Unsupervised Learning Approach

Unsupervised learning techniques were applied to explore long-term patterns in European weather data without predefined labels. This approach is appropriate given the absence of target variables and the goal of identifying natural groupings among weather observations. The analysis focuses on hierarchical clustering to examine similarities among weather stations and conditions, with dimensionality reduction using Principal Component Analysis (PCA) applied to assess whether reducing the feature space improves cluster interpretability.

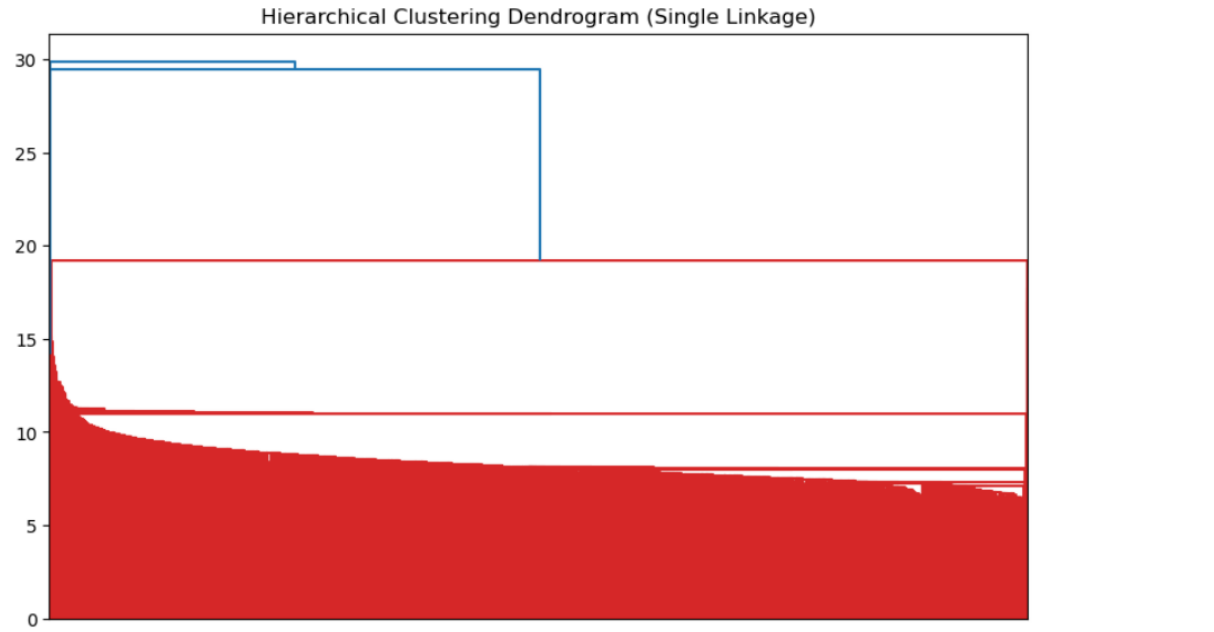
To manage computational complexity, a subset of the data corresponding to a single decade was selected, following guidance to reduce dendrogram runtime while preserving meaningful structure.

For this analysis, data from the 1980s were selected to reduce computational complexity while preserving representative weather patterns.

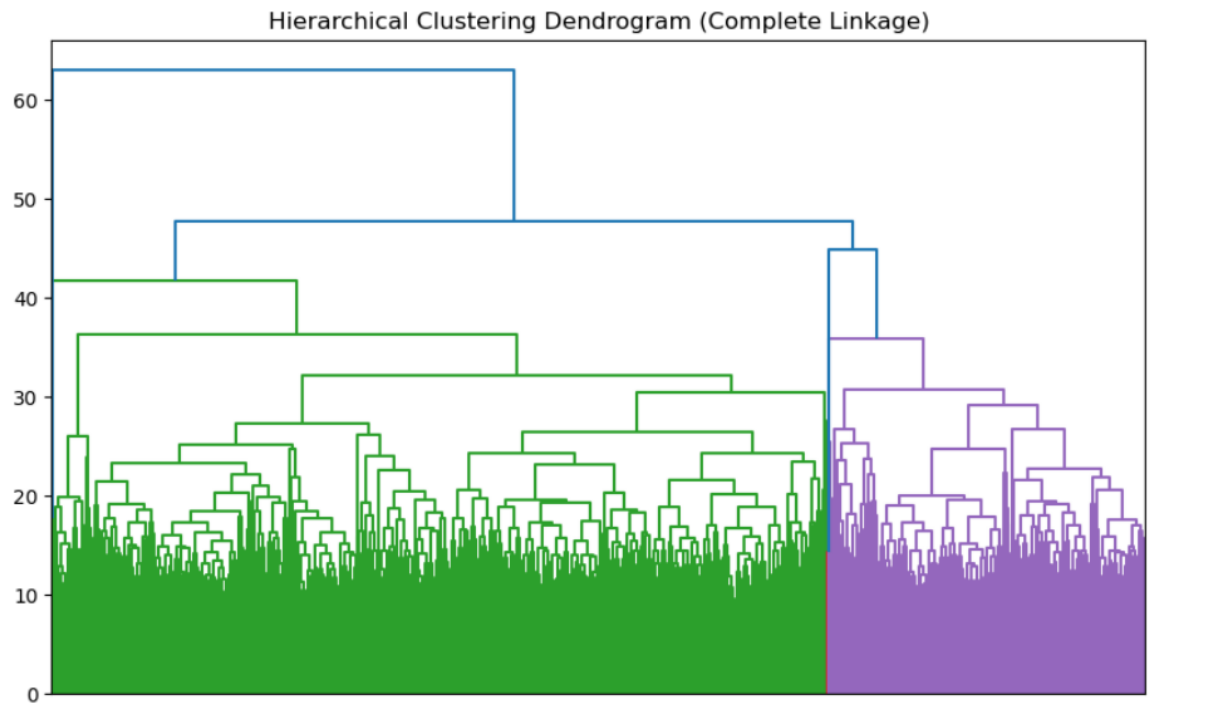
Hierarchical Clustering on Original Data (Scaled)

Hierarchical clustering was first applied to the scaled weather data using four linkage methods: single, complete, average, and ward. Scaling was performed using the StandardScaler to ensure that all meteorological variables contributed equally to distance calculations.

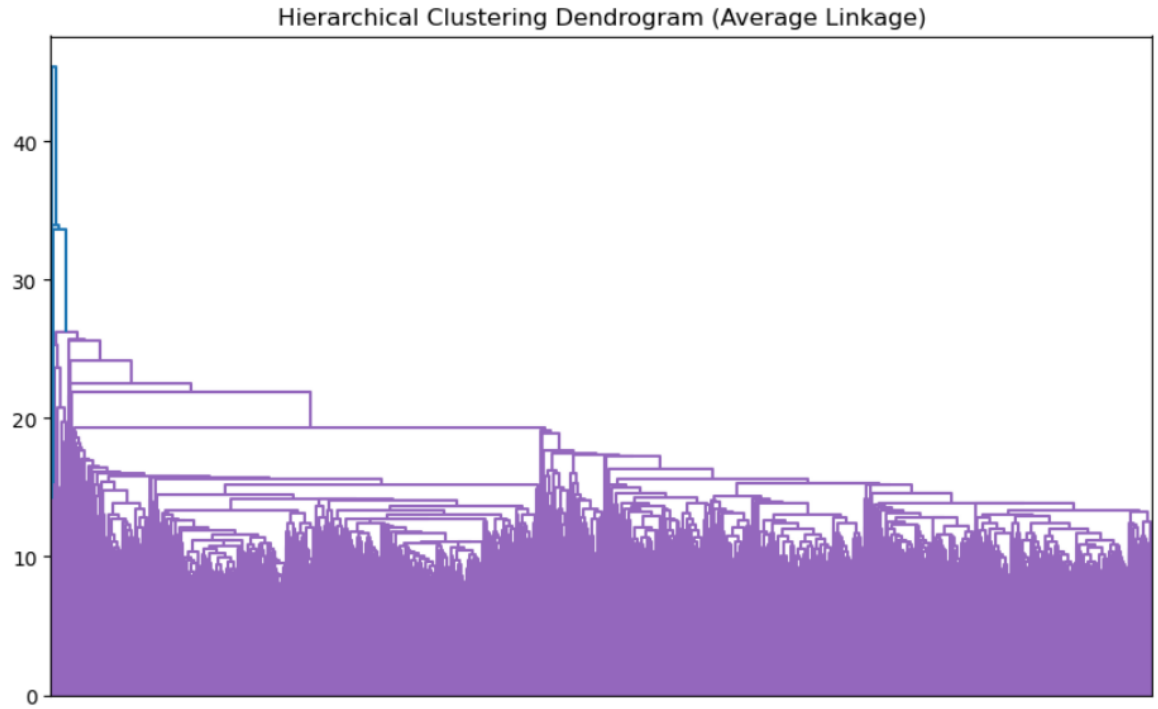
- Single linkage



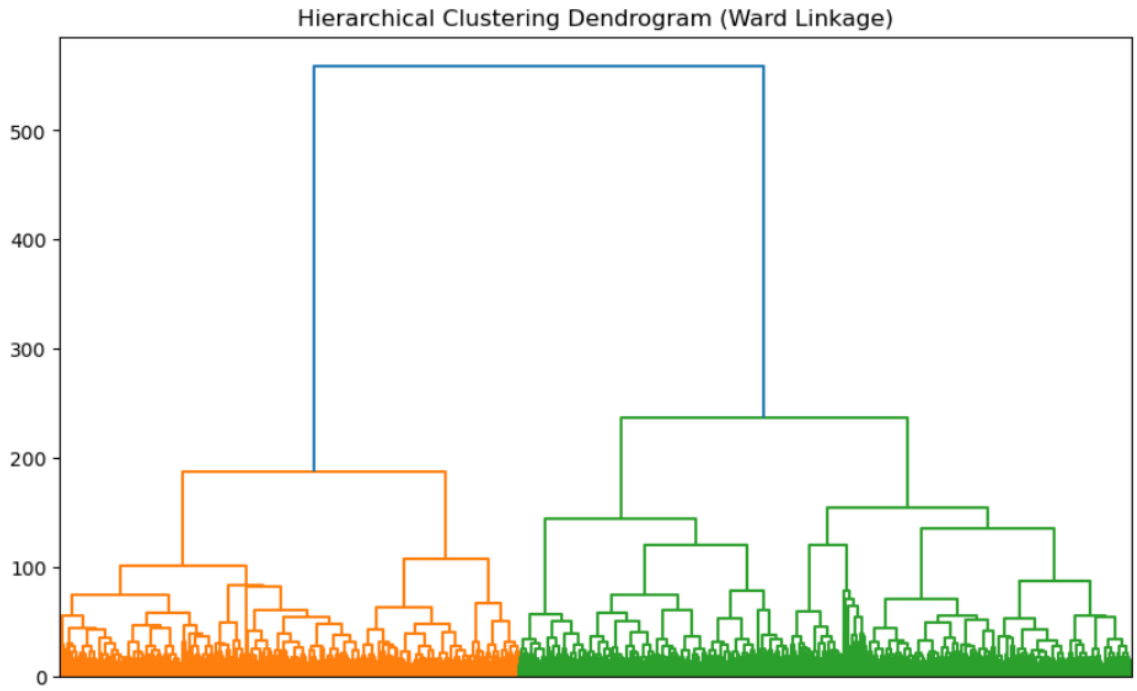
- Complete linkage



- Average linkage



- Ward linkage



Visual Interpretation of Dendrograms

Single linkage produced a pronounced chaining effect, with most observations merging gradually at low linkage distances. This made it difficult to identify clear cluster boundaries, rendering single linkage the least informative method for this dataset.

Complete linkage showed improved separation, with large groupings forming at higher linkage distances. While clearer than single linkage, the dendrogram still displayed substantial internal structure within clusters.

Average linkage resulted in gradual merges with fewer distinct break points, suggesting ambiguous cluster boundaries under this method.

Ward linkage produced the most interpretable dendrogram. Distinct, high-distance merges were visible near the top of the hierarchy, indicating that a small number of clusters captured major structural differences in the data.

Cluster Counts and Quantitative Comparison

To support visual interpretation, dendrograms were cut at multiple linkage distances and the number of resulting clusters was recorded. At low cut heights, all methods produced a very large number of clusters, reflecting fine-grained distinctions among observations. As cut height increased, the number of clusters decreased substantially, revealing higher-level groupings.

Complete linkage showed a reduction from dozens of clusters at moderate cut heights to approximately five clusters at higher cut heights, suggesting a small number of broad weather regimes. Average linkage aggregated observations more aggressively, producing as few as two to four clusters at higher cut heights. Ward linkage exhibited many clusters at low cut heights but progressively merged observations as distance increased, consistent with its variance-minimizing behavior.

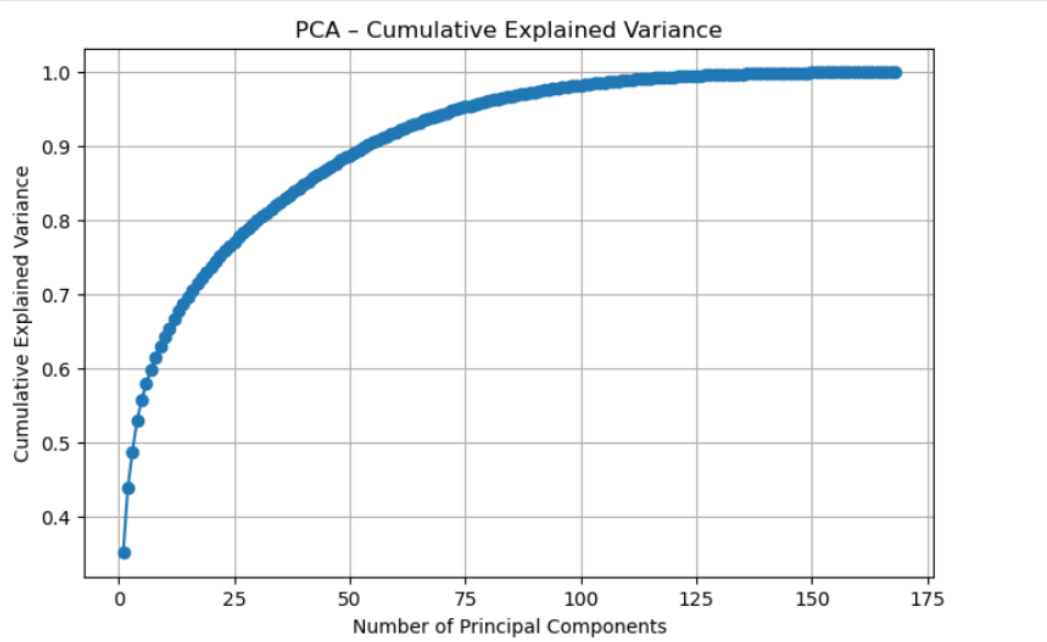
For example, under complete linkage, increasing the cut height reduced the number of clusters from dozens at moderate distances to approximately five clusters at higher distances, indicating the emergence of broad weather regimes.

Overall, ward linkage provided the most interpretable balance between cluster separation and internal cohesion when considered alongside the dendrogram visualizations.

Dimensionality Reduction Using PCA

Because ClimateWins has limited computational capacity, PCA was applied to reduce the dimensionality of the dataset and assess whether clustering results improved when noise and correlated variables were removed.

Figure 1. PCA Cumulative Explained Variance



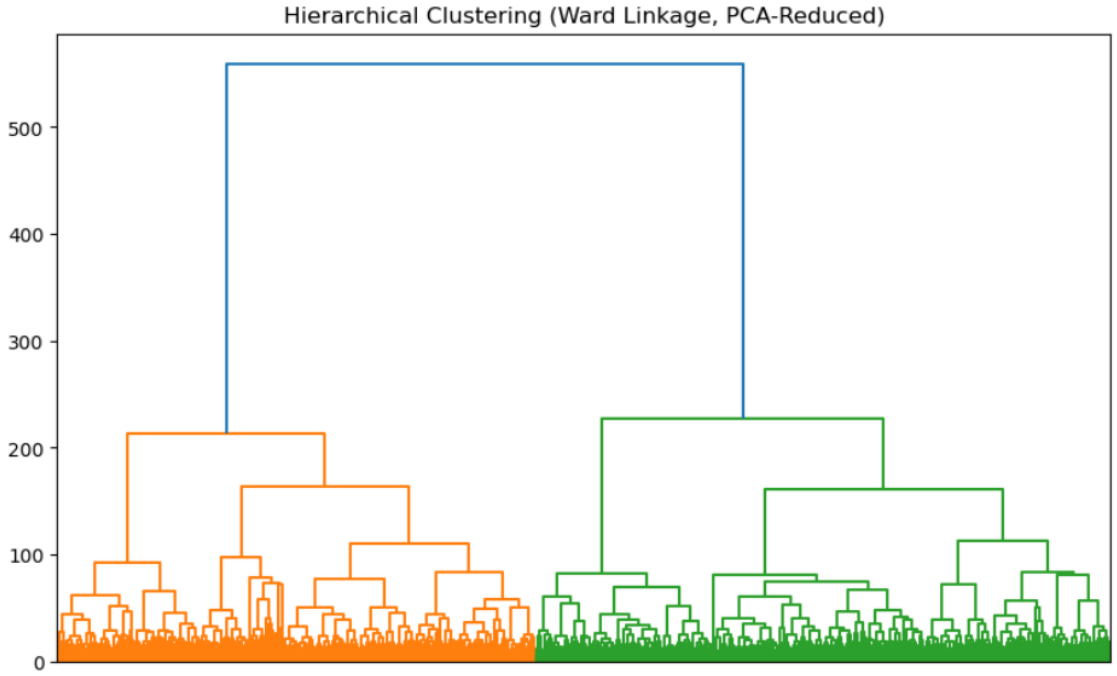
Based on the cumulative explained variance, 50 principal components were selected, capturing approximately 90% of the total variance. This reduction preserved most of the informational content while substantially reducing the feature space.

The PCA-reduced dataset was exported to a new CSV file for downstream analysis.

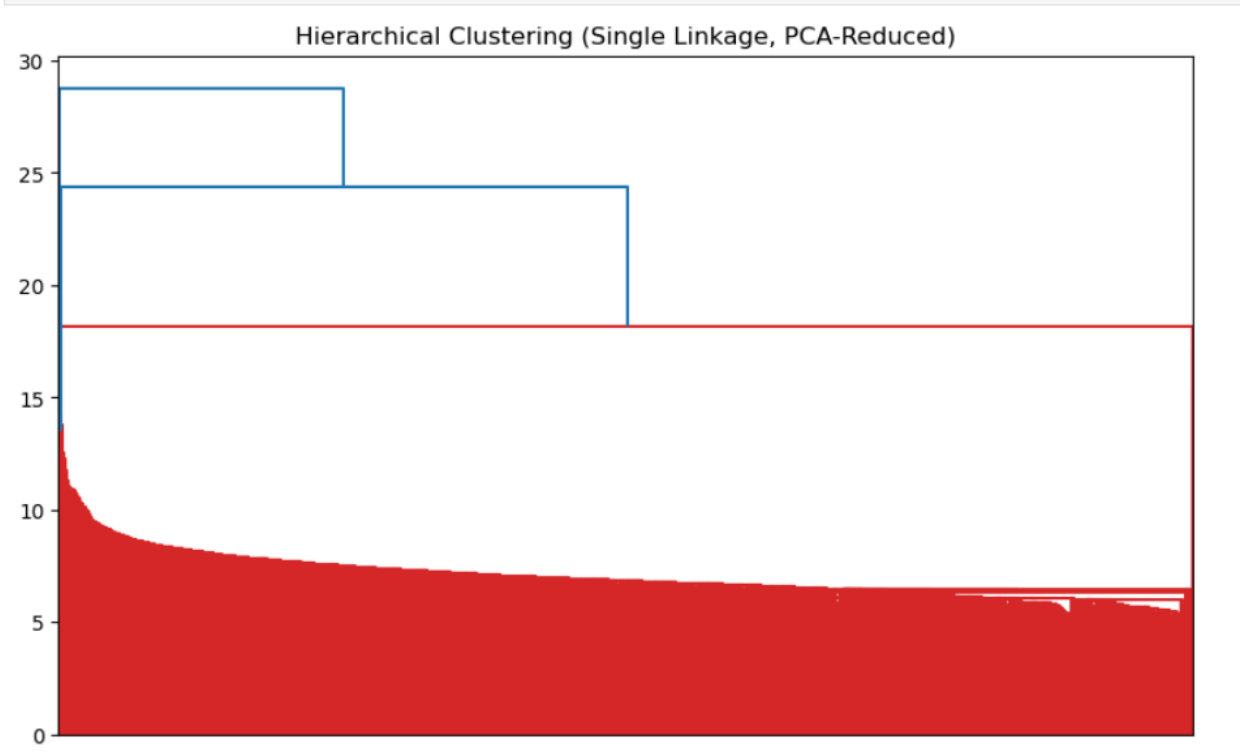
Hierarchical Clustering on PCA-Reduced Data (Rerun)

Hierarchical clustering was then rerun on the PCA-reduced data using the same four linkage methods (single, complete, average, and ward), enabling a direct comparison with clustering on the original scaled dataset.

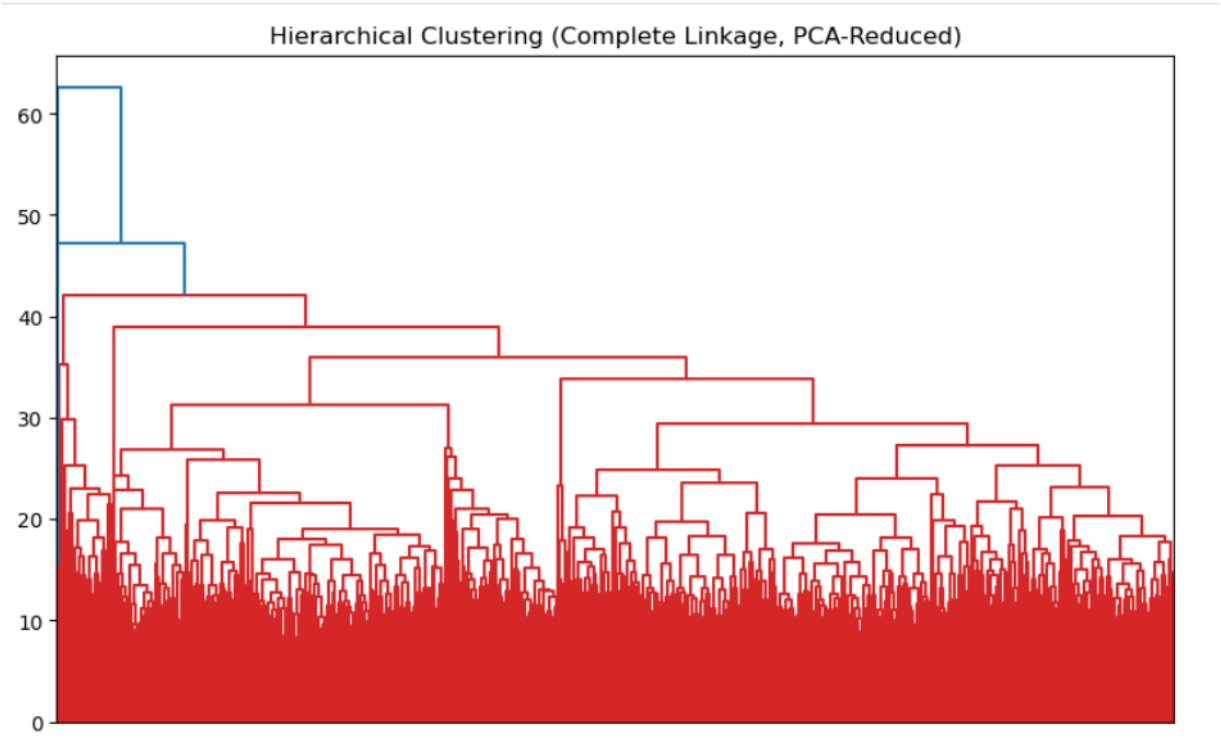
Ward Linkage



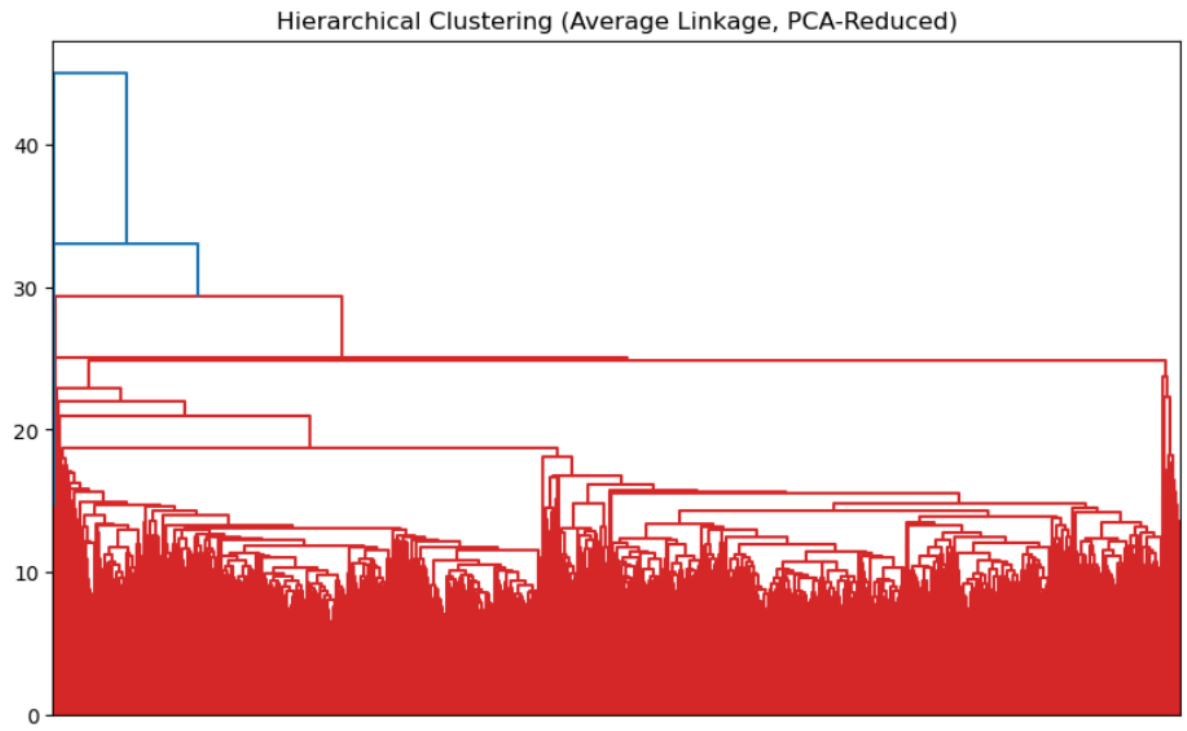
Single Linkage



Complete Linkage



Average Linkage



Compared with the original data, the PCA-reduced dendrograms showed cleaner high-level separation, particularly under ward linkage. Reducing dimensionality simplified the hierarchical structure and reduced noise, making dominant cluster groupings easier to interpret. Single linkage remained sensitive to chaining effects, while complete and average linkage showed modest improvements in clarity.

Comparison to Pleasant Weather Labels

When comparing hierarchical clusters to ClimateWins' existing "pleasant weather" labels, partial alignment was observed. Some clusters corresponded to clearer, drier, and higher-radiation conditions, while others reflected cloudier, wetter, or colder regimes. However, the alignment was not exact, suggesting that pleasant weather classifications depend on a subset of variables rather than the full meteorological profile captured by clustering.

Conclusion and Limitations

This analysis demonstrates the value of unsupervised learning for uncovering latent structure in complex climate data. Hierarchical clustering revealed meaningful groupings of weather observations, with ward linkage consistently providing the most interpretable

results. PCA improved clustering clarity by reducing dimensionality and noise, supporting its use under computational constraints.

A limitation of this approach is that hierarchical clustering outcomes depend on distance metrics, linkage methods, and cut heights, which introduces subjectivity in determining the number of clusters. Additionally, simplifying weather patterns into clusters may obscure localized or seasonal variability.