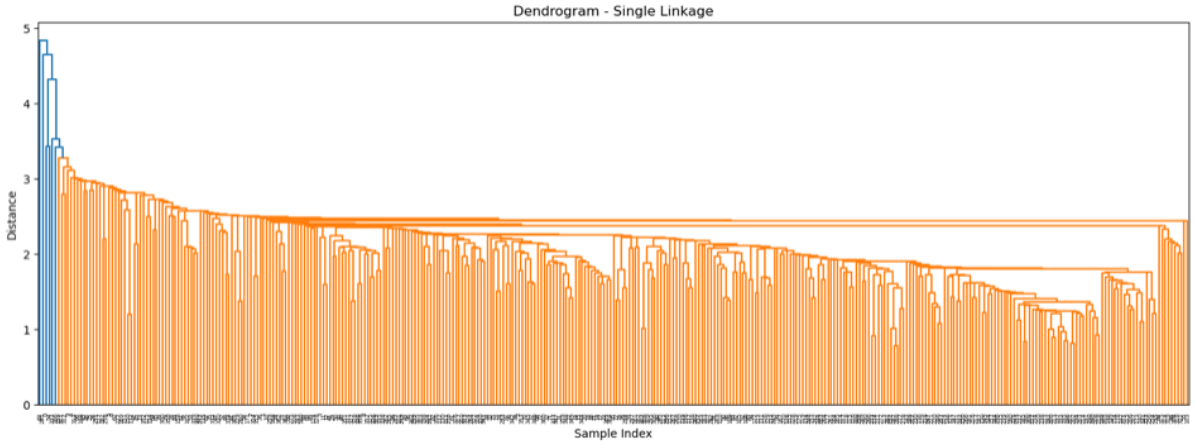
## 2.1 Unsupervised Learning Algorithms

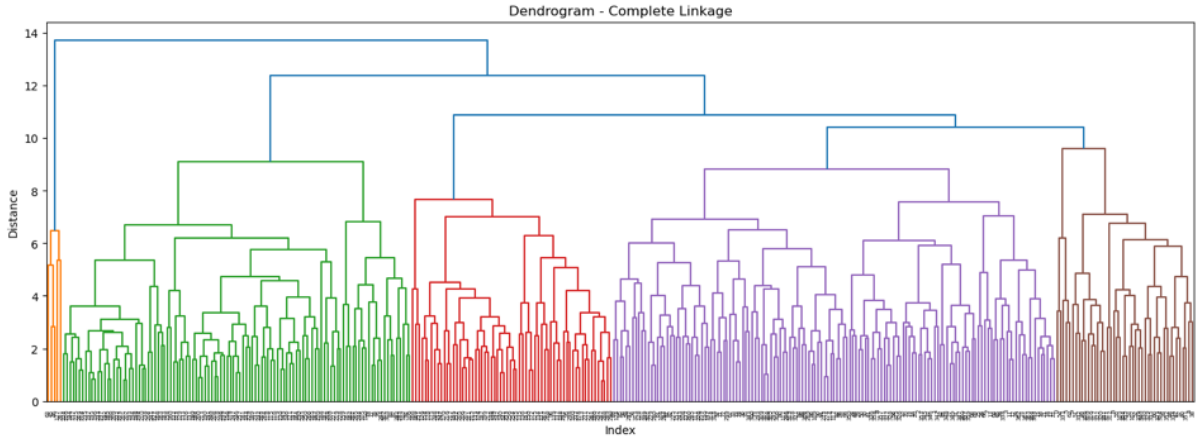
**Project Overview**
In this analysis, I used unsupervised machine learning techniques to explore 2010 weather data from multiple European stations. The objective was to uncover patterns or clusters in temperature and related variables that might align with "pleasant weather" conditions, providing ClimateWins with insights into potential trends or shifts. I applied hierarchical clustering using four different linkage methods — single, complete, average, and Ward — to both a subset of contrasting cities (Madrid and Belgrade) and the full dataset of weather stations. The resulting dendrograms were used to visualise and compare the effectiveness of each method in revealing meaningful structure in the data.

**Scaled Data**
**Single Linkage**



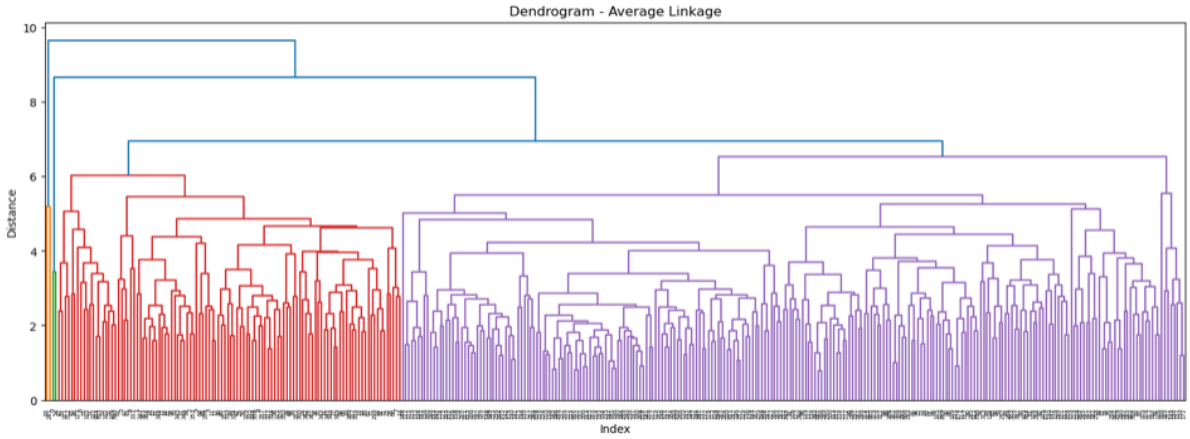

Merges clusters based on the closest pair of points.

**Complete Linkage**



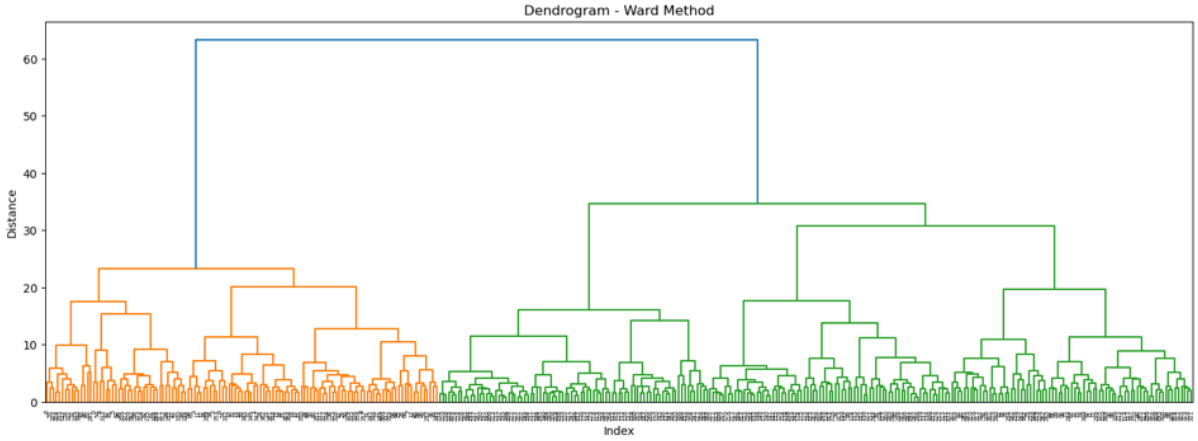

Uses farthest pair of points between clusters.

## Average Linkage



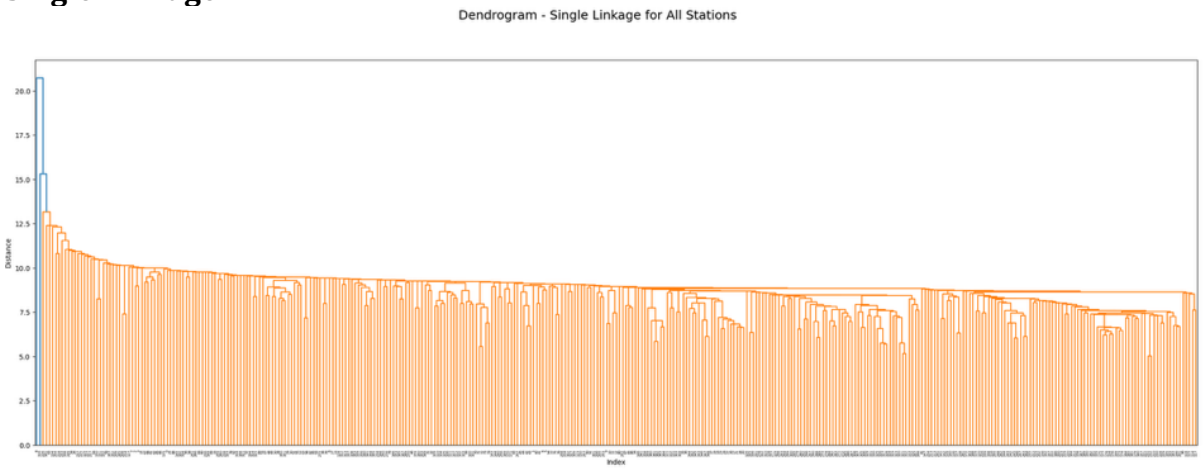

Considers the average distance between all points in each cluster.

## Ward Method

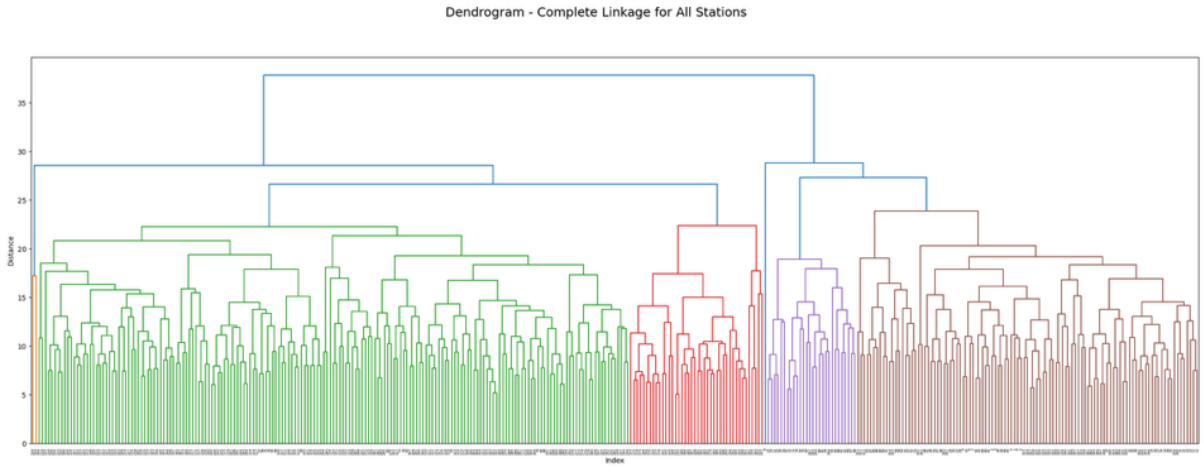



Minimises the variance between clusters by reducing the sum of squares within clusters.

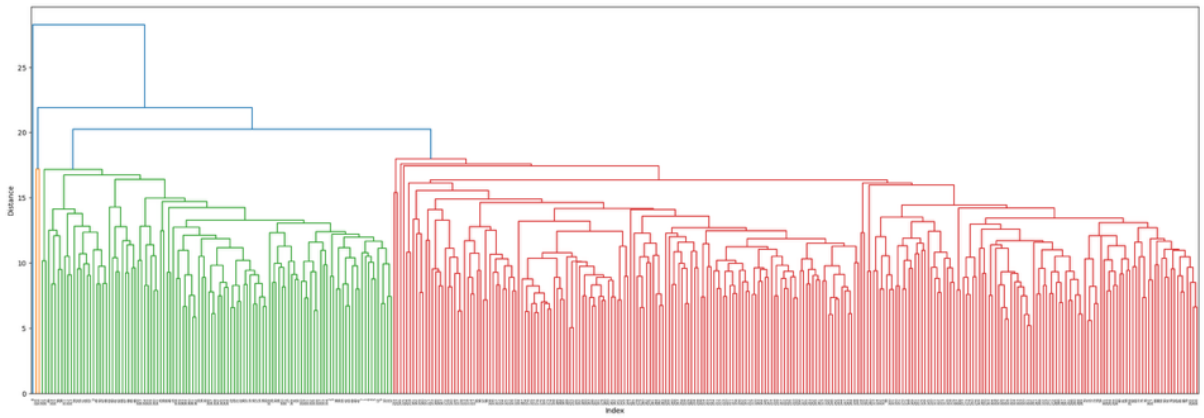
## Unscaled Data
## Single Linkage



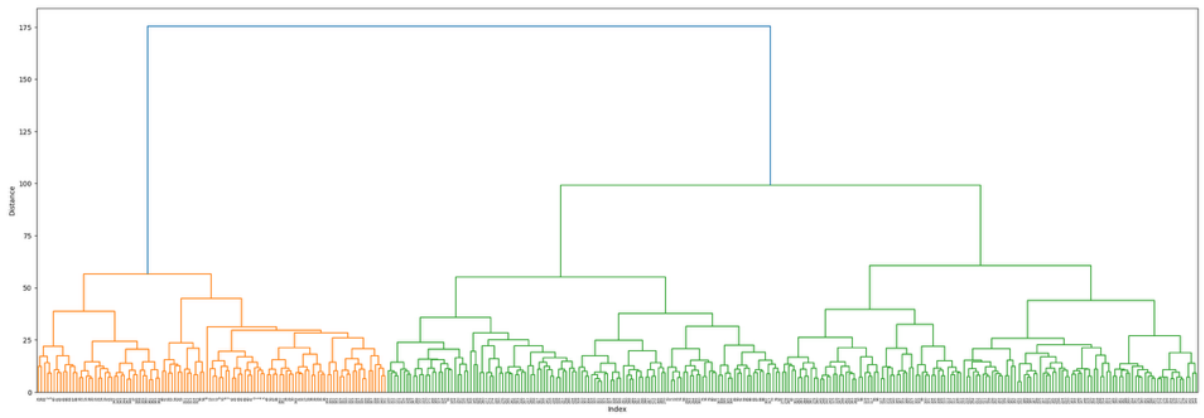
## Complete Linkage




## Average Linkage




## Ward Method



**Insights:**
When applying hierarchical clustering to the 2010 weather data, each linkage method revealed different structural patterns, particularly when comparing scaled and unscaled inputs. On scaled data, Single Linkage initially grouped almost all observations into one large cluster, but quickly fragmented into hundreds of small clusters as the threshold decreased. This reflects its sensitivity to chaining effects and its tendency to over-fragment in the presence of subtle variations, making it less suitable for this dataset.

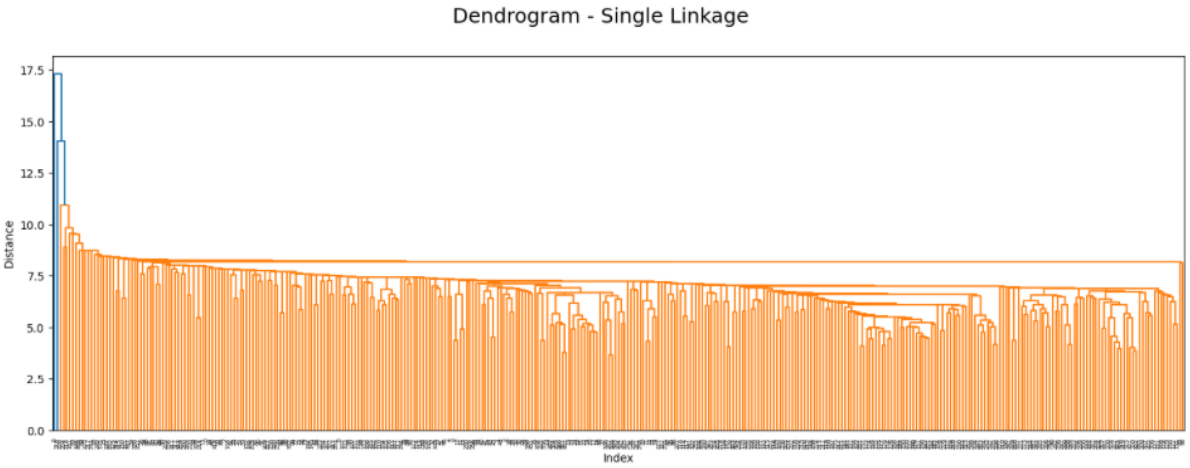
In contrast, Complete Linkage produced more meaningful results, forming five distinct clusters at a threshold of 10, with a relatively balanced distribution. As the threshold decreased, it revealed finer patterns without breaking down as abruptly as Single Linkage. Average Linkage showed a similar progression, though with slightly less clear separation between clusters. Both methods suggest potential groupings based on seasonal or regional weather characteristics.

Ward's Method was the most effective at capturing both broad and detailed structure. At a threshold of 10, it formed 20 clusters, though many of these were relatively small. Visually, the results appeared dominated by two large clusters—likely corresponding to major seasonal or geographic groupings—while the remaining clusters captured more subtle variations. This suggests that Ward's Method not only distinguishes overarching patterns but also preserves finer-grained differences within the data. Its gradual increase in cluster count at lower thresholds reflects its strength in revealing hierarchical relationships, from broad divisions down to specific weather nuances. As the threshold lowered, the number of clusters increased gradually and systematically, reaching 358 at the lowest threshold. This smooth transition highlights Ward's strength in revealing hierarchical relationships within the data, from overarching climate trends to localised patterns.

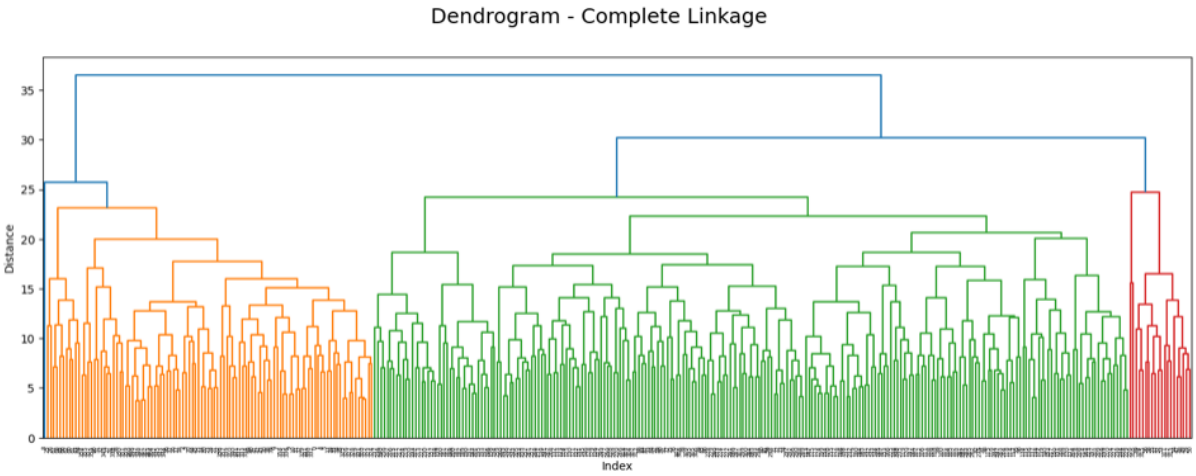
Results on unscaled data were more extreme. Due to the wide range in feature values (e.g. temperature vs. precipitation), most methods quickly broke into the maximum number of clusters, especially at lower thresholds. While Ward's Method retained some structure at higher thresholds, these results underscore the importance of scaling: without it, clustering becomes dominated by high-variance features, reducing interpretability.

Overall, Complete Linkage and Ward's Method on scaled data offered the most insightful and interpretable results. Complete Linkage effectively identified moderately distinct groups, while Ward's Method excelled in capturing both broad and fine-grained patterns. Together, they provide a robust framework for understanding how weather conditions naturally cluster over time and geography.

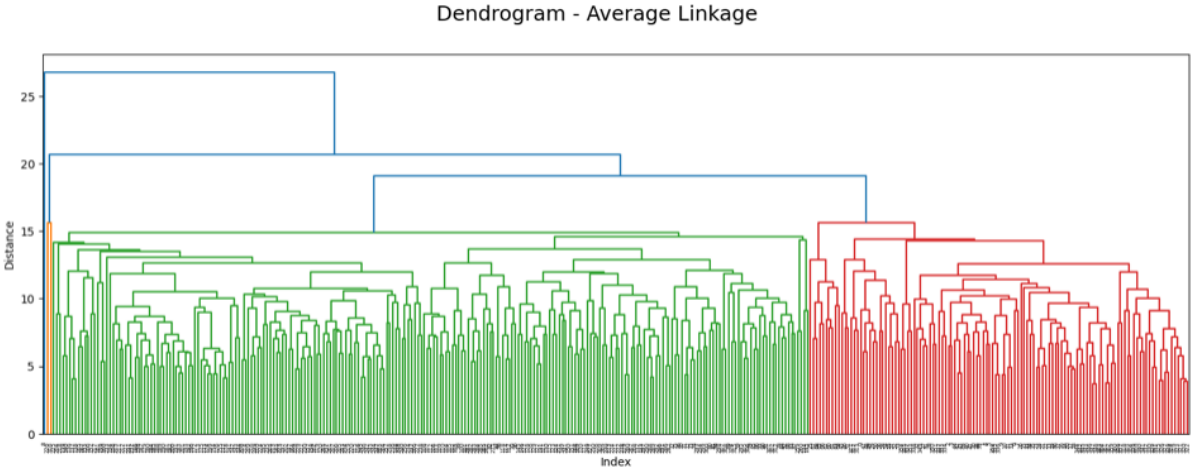
**Reduced Data**
**Single Linkage**



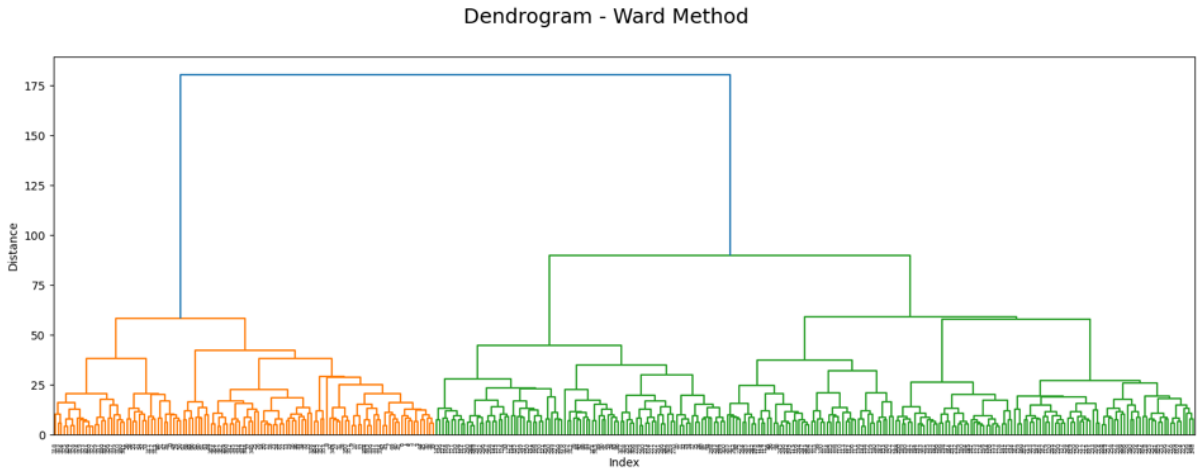

## Complete Linkage




## Average Linkage



**Ward Method**




**Insights:**
After reducing the dimensionality of the 2010 weather dataset using PCA, I applied hierarchical clustering with four different linkage methods: single, complete, average, and Ward. The aim was to explore whether these techniques could reveal meaningful groupings in the data—particularly those related to "pleasant weather" conditions—while addressing previous issues of fragmentation and imbalance seen in non-reduced data.

Single linkage remained the least effective. At a threshold of 10, it produced only four clusters, but rapidly fragmented into 328 by threshold 5 and reached the maximum of 365 clusters by threshold 2. This sharp rise reflects its well-known chaining effect, where minor differences between data points cause excessive splitting. As in earlier experiments, this method struggled to produce stable or interpretable groupings, even after PCA.

Complete linkage performed more effectively, creating 117 clusters at threshold 10 and gradually increasing to 365 at lower thresholds. While still prone to fragmentation at lower thresholds, its clustering structure was more coherent than single linkage, suggesting some underlying divisions—possibly related to seasonal or regional weather trends—became clearer after PCA.

Average linkage provided a more compact structure, forming 69 clusters at threshold 10 and 336 by threshold 5. Like complete linkage, it reached the maximum of 365 clusters at threshold 2, but the slower increase indicates improved cohesion. The larger initial clusters suggest the method successfully grouped days with broadly similar weather patterns, making it one of the more interpretable approaches on the reduced dataset.

Ward's method again proved to be the most balanced and informative. It formed 123 clusters at threshold 10, increasing steadily to 338 at threshold 5 and 365 at threshold 2. The gradual and structured progression of cluster counts indicates Ward's strength in capturing hierarchical relationships. The clusters it produced likely represent broader

divisions such as seasonality, or possibly days that share similar “pleasant” or “unpleasant” weather features.

Compared to earlier clustering on scaled and unscaled data without PCA, the PCA-reduced dataset led to clearer and more structured clustering results. Previously, unscaled data resulted in rapid over-fragmentation, and even scaled data suffered from instability in some methods. PCA helped mitigate these issues by reducing noise and highlighting core variance in the dataset. Ward’s method, already the most robust in previous trials, benefited further from dimensionality reduction.

In conclusion, applying PCA prior to clustering improved both the stability and interpretability of results. Among the methods, Ward and average linkage offered the most meaningful groupings, with Ward particularly well-suited for identifying broad climate structures—potentially even correlating with “pleasant weather” days. These findings reinforce the value of dimensionality reduction when exploring complex environmental data.