GHCN Data Analysis using Spark

DATA420

Scalable Data Science

Thomas Waldin

University of Canterbury

Date Submitted: 11 April 2025

TABLE OF CONTENTS

1. BACKGROUND

The Global Historical Climatology Network (GHCN) is a large database of climate data spanning 250 years. 200 countries and territories provide data collected from over 100,000 weather stations. In this assignment this data was processed and analysed in a distributed way, and visualisations were generated to show key information in the dataset, in a local and international context. A significant effort was made to process and analyse the data in an efficient manner, such that resources were preserved, and the learnings could be applied to larger datasets in real-world problems. The PySpark library was utilised to achieve this. A variety of visualisations were generated using multiple tools, giving experience in plotting and geospatial data representation.

2. PROCESSING

The data used in this assignment was stored in Azure Blob Storage, and consisted of one large, distributed dataset and four metadata tables. Azure handles the distribution and replication internally, so the focus was only to interact with the dataset in an efficient manner. The dataset, ‘daily’, represents daily outputs of all weather stations in the network, where a single row contains an observation for a single day and station, and each element collected is in a separate row also. The ‘stations’ table contains information about each station in the network, including location through latitude, longitude, and elevation. The ‘inventory’ table contains information about the elements that each station collects, including what elements are collected, and the first and last year the elements were collected. The tables ‘states’ and ‘countries’ each contain reference I.D.s and names of the states and countries. The data was found to be structured as follows.

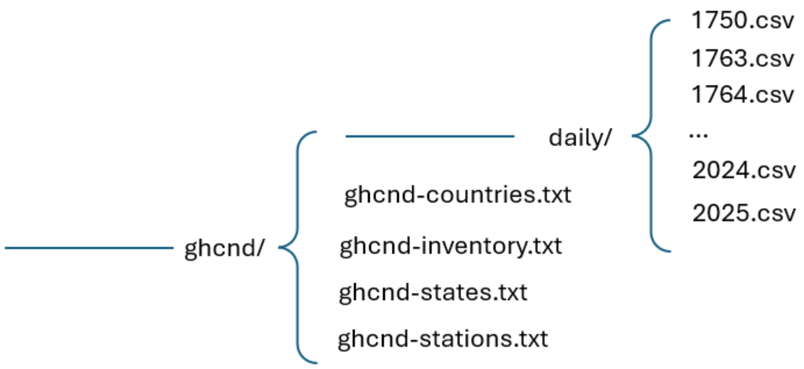


Figure 1: *A visualisation of the directory tree of the data.*

The data is all stored in a directory called ghcnd and each of the metadata tables are stored here as .txt files. In the ghcnd directory there is a directory called daily that contains the 'daily' data, stored in multiple .csv files, one for each year of data.

The years of 'daily' are from 1763 to 2025, with a single file for 1750 as well. The sizes of these files generally increase through the years, except for 2025, as at the time of this analysis, it is only part way through 2025.

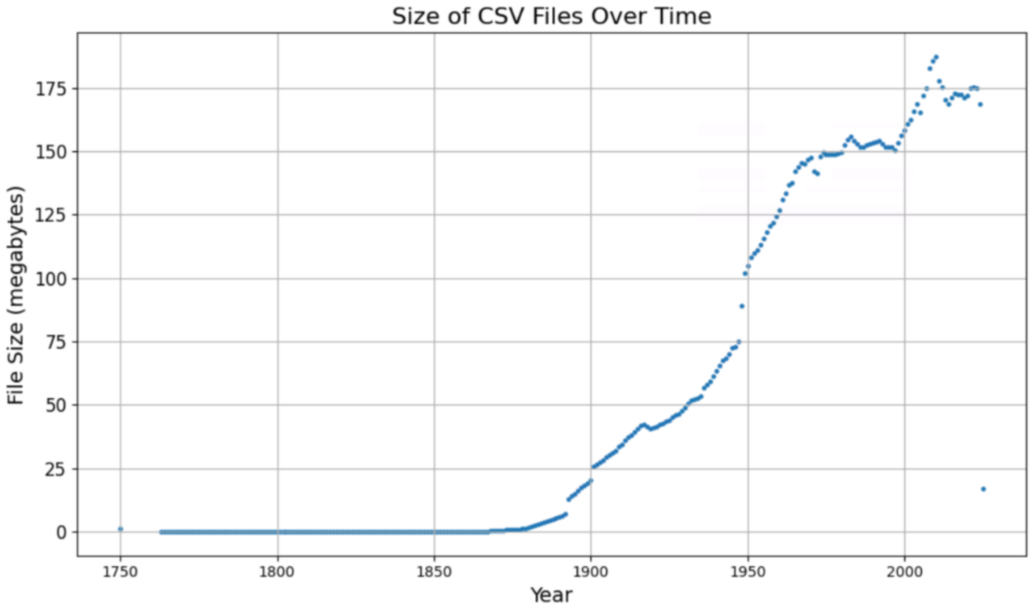


Figure 2: *A visualization of the change in size in 'daily' over time.*

It is expected that the increase in the size of 'daily' is likely due to more stations being added to the network over time, and perhaps stations being updated to report more elements each day.

The sizes of the data were also investigated. Shown as follows in Table 1, 'daily' is significantly larger than the rest of the data.

Table 1: *A table containing the names and sizes of each of the datasets.*

Table Name	Size (Bytes)
countries	3659
states	1086
stations	11150502
daily	13975887693
inventory	35218290
Total	**14022261230**

The dataset ‘daily’ makes up 99.7% of the dataset and is nearly 14 GB in total. The ‘inventory’ table is the next largest at only 0.035 GB.

Each of the datasets were loaded into PySpark, and schemas were applied. The dataset ‘daily’ had eight columns. ‘DATE’, ‘VALUE’, and ‘OBS-TIME’ were loaded as IntegerType as the columns contained whole numbers and the rest were loaded as StringType. The dates and times were not converted to date-time objects as the integer was functional for ordering and at this stage no calculations were expected, so changing it would be a waste of resources. A subset of ‘daily’ was successfully loaded into Spark and the schema applied as expected.

As the metadata tables were stored as .txt files instead of .csv, some extra processing was required to apply a schema. Column values were extracted using the PySpark substring function, parsing the text manually based on the range of characters of each. The column with the original data needed to be dropped from each table once the schema was applied correctly.

Once the metadata tables were loaded correctly, the number of rows of each could be counted, and these are presented as follows in Table 2.

Table 2: *A table containing the number of rows of each of the metadata tables.*

Metadata Table Name	Number of Rows
stations	129657
states	74
countries	219
inventory	765615

The relevant information from the metadata tables were then combined to gain an enriched version of the 'stations' table.

The two character country code in each station I.D. was extracted and added to 'stations' as a new column using the withColumn method. A left join was performed on 'stations' and 'countries' on the new country I.D. column, enriching 'stations' with a column containing the country that the station operates in. A similar left join was done with 'states', to give a column of state names for stations in countries that had states or provinces.

Five of the collected elements were defined by the GHCN as 'core' elements. These were the minimum and maximum temperatures, precipitation, snowfall, and snow depth. A new column was added to the 'inventory' table, designating if the collected value was a core element or not. The 'inventory' table was further enriched, by grouping by the station ID, then taking the minimum start date and maximum end date to give the first and last year each station collected any element. A list of the distinct core elements and the counts of core and non-core elements were also added as new columns.

By filtering the 'inventory' table it was found that 20504 stations collect all five core elements, and 16348 collect precipitation and no other elements.

Table 3 below shows the columns of the enriched 'inventory' table.

Table 3: *A table naming each column in the enriched 'inventory' and a description of its contents.*

Column Name	Type	Description
ID	StringType	Unique station ID
first_year	IntegerType	The first year the station was active collecting any element
last_year	IntegerType	The last year the station was active collecting any element
distinct_elements_count	IntegerType	Count of how many elements the station collects
core_elements_count	IntegerType	Count of how many core elements the station collects
other_elements_count	IntegerType	Count of how many non-core elements the station collects

This table was left joined with the 'stations' table and saved to storage. From this point onwards, 'stations' refers to this enriched table.

A subset of 'daily' was left joined with 'stations' and it was found that all stations in the subset of 'daily' were present in 'stations'. To find if this is true for the full 'daily' dataset a different method is required.

Left joining 'daily' and 'stations' would involve matching every row in 'daily' (~14GB) to each row in 'stations' across the partitions. This requires a shuffle, meaning Spark would redistribute both datasets across partitions, having to read the data, serialize it, and then writing it again after sending it over the network. This is very costly and there are more efficient ways to find out if all stations in 'daily' and present in 'stations'.

Using the subtract method still requires a shuffle but is much more efficient as the amount of shuffling required can be greatly reduced. By first filtering the two tables to only include the ID column first, the cost becomes manageable, and the unique IDs can be subtracted from each other. Through this method it was found that all stations in 'daily' were present in 'stations'.

3. ANALYSIS

Before looking at the 'daily' dataset in more detail, it is useful to know more about the stations themselves. It was found that there were 129657 stations in total, and there were 30735 active so far in 2025.

1218 stations were in the US Historical Climatology Network (HCN), 991 were in the GCOS Surface Network (GSN), and 15 of these were included in both networks (HCN and GSN). 234 were in the US Climate Reference Network (CRN).

There were 25316 stations in the Southern Hemisphere (Latitude < 0).

To count how many stations there were in territories of the United States, the table was filtered by the country column for countries that contained the string 'United States' but was not the exact string 'United States'. This was because the territories were listed as 'Territory Name [United States]'. It was found that there were 414 stations in United States territories.

The total number of stations in each state and country were also calculated. These tables were saved for potential later use.

To compute geographic distance between stations, the Haversine function was chosen, and coded as a user defined function in Spark. The Haversine function works by calculating the distance between two points on the outside of a sphere (Mendoza y Rios, 1796).

This function was briefly tested on a subset of stations, then was used to find the stations in New Zealand that are closest to each other via a cross-join. These stations are the Paraparaumu ASW station (NZ000093417) and the Wellington Aero AWS station (NZM00093439) and they were calculated to be 50.53 km apart.

Next the 'daily' climate dataset was studied in more detail. The count method was applied and the total number of rows in the dataset was found to be 3139143397.

The 'daily' dataset was filtered to gain a subset of only the core elements defined earlier. Table 4 shows the counts of these elements.'

Tale 4: *The counts of each of the core elements in 'daily'.*

Element	Count
PCRP (precipitation)	1079767077
TMIN (minimum temperature)	458928768
TMAX (maximum temperature)	460114659
SNOW (snowfall)	359249644
SNWD (snow depth)	300711620

Precipitation (PRCP) is reported the most, at 1079767077 observations, which is approximately a third of the dataset. As half of the stations only report precipitation, this makes sense.

Interestingly there are not equal observations for TMIN and TMAX, and the stations may not necessarily report both at the same time. It was investigated how many values of TMAX do not have a corresponding TMIN value.

The 'daily' dataset was filtered to only include the TMAX and TMIN values and then was grouped by ID and DATE. A collect set function was performed on ELEMENT to create a column with a list of distinct elements for each station and date. The dataset was filtered by the new column to find TMAX values that do not have a corresponding TMIN.

It was found that 10660214 values of TMAX don't have a corresponding TMIN value. It was found that 28754 stations contribute to these observations, which is approximately one fifth of all stations. This seems an unusually high proportion, given that the non-simultaneous reporting is due to issues with data collection or coverage.

4. VISUALISATIONS

The 'daily' dataset was filtered to just include observations of TMIN and TMAX and only for stations in New Zealand. This filtered table was saved to the output directory. It was found that there were 491441 observations spanning 86 years.

A time series for the average daily TMIN and TMAX for each station in New Zealand was generated. These plots are all included in Appendix 1, but one has been included here as an example.

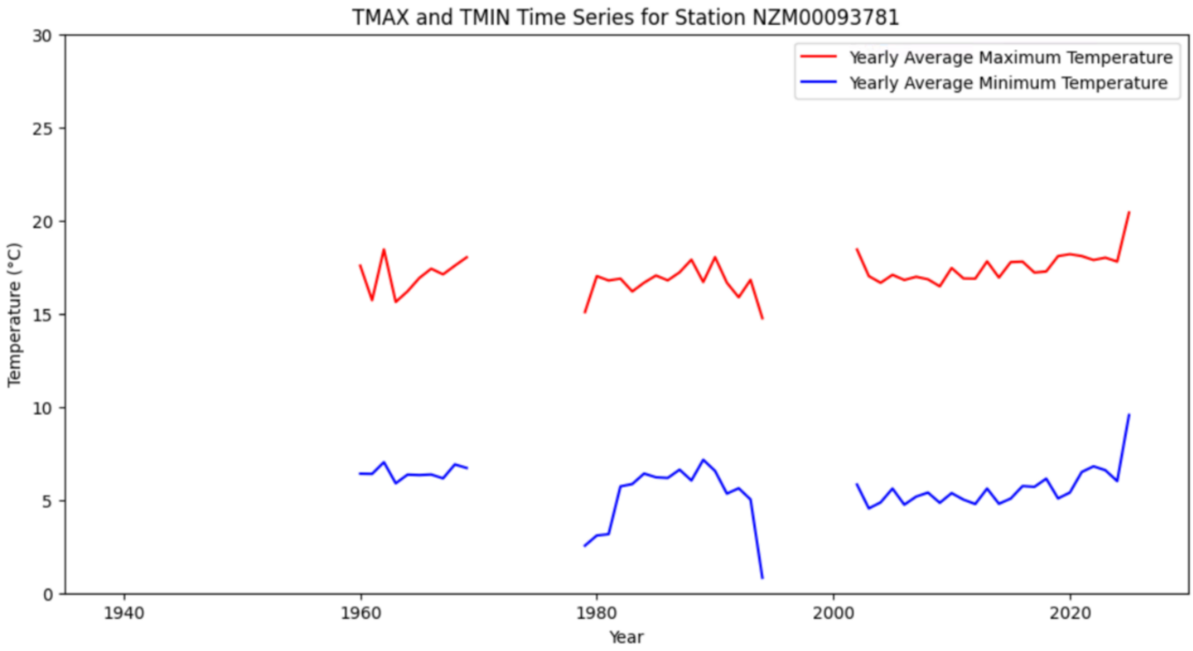


Figure 3: *A plot of observations of average daily TMIN and TMAX for station NZM00093781.*

The average of the values across all stations was also plotted.

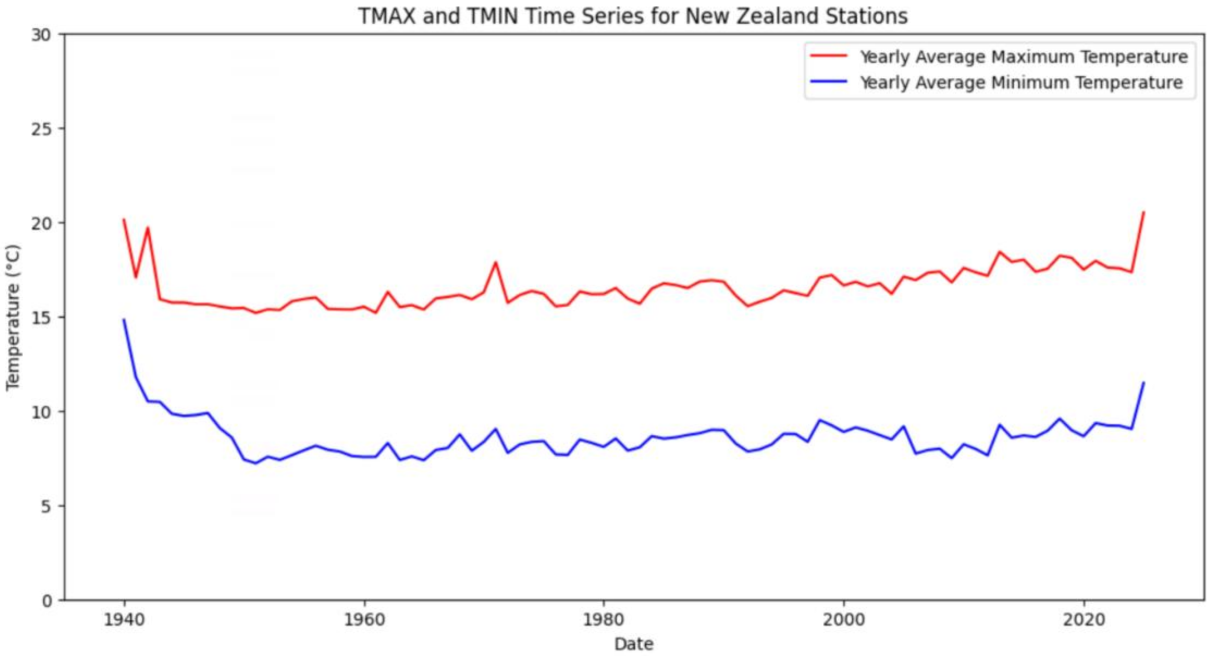


Figure 4: *A plot of observations of average daily TMIN and TMAX for all stations in New Zealand.*

To generate these plots, first the Spark pivot method was used to create a column for both TMAX and TMIN and then the data was grouped by year. For the plot of all stations (Figure 4), the values of each station were averaged. Both plots were smoothed by taking the average over each year. This method removes all seasonal variation, which may or may not be desired. In this case it was selected as a simple temporal smoothing method, as no further analysis of the plot is required. Initially the gaps were filled with interpolated values, so NaN values were generated for missing years, so that the gaps are evident in the plots. All plots used the same axis limits and scales, so they were easily compared, even though some stations only had a few years of data. Sensible colours of red for the warmer TMAX, and blue for the cooler TMIN were used.

Precipitation observations for stations around the world were considered next. 'daily' was filtered to only include precipitation. The data was grouped by year, and the aggregated data was joined with countries. This table was saved to an output directory. The descriptive statistics were generated and are shown in Table 5.

Table 5: *The descriptive statistics of the aggregated precipitation observation.*

Summary Statistic	Precipitation
Count	17712
Mean	43.8
Standard Deviation	88.8
Minimum	-1
Maximum	4361

The highest average daily rainfall over a year was 436.1 mm which was for Equatorial Guinea. Equatorial Guinea does receive a large annual rainfall but an average daily rainfall this large is unrealistic, suggesting there may be an issue with reporting. It also doesn't make sense that the minimum value is negative, as rainfall can only be positive. The mean of 4.4 mm is sensible.

A choropleth colouring a map was generated using plotly to show the average daily rainfall in 2024 for each country.

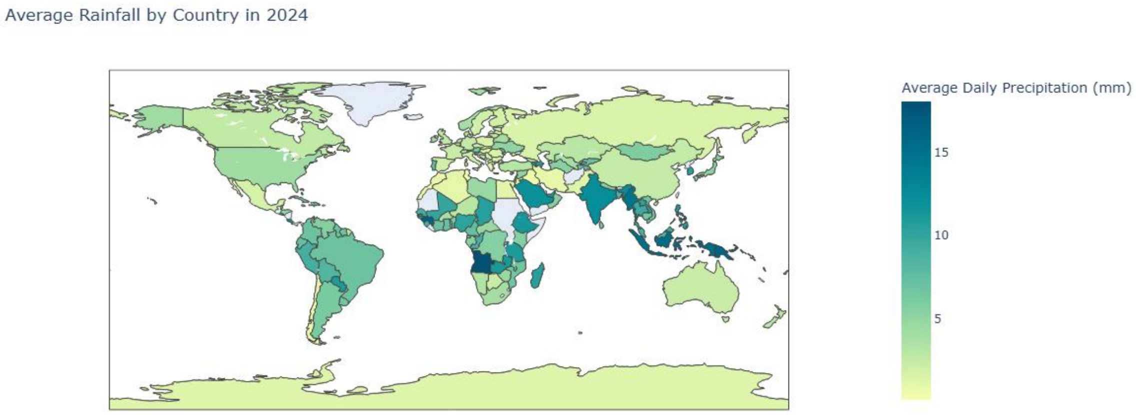

Average Rainfall by Country in 2024
Average Daily Precipitation (mm)
15
10
5

Figure 5: *A choropleth colouring a map based on the average daily rainfall in 2024.*

The territories of other nations were removed from the dataset, as they caused issues with how plotly interpreted the country of each row. The dataset was filtered to only include 2024 values and plotly was used to create the choropleth map. There are a few countries missing from the dataset in central Africa. Greenland was also missing which makes sense as it is not considered a country. A yellow to blue colour scale was used to associate more blue countries with higher rainfall. A blue scale alone made the countries without data indistinguishable from those with zero rainfall, so yellow was introduced at the lower end of the scale. Angola appears to have the greatest rainfall in this map, which does not make much sense, as a large proportion of the country has an arid climate. This may also be due to a reporting issue.

5. CONCLUSIONS

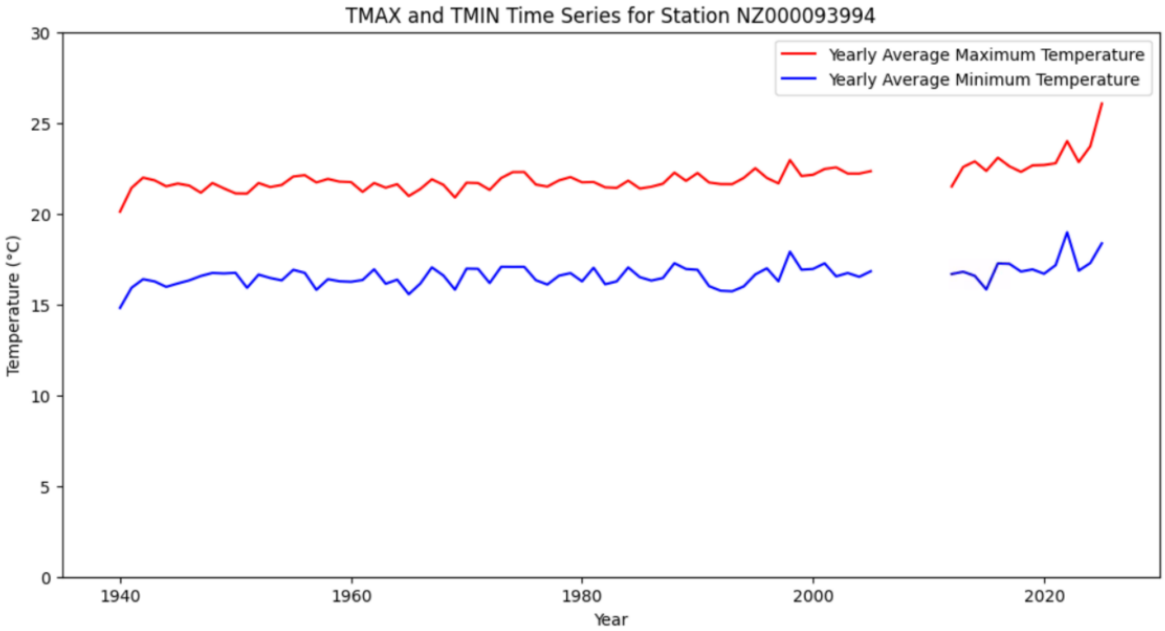
The GHCN dataset was successfully explored, providing practice in accessing distributed data and the associated meta-data tables from cloud storage. The dataset was processed in an efficient manner, using techniques that can be scaled to large datasets in industry. The analysis was performed in a similar manner, with care taken to be efficient with use of resources. Visualisations were generated that were successful and informative, giving the data real-world context while practising plotting techniques and the effective preparation of data for graphing.

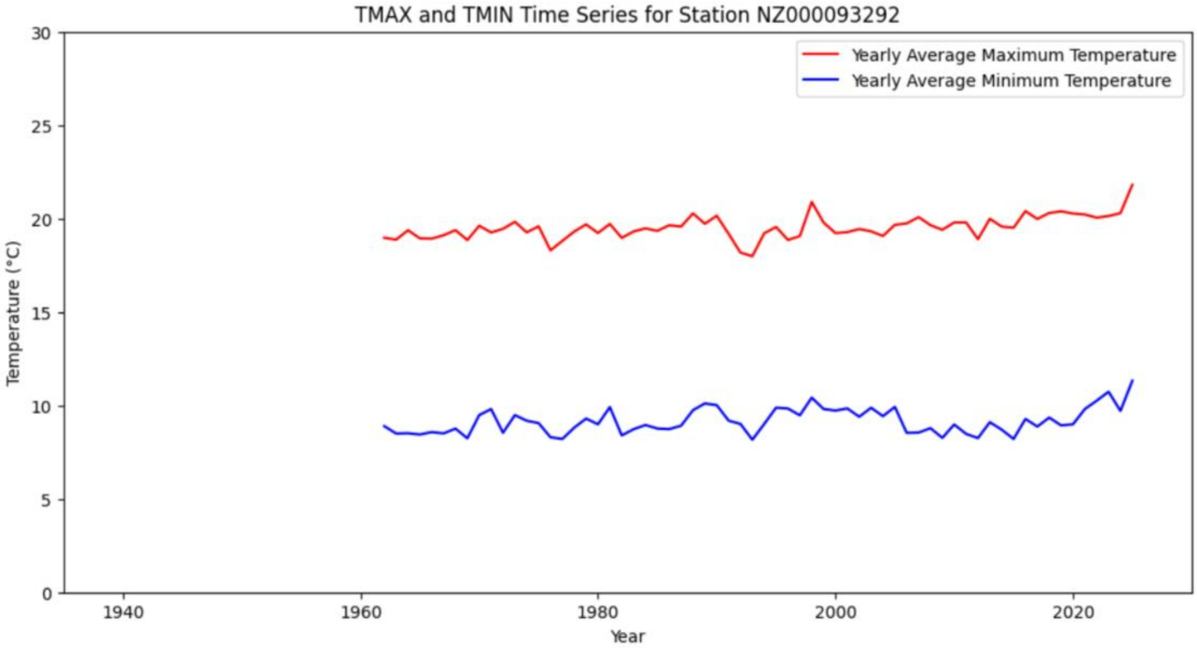
6. REFERENCES

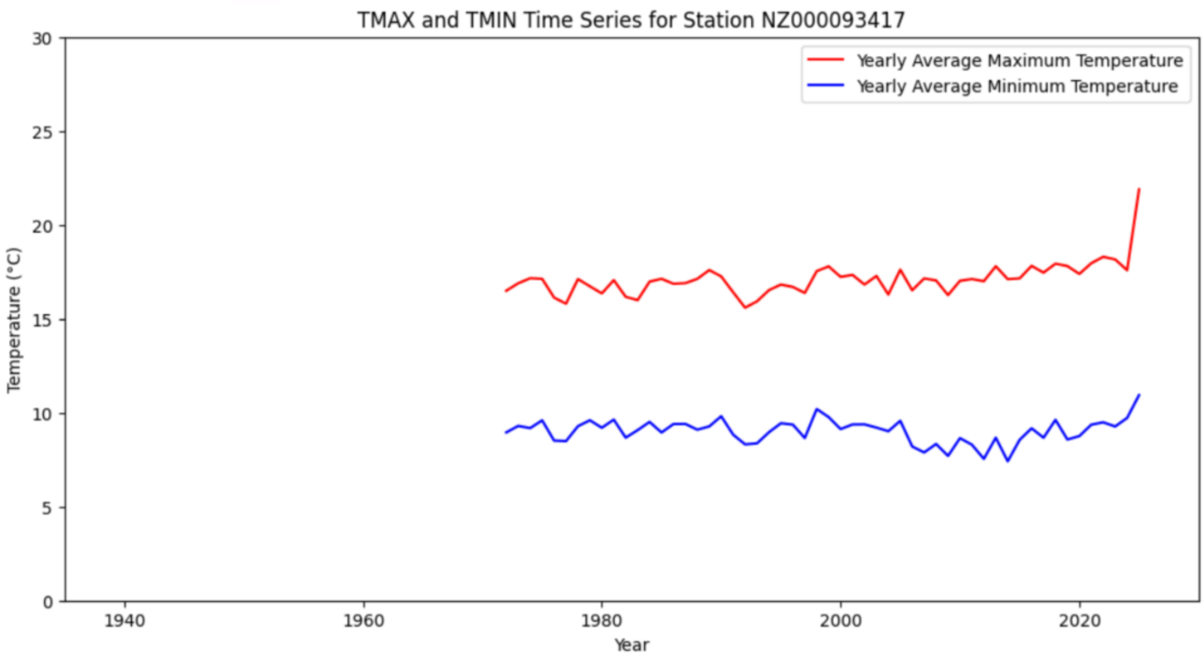
Don Josef de Mendoza y Rios, F.R.S. Recherches sur les principaux Problemes de l'Astronomie Nautique, Proc. Royal Soc., Dec 22, 1796.

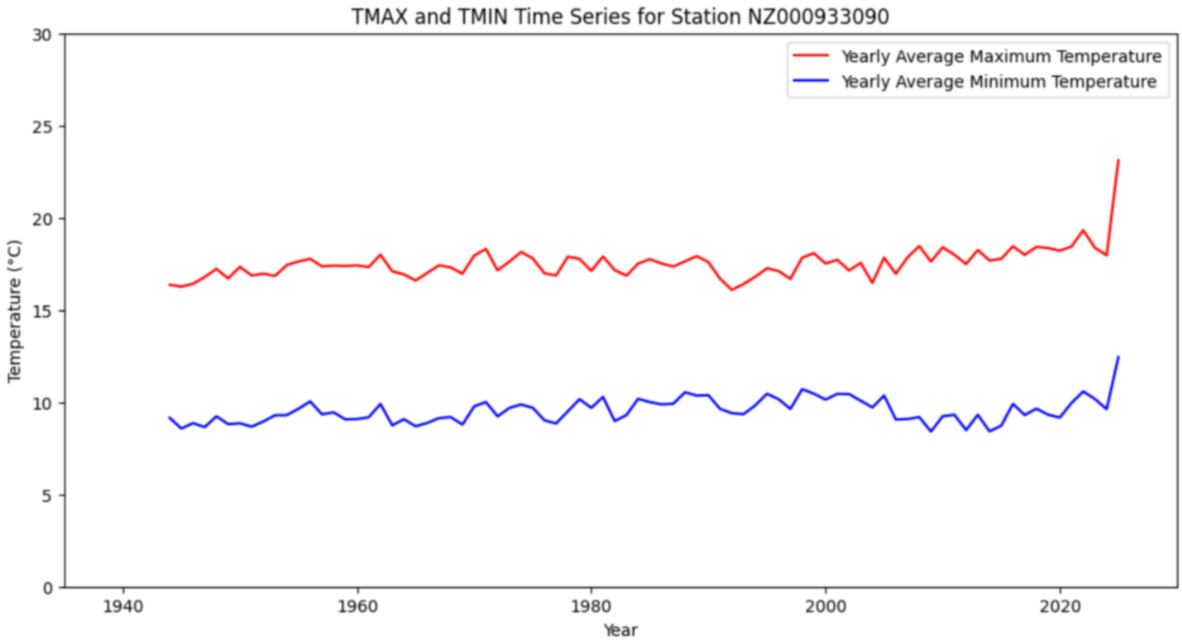
Data obtained from the Global Historical Climatology Network, 2025.

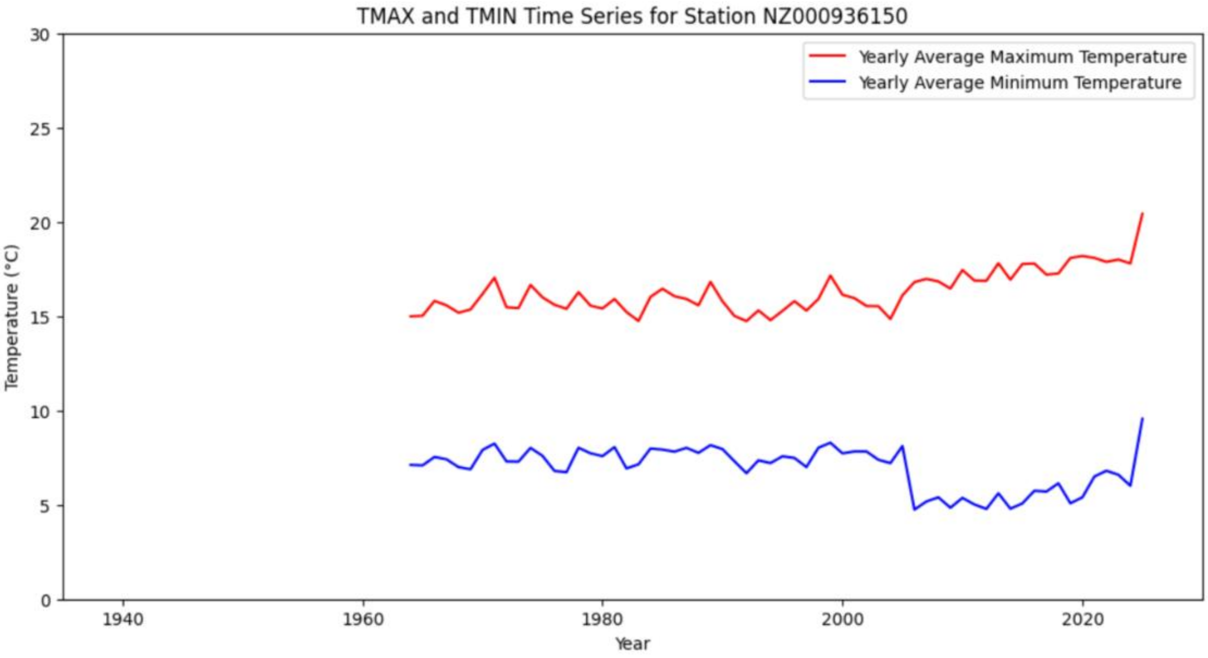
APPENDIX A. SUPPLEMENTARY VISUALISATIONS

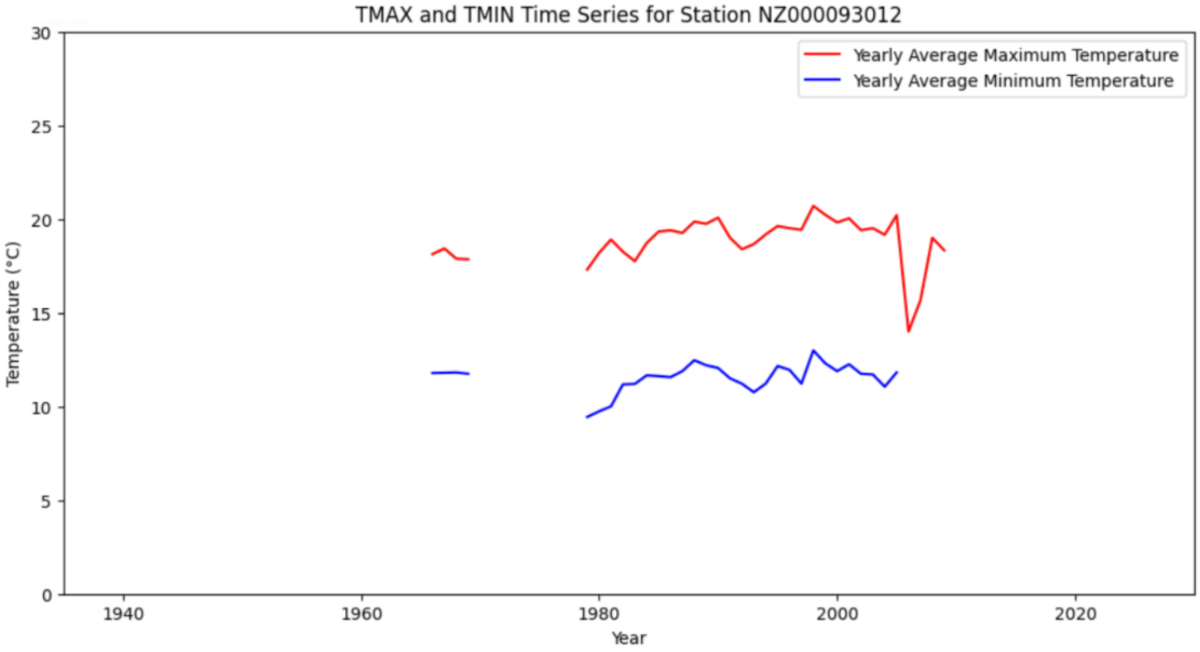


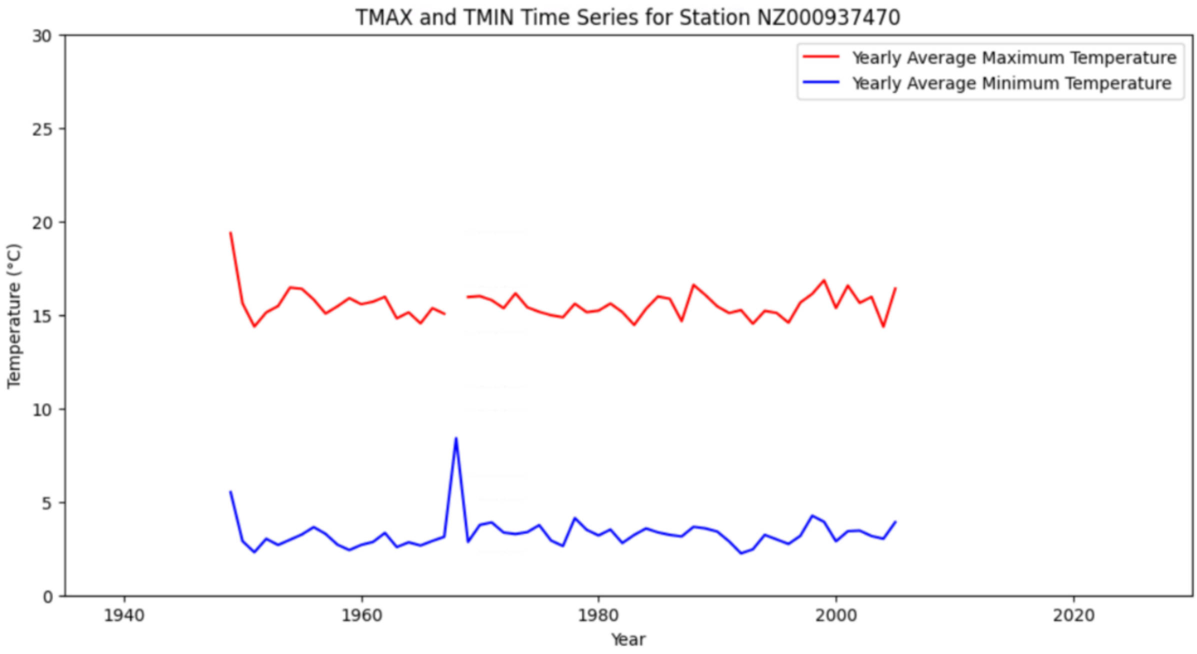


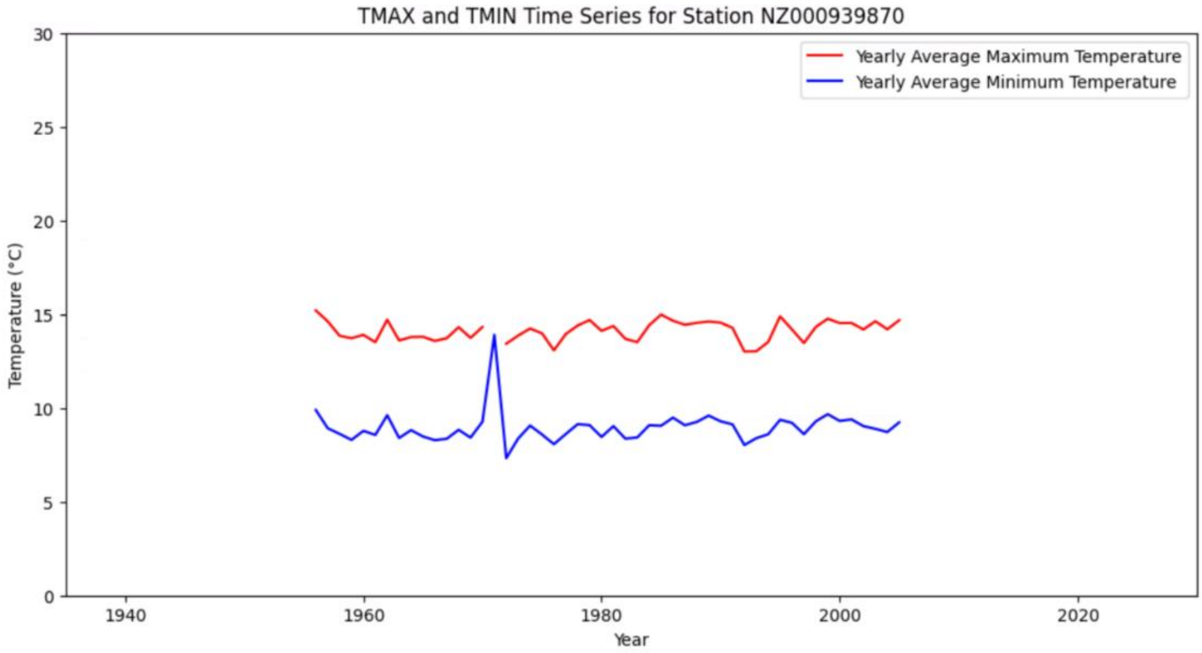

TMAX and TMIN Time Series for Station NZ000093417
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
0
5
10
15
20
25
30
1940
1960
1980
2000
2020
Year

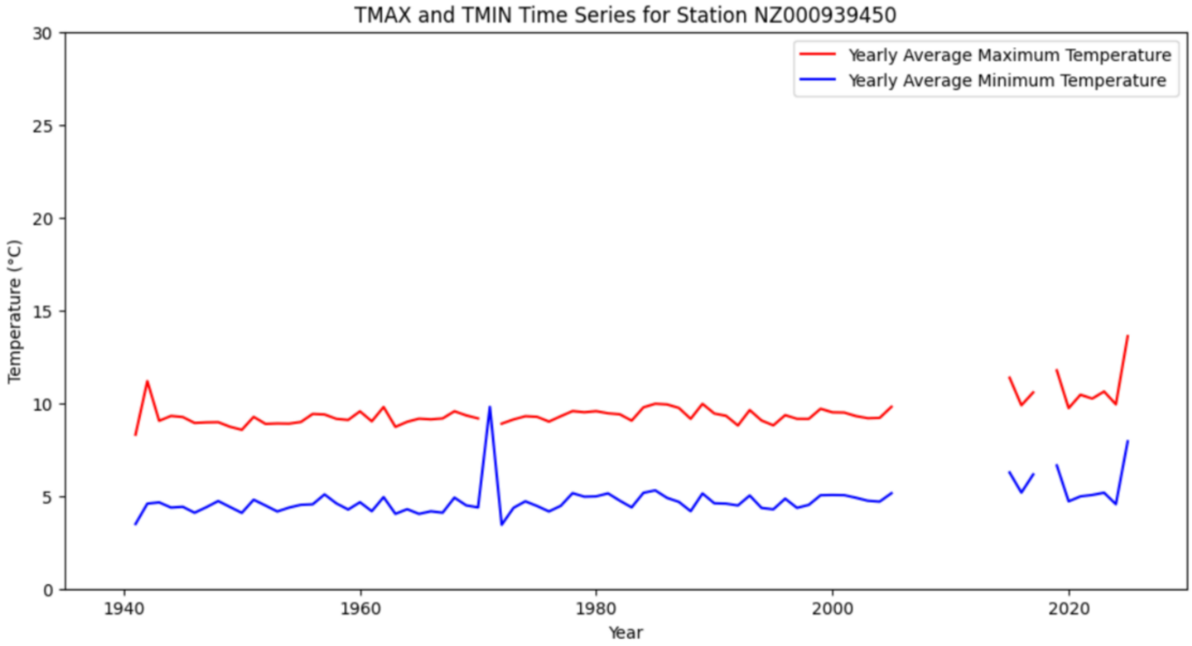

TMAX and TMIN Time Series for Station NZ000933090
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
0
5
10
15
20
25
30
1940
1960
1980
2000
2020
Year

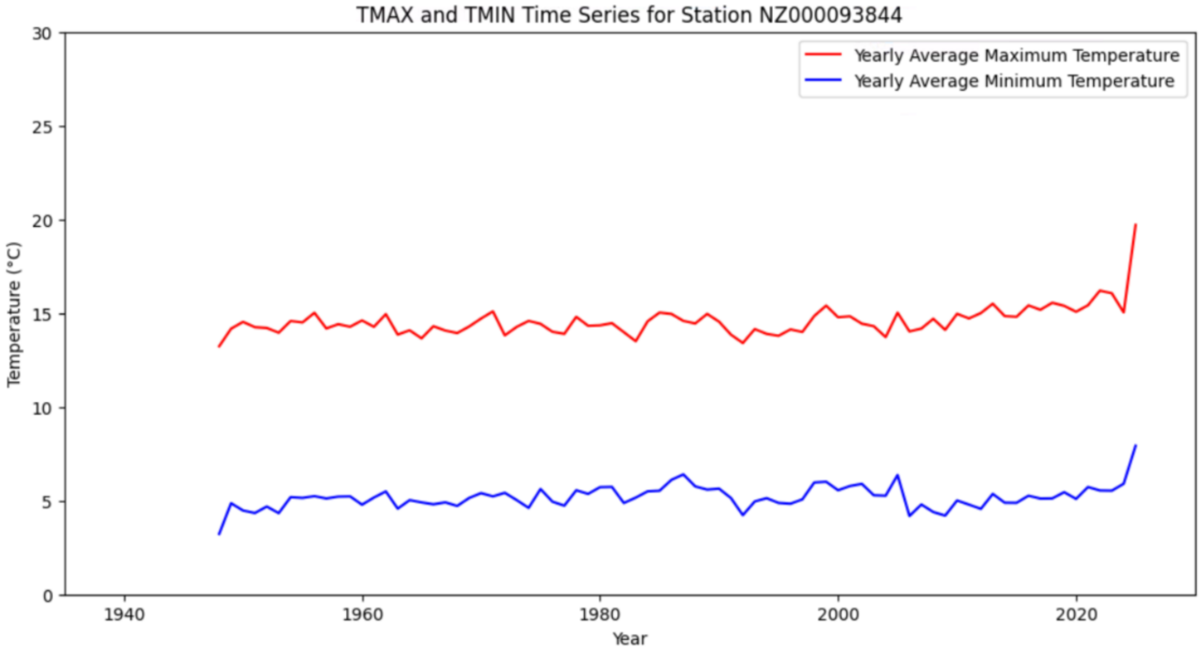

TMAX and TMIN Time Series for Station NZ000936150
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
0
5
10
15
20
25
30
1940
1960
1980
2000
2020
Year

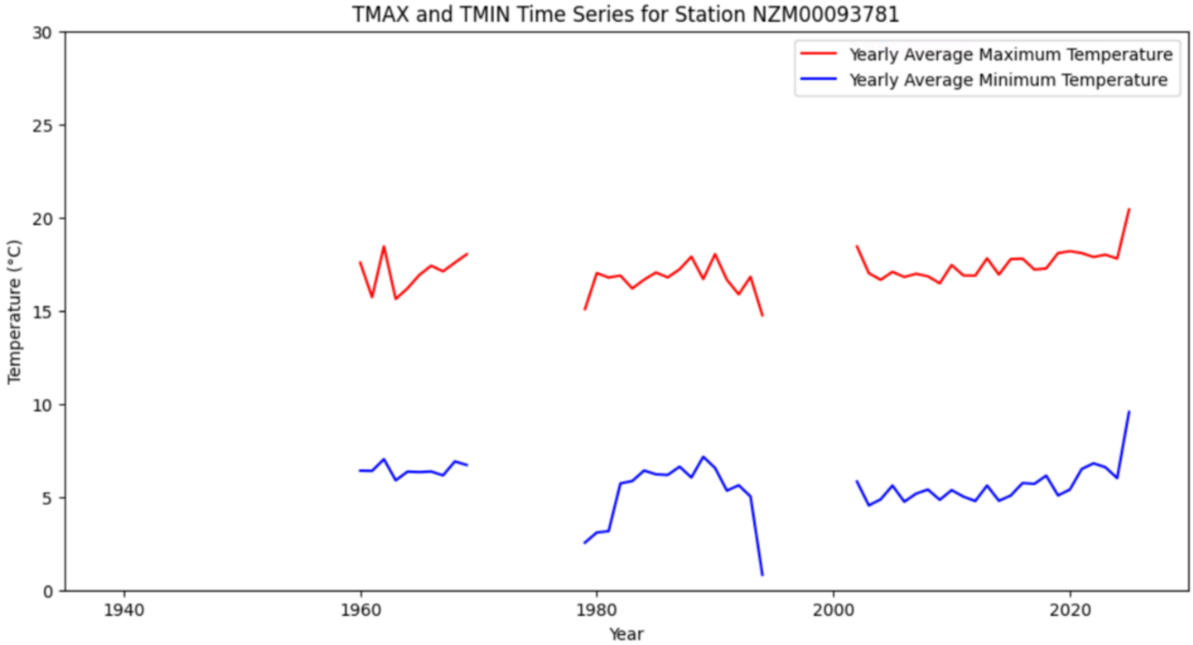

TMAX and TMIN Time Series for Station NZ000093012
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
0
5
10
15
20
25
30
1940
1960
1980
2000
2020
Year

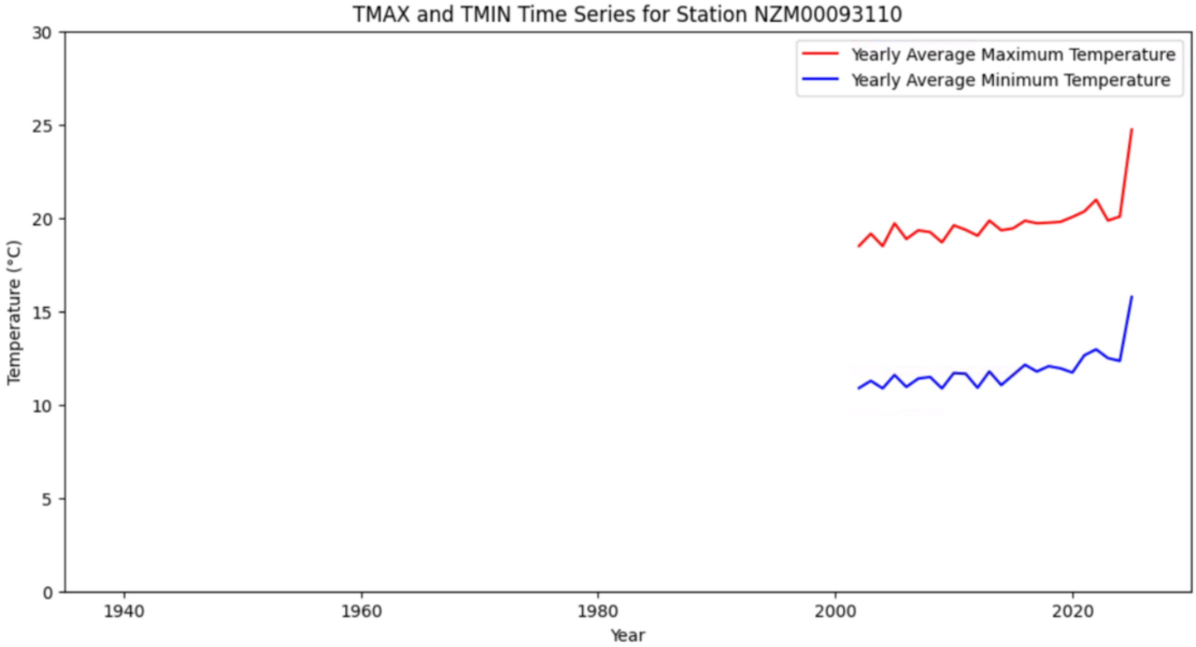

TMAX and TMIN Time Series for Station NZ000937470
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Year

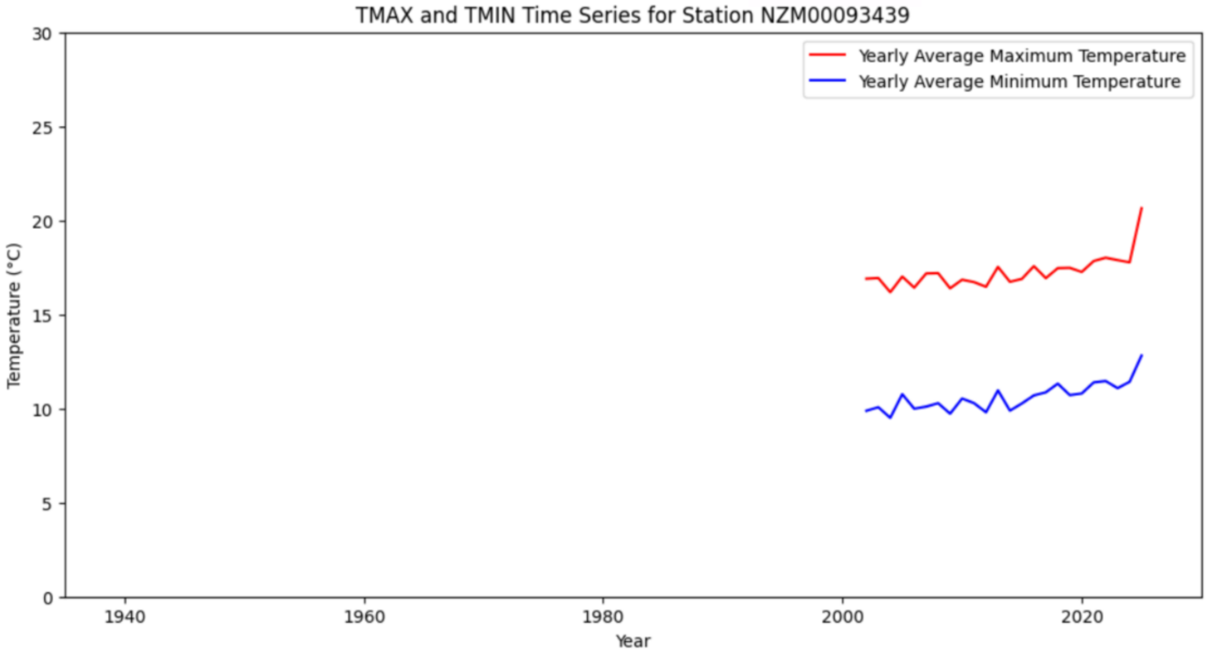

TMAX and TMIN Time Series for Station NZ000939870
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Year

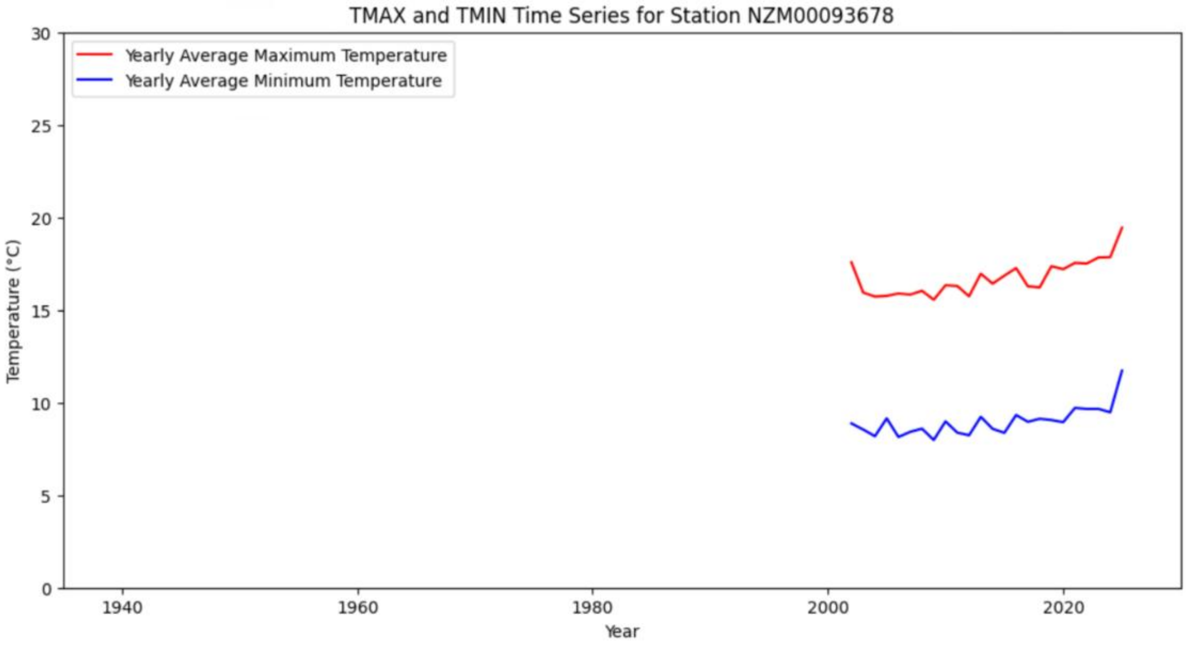

TMAX and TMIN Time Series for Station NZ000939450
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Year

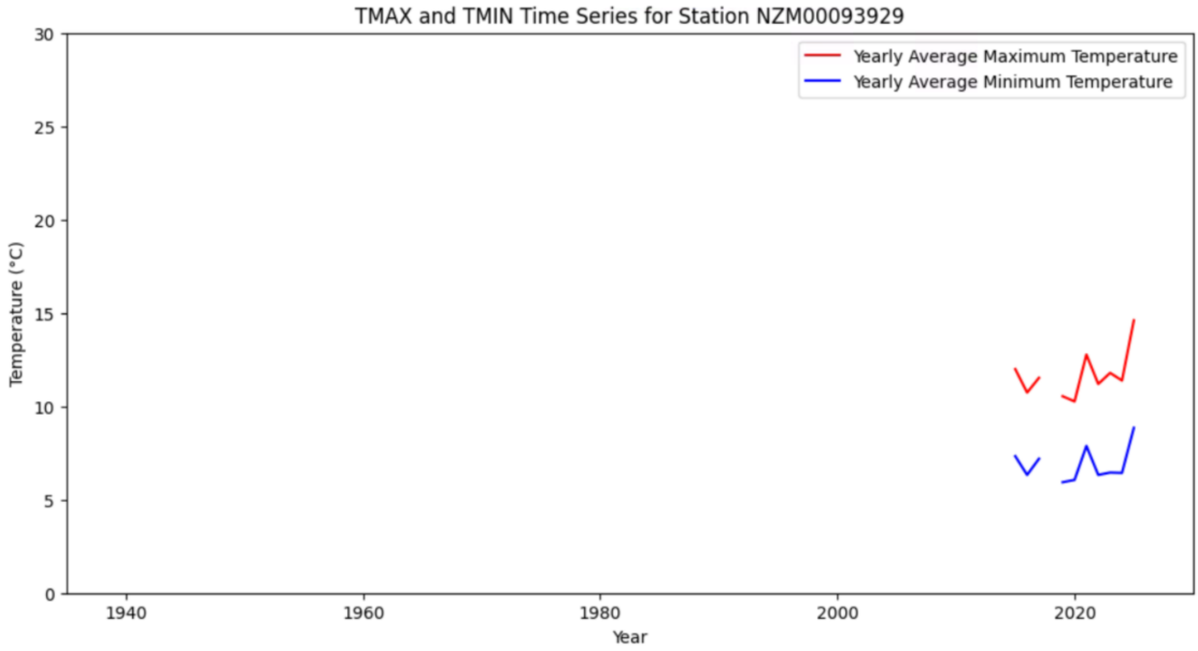

TMAX and TMIN Time Series for Station NZ000093844
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Year

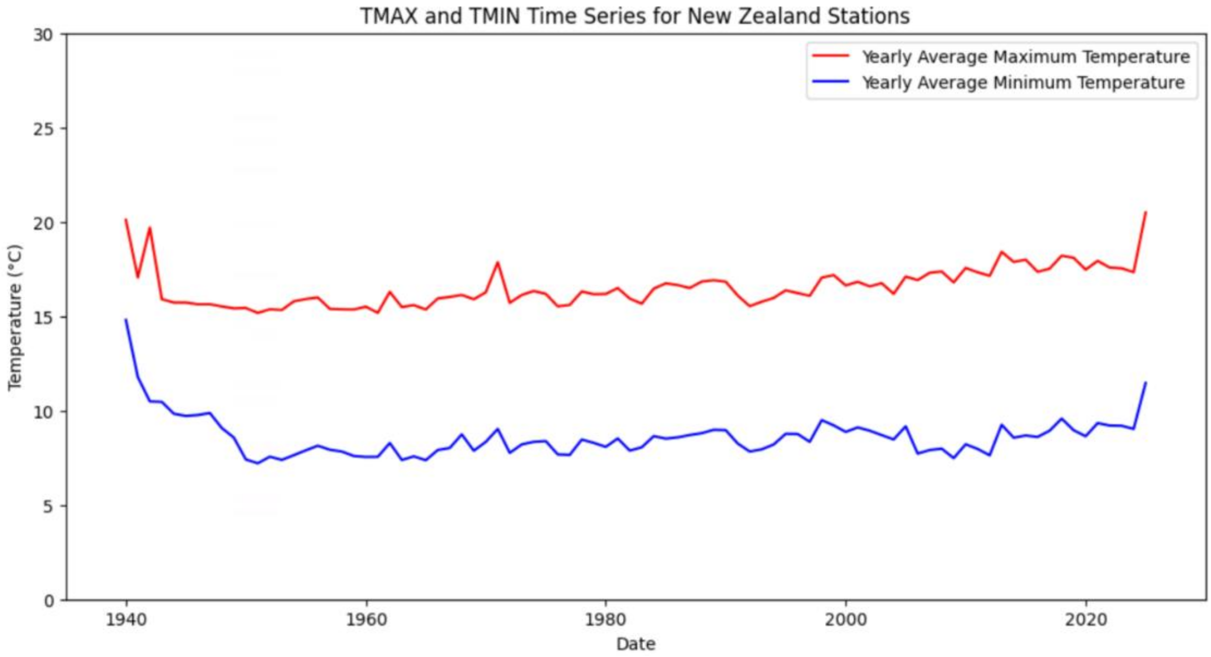

TMAX and TMIN Time Series for Station NZM00093781
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
0
5
10
15
20
25
30
1940
1960
1980
2000
2020
Year


TMAX and TMIN Time Series for Station NZM00093110
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
0
5
10
15
20
25
30
1940
1960
1980
2000
2020
Year


TMAX and TMIN Time Series for Station NZM00093439
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Year


TMAX and TMIN Time Series for Station NZM00093678
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Year


TMAX and TMIN Time Series for Station NZM00093929
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Year


TMAX and TMIN Time Series for New Zealand Stations
Yearly Average Maximum Temperature
Yearly Average Minimum Temperature
Temperature (°C)
30
25
20
15
10
5
0
1940
1960
1980
2000
2020
Date