

Weather Typing and Synoptic Climatology

CEVE 543 - Fall 2025

Dr. James Doss-Gollin

2025-10-27

i Today

- Weather typing: motivation and applications
- Clustering methods (K-means and alternatives)
- Dimension reduction via PCA
- Paper prep for Wednesday

1 Weather Typing and Analogs

1.1 The Power of Analogs

The fundamental problem:

Recall from 10/20 and 10/22 lectures:

- Climate models: systematic biases in local variables (drizzle bias, extremes)
- Coarse resolution: ~100-300 km gridcells
- “One gridcell = San Francisco to Sacramento”
- Too coarse for local impacts

Key insight: Models may get local precipitation wrong but large-scale patterns right!

The solution: Use analogs

Match large-scale patterns (models do well) to historical local impacts (what we observe).

Multiple analog methods:

- **LOCA:** Find matching historical days, select best analog per location [1]
- **Weather typing:** Group states into discrete “types,” use conditional statistics (today’s focus)
- **Archetype analysis:** Identify extreme “corner” patterns [2]
- **Self-organizing maps:** Neural network preserving topology [3], [4]

We’ll focus on weather typing, but the core idea generalizes.

1.2 Weather Typing Approach

Goal: Group similar atmospheric states into discrete “types”

Workflow:

1. Identify recurring large-scale circulation patterns
2. Link each weather type to observed local impacts
3. When model simulates a weather type, apply local statistics
4. Track weather type frequency changes under climate change

Applications:

- Statistical downscaling
- Bias correction conditioned on weather type
- Extreme event attribution
- Understanding circulation changes

Widely used in synoptic climatology and hydrology [5], [6], [7], [8], [9].

2 Clustering

2.1 Motivation

Goal: Partition observations into groups (clusters) with similar characteristics

Why cluster?

- Identify patterns in high-dimensional data
- Reduce continuous variability to discrete states
- Enable conditional statistics (e.g., $P(\text{rain} \mid \text{weather type})$)

2.2 K-means Clustering

Objective: Partition n observations into k clusters

Each observation is assigned to the cluster with the nearest mean (centroid).

Algorithm:

1. Initialize k centroids randomly
2. Assign each observation to the nearest centroid
 - Uses a distance metric (typically Euclidean distance)
3. Update the centroids by computing the mean of all assigned observations
4. Repeat steps 2-3 until convergence (assignments no longer change)

Key choices:**Number of clusters k :**

Use the “elbow method” to select k :

- Plot within-cluster sum of squares (WCSS) versus k
- WCSS measures total distance from points to their cluster centroids
- Look for an “elbow” where adding more clusters gives diminishing returns

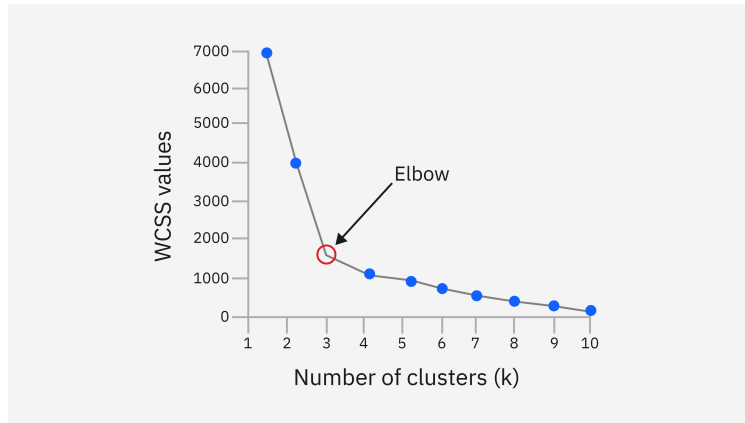


Figure 1: Elbow plot showing within-cluster variance decreasing as K increases, with optimal K at the elbow. Source: IBM

Distance metric:

- Euclidean distance is most common
- Other options: Manhattan, cosine similarity

Initialization:

- Algorithm is sensitive to initial centroid placement
- Common practice: run multiple times with different initializations
- Keep the solution with lowest WCSS

Limitations:

- “Hard” classification (each point assigned to exactly one cluster)
- Assumes spherical clusters of similar size
- Sensitive to initialization
- Must specify k in advance
- Can struggle with outliers
- **Sensitive to data scaling:** K-means uses Euclidean distance, so variables with large variances will dominate the clustering. This motivates dimension reduction (PCA) before clustering. If you cluster on PCs, should you scale them to equal variance first? There’s no “right” or “wrong” answer - just be aware of the choice!

2.3 Alternatives

Gaussian Mixture Models (GMMs):

- Probabilistic approach
- Each point has probability of belonging to each cluster
- “Soft” classification
- More flexible than K-means

Self-Organizing Maps (SOMs):

- Neural network approach
- Maps high-D data to low-D grid
- Preserves topological structure

- Good for visualization

Hierarchical clustering:

- Builds tree of clusters (dendrogram)
- Agglomerative (bottom-up) or divisive (top-down)
- Don't need to specify k in advance
- Can cut dendrogram at different levels

3 Dimension Reduction

Problem: Clustering algorithms work poorly in high dimensions

Why?

- Curse of dimensionality: points become equidistant
- Distance metrics become less meaningful
- More noise, slower computation

Solution: Reduce dimensionality first, then apply clustering

3.1 PCA

Also called: Empirical Orthogonal Functions (EOFs), Singular Value Decomposition (SVD)

Goal: Find orthogonal directions of maximum variance

3.1.a Mathematical Setup

Data matrix: $\mathbf{X}$ of size $n \times p$

- n = observations (e.g., time steps)
- p = variables (e.g., spatial locations)
- Each row = one observation

Objective: Find directions that maximize variance

3.1.b First Principal Component

$$\mathbf{z}_1 = \mathbf{X}\phi_1$$

- ϕ_1 = loading vector (size $p \times 1$)
- $\|\phi_1\|_2 = 1$ (normalized)
- $\mathbf{z}_1$ = PC scores (size $n \times 1$)

For observation i :

$$z_{i1} = \phi_{11}X_{i1} + \phi_{21}X_{i2} + \dots + \phi_{p1}X_{ip}$$

Geometric interpretation: ϕ_1 is the direction of maximum variance

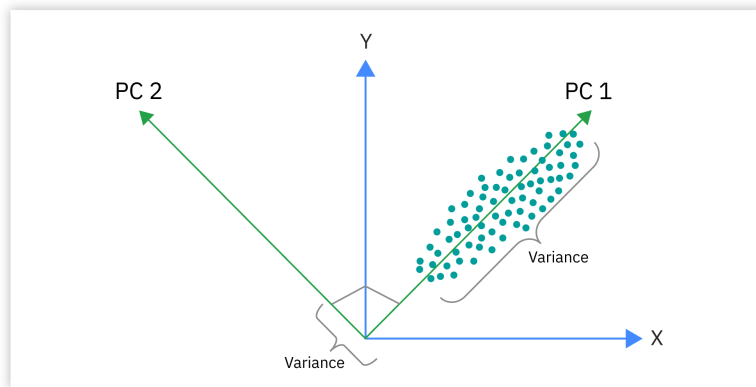


Figure 2: PCA finds orthogonal directions (PC1, PC2) that successively maximize variance in the data.
Source: IBM

Subsequent PCs: $z_2, z_3, \dots$

- Maximize the remaining (residual) variance after removing previous PCs
- Orthogonal to all previous PCs
- Each PC captures less variance than the previous one

3.1.c Solution via Eigendecomposition

Step 1: Center the data

$$\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{X}}$$

Subtract column means so each variable has mean zero.

Step 2 (Optional): Standardize

Divide each column by its standard deviation.

Each variable then has unit variance.

This is a key decision:

- If you only center (Step 1), PCA is performed on the **covariance matrix**. Variables with higher variance will dominate the first PCs.
- If you standardize (Step 2), PCA is performed on the **correlation matrix**. All variables are treated as having equal variance. This is the right choice if your variables have different units.

Step 3: Compute the covariance matrix

$$\mathbf{C} = \frac{1}{n-1} \mathbf{X}_c^T \mathbf{X}_c$$

This is a $p \times p$ matrix.

Step 4: Solve the eigenvalue problem

$$\mathbf{C} \mathbf{v}_j = \lambda_j \mathbf{v}_j$$

- λ_j = eigenvalue
- v_j = eigenvector

Step 5: Sort the eigenvalues

Order them as: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$

The corresponding eigenvectors are the loading vectors (principal components).

Step 6: Project the data

$$\mathbf{Z} = \mathbf{X}_c \Phi$$

- Φ = matrix of first k eigenvectors (size $p \times k$)
- $\mathbf{Z}$ = PC scores (size $n \times k$)

3.1.d Variance Explained

The eigenvalue λ_j equals the variance explained by PC j .

The fraction of total variance explained by the first k PCs is:

$$\frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^p \lambda_j}$$

Key assumption: Variance equals variability

This may not hold if:

- Data contain outliers
- Non-linear relationships are important

3.1.e Climate Data Interpretation

Setup: Data matrix has time in rows, space in columns

Eigenvectors ϕ_j represent **spatial patterns** (EOFs)

Scores z_{ij} represent **time series** showing when each pattern is active

Reconstruction:

$$\mathbf{X}_i \approx \sum_{j=1}^k z_{ij} \phi_j^T$$

This reconstructs the data at time i using the first k EOFs.

3.1.f Practical Considerations

Centering and anomalies:

Remove the time-mean at each location:

$$x_{(ij)'} = x_{ij} - \bar{x}_j$$

For seasonal data, remove the seasonal climatology instead.

Standardization:

When should you standardize?

- Variables have different units → standardize
- Variables have similar units and scales → maybe not
- Standardization prevents one variable from dominating the PCA

Spatial weighting:

Latitude-longitude grids have unequal areas.

Grid cell area is proportional to $\cos(\phi)$ where ϕ is latitude.

Two approaches:

- Weight each data point by $\sqrt{\cos(\phi)}$ before computing the covariance matrix
- This makes the covariance matrix effectively weighted by $\cos(\phi)$ (since covariance involves products)

Why $\sqrt{\cos(\phi)}$ and not $\cos(\phi)$?

The covariance matrix involves products of data values: $X^T X$.

If we weight data by $\sqrt{\cos(\phi)}$, the covariance becomes weighted by $\cos(\phi)$.

This gives each unit area equal influence in the analysis.

Selecting the number of components:

Use a **scree plot**: plot variance explained versus component number.

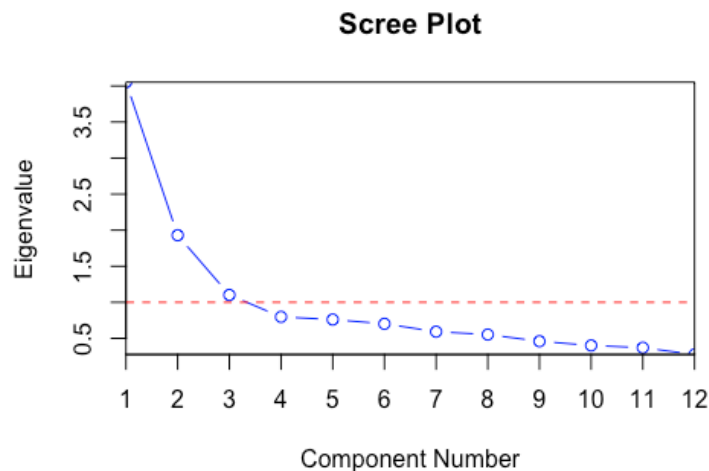


Figure 3: Example scree plot showing variance explained by each principal component. Look for an “elbow” but recognize this is subjective. Source: Wikipedia

How to choose k ?

There is no magical rule - it's a **trade-off between dimensionality and variance preserved**.

Common approaches:

- Look for an “elbow” where additional components add little variance (subjective!)
- Choose k to explain a target fraction of variance (e.g., 80% or 90%)

- **Rule of thumb** [10]: Check for separation in the eigenvalues. If eigenvalues are too “close” together, their corresponding EOFs may be mixed and not physically distinct.
- Consider the application: how much compression do you need?
- Try different values and see what works for your downstream task

Further Reading:

- SVD videos by Steven Brunton (watch Video 2)
- MIT PCA lecture
- G. James, D. Witten, T. Hastie, and R. Tibshirani [11] Chapter 10.2
- NCAR EOF tutorial

4 Putting It All Together: The PCA-then-Cluster Pipeline

Typical workflow for weather typing:

1. **Start with high-dimensional data $\mathbf{X}$** (size $n \times p$)
 - Rows = time (days)
 - Columns = space (grid cells) or other variables
2. **Perform PCA** to get low-dimensional PC scores $\mathbf{Z}$ (size $n \times k$)
 - Each row is one day represented by k PC scores
 - $k \ll p$ (much fewer dimensions)
3. **(Optional but important) Scale the PC scores**
 - Should you scale PCs to equal variance before clustering?
 - If yes: each PC weighted equally in clustering
 - If no: PC1 (highest variance) dominates the clustering
 - **This is the choice made in Wednesday’s paper**
4. **Run K-means clustering** on the (possibly scaled) PC scores $\mathbf{Z}$
 - Get k weather types
 - Each day assigned to one weather type

Key decisions in this pipeline:

- How many PCs to retain? (Step 2)
- Should you scale the PCs before clustering? (Step 3)
- How many clusters k ? (Step 4)

These choices are **subjective** but should be **justified** based on your scientific goals.

Preparing for Wednesday

Paper: J. Doss-Gollin, Á. G. Muñoz, S. J. Mason, and M. Pastén [12]

Reading instructions:

- Read carefully: Sections 2-4, 6
- **Skip:** all of Section 5 (ENSO/MJO)
- **Skip:** ENSO/MJO discussion in Section 6

Discussion questions:

1. **Subjective choices (Section 3a):**
 - Why scale PCs to unit variance?

- Why choose $K=6$ clusters?
 - How would $K=8$ change results?
2. **Hard vs. soft classification:**
- K-means assigns each day to exactly one weather type
 - What are limitations of “hard” classification?
 - Atmospheric flow is continuous and can transition smoothly
 - How does forcing into discrete “boxes” affect results?
3. **Bias correction application:**
- You have a climate model that simulates 850-hPa streamfunction and precipitation
 - How would you use the 6 weather types to bias correct precipitation?
 - What biases would this method fix?
 - What biases would it miss?

5 References

Bibliography

- [1] D. W. Pierce, D. R. Cayan, E. P. Maurer, J. T. Abatzoglou, and K. C. Hegewisch, “Improved Bias Correction Techniques for Hydrological Simulations of Climate Change,” *Journal of Hydrometeorology*, vol. 16, no. 6, pp. 2421–2442, 2015, doi: 10.1175/jhm-d-14-0236.1.
- [2] S. Steinschneider and U. Lall, “Daily Precipitation and Tropical Moisture Exports across the Eastern United States: An Application of Archetypal Analysis to Identify Spatiotemporal Structure,” *Journal of Climate*, vol. 28, no. 21, pp. 8585–8602, 2015, doi: 10.1175/jcli-d-15-0340.1.
- [3] L. Agel, M. Barlow, S. B. Feldstein, and W. J. J. Gutowski, “Identification of Large-Scale Meteorological Patterns Associated with Extreme Precipitation in the US Northeast,” *Climate Dynamics*, vol. 50, no. 5–6, pp. 1819–1839, Mar. 2018, doi: 10.1007/s00382-017-3724-8.
- [4] C. H. R. Lima, A. AghaKouchak, and U. Lall, “Classification of Mechanisms, Climatic Context, Areal Scaling, and Synchronization of Floods: The Hydroclimatology of Floods in the Upper Parana River Basin, Brazil,” *Earth System Dynamics*, vol. 8, no. 4, pp. 1071–1091, 2017, doi: 10.5194/esd-8-1071-2017.
- [5] V. Moron, A. W. Robertson, J.-H. Qian, and M. Ghil, “Weather Types across the Maritime Continent: From the Diurnal Cycle to Interannual Variations,” *Frontiers in Environmental Science*, vol. 2, p. 163, Jan. 2015, doi: 10.3389/fenvs.2014.00065.
- [6] C. C. Lee and S. C. Sheridan, “Trends in Weather Type Frequencies across North America,” *npj Climate and Atmospheric Science*, vol. 1, no. 1, p. 41, Nov. 2018, doi: 10.1038/s41612-018-0051-7.
- [7] A. Rueda, P. Camus, A. Tomás, S. Vitousek, and F. J. Méndez, “A Multivariate Extreme Wave and Storm Surge Climate Emulator Based on Weather Patterns,” *Ocean Modelling*, vol. 104, pp. 242–251, Aug. 2016, doi: 10.1016/j.ocemod.2016.06.008.
- [8] L. Cagigal *et al.*, “A Multivariate, Stochastic, Climate-Based Wave Emulator for Shoreline Change Modelling,” *Ocean Modelling*, vol. 154, p. 101695, Oct. 2020, doi: 10.1016/j.ocemod.2020.101695.
- [9] V. Moron, A. W. Robertson, M. N. Ward, and O. Ndiaye, “Weather Types and Rainfall over Senegal. Part II: Downscaling of GCM Simulations,” *Journal of Climate*, vol. 21, no. 2, pp. 288–307, Jan. 2008, doi: 10.1175/2007JCLI1624.1.

- [10] G. R. North, T. L. Bell, R. F. Cahalan, and F. J. Moeng, “Sampling Errors in the Estimation of Empirical Orthogonal Functions,” *Monthly Weather Review*, vol. 110, no. 7, pp. 699–706, 1982, doi: 10.1175/1520-0493(1982)110<0699:seiteo>2.0.co;2.
- [11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. in Springer Texts in Statistics, vol. 103. New York, NY: Springer New York, 2013.
- [12] J. Doss-Gollin, Á. G. Muñoz, S. J. Mason, and M. Pastén, “Heavy Rainfall in Paraguay during the 2015-2016 Austral Summer: Causes and Sub-Seasonal-to-Seasonal Predictive Skill,” *Journal of Climate*, vol. 31, no. 17, pp. 6669–6685, Jan. 2018, doi: 10.1175/jcli-d-17-0805.1.