% Checks for outliers in air carbon dioxide content


index_Poutliers = find(detection_P);
% Indexes the position of outliers in atmospheric pressure and outputs the
respective row number
index_Toutliers = find(detection_T);
% Indexes the position of outliers in ocean surface temperature and outputs
the respective row number
index_xoutliers = find(detection_x);
% Indexes the position of outliers in water carbon dioxide content and
outputs the respective row number
index_youtliers = find(detection_y);
% Indexes the position of outliers in air carbon dioxide content and outputs
the respective row number


cleanPvalues = Pvalues;
cleanPvalues(index_Poutliers,:) = [];

cleanTvalues = Tvalues;
```

```
cleanTvalues(index_Toutliers,:) = [];

cleanxvalues = xvalues;
cleanxvalues(index_xoutliers,:) = [];

cleanyvalues = yvalues;
cleanyvalues(index_youtliers,:) = [];


cleanPvalues.year = year(cleanPvalues.dclipped);
cleanTvalues.year = year(cleanTvalues.dclipped);
cleanxvalues.year = year(cleanxvalues.dclipped);
cleanyvalues.year = year(cleanyvalues.dclipped);
% Plotting Atmospheric Pressure (Days vs Pressure)
figure;
hold on;
plot(d_clipped(~detection_P), Pclipped(~detection_P), 'm.', 'DisplayName',
'Original Data');
scatter(d_clipped(detection_P), Pclipped(detection_P), 'rx', 'DisplayName',
'Outliers');
xlabel('Time since start (days)');
ylabel('Atmospheric pressure (kPa)');
title('Atmospheric pressure vs Time since start');
legend('show');
hold off;
```

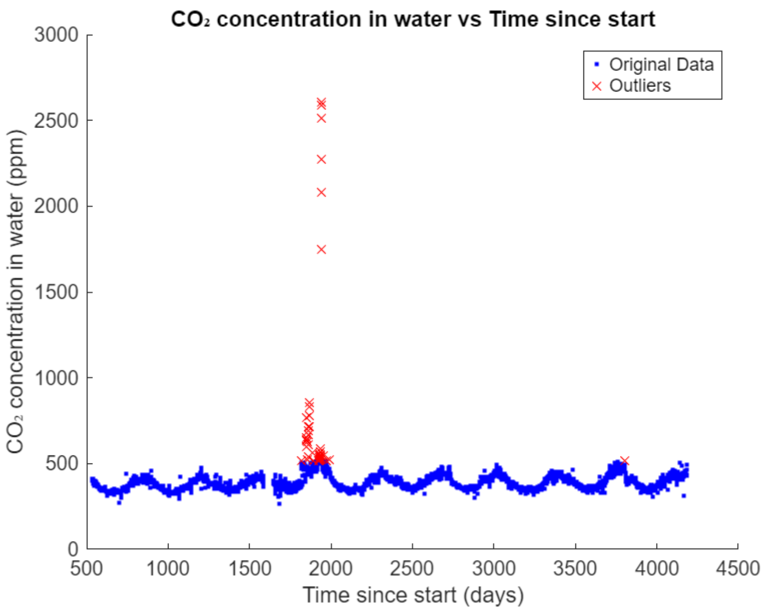



```

% Plotting Temperature (Days vs Temperature)
figure;
hold on;
plot(d_clipped(~detection_T), Tclipped(~detection_T), 'c.', 'DisplayName',
'Original Data');
```

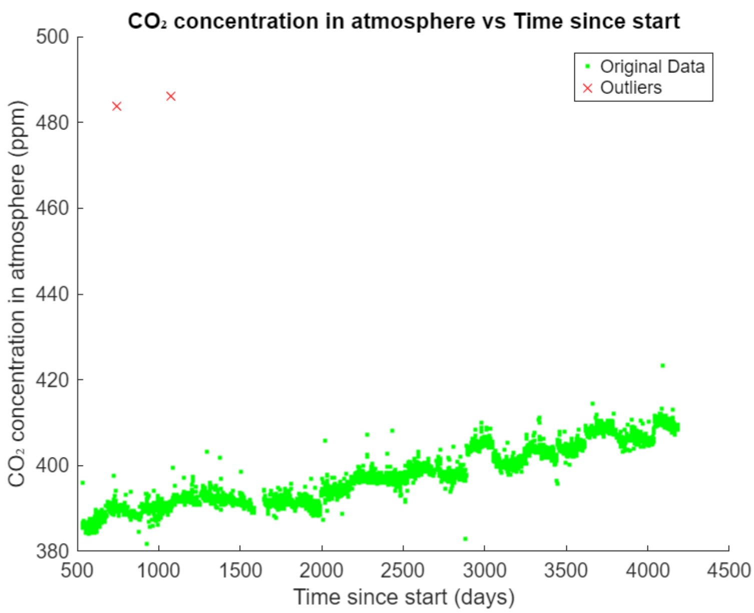
```
scatter(d_clipped(detection_T), Tclipped(detection_T), 'rx', 'DisplayName',
'Outliers');
xlabel('Time since start (days)');
ylabel('Temperature (°C)');
title('Temperature vs Time since start');
legend('show');
hold off;
```



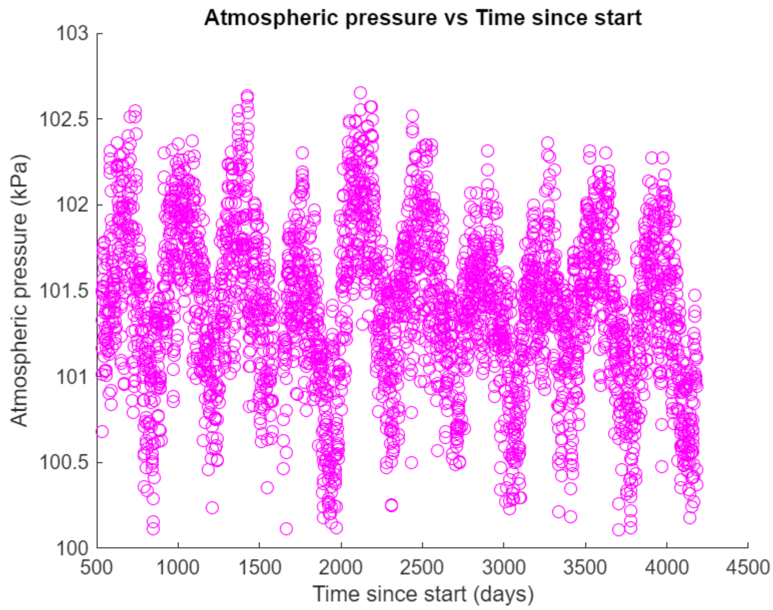

```
% Plotting CO₂ concentration in water (Days vs CO₂ in water)
figure;
hold on;
plot(d_clipped(~detection_x), xclipped(~detection_x), 'b.', 'DisplayName',
'Original Data');
scatter(d_clipped(detection_x), xclipped(detection_x), 'rx', 'DisplayName',
'Outliers');
xlabel('Time since start (days)');
ylabel('CO₂ concentration in water (ppm)');
title('CO₂ concentration in water vs Time since start');
legend('show');
hold off;
```



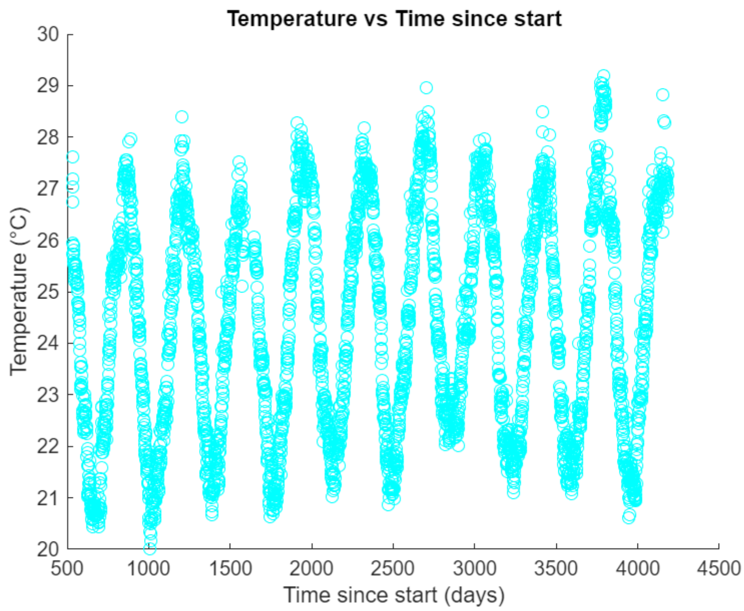

```matlab
% Plotting CO₂ concentration in atmosphere (Days vs CO₂ in atmosphere)
figure;
hold on;
plot(d_clipped(~detection_y), yclipped(~detection_y), 'g.', 'DisplayName',
'Original Data');
scatter(d_clipped(detection_y), yclipped(detection_y), 'rx', 'DisplayName',
'Outliers');
xlabel('Time since start (days)');
ylabel('CO₂ concentration in atmosphere (ppm)');
title('CO₂ concentration in atmosphere vs Time since start');
legend('show');
hold off;
```



```
scatter(cleanPvalues.d_clipped,cleanPvalues.Pclipped,"magenta") % Plots with clean data (days)
xlabel('Time since start (days)');
ylabel('Atmospheric pressure (kPa)');
title('Atmospheric pressure vs Time since start');
```

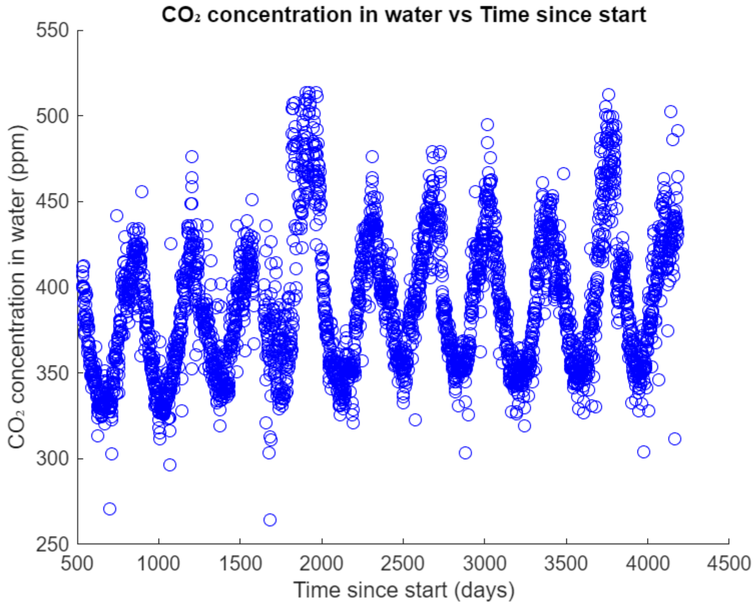



```
scatter(cleanTvalues.d_clipped,cleanTvalues.Tclipped,"cyan")
xlabel('Time since start (days)');
ylabel('Temperature (°C)');
title('Temperature vs Time since start');
```

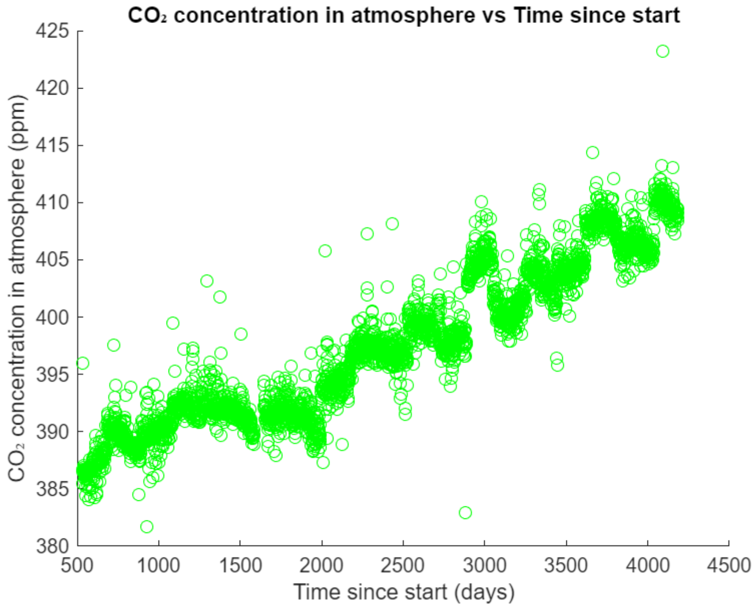



```
scatter(cleanxvalues.d_clipped,cleanxvalues.xclipped,"blue")
```

```
xlabel('Time since start (days)');
ylabel('CO₂ concentration in water (ppm)');
title('CO₂ concentration in water vs Time since start');
```




```
scatter(cleanyvalues.d_clipped, cleanyvalues.yclipped,"green")
xlabel('Time since start (days)');
ylabel('CO₂ concentration in atmosphere (ppm)');
title('CO₂ concentration in atmosphere vs Time since start');
```




```
Pinitialstats = struct2cell(datastats(cleanPvalues.Pclipped));
Tinitialstats = struct2cell(datastats(cleanTvalues.Tclipped));
xinitialstats = struct2cell(datastats(cleanxvalues.xclipped));
yinitialstats = struct2cell(datastats(cleanyvalues.yclipped));
```

```
 Pquartiles = num2cell(quantile(cleanPvalues.Pclipped,[0.25;0.75;]));
 Tquartiles = num2cell(quantile(cleanTvalues.Tclipped,[0.25;0.75;]));
 xquartiles = num2cell(quantile(cleanxvalues.xclipped,[0.25;0.75;]));
 yquartiles = num2cell(quantile(cleanyvalues.yclipped,[0.25;0.75;]));


 P_stats = [Pinitialstats; Pquartiles; iqr(cleanPvalues.Pclipped);
var(cleanPvalues.Pclipped)];
 T_stats = [Tinitialstats; Tquartiles; iqr(cleanTvalues.Tclipped);
var(cleanTvalues.Tclipped)];
 x_stats = [xinitialstats; xquartiles; iqr(cleanxvalues.xclipped);
var(cleanxvalues.xclipped)];
 y_stats = [yinitialstats; yquartiles; iqr(cleanyvalues.yclipped);
var(cleanyvalues.yclipped)];


 statnames = ["Total"; "Maximum"; "Minimum"; "Mean"; "Median";
"Range";"Standard Deviation";"Lower Quartile"; "Upper Quartile";
"Interquartile Range"; "Variance";];


 allStats = table(statnames,P_stats,T_stats,x_stats,y_stats);


 PyearlySummary = groupsummary(cleanPvalues, "year", {"mean", "std"},
'Pclipped');
 TyearlySummary = groupsummary(cleanTvalues, "year", {"mean", "std"},
'Tclipped');
 xyearlySummary = groupsummary(cleanxvalues, "year", {"mean", "std"},
'xclipped');
 yyearlySummary = groupsummary(cleanyvalues, "year", {"mean", "std"},
'yclipped');

 allYearlySummary = table(PyearlySummary, TyearlySummary, xyearlySummary,
yyearlySummary);
 linreg_P = fitlm(cleanPvalues.d_clipped, cleanPvalues.Pclipped, 'linear');
 coef_P = corrcoef(cleanPvalues.d_clipped, cleanPvalues.Pclipped);
 linreg_T = fitlm(cleanTvalues.d_clipped, cleanTvalues.Tclipped, 'linear');
 coef_T = corrcoef(cleanTvalues.d_clipped, cleanTvalues.Tclipped);
 linreg_x = fitlm(cleanxvalues.d_clipped, cleanxvalues.xclipped, 'linear');
 coef_x = corrcoef(cleanxvalues.d_clipped, cleanxvalues.xclipped);
 linreg_y = fitlm(cleanyvalues.d_clipped, cleanyvalues.yclipped, 'linear');
 coef_y = corrcoef(cleanyvalues.d_clipped, cleanyvalues.yclipped);

 disp(linreg_P);
```

```
Linear regression model:
    y ~ 1 + x1
```

```
Estimated Coefficients:
                    Estimate         SE        tStat       pValue
                   ___________    __________    _______    __________

    (Intercept)         101.63      0.018592     5466.4             0
    x1             -9.9496e-05    6.9418e-06    -14.333    2.2473e-45


Number of observations: 3666, Error degrees of freedom: 3664
Root Mean Squared Error: 0.459
R-squared: 0.0531,  Adjusted R-Squared: 0.0528
F-statistic vs. constant model: 205, p-value = 2.25e-45
```

```
disp(coef_P);
```

```
    1.0000   -0.2304
   -0.2304    1.0000
```

```
disp(linreg_T);
```

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                    Estimate         SE        tStat       pValue
                   __________    __________    ______    __________

    (Intercept)        23.492       0.08575    273.96             0
    x1             0.00038675    3.2048e-05    12.068    6.4739e-33


Number of observations: 3684, Error degrees of freedom: 3682
Root Mean Squared Error: 2.12
R-squared: 0.038,  Adjusted R-Squared: 0.0378
F-statistic vs. constant model: 146, p-value = 6.47e-33
```

```
disp(coef_T);
```

```
    1.0000    0.1951
    0.1951    1.0000
```

```
disp(linreg_x);
```

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ________    __________    ______    __________

    (Intercept)      365.03        1.5408    236.91             0
    x1             0.010103    0.00057434     17.59    1.5427e-66


Number of observations: 3642, Error degrees of freedom: 3640
Root Mean Squared Error: 38
R-squared: 0.0783,  Adjusted R-Squared: 0.0781
F-statistic vs. constant model: 309, p-value = 1.54e-66
```

```
disp(coef_x);
```

```
    1.0000    0.2799
```

```
    0.2799    1.0000
```

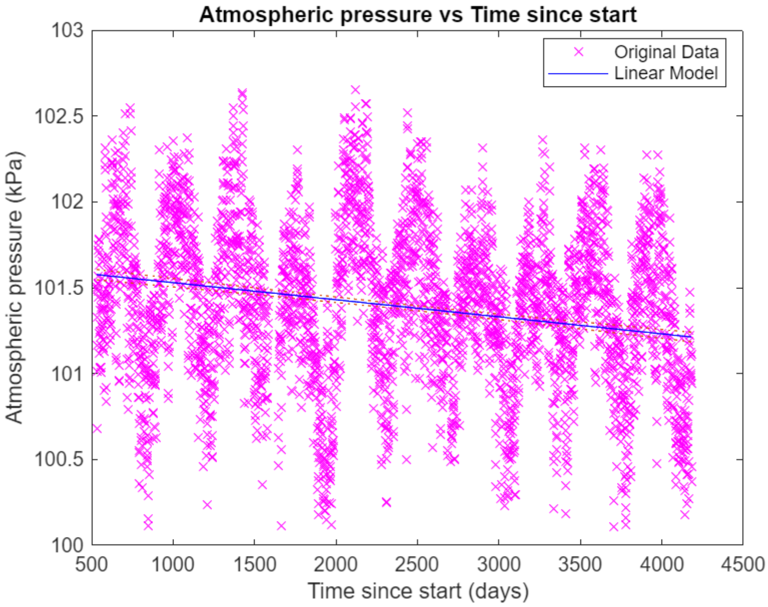
```
disp(linreg_y);
```

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate        SE        tStat     pValue
                   _________    __________    ______    ______

    (Intercept)       383.02      0.091674    4178.1      0   
    x1             0.0061743    3.4254e-05    180.25      0   


Number of observations: 3682, Error degrees of freedom: 3680
Root Mean Squared Error: 2.27
R-squared: 0.898,  Adjusted R-Squared: 0.898
F-statistic vs. constant model: 3.25e+04, p-value = 0
```

```
disp(coef_y);
```

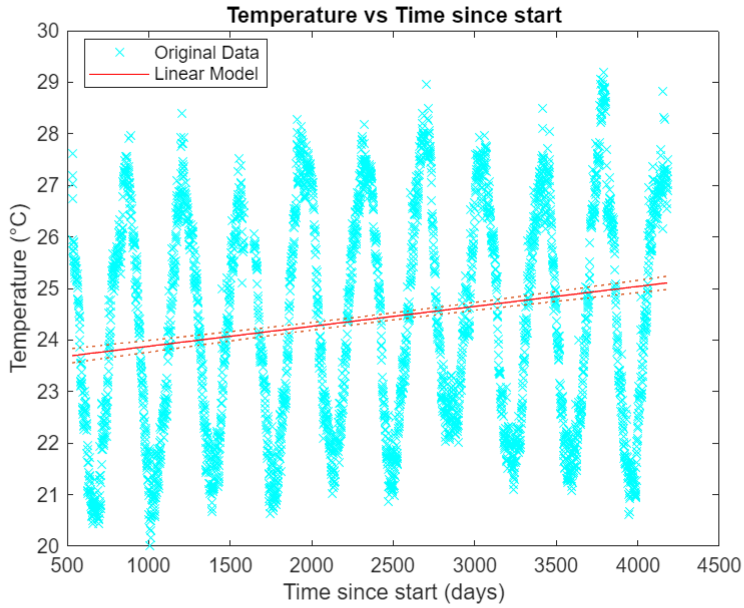
```
    1.0000    0.9478
    0.9478    1.0000
```

```
plotHandleP = plot(linreg_P);
set(plotHandleP(1), 'Color', 'm');
set(plotHandleP(2), 'Color', 'b');
xlabel('Time since start (days)');
ylabel('Atmospheric pressure (kPa)');
title('Atmospheric pressure vs Time since start');
legend('Original Data', 'Linear Model');
```

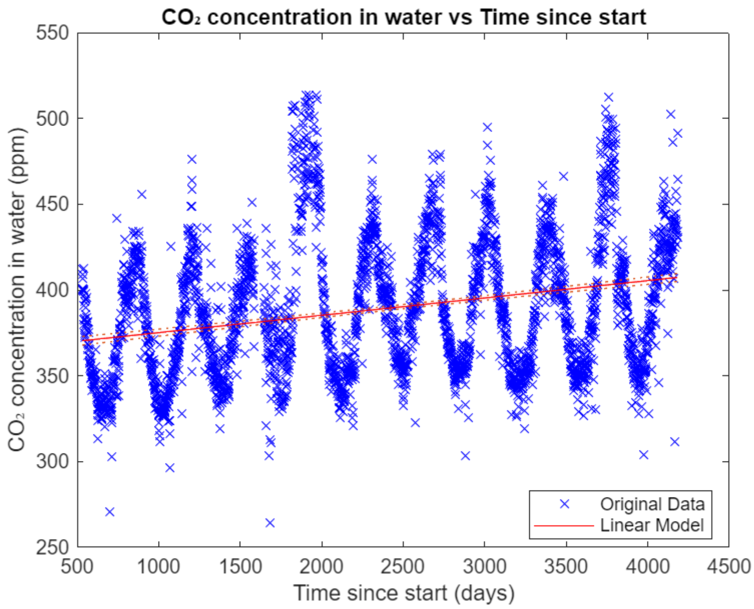



```
plotHandleT = plot(linreg_T);
set(plotHandleT(1), 'Color', 'c');
```

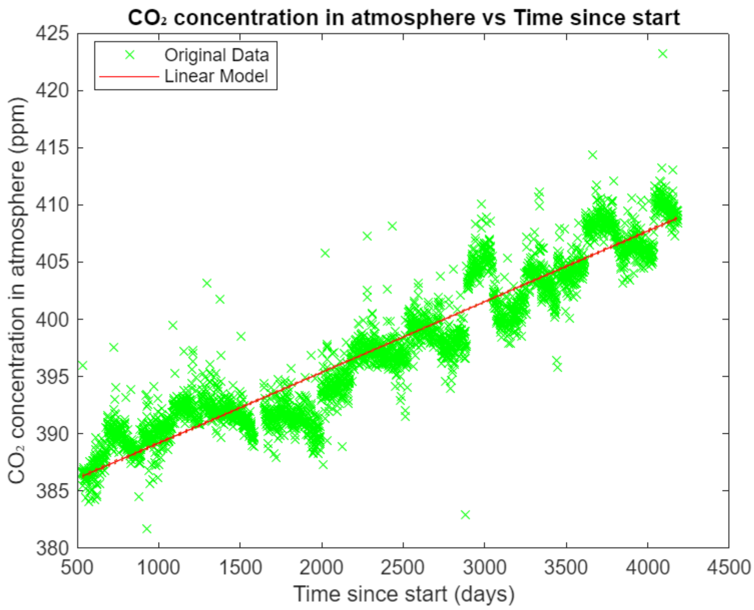
```
set(plotHandleT(2), 'Color', 'r');
xlabel('Time since start (days)');
ylabel('Temperature (°C)');
title('Temperature vs Time since start');
legend('Original Data', 'Linear Model');
```




```

plotHandlex = plot(linreg_x);
set(plotHandlex(1), 'Color', 'b');
set(plotHandlex(2), 'Color', 'r');
xlabel('Time since start (days)');
ylabel('CO₂ concentration in water (ppm)');
title('CO₂ concentration in water vs Time since start');
legend('Original Data', 'Linear Model');
```

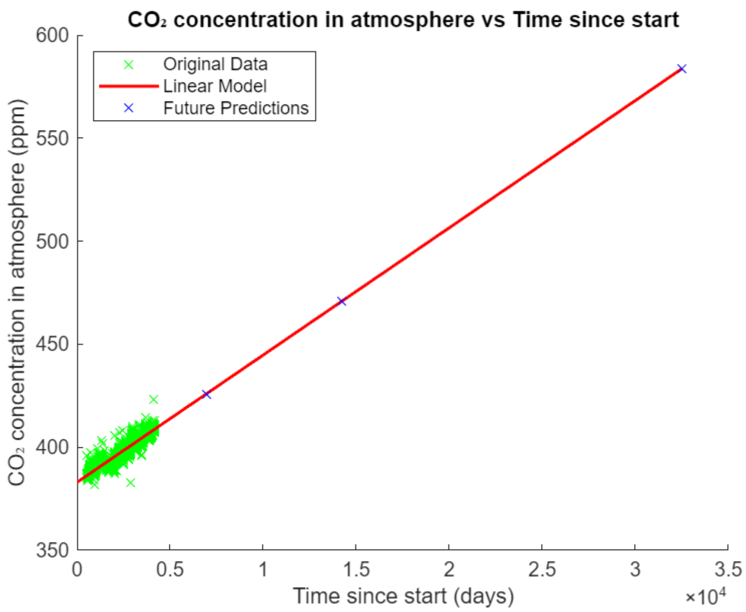


```
plotHandley = plot(linreg_y);
set(plotHandley(1), 'Color', 'g');
set(plotHandley(2), 'Color', 'r');
xlabel('Time since start (days)');
ylabel('CO₂ concentration in atmosphere (ppm)');
title('CO₂ concentration in atmosphere vs Time since start');
legend('Original Data', 'Linear Model');
```




```
startDate = datetime('24-Mar-2011');

futureDates = datetime({'24-Mar-2030'; '24-Mar-2050'; '24-Mar-2100';});

future_d_ = days(futureDates - startDate);
future_d_ = future_d_(:);

linreg_ytest = fitlm(cleanyvalues.d_clipped(:), cleanyvalues.yclipped(:));

predicted_futurey = predict(linreg_ytest, future_d_);

endDate = futureDates(end);
all_dates = startDate:caldays(1):endDate;
all_d_ = days(all_dates - startDate);
all_d_ = all_d_(:);

predicted_all_y = predict(linreg_ytest, all_d_);

figure;
scatter(cleanyvalues.d_clipped, cleanyvalues.yclipped, 'gx');
hold on;
plot(all_d_, predicted_all_y, 'r-', 'LineWidth', 1.5);
scatter(future_d_, predicted_futurey, 'bx');
```

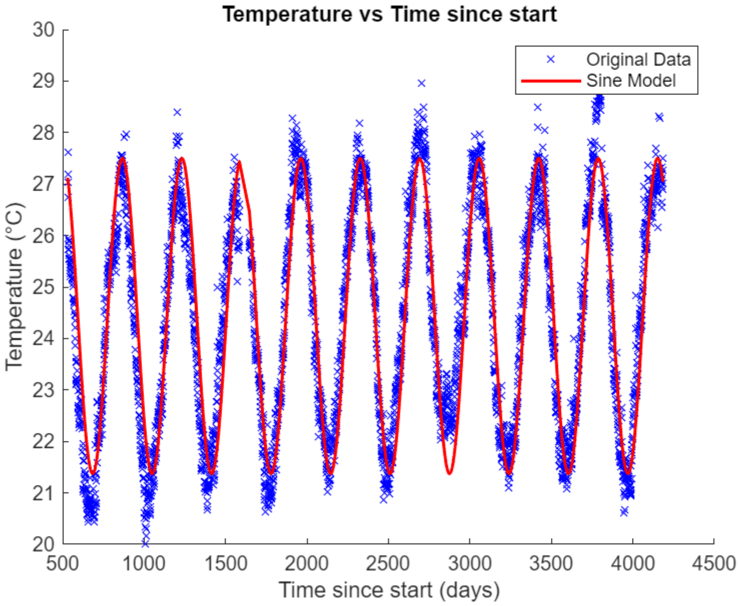
```
xlabel('Time since start (days)');
ylabel('CO₂ concentration in atmosphere (ppm)');
title('CO₂ concentration in atmosphere vs Time since start');
legend('Original Data', 'Linear Model', 'Future Predictions', 'Location',
'NorthWest');
hold off;
```




```
Trms = rms((cleanTvalues.Tclipped)-mean(cleanTvalues.Tclipped));
T_Amp = Trms*sqrt(2);

k1 = T_Amp;
k2 = 1/365;
k3 = -pi/4
```

```
k3 = -0.7854
```

```
k4 = mean(cleanTvalues.Tclipped);


T_model = k1 * sin((2*pi*k2*cleanTvalues.d_clipped)+k3) + k4;

figure;
hold on;
plot(cleanTvalues.d_clipped, cleanTvalues.Tclipped, 'bx', 'DisplayName',
'Original Data', 'MarkerSize', 5);
plot(cleanTvalues.d_clipped, T_model, 'r-', 'DisplayName', 'Sine Model',
'LineWidth', 1.5);
xlabel('Time since start (days)');
ylabel('Temperature (°C)');
title('Temperature vs Time since start');
legend;
hold off;
```




```
Q3model = @(params, d) params(1) * sin(2 * pi * params(2) * d + params(3)) +
params(4);

%insert your initial guesses for initial parameters in this vector by
replacing
%the '_' elements
initial_parameters = [k1 , k2, k3, k4];
%insert your initial guesses for the lower bound of parameters in this
vector
lower_bounds = [ k1-2, 1/370, -pi, k4-3];
%insert your initial guesses for the upper bound of parameters in this
vector
upper_bounds = [k1+2 , 1/350, 0, k4+3];


[fitParams, resnorm, residual, exitflag, output] = ...
    lsqcurvefit(Q3model, initial_parameters, d_, T, lower_bounds,
upper_bounds);
```

```
Local minimum possible.

lsqcurvefit stopped because the final change in the sum of squares relative to
its initial value is less than the value of the function tolerance.

<stopping criteria details>
```

```
% Write code to access the 'fitParams' variable to find out the values
fitted for
%each parameter
disp(fitParams)
```

```
    2.9685    0.0027   -0.5302   24.4225
```

```
% Calculate the fitted model
T_fit = Q3model(fitParams, d_);

% plot a comparison of the data and your results
figure;
plot(d, T, 'b.', 'MarkerSize', 12); hold on;
plot(d, T_fit, 'r-', 'LineWidth', 2);
xlabel('Time since start (days)');
ylabel('Temperature (°C)');
title('Temperature vs Time since start');
legend('Original Data', 'Optimised Model');
```