Climalytics AT

David Kalteis
s2410455001@fhooe.at
FH Hagenberg
Mobile Computing
Austria

Dominik Forsthuber
s2410455011@fhooe.at
FH Hagenberg
Mobile Computing
Austria

Michael Kerscher
s2410455014@fhooe.at
FH Hagenberg
Mobile Computing
Austria

Abstract

This project analyzes long-term climate trends and extreme weather events in Austria using scalable big data technologies. By processing decades of meteorological data with Apache Spark and visualizing results through Matplotlib, the study investigates temperature shifts, spatial patterns of hot and frost days, data coverage gaps, and optimal environmental conditions across elevation zones.

ACM Reference Format:
David Kalteis, Dominik Forsthuber, and Michael Kerscher. 2025. Climalytics AT. In . ACM, New York, NY, USA, 5 pages.

1 Introduction

Our project focuses on analyzing long-term weather trends and extreme climate patterns in Austria using scalable Big Data technologies. We aim to uncover patterns in temperature, precipitation, and extreme events across regions and over time.

2 Dataset

This dataset contains monthly aggregated climate data from various weather stations across Austria. Each entry includes numerous meteorological measurements (e.g., temperature extremes, precipitation, humidity, sunshine duration, frost days, wind data) over many years, making it suitable for large-scale time series and spatial weather analysis. Time span: January 1970 – April 2025 (monthly resolution)
The dataset used in this study was downloaded from GeoSphere Austria's climate data portal [1].

Downloaded datafiles:

• `climate_all_stations.csv`	676 MB
• `parameter_metadata.csv`	58 KB
• `stations_metadata.csv`	168 KB

3 Technology Stack

Our data processing and analysis pipeline is built upon a suite of robust big-data and visualization technologies. Each component plays a specific role in efficiently ingesting, transforming, querying, and visualizing the weather station data.

- **Apache Spark** We use Spark as the core distributed-compute engine for all large-scale data operations. Its resilient distributed datasets (RDDs) and DataFrame APIs enable parallel processing across our Docker-hosted Spark cluster, allowing us to scale computations over the full 800 MB weather dataset with fault tolerance and in-memory acceleration.
- **PySpark** PySpark serves as the Python interface to Spark, offering the flexibility of Python for data wrangling and the performance of Spark's JVM runtime. All transformation logic—such as filtering by station, aggregating daily summaries, and computing derived meteorological metrics—is implemented using PySpark DataFrame operations.
- **Apache Parquet** To optimize on-disk storage and I/O throughput, we convert the original CSV files into the columnar Parquet format. Parquet's efficient compression and column pruning dramatically reduce the storage footprint and accelerate both full-table scans and selective queries. We partition the Parquet dataset by date and station region to further improve read performance for common query patterns.
- **Spark SQL** Spark SQL provides a familiar SQL dialect for ad-hoc exploration and for expressing complex joins and window functions. We register our Parquet datasets as temporary views, enabling concise, declarative queries to compute historical trends (e.g., rolling averages of temperature) and to filter by metadata attributes such as station elevation.
- **Matplotlib** After extracting query results from Spark into Pandas DataFrames (for manageable result sizes), we employ Matplotlib to generate publication-quality visualizations. Line charts, histograms, and heatmaps illustrate temporal patterns, spatial distributions, and correlations in the meteorological parameters.

4 Research questions

4.1 Long-Term Temperature Trends (RQ1)

4.1.1 Objective. This research question investigates whether long-term temperature trends differ by elevation. To illustrate the phenomenon of elevation-dependent warming, this section focuses specifically on the highest elevation category: **2000+ m (High Alpine)**.

4.1.2 Methodology. Using Apache Spark, monthly average temperature values (`tl_mittel`) from the full dataset `climate_all_stations` were grouped by station and year. Station metadata was joined and used to classify each station into one of five predefined elevation bands:

- 0–499 m (Lowland)
- 500–999 m (Upland)
- 1000–1499 m (Lower Alps)

Big Data, FH Hagenberg

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

- 1500–1999 m (Alpine)
- 2000+ m (High Alpine)

Two additional plots were generated for context:

- A distribution plot of all stations by elevation zone and region.
- A labeled diagram of the highest station per region to confirm extreme values.

4.1.3 *Results and Interpretation.*

Figure 1: Temperature Trends in the High Alpine Zone. The line plot below shows annual average temperatures from 1970 to 2025 for each region in the High Alpine zone. A general increase is evident. The curves show irregularities, stagnation phases, and sudden drops—indicating either microclimatic variability or data limitations.

In addition, abrupt changes and missing values are visible at several lines. These are most likely caused by:

- sensor malfunctions,
- missing or incomplete historical data,
- or gaps due to station shutdowns.

These patterns emphasize the importance of pairing trend analysis with a careful check of data completeness.

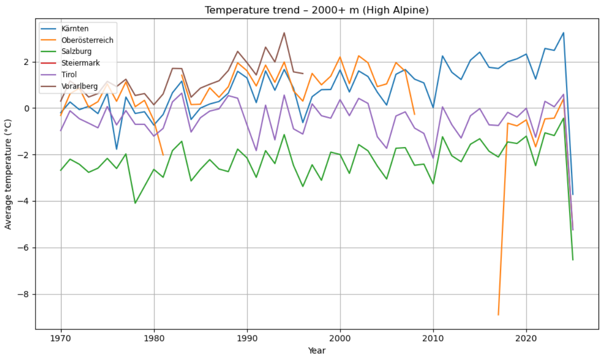

Figure 1: Average annual temperature trends (1970–2025) in 2000+ High Alpine zone

Figure 2: Altitude Distribution of Weather Stations. This scatterplot validates that stations assigned to the High Alpine zone are located well above 2000+. The distribution also shows regional diversity, which supports cross-regional comparisons.

Figure 3: Highest Stations per Region. This annotated graphic confirms the locations and altitudes of the weather stations in the two highest altitude zones per federal state.

4.1.4 Conclusion. The high alpine zone (2000+) in Austria shows a clear warming trend over the last five decades. Regional fluctuations are present, but do not affect the overall pattern. Exactly the same analysis was carried out for the four lower altitude zones; these analyses are included in the Jupyter Notebook.

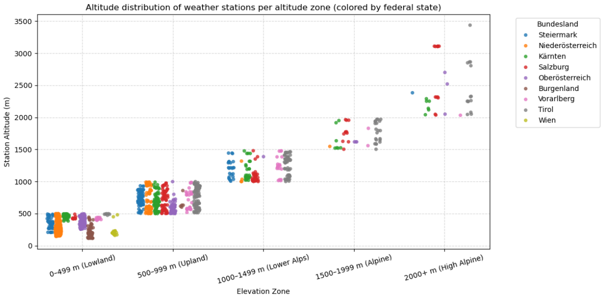

Figure 2: Elevation distribution of stations by zone and region

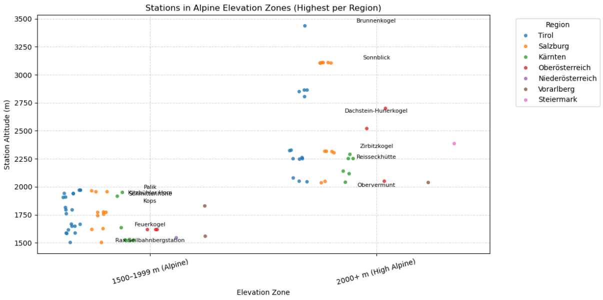

Figure 3: Highest station per region in Alpine zones with labels

4.2 Spatial Patterns of Extremes (RQ2)

This question examines how the frequency of very hot days (>= 30 °C) and frost days (<= 0 °C) has changed since 1970 in three elevation-based geographic zones: *valley* (<= 700 m), *plateau* (701–1500 m), and *alpine* (> 1500 m).

4.2.1 *Data Processing.*

1. **Load and transform:** Raw CSV data were ingested into a Spark DataFrame, with parsing of the "date" field and extraction of `year` for each station.
2. **Parquet conversion:** The joined DataFrame (including elevation and zone labels) was repartitioned by `year` and `zone` and written to Parquet for efficient subsequent queries.

4.2.2 *Data Analysis.*

Full Time Series. We computed the annual, per-station average counts of hot days (see Figure 4) and frost days (see Figure 5) for each zone.

End–Minus–Start Difference. To highlight net shifts, we compared the mean of 2014–23 vs. 1970–79 per zone, yielding the change in average hot- and frost-days (see Figure 6).

Linear Regression Trend. Finally, we fitted a simple linear model (OLS) of day-count vs. year for each zone, extracting the slope (days per year), R^2, and p–value to quantify rate and significance of change (see Figure 7).

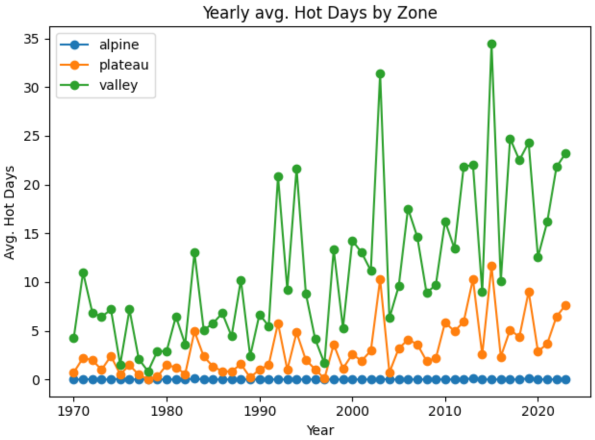

Figure 4: Yearly average hot-day counts per station, by zone (1970–2023).

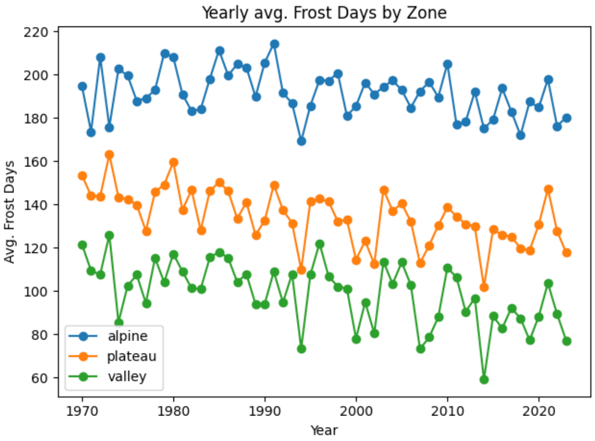

Figure 5: Yearly average frost-day counts per station, by zone (1970–2023).

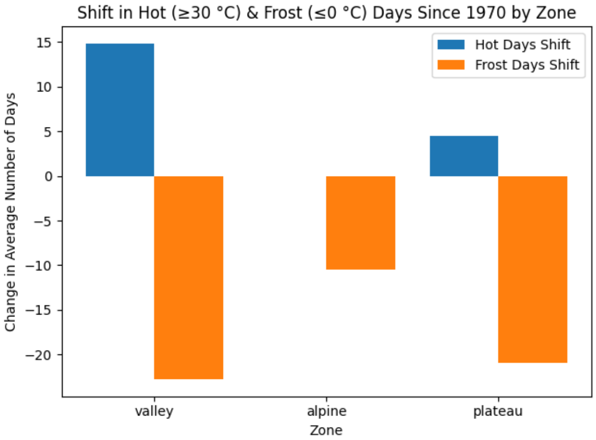

Figure 6: Change in average hot and frost days (2014–23 minus 1970–79) by zone.

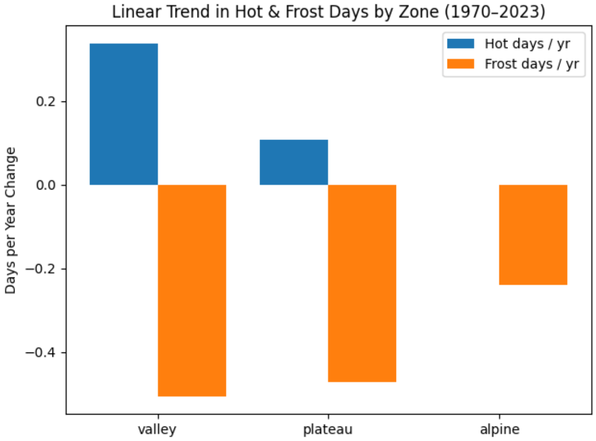

Figure 7: Estimated trend slopes in hot- and frost-day counts (days per year) by zone.

4.3 Spatio-Temporal Data Coverage Profiling (RQ3)

4.3.1 Objective. The third research question investigates how the installation dates and operational periods of weather stations affect data availability over time and across Austrian regions and elevation zones. The goal is to identify spatial and temporal coverage gaps, also referred to as "data deserts," that could affect the interpretation of long-term climate analyses.

4.3.2 Methodology. To address this question, the implementation proceeded in two main steps:

(1) **Metadata-Based Coverage Matrix:** Using station metadata (installation and deactivation dates), a year-by-year activity matrix from 1970 to 2025 was constructed via a cross join with the full year range. Only periods when stations were active were retained. Each station was then categorized into one of five elevation zones (as in RQ1). Aggregated counts by year, elevation zone, and federal state were used to provide a theoretical view of station availability.

(2) **Real Measurement-Based Coverage:** To assess actual data availability, the climate dataset was filtered to include only records with at least one valid measurement. These were grouped using the same year–elevation–region schema as the metadata-based matrix.

In this report, the results are shown for the Lower-Alps zone (1000–1499 m). Results for the other four zones are available in the accompanying Jupyter Notebook.

4.3.3 Results and Interpretation.

Figure 8: Metadata-Based Coverage (Lower Alps). The heat map shows that Carinthia, Salzburg, and Styria have maintained a consistent network of 7 to 16 active stations per year since the 1970s. In contrast, regions such as Lower Austria and Upper Austria are either absent or only sporadically represented in this altitude zone.

Vorarlberg shows moderate coverage from around 2008 onward, while Tyrol stands out with a well-developed and continuous station network throughout the entire period.

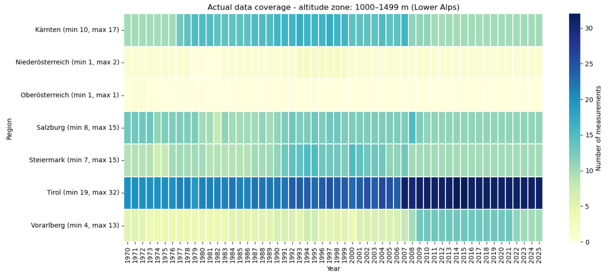

Figure 8: Metadata-based station coverage in elevation zone: 1000–1499 m (Lower Alps)

Figure 9: Actual Measurement-Based Coverage (Lower Alps). The second heatmap confirms that actual measurement coverage aligns closely with the metadata. Again, Tyrol is best represented, while several eastern federal states exhibit minimal or no active data-producing stations in this elevation zone.

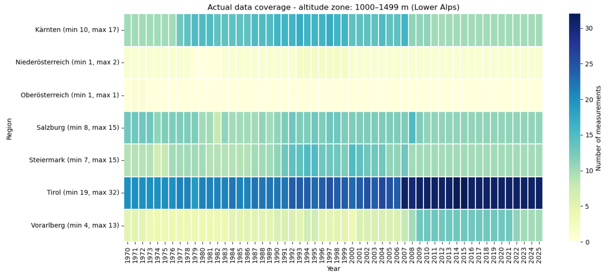

Figure 9: Actual measurement-based coverage in elevation zone: 1000–1499 m (Lower Alps)

4.3.4 Conclusion. Both the metadata-based and measurement-based heat maps confirm persistent long-term data gaps in the Lower Alps (1000–1499 m), particularly in Upper and Lower Austria.

In other elevation zones as well, the two coverage representations generally agree. However, minor differences in shading between the two heat maps reveal important nuances: lighter shading in the measurement maps compared to metadata suggests technically active stations with little or incomplete data. Conversely, darker shading in the measurement maps may indicate incorrect metadata or other data quality issues.

These findings underline the importance of validating metadata-based assumptions against actual measurement data—especially for high-resolution analyses or policy-relevant climate studies.

4.4 Operational Window Optimization (RQ4)

This section addresses the research question: *Which seasonal windows and locations optimize safety—combining sunshine hours, wind-gust flags, and frost/heat indicators?*

4.4.1 Calculation Method.

The safety score is a composite metric designed to evaluate environmental safety conditions for each weather station across Austria, using data from the last five years only. It combines the following variables:

- **Average Sunshine Hours**: Higher values are considered favorable.
- **Frequency of Wind Gusts**: Higher values are penalized due to increased hazard potential.
- **Frequency of Frost Days**: Higher values indicate harsher conditions and are penalized.
- **Frequency of Heat Days**: Moderately penalized to reflect discomfort and potential risk.

The score is calculated as shown in equation 1.

$$\text{Score} = \text{Sunshine} - \text{Wind Gust Freq} - \text{Frost Freq} - \text{Heat Freq} \quad (1)$$

A higher safety score indicates more favorable and stable environmental conditions. The score is then normalized and scaled to 0–100 for comparability.

4.4.2 Top Station Scores.

Number of stations used: 1134. Table 1 lists the top 10 station-season combinations ranked by safety score within the last five years. These stations span multiple regions and elevation zones.

ID	**Season**	**Region**	**Elevation**	**Avg Sun**	**Safety Score**
25	Winter	Stmk	1000–1499 m	38.0 h	100.0
205	Spring	OÖ	1500–1999 m	85.94 h	100.0
82	Winter	Sbg	2000+ m	177.59 h	100.0
66	Spring	T	1500–1999 m	75.13 h	100.0
173	Winter	Vlg	1000–1499 m	38.0 h	100.0
82	Spring	Sbg	2000+ m	228.75 h	100.0
88	Winter	NÖ	500–999 m	41.81 h	100.0
68	Winter	Sbg	1500–1999 m	141.25 h	100.0
213	Winter	Sbg	2000+ m	221.76 h	100.0
82	Summer	Sbg	2000+ m	43.20 h	100.0

Table 1: Top 10 Safety Scores by Station and Season (Last 5 Years)

4.4.3 Visual Interpretation.

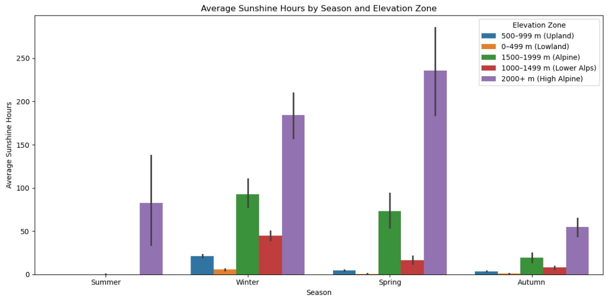

Figure 10: Average Sunshine Hours by Season and Elevation Zone (Last 5 Years)

Figure 10 shows the high alpine zone consistently records higher average sunshine hours across all seasons, especially in spring. This seems plausible because they are above the cloud cover.

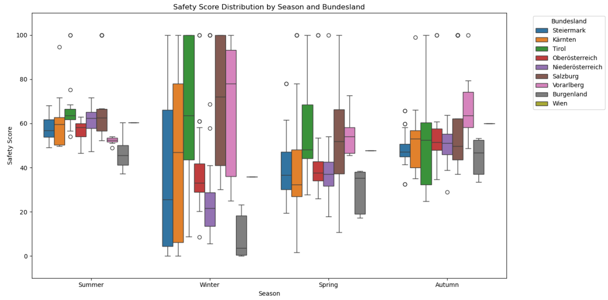

Figure 11: Safety Score Distribution by Season and Region (Last 5 Years)

Figure 11 presents the distribution of safety scores filtered to recent years. Salzburg, Tirol, and Vorarlberg show the highest spread and highest values in winter and spring, corresponding to high-elevation stations.

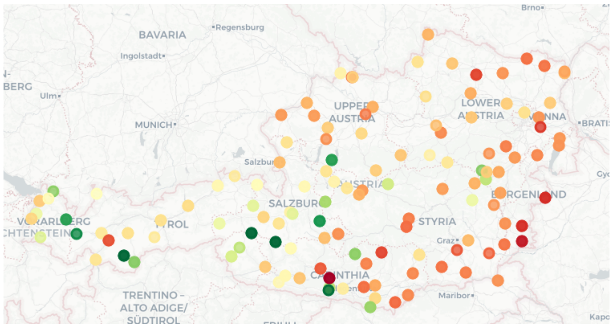

Figure 12: Map of Stations by Safety Score (Green: High, Red: Low, Last 5 Years)

Figure 12 visualizes the spatial distribution of safety scores over the last five years. Green points indicate stations with relatively high safety scores, predominantly in Alpine regions. Red points represent stations with lower scores, more frequent in lower elevation areas.

References

[1] GeoSphere Austria. 2025. klima-v2-1m: Monthly Weather Station Data for Austria. https://dataset.api.hub.geosphere.at/app/frontend/station/historical/klima-v2-1m. Accessed: 2025-05-17. Licensed under CC BY 4.0.