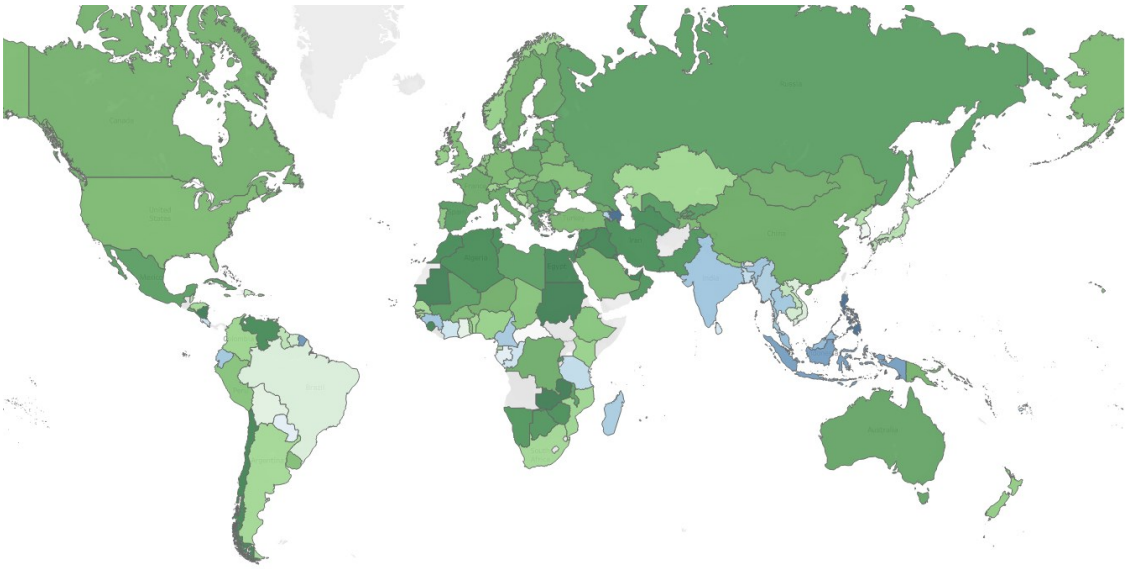
DATA420-24S1
Scalable Data

Assignment 1
GHCN Data Analysis using Spark



Student Name: Wijesingha Mayadunnage Uditha Iresh Mayadunna

Student ID: 39768450

1 Contents

1 Contents 2
2 Table of Figures 3
3 Background 5
4 Processing 6
4.1 Question 1 6
4.1.1 Data Structure 6
4.2 Question 2 7
4.2.1 Define a schema for daily. 7
4.2.2 Loading other metadata files 8
4.3 Question 3 9
4.3.1 Inventory Analysis 9
4.3.2 Merging a subset of daily and enriched stations table 9
5 Analysis 10
5.1 Question 1 10
5.1.1 Analysing Stations Data Table 10
5.2 Question 2 11
5.2.1 Geographical distance between two stations 11
5.2.2 New Zealand stations and distances 11
5.3 Question 3 12
5.3.1 Block Sizes 12
5.3.2 Number of observations 12
5.4 Question 04 13
5.4.1 Daily Climate Summaries 13

5.4.2 TMIN & TMAX in New Zealand Stations 14
5.4.3 Precipitation observations 15
6 Conclusion 17
7 References 18
8 Appendix 19
8.1 Appendix 1 - Time series for TMIN and TMAX for station in New Zealand 19
8.2 Appendix 2 - Average time series for TMIN and TMAX for New Zealand 26
8.3 Appendix 3 - Average Rainfall in 2023 for each country 28
8.3.1 Asia 28
8.3.2 Africa 29
8.3.3 Europe 30
8.3.4 Australia and New Zealand 30
8.3.5 North America 31
8.3.6 South America 32
8.3.7 Antarctica 32

2 Table of Figures

Figure 1: The Data Structure 6
Figure 2: Daily file size over time 7
Figure 3 : Time Series Plot for Station NZ000093994 14
Figure 4 : Average Daily Temperature in New Zealand 15
Figure 5: Average Rainfall by Country in 2023 World Map 16
Figure 6 : Time Series for Station Kaitaia 19
Figure 7: Time Series for Station Gisborne Aerodrome 19
Figure 8:Time Series for Station Paraparaumu AWS 20
Figure 9: Time Series for Station Invercargill Airpor 20
Figure 10:Time Series for Station Raoul Isl/Kermadec 21
Figure 11: Time Series for Station New Plymouth AWS 21
Figure 12: Time Series for Station Hokitika Aerodrome 22
Figure 13: Time Series for Station Tara Hills 22
Figure 14: Time Series for Station Campbell Island AWS 23

Figure 15: Time Series for Station Chatham Island AWS 23
Figure 16: Time Series for Station Auckland Aero AWS 24
Figure 17: Time Series for Station Wellington Aero AWS 24
Figure 18: Time Series for Station Kaikoura 25
Figure 19: Time Series for Station Christchurch Intl 25
Figure 20: Time Series for Station Enderby Island AWS 26
Figure 21: Monthly Average time series for Maximum and Minimum temperature in New Zealand 26
Figure 22 : Quarterly Average time series for Maximum and Minimum temperature in New Zealand 27
Figure 23 : Yearly Average time series for Maximum and Minimum temperature in New Zealand 27
Figure 24 : 2023 Average Rainfall in Asia 28
Figure 25: 2023 Average Rainfall in Africa 29
Figure 26: 2023 Average Rainfall in Europe 30
Figure 27: Average Rainfall in Australia and New Zealand 30
Figure 28: 2023 Average Rainfall in North America 31
Figure 29: 2023 Average Rainfall in South America 32
Figure 30: 2023 Average Rainfall in Antarctica 32

Table 1 : GHCN Sub Folder Size 6
Table 2 : Schema for Daily 8
Table 3 : Number of Rows in Meta Data Files 8
Table 4 : Core Elements 9
Table 5 : Stations belong to Flags 10
Table 6 : US territories excluding US 11
Table 7 : Block Size 12
Table 8 : Core Element Observation Count 13

3 Background

This is the 1st Assignment for the course DATA420 Scalable data and in this assignment, weather data from the Global Historical Climate Network (GHCN), a vast collection of climate summaries from land-based stations throughout the world, were examined. The following links provided a summary of the data which were used, including all the necessary information.

Link1: Global Historical Climatology Network (GHCN) (National Oceanic and Atmospheric Administration, n.d.) link contains all important information for the Meta Data tables Stations, Inventory, Country, and States.

Link2: GHCN Daily (Menne, 2012) link provides all relevant information about Daily Climate Summary files.

This assignment comprised three main parts. The first part involved exploring the metadata tables and performing preprocessing to create an enriched metadata table using all available tables. This step helped gain an overview of all weather stations and understand how the various tables were interconnected. Also, in this part only 1,000 rows from the daily climate were loaded and joined it with the metadata tables to get a preliminary understanding of the data structure and relationships.

In the second step, the enriched metadata table created in the first phase was used to gain insights into weather stations across the globe. This phase explored the total number of stations, including active and inactive ones, the number of stations per country, and other notable facts. Given the relatively small size of the enriched metadata table, counting, and analysing the data in this step was not time or resource intensive. This allowed for efficient exploration and understanding of the distribution and status of weather stations worldwide.

The final phase of the assignment involved analysing the daily data and visualizing key findings. Given that the daily data comprised approximately 250 compressed CSV files, loading and analysing them required significant time and computational resources. All daily data from the beginning was integrated with the enriched stations table, examined in accordance with the assignment's questions, and the results were recorded.

All the preprocessing and analysis were conducted using the Apache Spark distributed computing framework, where the allocation of resources, creation of partitions, and task assignments for each stage were also explored. For the majority of the visualizations, Tableau software was utilized. Additionally, Python coding, grammar corrections, and other information were supported by ChatGPT (OpenAI, 2024), Stack Overflow (Stack Exchange Inc;, 2024), and Quill Bot (QuillBot, 2024).

4 Processing

4.1 Question 1

4.1.1 Data Structure

The GHCN data was organized into one folder for daily data and four .txt files for country, inventory, state, and station data. Within the daily data folder, 263 csv files were contained, providing daily data for each year from 1750 to 2024. Files from 1751 to 1762 were missing, however, csv.gz files for each other year were present. The directory structure is presented as follows.

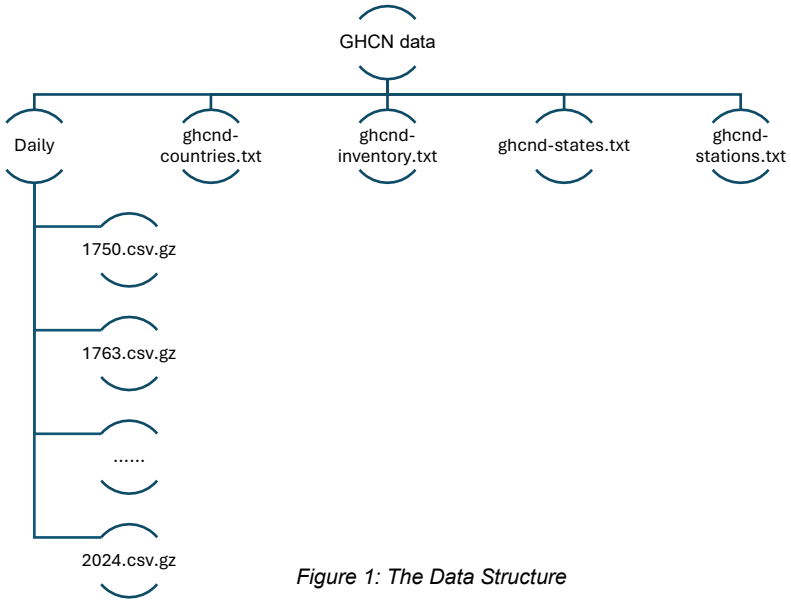


Figure 1: The Data Structure

The actual size of the GHCN dataset was approximated at 100 GB, while the space it occupied on the HDFS was roughly 12.5 GB. Table 1 shows the sizes of the original files and their compressed equivalents in the HDFS environment.

GHCN data		
Sub Folder/File	Actual Size	Compressed Size
Daily Folder	98.9 GB	12.4 GB
Countries.txt	28.6KB	3.6 KB
Inventory.txt	262.3 MB	32.8 MB
States.txt	8.5 KB	1.1 KB
Stations.txt	82.7 MB	10.3 MB
Total	**99.237 GB**	**12.442 GB**

Table 1 : GHCN Sub Folder Size

As previously noted, the daily folder was found to contain 263 csv files. The first csv file was reported to contain daily data for the year 1750, followed by daily data for the year 1763. Since 1763, each year was represented by a distinct csv file containing daily data up until 2024.

The file size of the 1750 csv file was recorded at 3.1 MB, which was 100 times greater than the subsequent csv file in 1763. This could be explained by the presence of more data prior to 1750 or the inclusion of data for the missing years 1751-1762. This anomaly was noted as abnormal and should be investigated further.

In 1763, the file size was 26.2 KB, but it grew rapidly year after year, and by 2023, it was 1.2 GB. The file size for 2024 is still 187.3 MB, as the data is only saved until March 2024. As the original file size expands, the storage used on the HDFS increases correspondingly. Figure 2 illustrates the progression of the daily file size over time.

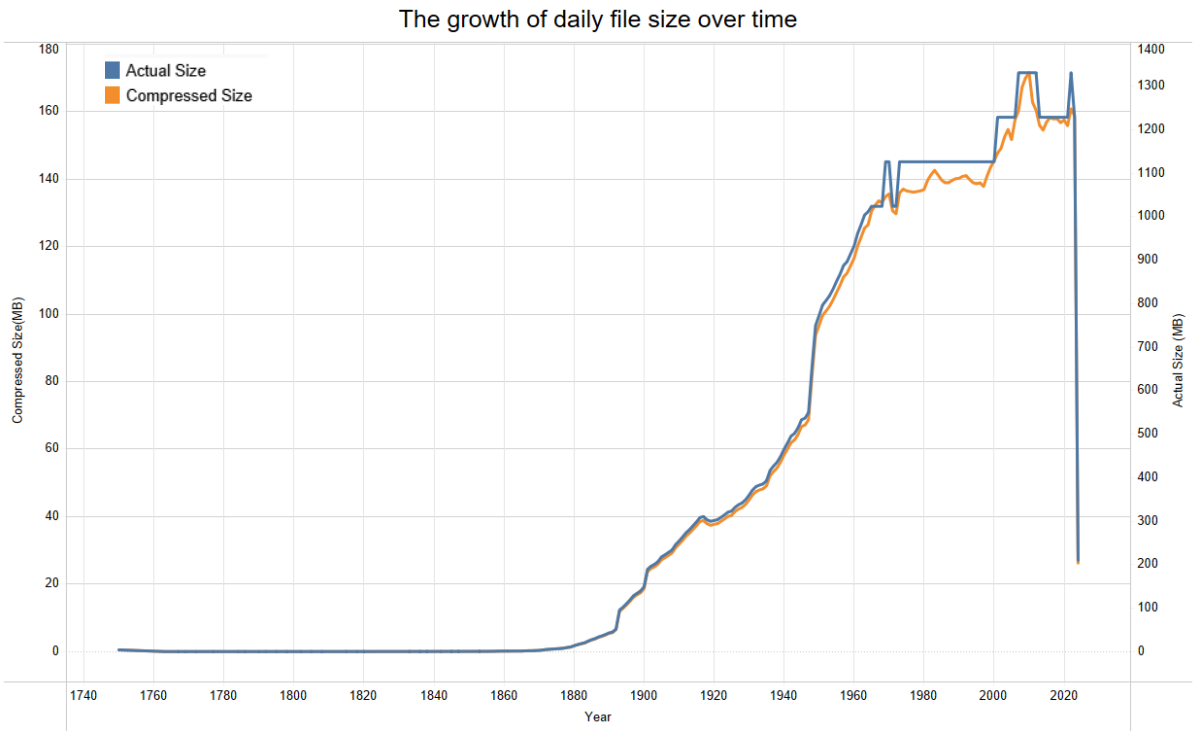


Figure 2: Daily file size over time

4.2 Question 2

4.2.1 Define a schema for daily.

Following a review of the data, a schema was created in order to load a thousand rows from the daily data file for 2023. The daily climate summary files included eight columns, and the following data types were utilised to define the structure.

COLUMN NAME	DATA TYPE
ID	String
DATE	String
ELEMENT	String
VALUE	Float
M_FLAG	String
Q_FLAG	String
S_FLAG	String
OBS_TIME	String

Table 2 : Schema for Daily

The readme file for the daily data states that the OBS_TIME data type should be Time and the Date data type should be Date. However, the DATE and OBS_TIME fields were filled with null values when such data types were applied. So, tried using the string data type, and DATE and OBS_TIME in string format loaded. Hence, these columns' data types should be changed before conducting additional analysis.

4.2.2 Loading other metadata files

Although the daily climate summary files were in csv.gz format, all other metadata files, including Stations, Countries, States, and Inventory files, were in .txt format. After these .txt files were imported into Spark, the pyspark.sql.functions.substring function was used to separate the columns of these specific data frames in the Spark environment. Upon separation, the data types of all columns in each data frame were found to be String, and the data types of some attributes needed to be changed before proceeding with the data analysis. The count of rows in each metadata table was then computed, and the results are displayed below.

METADATA TABLE	NUMBER OF ROWS
STATIONS	125,983
COUNTRIES	219
STATES	74
INVENTORY	747,382

Table 3 : Number of Rows in Meta Data Files

According to table 3, the stations metadata file contains data on 125,983 stations in 219 countries. The states table only contains the names of 74 separate states, however the inventory data frame comprises 747,382 rows.

The WMO ID is a number assigned by the World Meteorological Organisation (WMO, 2020) to stations and in this station data frame, only 7960 stations have a WMO ID, whereas the remaining 118023 stations do not.

4.3 Question 3

The left join function (Apache Spark, n.d.) in Spark was used to merge the data frames for stations, countries, and states. This method ensured the inclusion of all entries from the 'stations' data frame, along with corresponding entries from the 'countries' and 'states' data frames where applicable.

4.3.1 Inventory Analysis

According to the readme file for daily global historical climatology network, there are five core elements in the inventory and the descriptions for those core elements are as follows.

ELEMENT	DESCRIPTION
PRCP	Precipitation (tenths of mm)
SNOW	Snowfall (mm)
SNWD	Snow depth (mm)
TMAX	Maximum temperature (tenths of degrees C)
TMIN	Minimum temperature (tenths of degrees C)

Table 4 : Core Elements

According to the Inventory data frame, 20,467 stations gather all five core elements, whereas 35,662 stations collect only precipitation data. Also, when analysing Precipitation, Maximum Temperature and Minimum Temperature, the values should be divided by 10 to get the exact value.

All inventory data were grouped by station and merged with station data to create the enriched stations table.

4.3.2 Merging a subset of daily and enriched stations table

The 1000-row subset of daily was left joined to the enriched stations table, and it was discovered that there are 7 stations that are only in the daily subset and not in the stations data table.

In Spark, joining large datasets such as "daily" and "stations" using a LEFT JOIN has been found to be computationally expensive. This cost primarily arises from the shuffle operation that Spark performs in order to distribute the data across different nodes, aligning matching keys (e.g., station IDs). The volume of data that needs to be shuffled, can significantly affect both performance and resource usage.

An ANTI-JOIN can be used to assess whether there are stations in "daily" that are not in "stations" without the need for a LEFT JOIN. An anti-join returns rows from the first data frame that do not match any rows in the second data frame.

5 Analysis

5.1 Question 1

5.1.1 Analysing Stations Data Table

According to an analysis of the stations data table, there are 125,983 unique stations in total and only 31,837 of these stations are operational as of 2024 out of all of them.

The stations data table has three types of flags, and the number of stations in each flag is as follows.

FLAG TYPE	DESCRIPTION	STATIONS COUNT
GSN	Station is part of the GCOS Surface Network	991
HCN	Station is part of the U.S. Historical Climatology Network	1,218
CRN	Station is part of the U.S. Climate Reference Network	234

Table 5 : Stations belong to Flags

Also, there are 15 stations which are in more than one of these networks.

The "Latitude" can be used to filter out any stations that are in the Southern Hemisphere, and any station with a Latitude < 0 is included in this category. After filtering, it is discovered that only 25,357 stations are in the Southern Hemisphere, while the remaining 100,626 are in the Northern Hemisphere. There are no stations right at the equator.

After counting the total number of stations belonging to each country, it is discovered that the United States has the most stations, 72,289. Texas has the most stations in the United States, with 6154 total.

Although the United States has 72,289 stations, some of them are not located within the country itself. In all, there are 386 stations in US territory across the world, excluding the United States itself. The country code was utilised to filter out those stations, because their country codes are different from "US". A summary of these stations are as follows.

TERRITORY NAME	NUMBER OF STATIONS
PUERTO RICO	243
VIRGIN ISLANDS	71
GUAM	29
AMERICAN SAMOA	21

NORTHERN MARIANA ISLAND	11
JOHNSTON ATOLL	4
MIDWAY ISLANDS	3
PALMYRA ATOLL	3
WAKE ISLAND	1
TOTAL	**386**

Table 6 : US territories excluding US

5.2 Question 2

5.2.1 Geographical distance between two stations

When determining the geographical distance between two stations, a user-defined function was developed using the Haversine formula.

5.2.1.1 Haversine Formula

The Haversine formula is one method for calculating the distance between two points. It determines the distance between two latitude and longitude coordinates by measuring the length of a straight line connecting them. When computing a large circle distance between two points on Earth, the Haversine formula is frequently used in navigation, without considering the height of hills or depth of valleys (Rezania Agramanisti Azdy, 2019).

R = Radius of the earth = 6,371 km

Δ lat = latitude 1 - latitude 2

Δ long = longitude1 - longitude2

$$\text{Distance} = 2\text{R} * \arcsin(\sqrt{sin^2\left(\Delta\frac{\text{lat}}{2}\right) + \cos(latitude2) * \cos(latitude1) * sin^2\left(\Delta\frac{\text{long}}{2}\right)}\ )$$

By substituting the latitude and longitude values, the distance between two points can be calculated.

5.2.2 New Zealand stations and distances

Using the country name as a filter, the stations data frame was first filtered to determine the distance between stations in New Zealand. New Zealand belongs 18 stations in total, three of which are not in New Zealand but are situated in Tokelau, the Cook Islands, and Niue.

After filtering the data, the user-defined function to compute geographical distances was used to determine the pairwise distances between all stations in New Zealand, which revealed that WELLINGTON AERO AWS and PARAPARAUMU AWS are the nearest stations inside New Zealand, with a distance of 50.53 km.

5.3 Question 3

5.3.1 Block Sizes

A HDFS block's default size is 134,217,728 bytes, or 128 megabytes. The compressed file size of the daily climate summary 2024 is 27,492,832 bytes (26.2 MB), which is smaller than the size of one HDFS block. So, 1 HDFS block is adequate for the daily climate summary in 2024.

The file size of the daily climate summary 2023 is 166,367,488 bytes, which exceeds the HDFS block capacity. So, the daily climate summary 2023 will require two blocks, with the first block being 134,217,728 bytes (128 MB) and the second block with the size of 32,149,760bytes (30.99MB).

File Name	**Block Size**	
	Block 1	Block 2
Daily climate summary 2024	27,492,832 bytes (26.2 MB)	-
Daily climate summary 2023	134,217,728 bytes (128 MB)	32,149,760 bytes(30.99MB)

Table 7 : Block Size

5.3.2 Number of observations

The number of observations for daily climate summary in 2023 is 37,395,852 and in 2024 it is 6,061,827. When counting observations in 2023 and 2024, each job had two stages, each with a single task. Also, each year's files had one partition. This implies that, although having two blocks for the Daily climate summary 2023, there was only one task, and the Daily climate summary 2024 which has one block also has one task in each stage.

From the year 2014 to 2023, there are total 10 files and 369,419,055 observations in total. The number of partitions for these input data is 10 which is equal to the number of distinct files. So, in each stage there could be 10 tasks in total.

When Spark reads a compressed input file, the way it partitions the data depends on the type of compression used. For compression formats like bzip2 that allow partitioning, Spark can create a task for each compressed block, allowing for parallel processing. This means that even if the data is compressed, Spark can use all available resources in the cluster to read and process the data simultaneously.

On the other hand, for compression formats that don't allow partitioning, like gzip, Spark treats the entire compressed file as a single block, resulting in a single task processing it. This limits parallelism because even if your cluster has significant resources, Spark can only use one task to process the entire compressed file.

Thus, the choice of compression format can greatly affect the performance of the Spark application. If the priority is to maximize parallelism, a partitionable compression format might be the best option. But the focus is on conserving storage space or reducing network bandwidth, a non-partitionable compression format could be more beneficial.

The daily climate summaries for 2023 and 2024 should ideally require 2 and 1 HDFS blocks, suggesting that Spark would execute 2 and 1 tasks when reading these files. However, since the files are compressed in .gz format, each file is treated as a single block. This results in Spark processing each file with just one task, regardless of its original block count.

To boost parallelism when processing large datasets, we can increase the number of partitions at each stage. This leads to more tasks running simultaneously, enhancing processing speed. By explicitly setting a higher number of partitions when reading data, Spark creates additional blocks, allowing more tasks to run in parallel. Using partitionable compression formats, like bzip2, supports this approach, as Spark can assign a task to each block within these compressed files, maximizing parallelism. Adjusting Spark's configuration can also increase parallelism. This can be done by increasing the number of tasks per core or adding more cores to the Spark application, allowing more tasks to run concurrently.

5.4 Question 04

5.4.1 Daily Climate Summaries

All daily climate summaries from 1750 to 2024 contain 3,103,954,141 rows. This indicates there are more than three billion observations made over those time periods.

As previously stated, the stations gathered five core elements, and the following table shows the total number of observations for each core element.

CORE ELEMENT	NUMBER OF OBSERVATIONS
PRCP (PRECIPITATION)	1,069,105,193
SNOW (SNOW FALL)	353,904,309
SNWD (SNOW DEPTH)	297,846,434
TMAX (MAXIMUM TEMPERATURE)	455,946,095
TMIN (MINIMUM EMPERATURE)	454,759,421

Table 8 : Core Element Observation Count

As the table above demonstrates, with 1,069,105,193 observations, the precipitation element has the most observations and, with 297,846,434 total observations, the snow depth element has the fewest observations overall.

When the maximum and lowest temperature elements are taken into account, the total number of observations differs. This means that there may be observations with simply TMAX or TMIN. There are 9,322,723 daily observations with just TMIN and no matching TMAX value, with 27,927 distinct stations contributing to these observations.

5.4.2 TMIN & TMAX in New Zealand Stations

After filtering the TMIN and TMAX data for 15 stations in New Zealand (excluding the New Zealand stations located outside), there are a total of 485,520 observations from 1940 to 2024, covering 85 years.

When the HDFS output file was moved to the local directory and decompressed, 84-part files were produced. These files collectively contain a total of 485,604 rows. Given that each file includes a header row, the actual number of data rows, after subtracting 84 for the headers, equals 485,520. This computed figure corresponds with the data row count previously retrieved by Spark.

5.4.2.1 Time Series Plots for TMIN & TMAX values in New Zealand Stations

Appendix 1 presents time series plots for TMIN and TMAX values across 15 stations in New Zealand. Among these, station NZ000093994 (RAOUL ISL/KERMADEC) stands out as the only station that has been active since 1940, with a brief period of inactivity between December 31, 2005, and August 17, 2022. Additionally, this station is the only one where all temperature readings were consistently above zero throughout the recorded period, with a minimum temperature of 7.40°C and a maximum of 31°C. The reason behind that would be the station is located on Raoul Island, situated 1,100 km northeast of New Zealand's North Island.

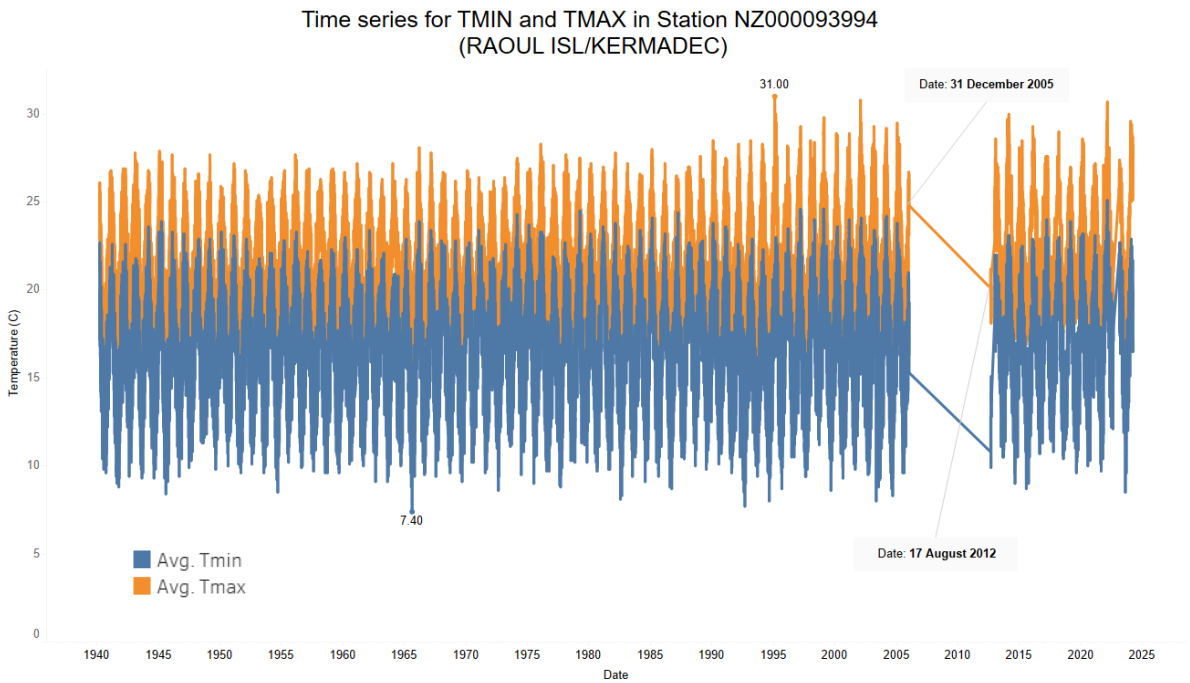


Figure 3 : Time Series Plot for Station NZ000093994

5.4.2.2 Average time series for TMIN and TMAX in New Zealand

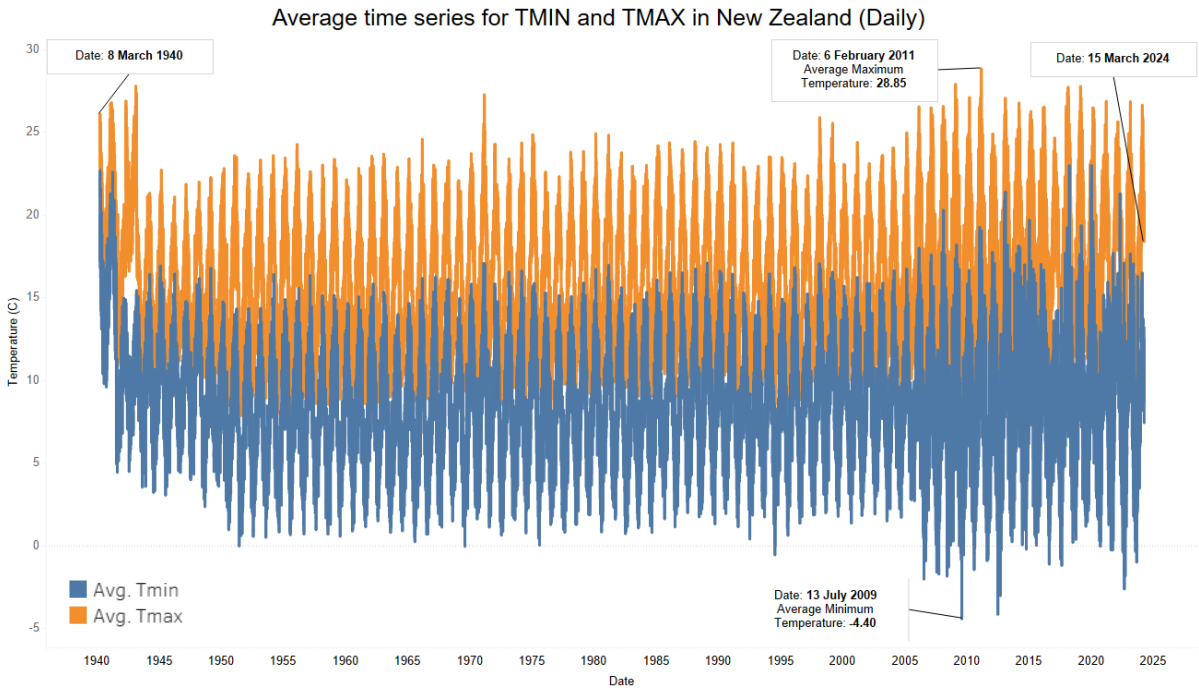


Figure 4 : Average Daily Temperature in New Zealand

Figure 3 depicts the average daily maximum and minimum temperatures in New Zealand from March 8, 1940, to March 15, 2024. During this time, the highest average temperature for the entire country was recorded on February 6, 2011, at 28.85°C. The lowest average temperature was recorded on July 13, 2009, at -4.40°C.

5.4.3 Precipitation observations

The annual average precipitation is determined by adding up all the daily precipitation values for a specific year. To find the average daily precipitation for a particular country, gather the total precipitation from all stations and then divide it by the total number of stations.

$$\text{Average Daily Precipitation for a Specific Country} = \frac{\text{Total Daily Precipitations in All the Stations in that country}}{\text{Number of unique stations in that country}}$$

$$\text{Average Annual Precipitation for a Specific Country} = \sum\left(\text{Average Daily Precipitation for a specific country}\right)$$

Upon analysing the average rainfall by country and year, it was found that New Caledonia had the highest average rainfall, reaching 17,674.4 mm in 1999. Additionally, many of the top average annual rainfalls were also from New Caledonia between 1950 and 1999, which seemed unusual. To investigate, the precipitation data for New Caledonia in 1999 was filtered and examined further. It revealed that only two stations were active in 1999 and that

some of the precipitation data collected throughout the year appeared excessively high and possibly inaccurate. This permits a more detailed examination to determine the cause of these anomalies.

5.4.3.1 Average Rainfall in 2023

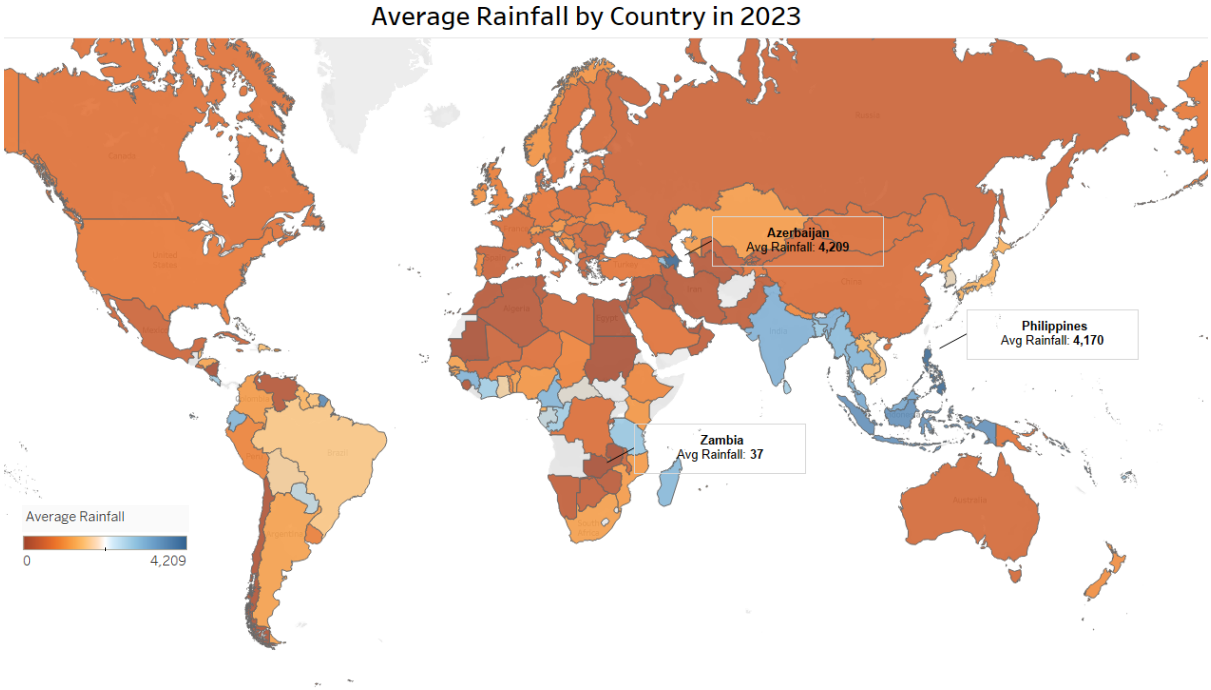


Figure 5: Average Rainfall by Country in 2023 World Map

The average rainfall for 2023 across various countries was analysed, revealing that Azerbaijan had the highest average rainfall, totalling 4,209 mm. In contrast, Zambia had the lowest average rainfall, with only 37 mm. The Philippines ranked second, with an average rainfall of 4,170 mm.

Given that Azerbaijan is situated near the Middle East, the unusually high average rainfall of 4,209 mm in 2023 seems suspicious, indicating a need for further investigation. According to the Climate Change Knowledge Portal (The World Bank Group), the average annual rainfall in Azerbaijan is approximately 500 mm for last 20 years. This significant discrepancy suggests there may be measurement errors or data anomalies that require further examination.

Figure 4, the world map, reveals that some countries lack precipitation data. In Asia, countries like Bhutan, Taiwan, Afghanistan, and Yemen are not depicted with any precipitation information. In Africa, Angola, South Sudan, Uganda, Somalia, Liberia, Eritrea, and Western Sahara also lack precipitation data. Similarly, in Europe, Albania and Kosovo are empty on the map, while all the other countries are coloured based on their average rainfall for 2023. The absence of precipitation data for some countries on the map could be due to two possible reasons: either these countries have no weather stations, or they do not have precipitation data recorded for the year 2023.

6 Conclusion

In this assignment, the climate summaries were analysed using the data gathered from Global Historical Climate Network (GHCN). The primary conclusion regarding this dataset is that it is large, approximately 100 GB for 250 years of data, and continues to grow. Consequently, analysing this data will require more computational resources and storage capacity in the future.

In the initial phase of the assignment, the enriched metadata table was created by joining all metadata tables, using the Apache Spark distributed computing framework. Since these metadata tables were relatively small, processing them was straightforward, requiring minimal resources and time. The most significant tables were the Stations data table and the Inventory data table. The Country and States tables were less complex, primarily providing the names of the countries and states where the stations were located. By constructing the enriched metadata table using these smaller tables, we were able to consolidate information about weather stations worldwide, making it easier to analyse and reference during the assignment.

The analysis phase began with an examination of the enriched stations data table, revealing that there has been a total of 125,983 weather stations, with only 31,837 still active in 2024. This finding permits further investigation to understand why so many stations are inactive and to identify any patterns or reasons behind this inactivity. It was also noted that the United States of America has 72,289 weather stations, constituting approximately 57% of the total stations, indicating that the USA has a substantial lead in weather data collection and analysis.

In the final phase of the assignment, all daily climate summary data were combined with the enriched stations metadata table to provide a comprehensive view of global weather. This combined dataset comprised 3,103,954,141 rows (observations), with precipitation being the most frequently collected element, accounting for 1,069,105,193 observations. Upon calculating the average annual precipitation for all countries each year, it was discovered that New Caledonia had the highest recorded precipitation of 17,674.4 mm in 1999. After sorting the countries with the highest precipitation, New Caledonia appeared 19 times in the top 20 places, raising suspicions. This anomaly indicates a need for further investigation to understand the reasons behind these unusually high precipitation values.

The world map, created to illustrate average rainfall in 2023, suggests that some precipitation data might be incorrect due to unusually high or low average values. The map reveals that data from many developed countries appears more accurate compared to data from developing countries. Additionally, when large countries have only a few weather stations, the average precipitation for the entire country can be biased or misleading. For example, many African countries are large in geographical size, but they have relatively few weather stations, leading to potentially skewed average precipitation values. This lack of data can result in less reliable averages compared to countries with more comprehensive weather station networks, like the USA, Canada, and Australia. As a result, data accuracy may vary significantly, requiring careful interpretation and further analysis.

7 References

Apache Spark. (n.d.). *JOIN*. Retrieved from Apache Spark - A Unified engine for large-scale data analytics: https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-join.html

Menne, M. I. (2012). *Daily (GHCN-Daily), Version 3*. Retrieved from Global Historical Climatology Network : https://www.ncei.noaa.gov/pub/data/ghcn/daily/readme.txt

National Oceanic and Atmospheric Administration. (n.d.). *Global Historical Climatology Network daily (GHCNd)*. Retrieved from National Centers for Environmental Information: https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily

OpenAI. (2024). *ChatGPT (4.0) [Software]*. Retrieved from https://www.openai.com/

QuillBot. (2024). *QuillBot Paraphrasing Tool [Software]*. Retrieved from https://www.quillbot.com/

Rezania Agramanisti Azdy, F. D. (2019). Use of Haversine Formula in Finding Distance Between. *Journal of Physics: Conference Series*, 6.

Stack Exchange Inc;. (2024). Retrieved from https://stackoverflow.com/

The World Bank Group. (n.d.). *Climate Change Knowledge Portal.* Retrieved from Climate Change Knowledge Portal - Azerbaijan: https://climateknowledgeportal.worldbank.org/country/azerbaijan/climate-data-historical#:~:text=Average%20rainfall%20in%20Azerbaijan%20follows,June%2C%20and%20again%20in%20October.

WMO. (2020). *Buoy WMO Identification numbers*. Retrieved from World Meteorological Organization - Community Platform: https://community.wmo.int/en/buoy-wmo-identification-numbers#:~:text=World%20Meteorological%20Organization,-Community%20Platform&text=WMO%20identification%20numbers%20are%20allocated,sites%2C%20and%20profiling%20floats).

8 Appendix

8.1 Appendix 1 - Time series for TMIN and TMAX for station in New Zealand

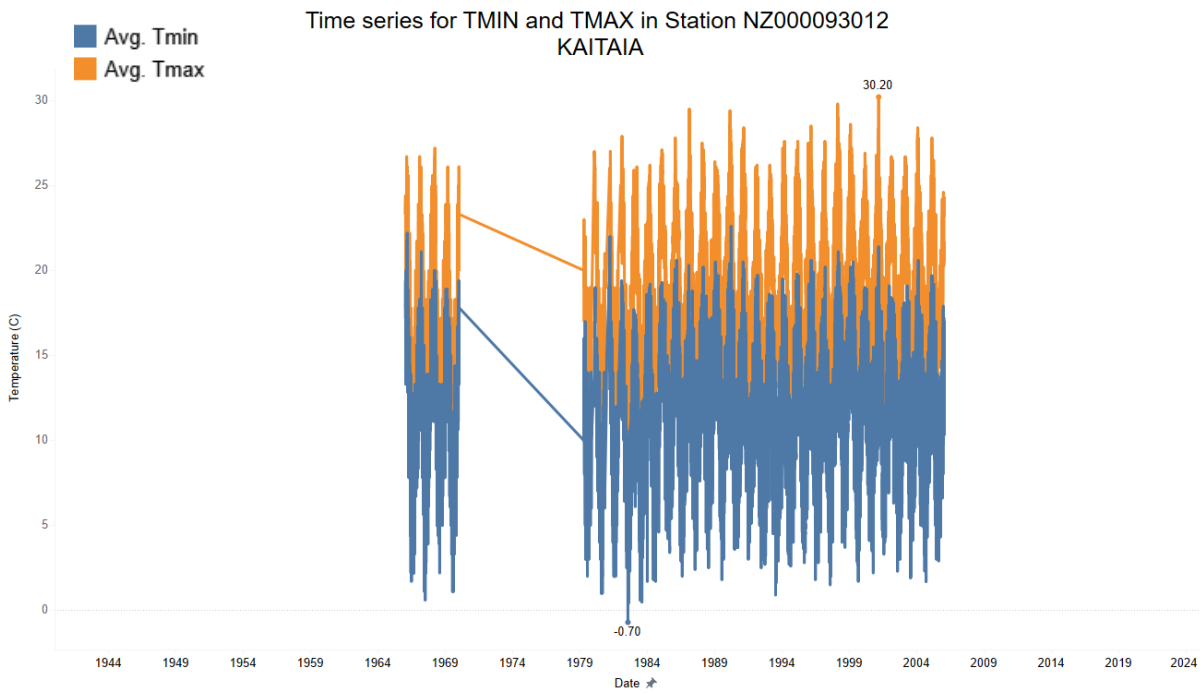


Figure 6 : Time Series for Station Kaitaia

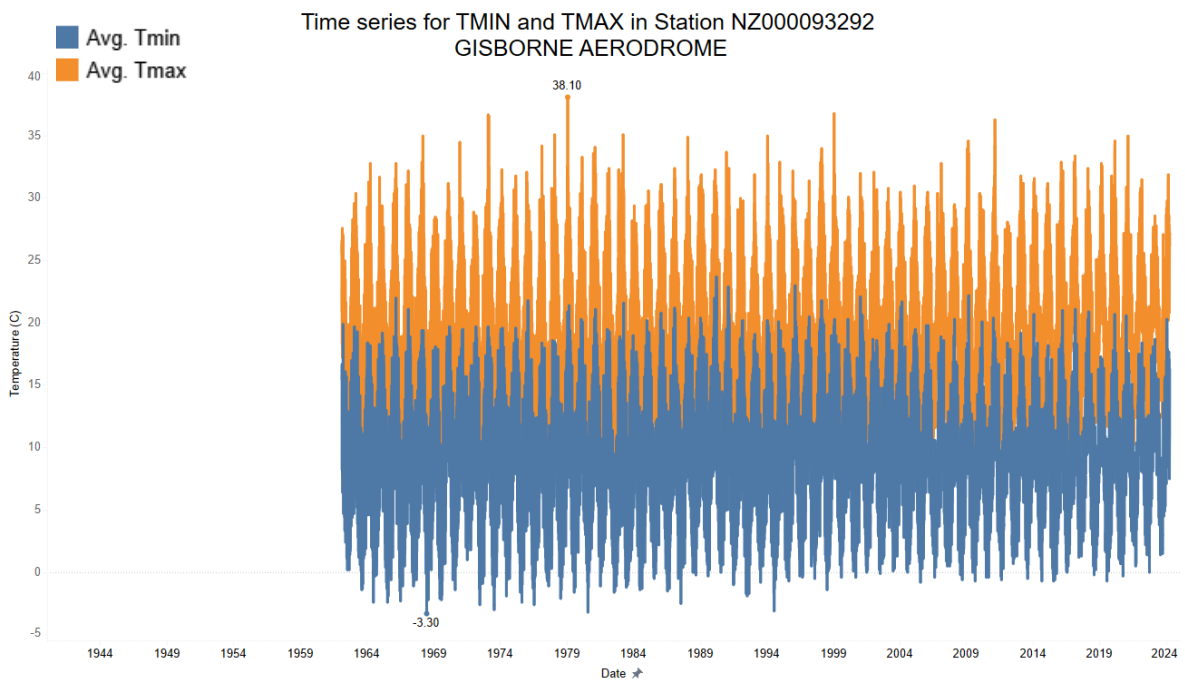


Figure 7: Time Series for Station Gisborne Aerodrome

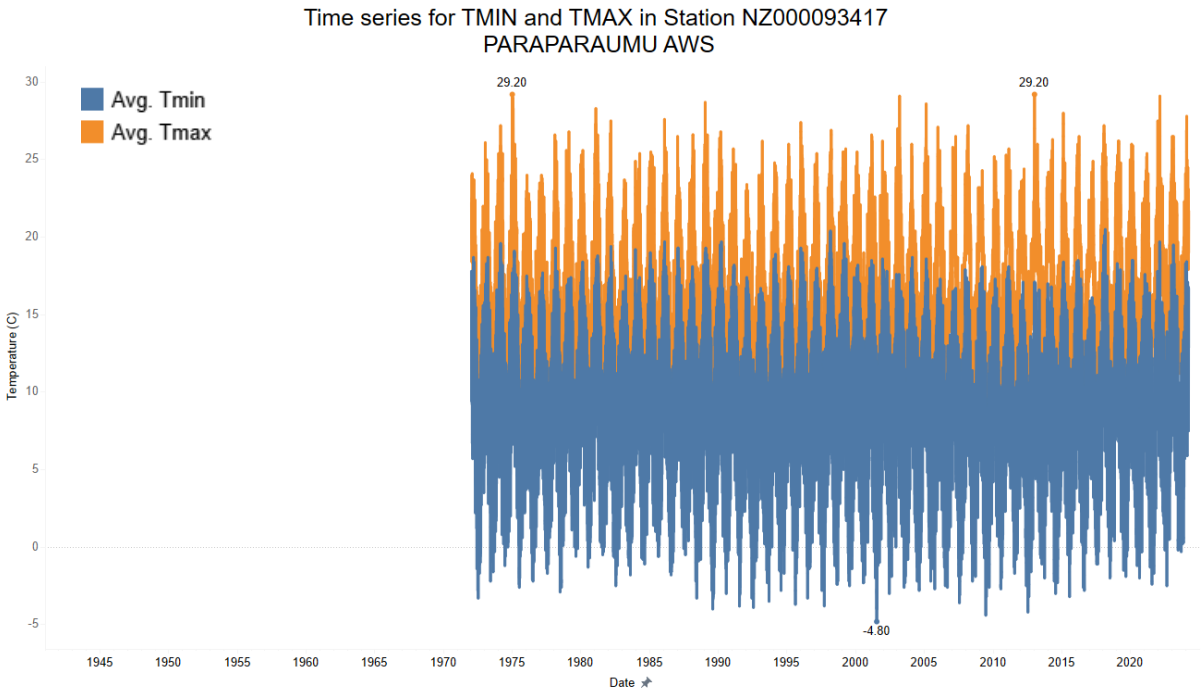


Figure 8:Time Series for Station Paraparaumu AWS

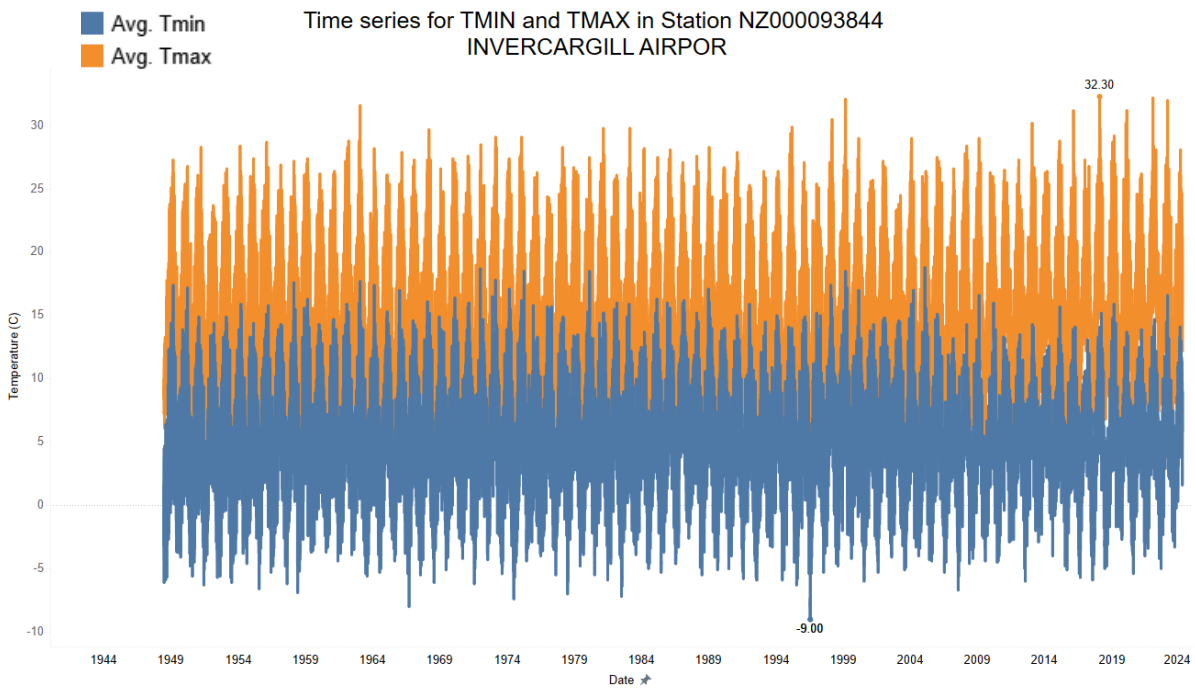


Figure 9: Time Series for Station Invercargill Airpor

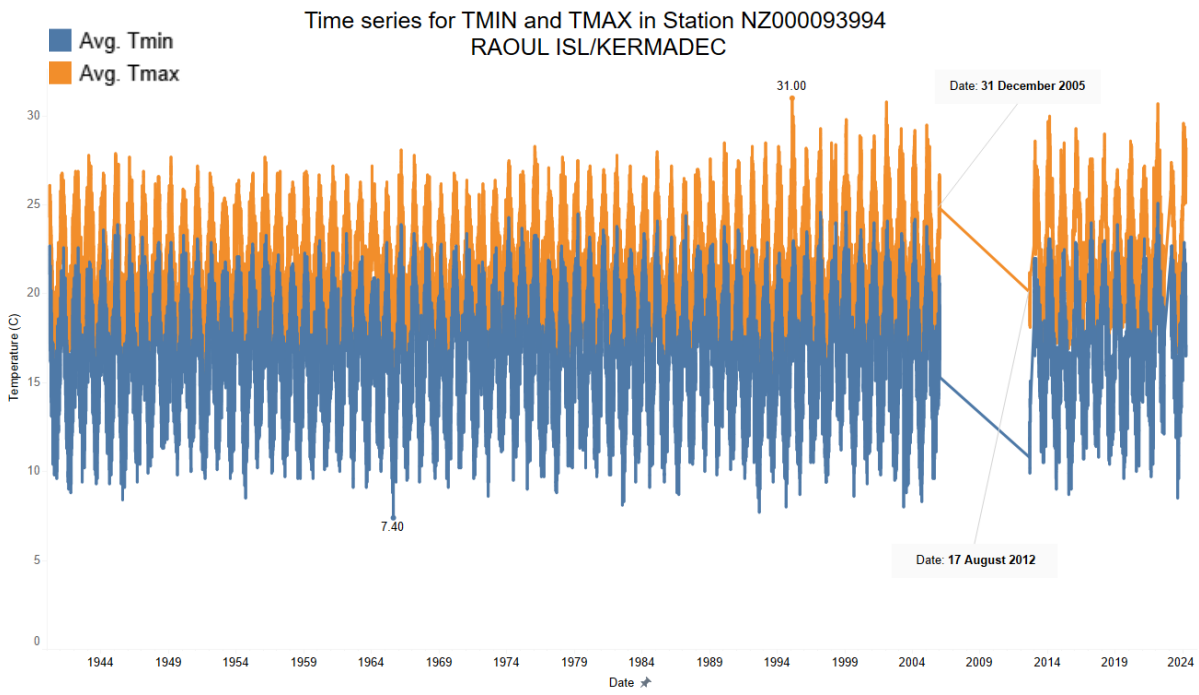


Figure 10:Time Series for Station Raoul Isl/Kermadec

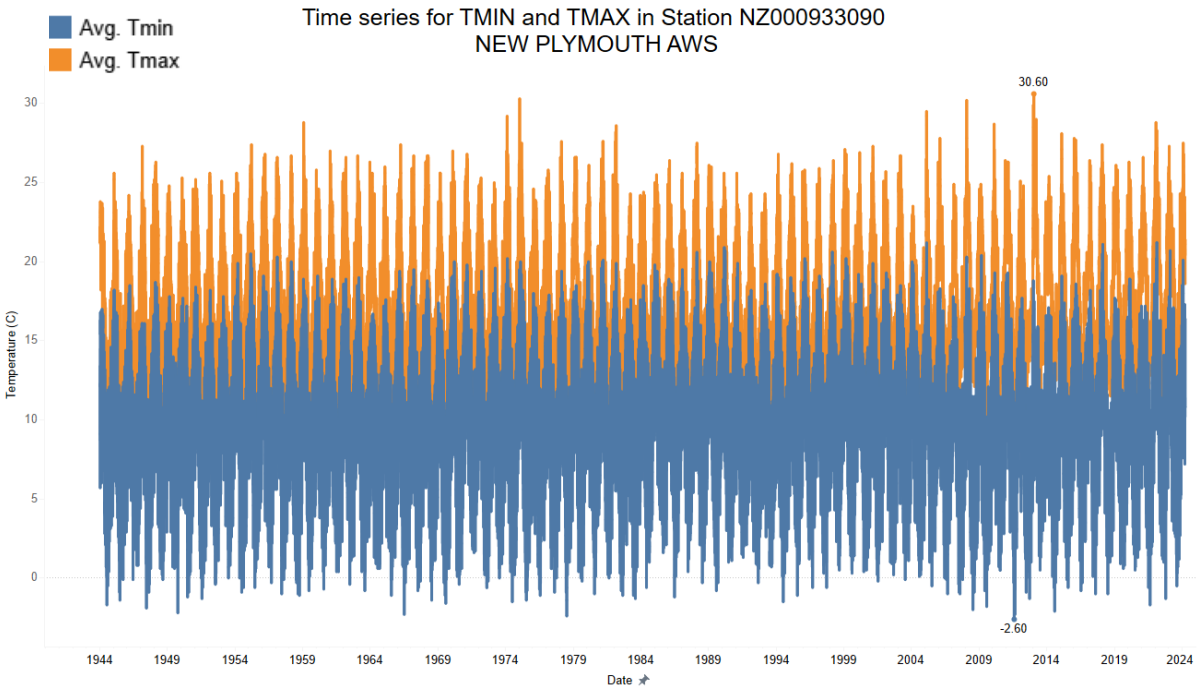


Figure 11: Time Series for Station New Plymouth AWS

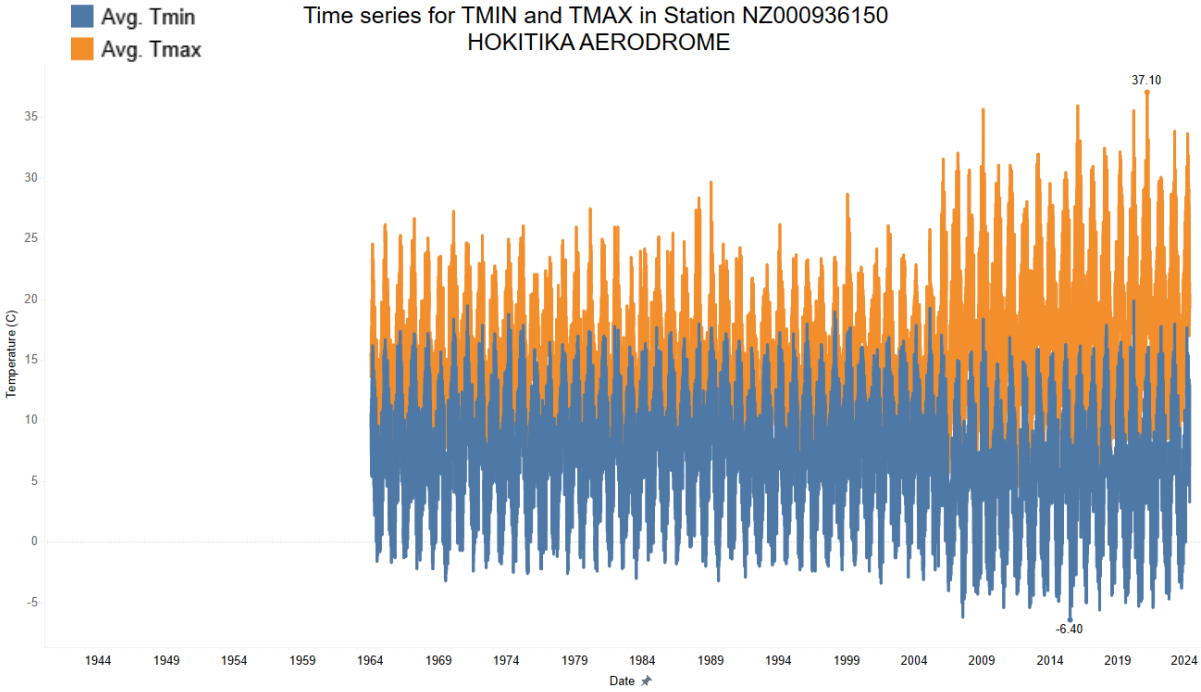


Figure 12: Time Series for Station Hokitika Aerodrome

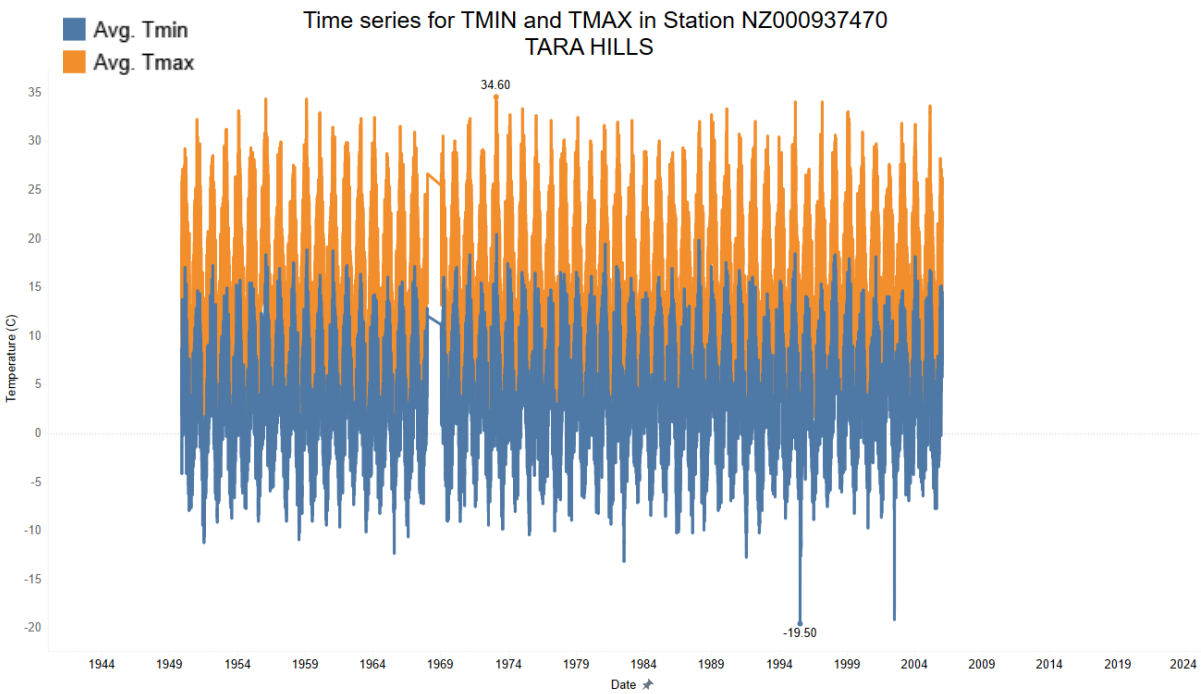


Figure 13: Time Series for Station Tara Hills

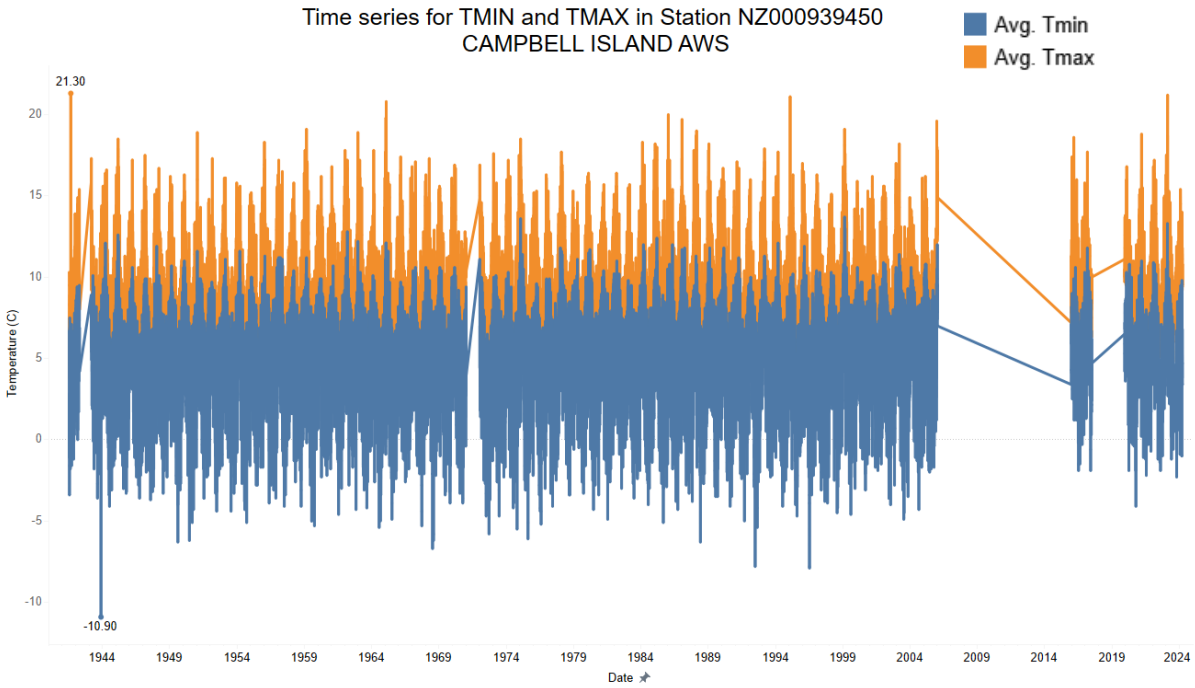


Figure 14: Time Series for Station Campbell Island AWS

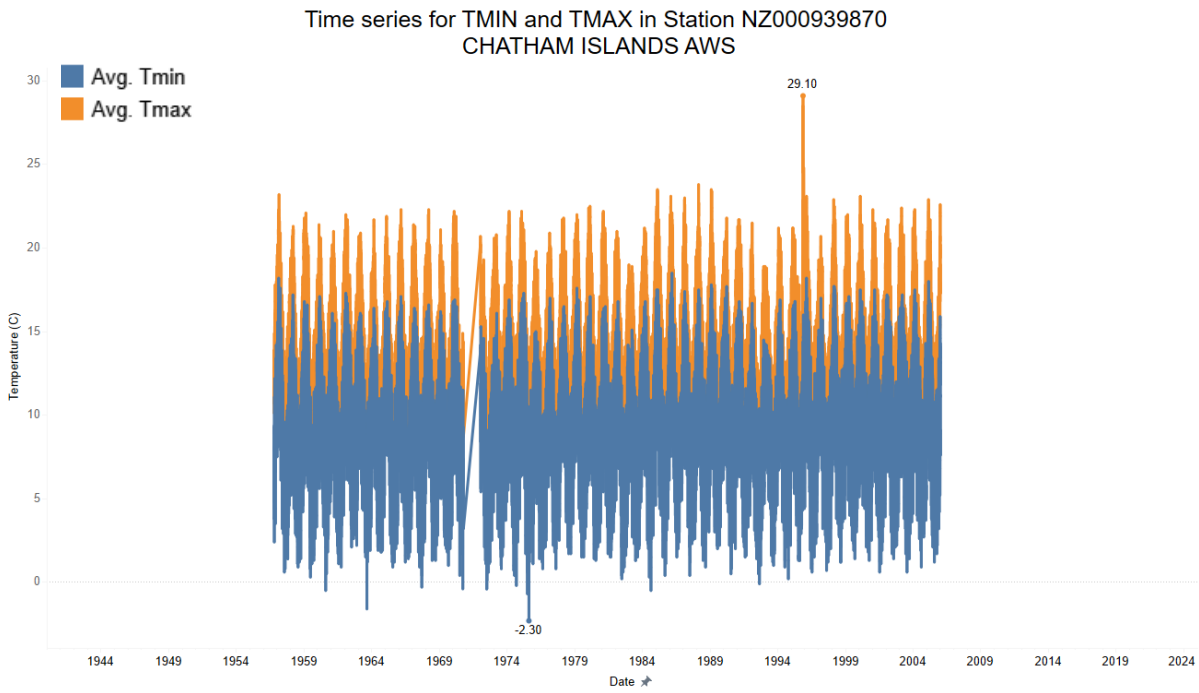


Figure 15: Time Series for Station Chatham Island AWS

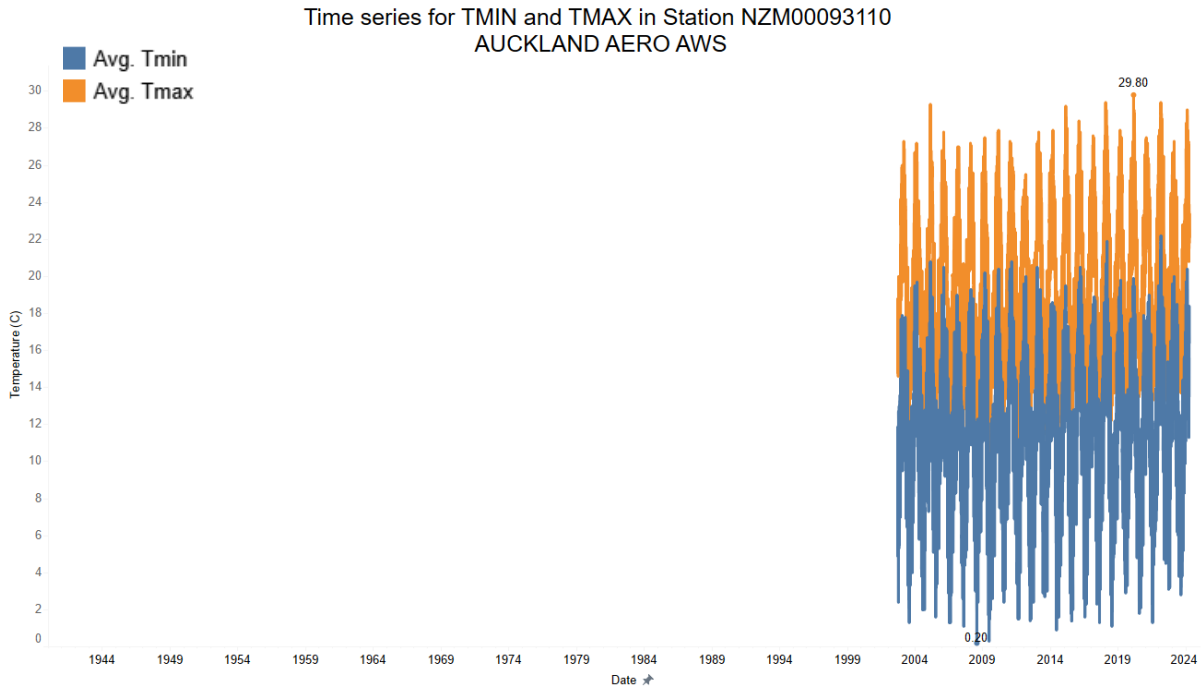


Figure 16: Time Series for Station Auckland Aero AWS

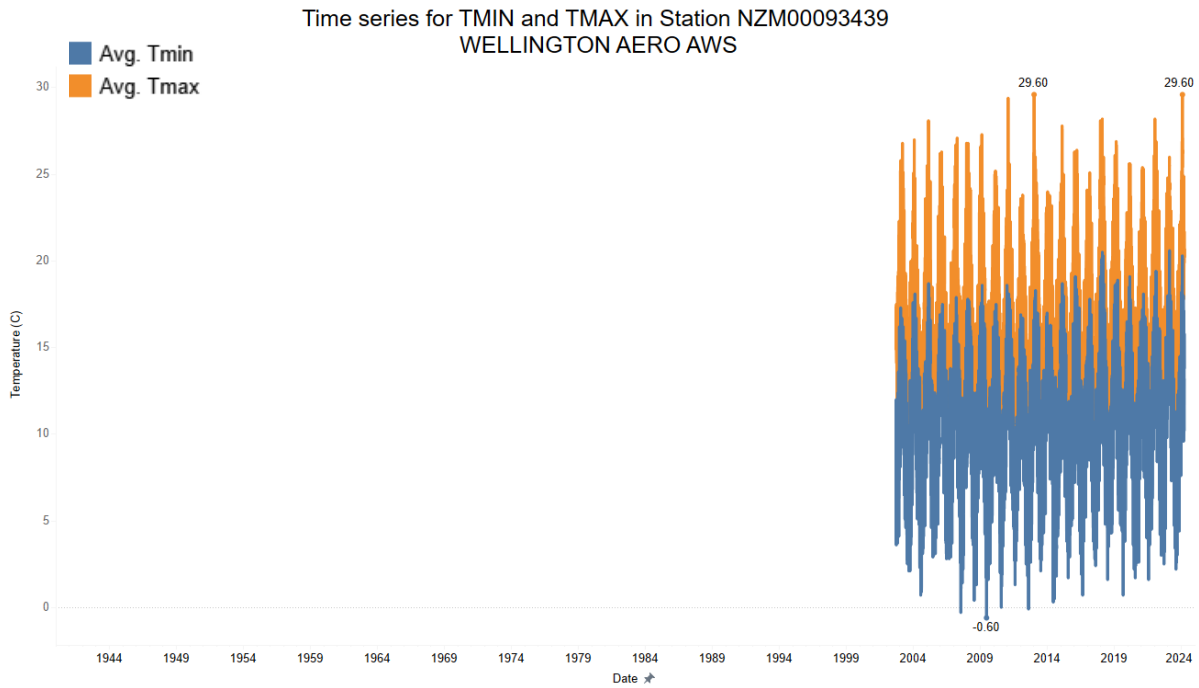


Figure 17: Time Series for Station Wellington Aero AWS

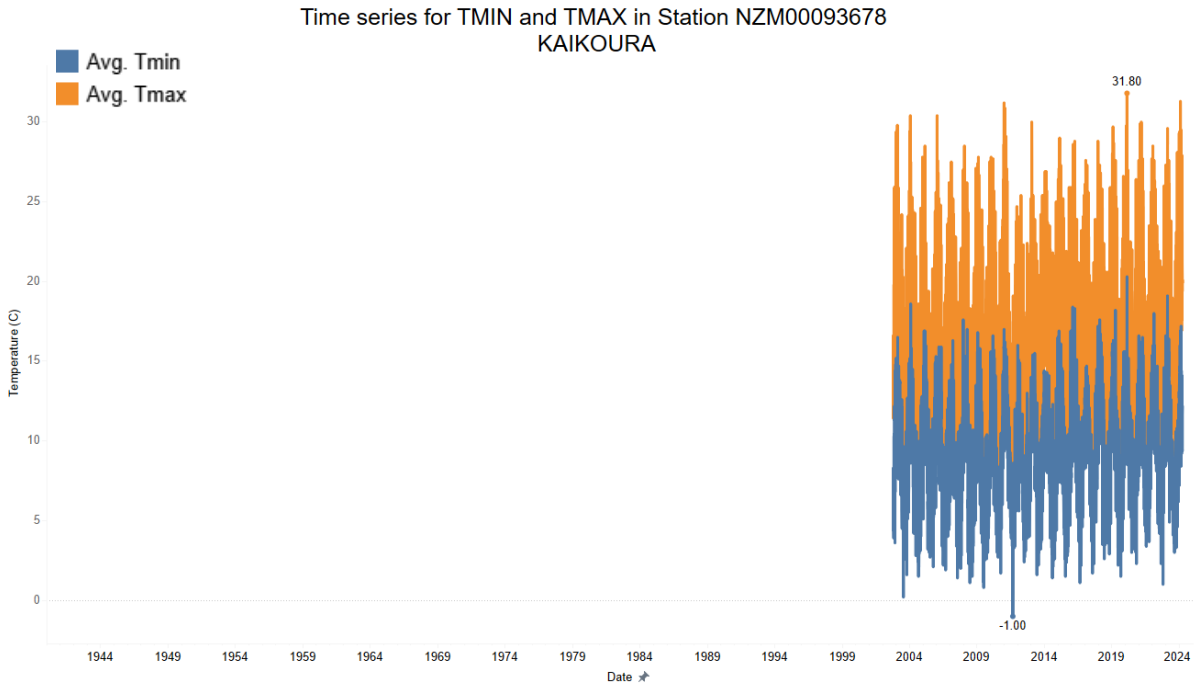


Figure 18: Time Series for Station Kaikoura

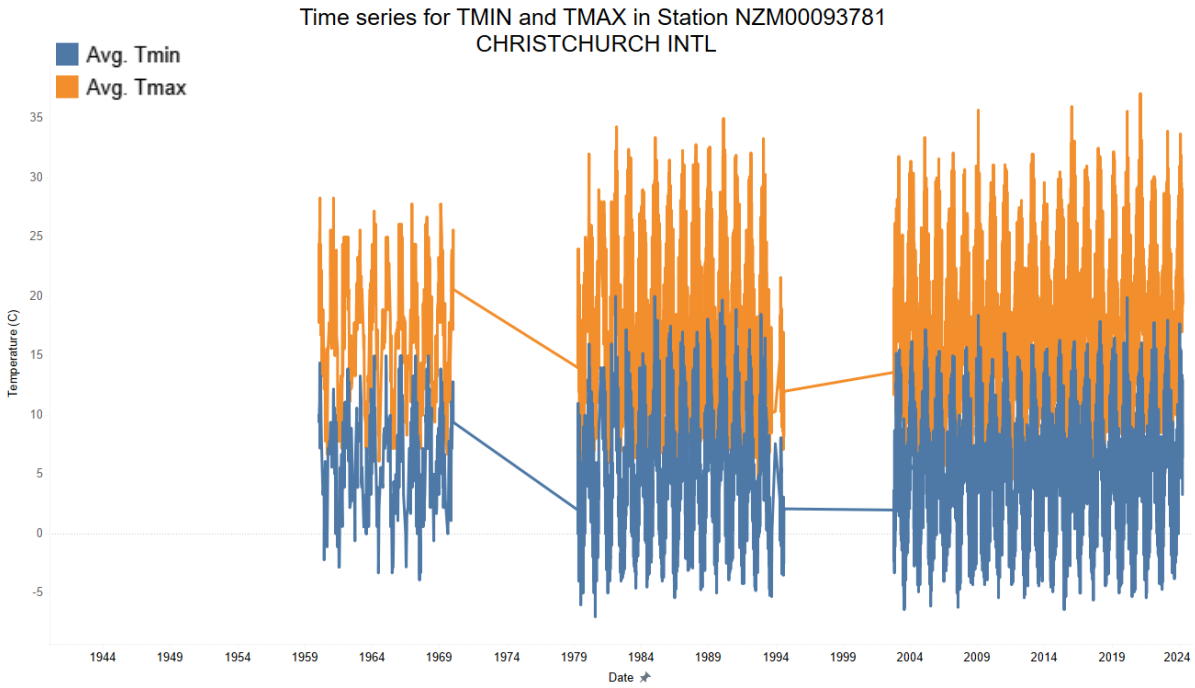


Figure 19: Time Series for Station Christchurch Intl

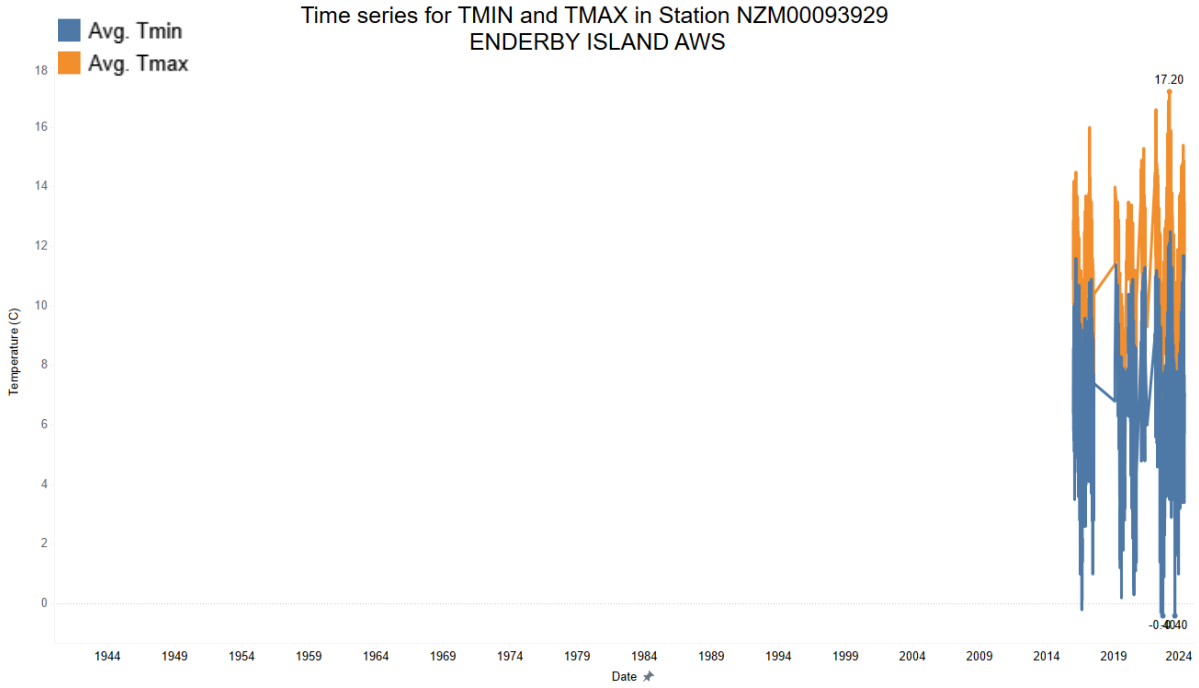


Figure 20: Time Series for Station Enderby Island AWS

8.2 Appendix 2 - Average time series for TMIN and TMAX for New Zealand

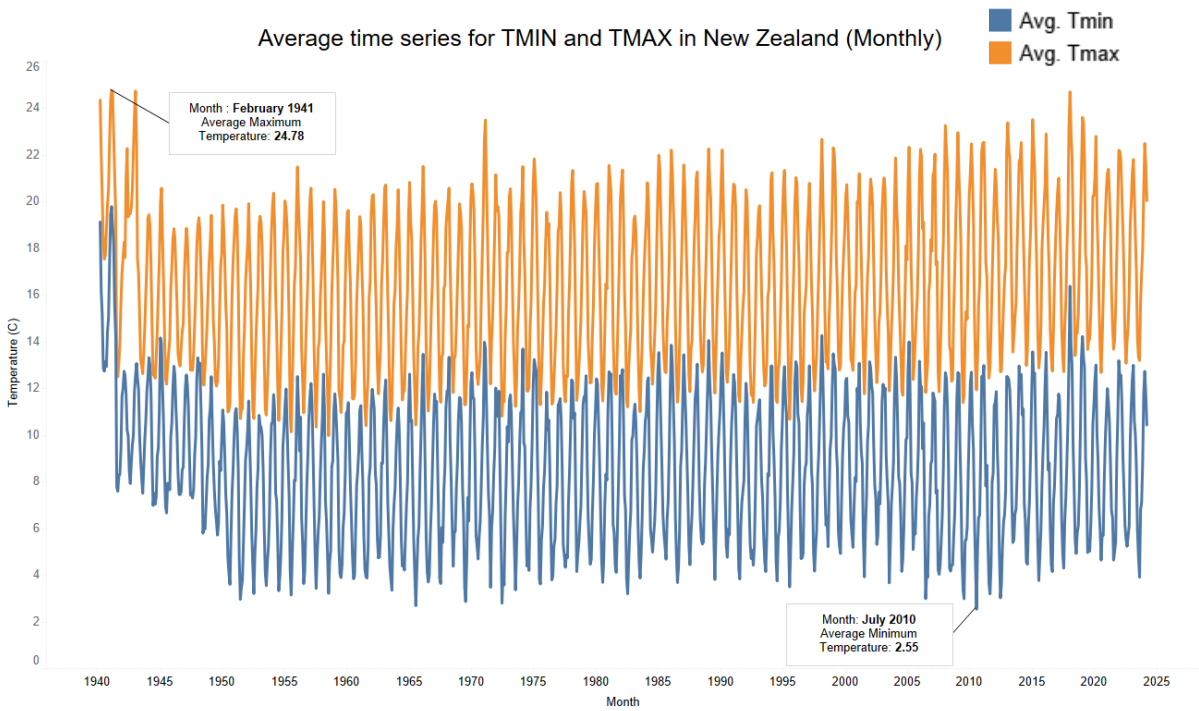


Figure 21: Monthly Average time series for Maximum and Minimum temperature in New Zealand

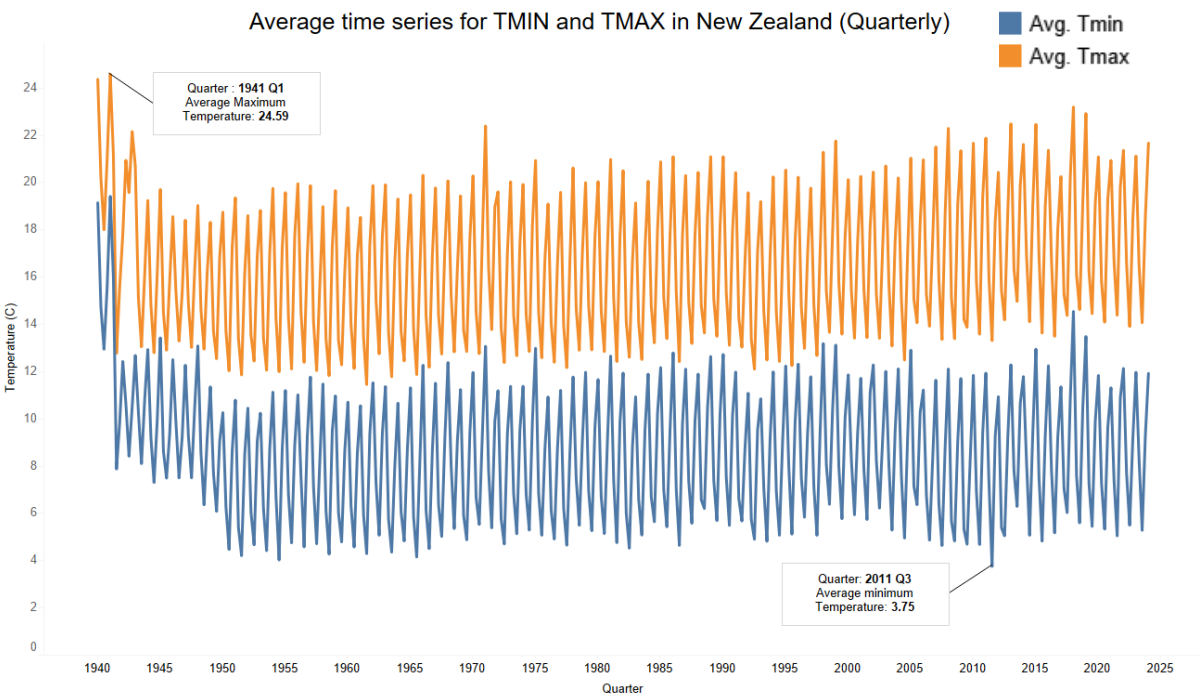


Figure 22 : Quarterly Average time series for Maximum and Minimum temperature in New Zealand

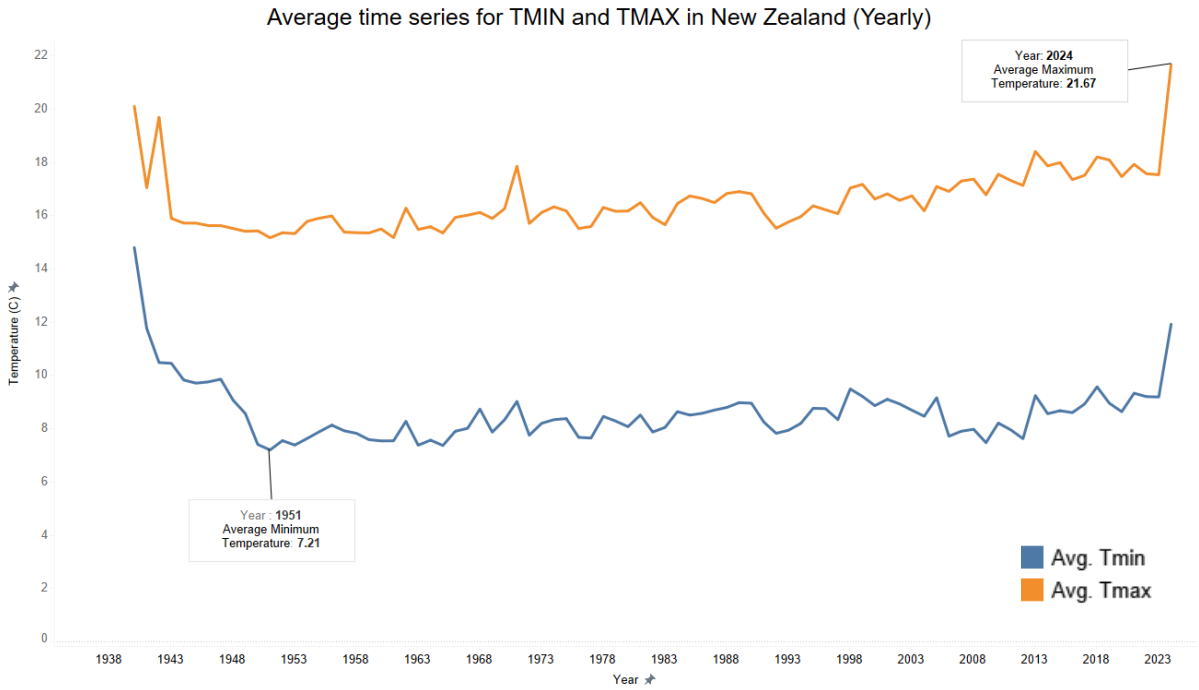


Figure 23 : Yearly Average time series for Maximum and Minimum temperature in New Zealand

8.3 Appendix 3 - Average Rainfall in 2023 for each country

8.3.1 Asia

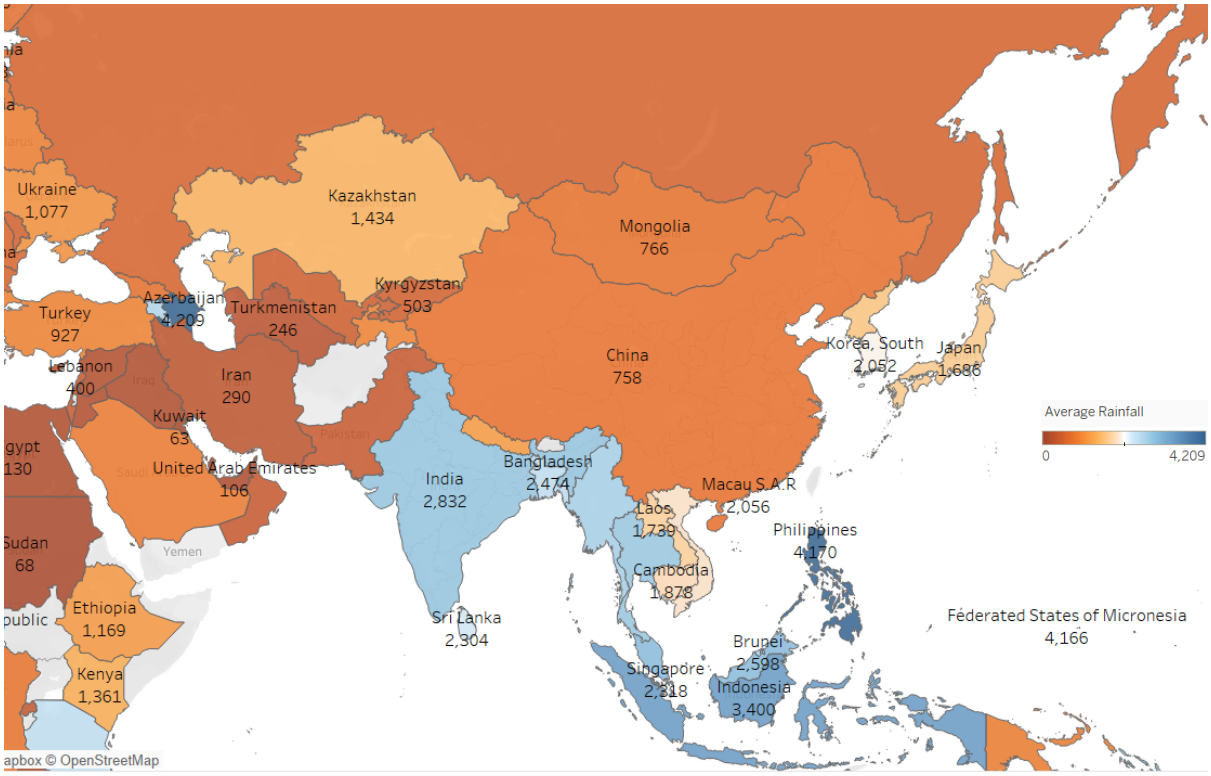


Figure 24 : 2023 Average Rainfall in Asia

8.3.2 Africa

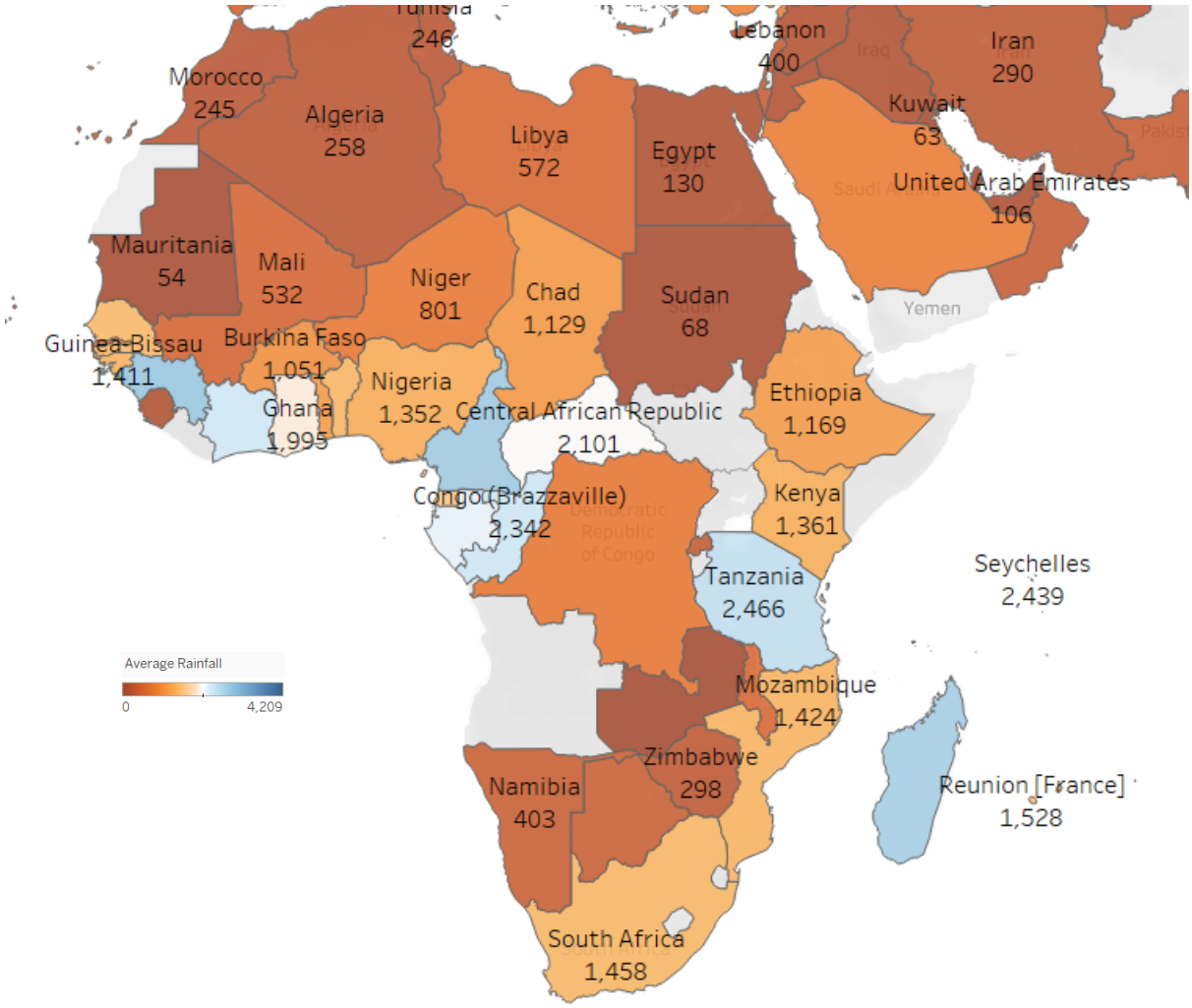


Figure 25: 2023 Average Rainfall in Africa

8.3.3 Europe

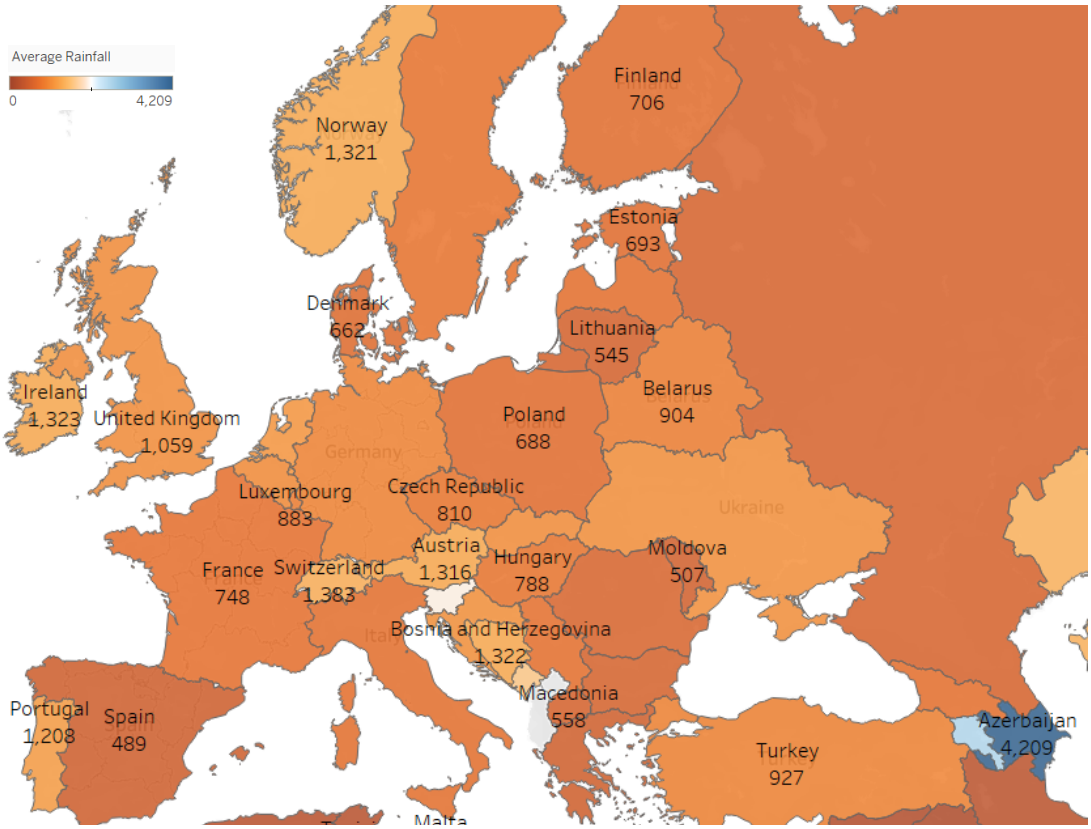


Figure 26: 2023 Average Rainfall in Europe

8.3.4 Australia and New Zealand

Figure 27: Average Rainfall in Australia and New Zealand

8.3.5 North America

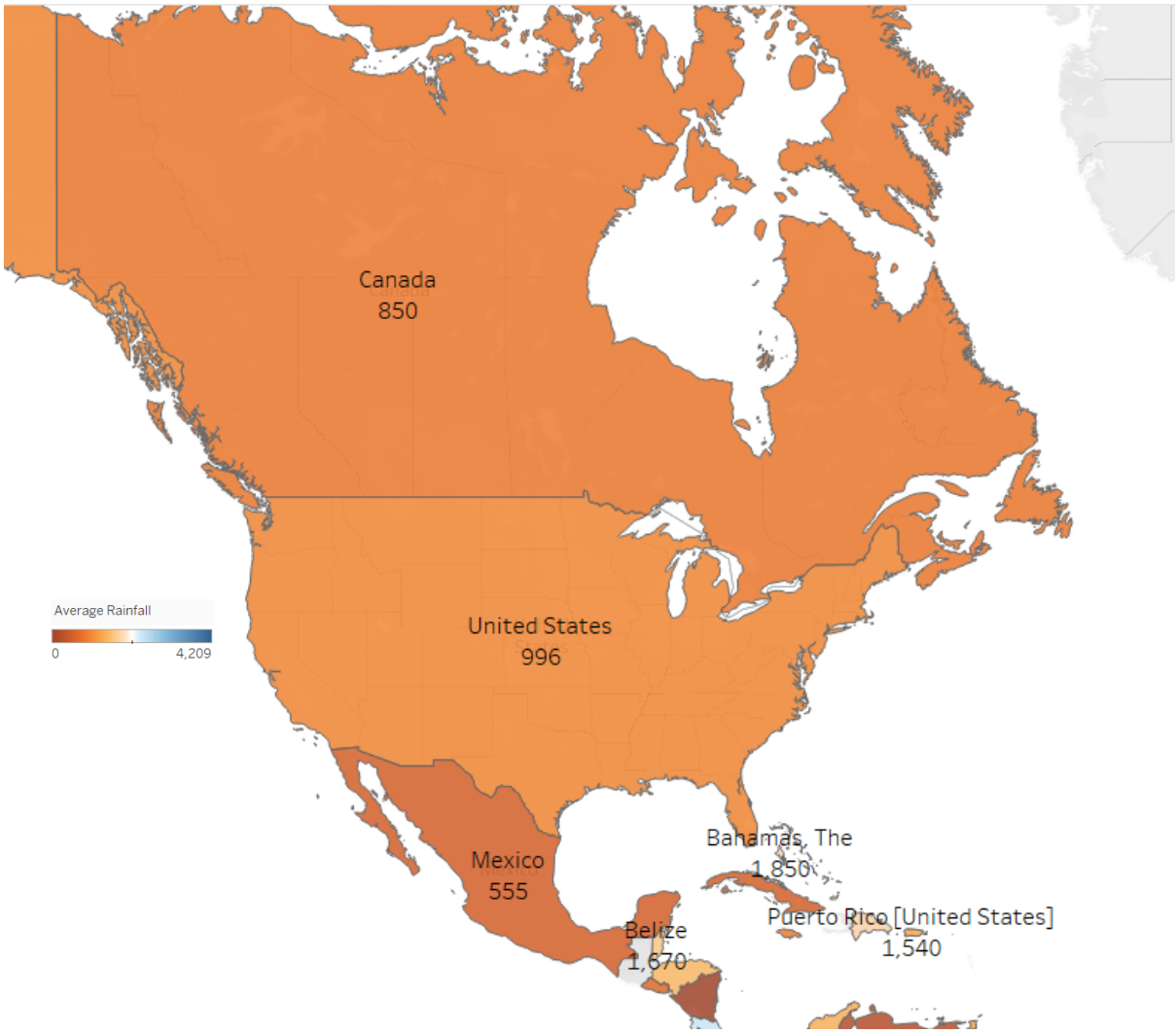


Figure 28: 2023 Average Rainfall in North America

8.3.6 South America

Figure 29: 2023 Average Rainfall in South America

8.3.7 Antarctica

Figure 30: 2023 Average Rainfall in Antarctica