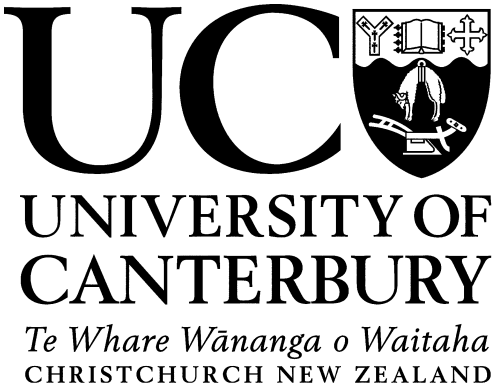


DATA420

Scalable Data Science

Assignment 1

GHCN Data Analysis using Spark

Yu Xia 62380486

September 14, 2025

1. Background

- Purpose of the assignment
- Overview of what was achieved (processing, analysis, visualization)
- High-level introduction of **GHCN Daily**: scope, scale, variables, metadata
- Context for why Spark was used (scale, distributed computing)

---

2. Processing

Description of Datasets

Five datasets from the GHCN-Daily archive stored in Azure Blob Storage were analysed using a simple EDA (Exploratory Data Analysis) procedure, resulting in the following dataset metadata summary.

- **daily**: contains weather observations in .csv.gz format, one record per station-day-element. Files are compressed as (.csv.gz) by year. Data spans 1750–2025, with 264 files found (12 years missing between 1751–1762).
- **stations**: fixed-width text, containing station ID, coordinates, elevation, country/state codes, station name, and network flags (GSN, HCN/CRN, WMO ID).
- **countries**: fixed-width text, mapping 2-character FIPS 10-4 country codes to full country names.
- **states**: fixed-width text, listing US states/territories and Canadian provinces.
- **inventory**: fixed-width text, listing each station-element pair with start and end years of available data, along with coordinates.

Schema definitions were implemented in PySpark using StructType for daily, and column parsing via substring for the fixed-width metadata files .

Steps Performed

1. **Loading datasets into Spark**

- **daily**: loaded with explicit schema (ID, DATE, ELEMENT, VALUE, MEASUREMENT FLAG, QUALITY FLAG, SOURCE FLAG, TIME). DATE was cast to DateType, TIME cleaned and padded to 4-digit strings, then converted to TimestampType for observation times.

- Metadata files (stations, countries, states, inventory) are parsed using substrings and cast to the proper types (e.g., LAT/LON as double, FIRSTYEAR/LAS).

2. **Exploration**
   - Verified presence of 264 compressed .csv.gz files for years 1750–2025, with 12 years missing (1751–1762).
   - File size trend analysis revealed growth from approximately 10^{-6} MB in the early years to about 10^{5} MB by 2024. As demonstrated in Figure 1: GHCN Daily File Size by Year.
   - Total size: ~13.1 GB, with daily contributions ~13.0 GB. The *daily* data dominates storage in the folder.

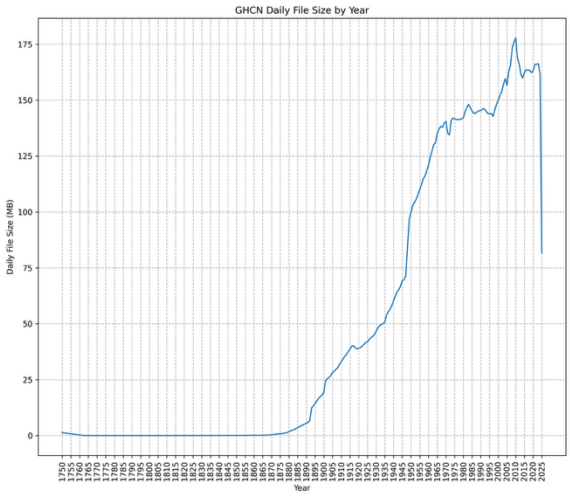


Figure 1: GHCN Daily File Size by Year

3. **Building the enriched stations table**
   - Extracted 2-character FIPS country code from station IDs.
   - Joined stations with countries (LEFT JOIN) to append country names.
   - Joined with states (for US stations only).
   - Derived activity range (FIRSTYEAR/LASTYEAR) and number of elements per station from inventory.
   - Pivoted core elements (TMAX, TMIN, PRCP, SNOW, SNWD) to indicator columns; identified 20,504 stations collecting all five measurements compared to 16,267 stations that collect only precipitation.
   - Final enriched stations table combined metadata + inventory, with 129,657 rows, stored in Parquet format (~43.8 MB) for efficiency.
4. **Data quality checks**
   - Stations shown in *enriched station table* but missing in *daily* dataset: found 89,941 stations missing in daily-2025, and 38 stations missing across all years.
   - Negative or zero precipitation values are flagged as placeholders, requiring cleaning for later analysis.

Results

- **Row counts**:
  - Stations: 129,657
  - Countries: 219
  - States: 74
  - Inventory: 766,784
  - Daily: 3,155,140,380
- **Core element coverage**:
  - 205,04 stations collect all 5 core elements.
  - 162,67 stations collect only precipitation.
- **Enriched stations table structure**:
  - See Figure 2: `enriched stations table schema`.
  - More field details refer to the README FILE FOR DAILY GLOBAL HISTORICAL CLIMATOLOGY NETWORK (GHCN-DAILY) Version 3.32

```
stations_enriched = spark.read.parquet(stations_enriched_savepath)
stations_enriched.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- STATE: string (nullable = true)
 |-- COUNTRY_CODE: string (nullable = true)
 |-- LATITUDE: double (nullable = true)
 |-- LONGITUDE: double (nullable = true)
 |-- ELEVATION: double (nullable = true)
 |-- NAME: string (nullable = true)
 |-- GSN_FLAG: string (nullable = true)
 |-- HCN_CRN: string (nullable = true)
 |-- WMO_ID: string (nullable = true)
 |-- COUNTRY_NAME: string (nullable = true)
 |-- STATE_NAME: string (nullable = true)
 |-- FIRSTYEAR_ANY: integer (nullable = true)
 |-- LASTYEAR_ANY: integer (nullable = true)
 |-- N_ELEMENTS: long (nullable = true)
 |-- TMAX: integer (nullable = true)
 |-- TMIN: integer (nullable = true)
 |-- PRCP: integer (nullable = true)
 |-- SNOW: integer (nullable = true)
 |-- SNWD: integer (nullable = true)
 |-- N_CORE_ELEMENTS: integer (nullable = true)
```

Figure 2: enriched stations table schema

3. Analysis

- **Stations**:
  - Total stations; active in 2025
  - Counts per network (GSN, HCN, CRN)
  - Stations in Southern Hemisphere
  - Stations in US territories (excluding mainland US)
  - Which stations in NZ are geographically closest (explain spherical distance calculation, haversine formula, cite sources)
- **Daily observations**:
  - Total number of rows in daily
  - Counts for each of the five core elements (TMAX, TMIN, PRCP, SNOW, SNWD)
  - Which element has the most observations
  - How many TMAX records lack a TMIN partner (methodology for efficient count)
- **Tables & Figures**:
  - Counts of stations table
  - Counts of observations per element table
  - Map of NZ stations

Awesome—let's turn your notebook outputs into a clean, Word-ready **Analysis** section that matches your rubric. I've kept it tight, evidence-driven, and clear about methods.

Stations

- **Totals and recency.**
  - Total stations in the enriched catalogue: 129,657.
  - Stations still active in 2025 (LASTYEAR ≥ 2025): 38,481.

  According to the NOAA Global Historical Climatology Network-Daily documentation, each station record includes FIRSTYEAR and LASTYEAR, indicating the period covered by observations (NOAA NCEI, 2012; Menne et al., 2012). The fact that only 38,481 of 129,657 stations remain active in 2025 reflects natural station turnover, temporary deployments, and permanent closures.

- **Network membership.**
  - By network flags: GSN = 991, HCN = 1,218, CRN = 234.
  - Multi-network membership is rare: 15 stations overlap, and all are GSN and HCN.

  The GHCN-Daily catalogue includes metadata flags indicating if a station belongs to the GSN, HCN, or CRN. These networks serve different purposes. The GSN, managed by WMO and NOAA, consists of globally distributed benchmark stations for long-term climate monitoring. The HCN includes U.S. stations with long, high-quality records for detecting climate trends. The CRN, established in the early 2000s, uses advanced instruments and standards for high-quality data in the 21st century. Few stations belong to multiple networks, with limited overlap between GSN and HCN (Menne et al., 2012; Diamond et al., 2013; NOAA NCEI, 2012).

- **Hemisphere split.**
  - Southern Hemisphere count reported by the query LATITUDE < 0: 25,357.

Southern Hemisphere stations account for only 19.6% of GHCND total stations, as most stations in GHCND are located in the United States, and around 70% of all values originate from North American stations. Although Figure 3: Global PRCP Stations by Country shows that It is not entirely based on the GHCND station count, it still includes 127606 out of 129657 stations, which is 98.42%. This also highlights that the deployment of GHCND climate stations is uneven.

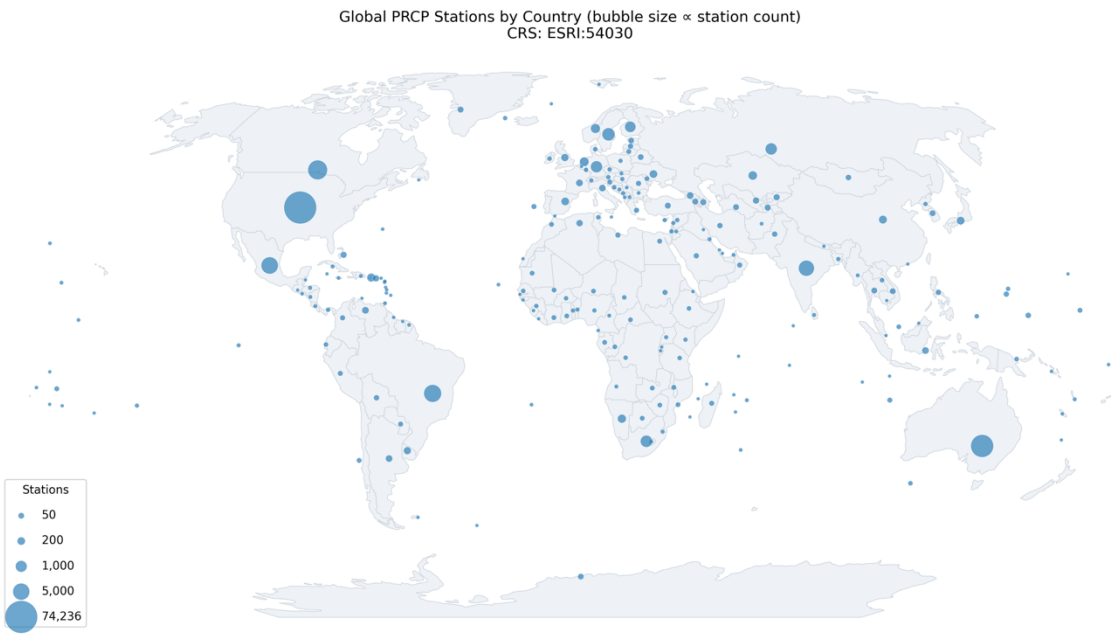


Figure 3: Global PRCP Stations Count by Country

- **U.S. territories (excluding mainland U.S.).**
  - Territory outside the U.S. mainland: Puerto Rico 260, U.S. Virgin Islands 77, Guam 34, American Samoa 21, Johnston Atoll 4, Palmyra 3, Wake Island 1; aggregate 414 stations.

  This phenomenon may introduce spatial bias when we apply spatial analysis; for example, some territorial land is included in its mother country, resulting in different longitude and latitude calculations rather than their true locations on Earth.

- **Closest stations in New Zealand**
  - Method: Haversine distance R=6371 km over latitude/longitude in radians; implemented as a Spark UDF and applied to New Zealand station pairs (crossJoin with ID_A < ID_B).
  - Result: NZ000093417 (PARAPARAUMU AWS) and NZM00093439 (WELLINGTON AERO AWS) are closest, 50.53 km apart.

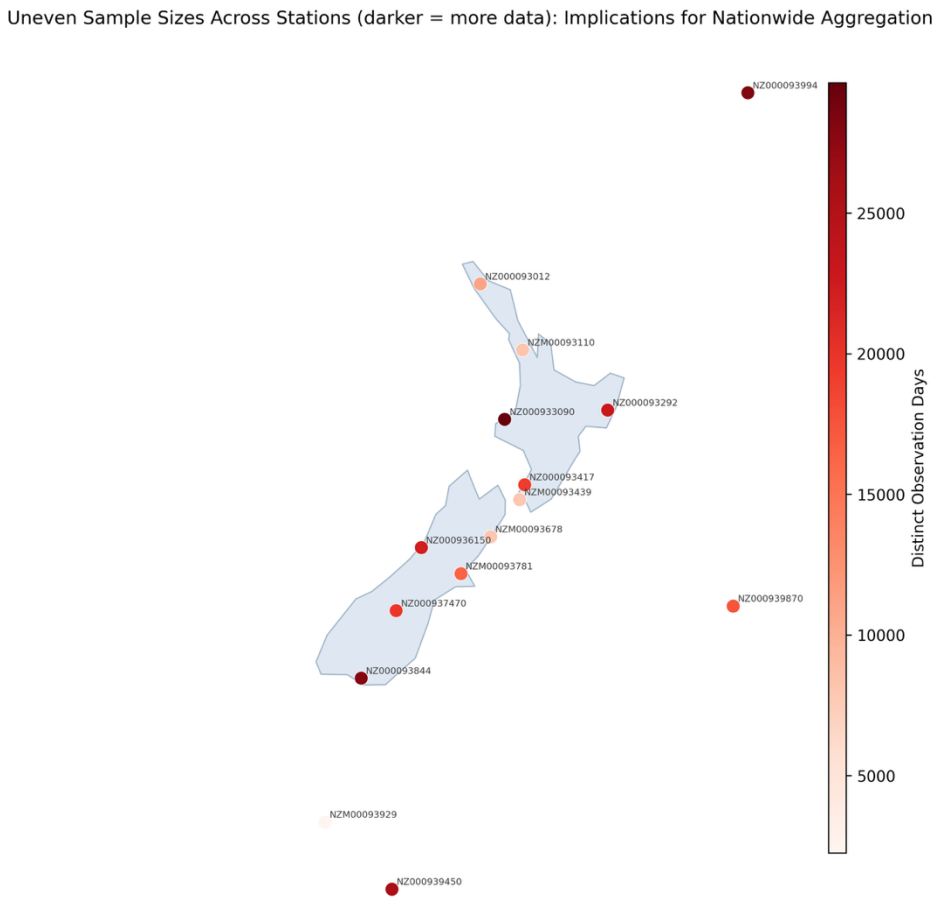


Figure 4: Uneven Sample Sizes Across Stations (darker = more data): Implications for Nationwide Aggregation

Haversine captures spherical geometry and is sufficiently accurate (Borisov et al., 2023) for continental-scale analysis; ellipsoidal geodesics (e.g., Vincenty/WGS84) would only be necessary for survey-grade accuracy.

The implications for nationwide aggregation also raise another question for me: how can a country's measurements be represented when the data is collected unevenly across both space and volume? This issue leads to my spatial-weighted and time-slice stratified sampling strategies algorithm in later New Zealand Monthly TMAX-TMIN plotting.

Daily observations

- **Scale.**

CORE ELEMENT	OBSERVATION COUNT
PRCP	1,084,610,240
TMAX	461,915,395
TMIN	460,752,965
SNOW	361,688,529
SNWD	302,055,219

Table 1: Daily Core Element Observation Count

PRCP (precipitation) has the most observations among the five core GHCN-Daily elements, due to its early and widespread measurement, affordability of rain gauges, and extensive data collection across regions. The maximum and minimum temperatures are considered together in order to ensure that the temperatures for a particular station and day always originate from the same source, and thus have been a data quality control method. (Menne et al., 2012). This explains our stats result that TMAX and TMIN have a very small difference.

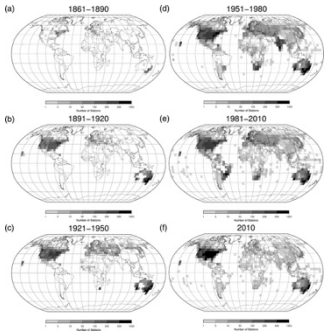


Figure 5: (a)–(f) Density of GHCN-Daily stations with daily precipitation

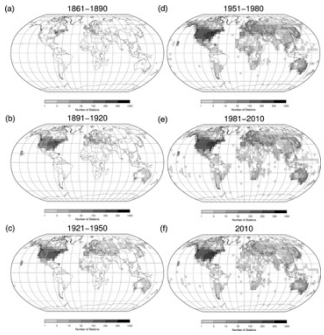


Figure 6:(a)–(f) Density of GHCN-Daily stations with daily maximum and minimum temperature

Pair completeness (TMAX vs. TMIN).

- Method: anti-join of (ID, DATE) pairs to find TMAX records lacking a TMIN partner.
- Result: 10,735,252 TMAX observations without a same-day TMIN; contributed by 28,751 distinct station IDs.

Methodology Deep Thinking

- **Spark SQL** was used for all large-scale group-by, joins, and anti-joins;
  - Arrow-accelerated to Pandas () only after heavy aggregation (e.g., per-country, per-element counts). This operation applies to the HDFS system's core data play strategy "*Moving Computation is Cheaper than Moving Data*"(*HDFS Architecture*, n.d.)
  - So I solidified my join strategy as:

Case 1: Table A (large) Table B (large) → minimise shuffle costs, referring to: (*Optimizing the Join Strateg*, n.d.):

✧ **Projection and filtering**: selecting only the necessary columns and rows before joining.

✧ **Pre-aggregation or distinct filtering**: reducing table cardinality when possible.

✧ **Sort-Merge Join (SMJ)** with Spark's Adaptive Query Execution (AQE), which can dynamically optimise skew handling and repartitioning.

✧ **Semi-joins** for existence checks to avoid carrying unnecessary columns from the second table.

✧ **Where applicable**, applying Bloom filters to prune non-matching keys before the full join.

Case 2: Table A (large) Table B (small) → Using F.broadcast to broadcast the smaller table B to all worker nodes, allowing each partition of the larger table A to perform the join locally without triggering a network-wide shuffle.

- **Network flags** were parsed into indicator columns and summed to count membership; overlap computed via net_count ≥ 2.
- **Haversine UDF** computed pairwise great-circle distances; for NZ, a self-join generated pairs with ID_A < ID_B to avoid duplicates.
- **Save and reload** frequent using data in **parquet** format and **reload** or **cache** if small enough so to make computation quicker. By spliting pythnb files into different charpters, this reload method allows me to run from the very beginning every time, e.g. some count result should be save to txt file as a int or float number or it will lead to duplicating time cost processing in spark.

3. Visualizations

New Zealand Temperature

Fifteen New Zealand stations were identified from the enriched stations table. Extracts all TMIN and TMAX observations from 1940 to 2025 and creates two levels of aggregation.

Monthly averages

Each station's daily values were grouped by (ID, year, month) to compute mean TMIN/TMAX, with missing months retained as NaN to visibly mark gaps. Then aligned all stations to a full monthly index (1940–2025). The resulting panel of 15 subplots highlights station-level variability and irregular data coverage. Some stations (e.g., NZ0000933090) show long continuous records, while others (e.g., NZM00093929) have large gaps.

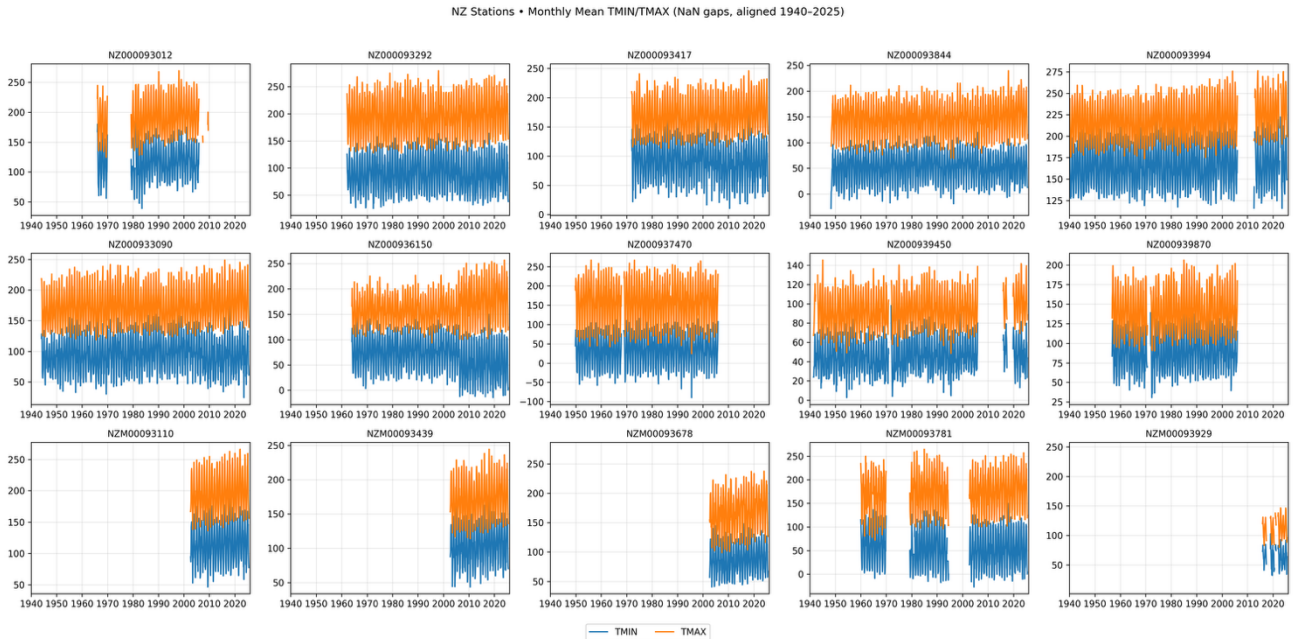


Figure 7: NZ Stations • Monthly Mean TMIN/TMAX (NaN gaps, aligned 1940-2025)

Yearly averages

Monthly means were further averaged to annual means, with a **quality threshold of at least 9 months** per year to prevent bias. This resulted in yearly TMIN/TMAX trajectories for each station (shown in a 15-subplot panel). Clear warming trends are apparent at many stations, although some early records are noisy due to sparse data coverage， referring to Figure 8: NZ Stations • Yearly Mean TMIN/TMAX (NaN gaps via month coverage threshold, aligned 1940-2025.

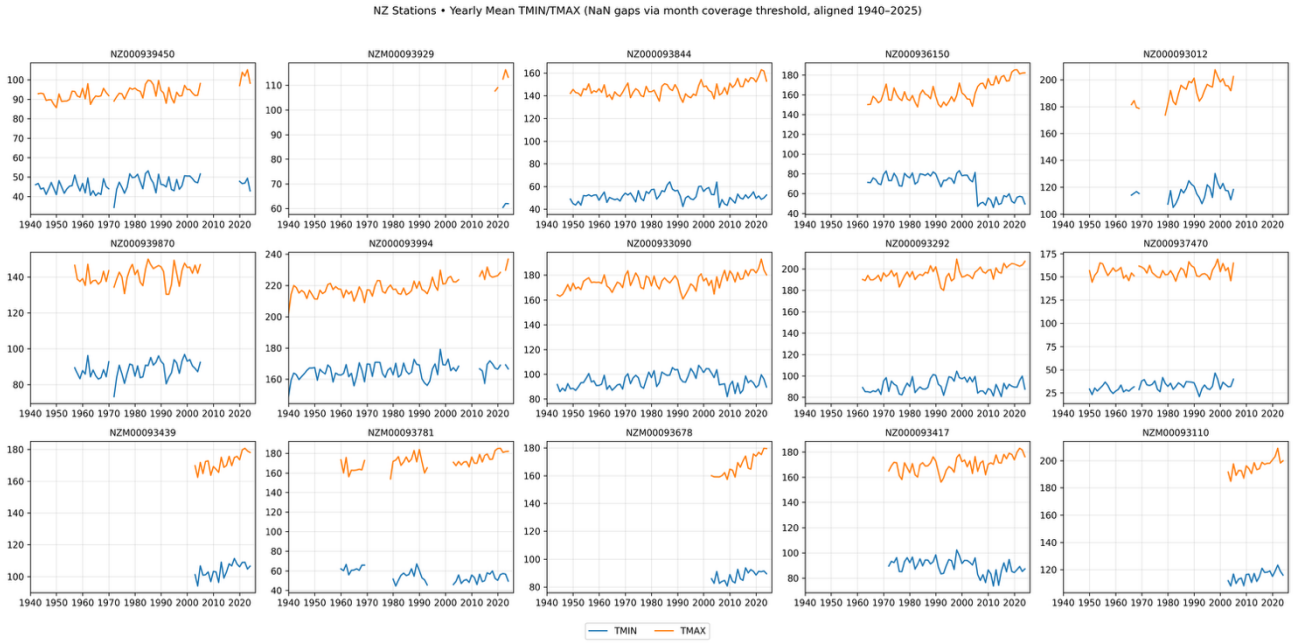


Figure 8: NZ Stations • Yearly Mean TMIN/TMAX (NaN gaps via month coverage threshold, aligned 1940-2025

Gap analysis.

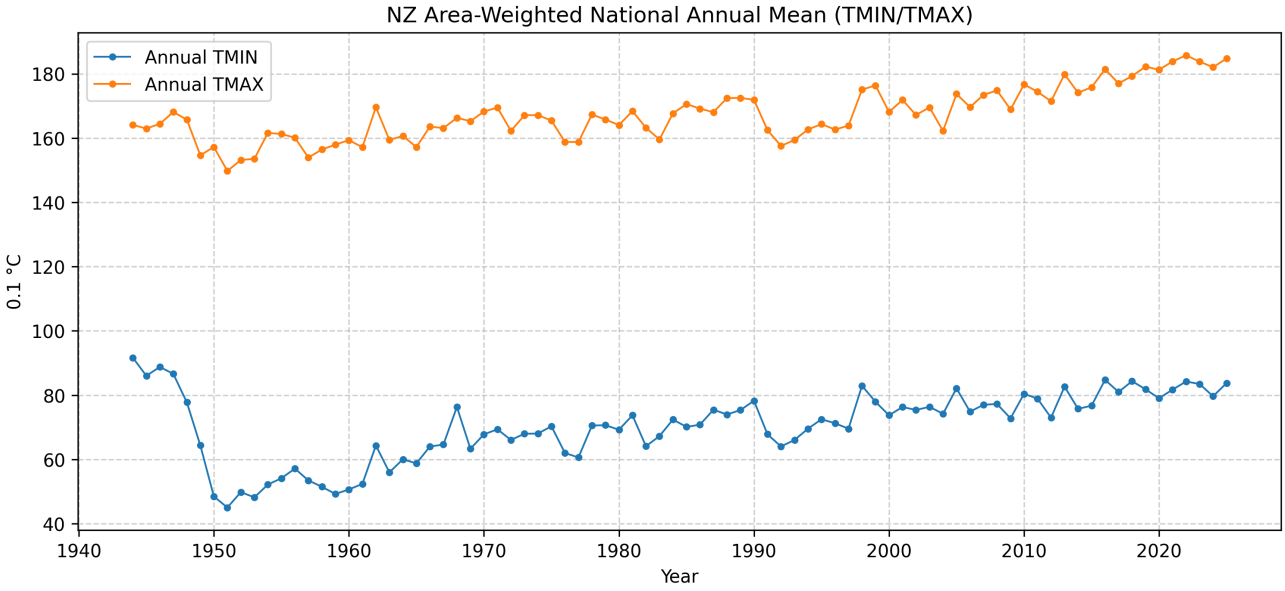
A gap-detection routine compared the expected (station × year × month) with the observed. Results show that certain stations missed entire decades. Others, like NZ000093844 Invercargill, missed ~100 months across its history. Data gaps are concentrated in early decades and highlight challenges in building long-term homogeneous series.

Station ID	missing_year_total	missing_month_toal
NZ000933090	5	48
NZ000093994	10	87
NZ000093844	10	100
NZ000939450	17	176
NZ000093292	24	265
NZ000936150	25	288
NZ000937470	31	364
NZ000093417	33	384
NZM00093781	40	448
NZ000939870	39	450
NZ000093012	56	650
NZM00093678	64	750
NZM00093110	64	752

Station ID	missing_year_total	missing_month_toal
NZM00093439	64	752
NZM00093929	81	937

Table 2: TMAX/TMIN Missing Data Stats

New Zealand National Annual Mean (TMIN/TMAX).



•Methodology: Spatial weighted plus time-slice QC controlled

1.Daily NZ observations F.pivoted to wide format (TMIN, TMAX).
2.For each station-month, the average TMIN/TMAX is calculated if there are at least 20 valid days.
3.Stations projected to NZTM (EPSG:2193); Voronoi tessellation constructed and clipped to New
4.Zealand's boundary.(Burrough et al., 1998)

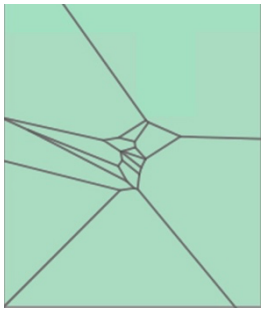


Figure 9: Voronoi Tessellation Shape

Figure 10: NZ boundary

5.Voronoi cell areas are normalized to define spatial weights w_i.
6.For each month, valid stations were **re-weighted and aggregated** to national TMIN/TMAX.

•Results:

Months with missing stations do not bias the averages. A spatial and stratified sampling smooth figure of national monthly and annual means was generated.

Observations are spatially and temporally heterogeneous

According to Figure 3: Uneven Sample Sizes Across Stations (darker = more data): Implications for Nationwide Aggregation, 15 stations are plotted on New Zealand's polygon, coloured by the number of distinct TMIN/TMAX days, showing that station coverage is uneven. Stations like *INVERCARGILL AIRPOR（NZ000093844）*and *NEW PLYMOUTH AWS（NZ000933090）* have long records, while remote stations (e.g., *ENDERBY ISLAND AWS（NZM00093929）*) have fewer observations. This emphasises the importance of weighting and normalisation of availability in national averages.

Global Precipitation

Aggregation to country-year means.

Daily precipitation (PRCP) records were filtered and aggregated to yearly means per station, then averaged across stations within each country. The resulting table contains 17,731 records, covering 218 countries and 17726 unique year–country pairs. A total of 127,610 stations recorded PRCP at least once.

Outlier detection.

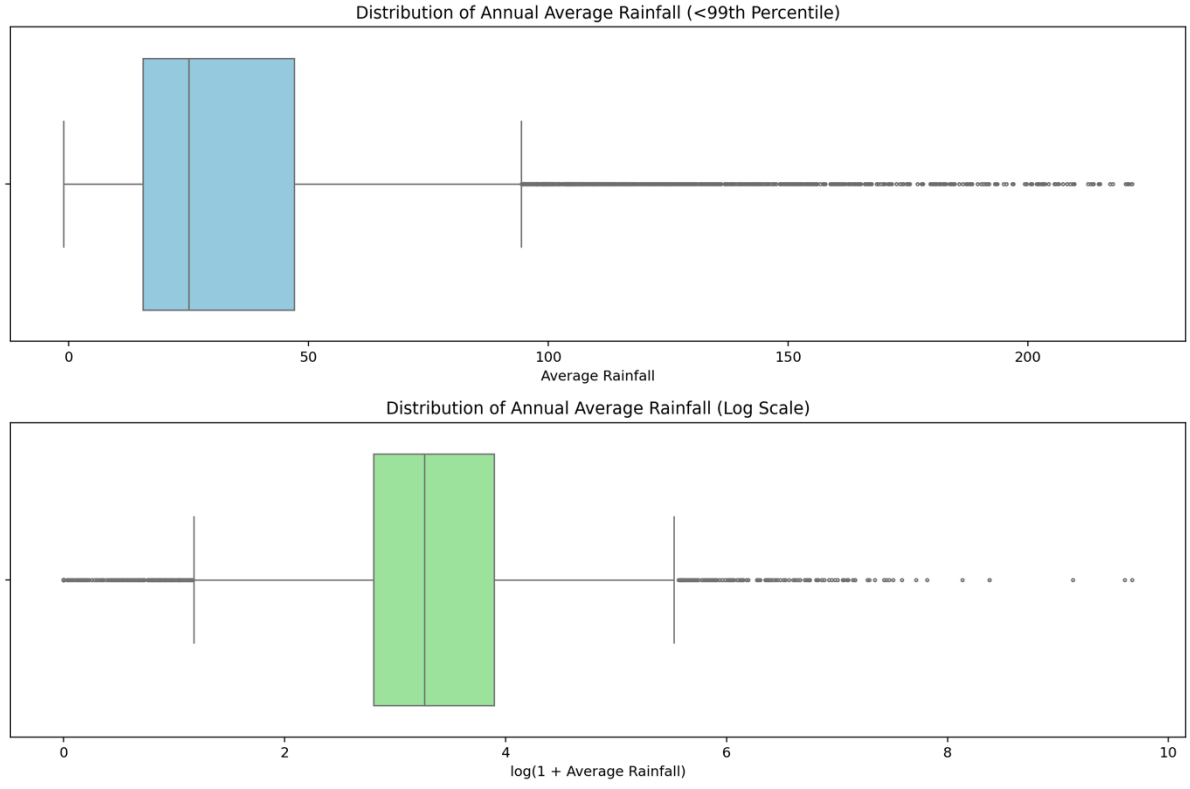


Figure 11: Distribution of annual mean Rainfall by boxplot

Inspection of the Figure 11: Distribution of annual mean Rainfall by boxplot revealed heavy skew. The upper clipped boxplot (<99th percentile) shows that most countries' or year's averages fall between 0–50 mm. The lower log-scale boxplot compresses the extreme outliers, revealing several anomalous points beyond 4000 mm. The United Kingdom recorded a negative observation in 1874, which is apparently illegal.

Descriptive statistics confirmed PRCP skew

Statistic Variale	Value
Count	17,731
Mean	44.39
Std	198.49
Median	25.20
Min	–1.08 (UK, 1874; clearly a placeholder error)
Max	15,875 (year 1952, unmapped country)

Table 3: Overall average rainfall descriptive statistics(units = average daily rainfall, mm/day equivalent)

Negative values and zeros are interpreted as missing-data codes rather than actual precipitation, consistent with GHCN-Daily documentation. These were flagged for cleaning.

Extreme maxima and missing country metadata

The highest recorded country-year rainfall average was in 1952 (15,875 mm), but its country metadata was missing. This suggests an unlinked FIPS country code is present in our dataset, emphasising the need for reliable country-code mapping.

2024 Choropleth.

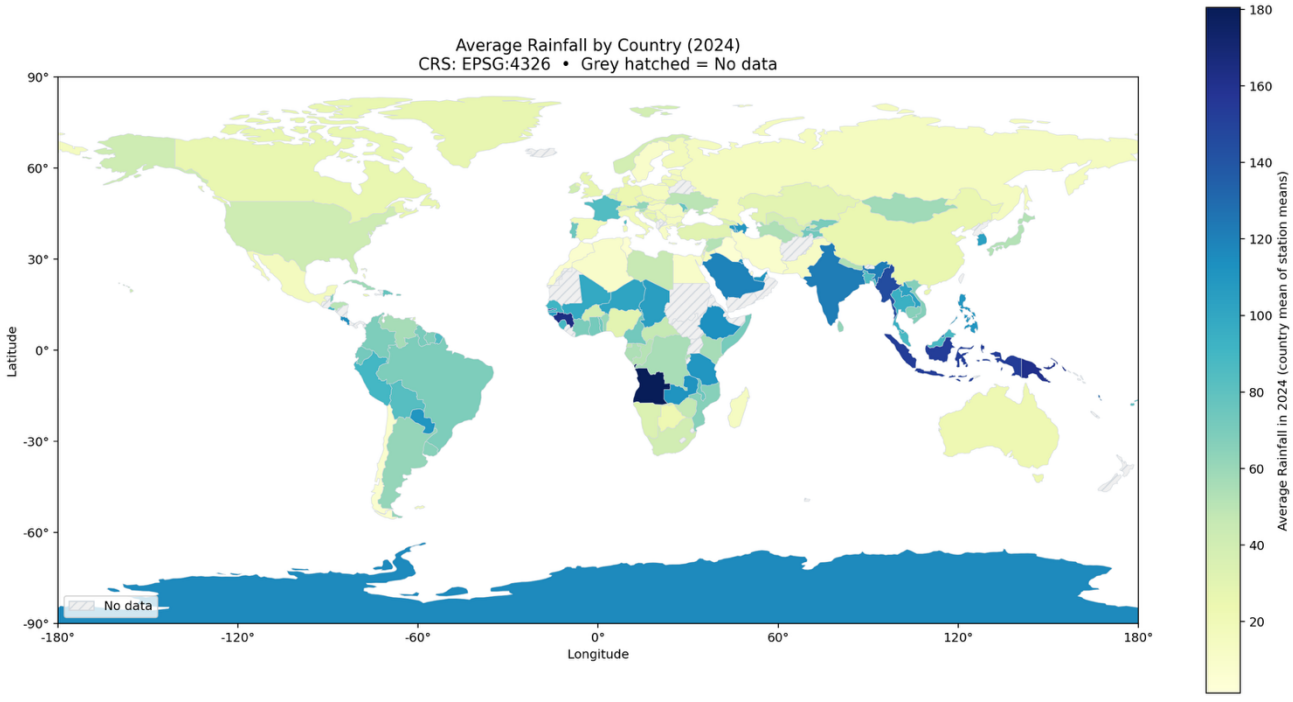


Figure 12: Average Rainfall by Country (2024)

Figure 12 shows patterns that align with climatological expectations. High rainfall in equatorial regions (e.g., Southeast Asia, Central Africa). Low rainfall in deserts (e.g., North Africa, Arabian Peninsula). Temperate regions (e.g., Europe, New Zealand) cluster near the global mean.

Trade-off decisions

I calculated the 2024 national averages using Spark aggregation, joined with station-country metadata, and mapped via FIPS→ISO3 translation. Where FIPS codes lacked direct mapping, we filled them in using spatial joins of average station coordinates to the Natural Earth polygon dataset. Choices and trade-offs I made as shown below:

Projection: Robinson (ESRI:54030) vs EPSG:4326, the former has more balanced aesthetics and lower distortion on a global scale, while the latter has a clear longitude and latitude axis. Failed in trial with cartopy.crs library to plot map in ESRI:54030 and applying EPSG:4326 axes, respectively.

No data Countries: The countries which have no 2024 PRCP observations are shown in grey with hatching rather than at zero level for clearance.

FIPS to ISO transformation: Address stations that failed to map FIPS to ISO_a3 by generating a centroid to represent the locations of those stations sharing the same country code. This partially alleviates the effect of out-of-mainland-territory bias in geolocations.

Colour Scale: YlGnBu was chosen from ColorBrewer as the primary colourmap because it offers an intuitive gradient from dry (yellow) to wet (blue), which corresponds well with rainfall semantics. As alternatives, Viridis (colour-blind friendly), Blues (water-focused), and Cividis (scientifically consistent) were evaluated to ensure readability across different audiences.

1. 5. Conclusions

- Summary of findings and what was learned
- Key insights from processing, analysis, visualizations
- Limitations (e.g., missing data, mapping issues, extreme outliers)
- Possible future improvements (more robust mapping, handling missing observations, larger scale Spark optimizations)

---

2. 6. References

- GHCN-Daily README
- Natural Earth (GeoPandas world dataset)
- PySpark, GeoPandas, Matplotlib/Seaborn/Cartopy docs
- Any external references (e.g., Haversine distance formula)
- AI usage acknowledgment (if required)

3. 7. Appendix (Optional)

- Supplementary tables/figures not included in main body
- Code excerpts (but bulk code should go to supplementary material zip, per grading rules)