Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.37835

Unsupervised Mining of Regional Warming Patterns from ERA5 Spatio-Temporal Data: K-means Clustering of 34 Major Chinese Cities, 1975–2024

Yuxin Tao

Soochow Academy, Suzhou High School of Jiangsu Province, Suzhou, China
floratao0111@163.com

Abstract. This paper treats regional warming in China as an unsupervised pattern-mining task on ERA5 spatio-temporal data for thirty-four provincial-level cities. First, about 1.8 million hourly values of six ERA5 variables for 2020 are reduced to five standardized features per city, and K-means groups the cities into five clusters with no predefined zoning labels. Second, the cluster labels are used to pool the 1975–2024 2-meter temperature series, and the regional decadal means are compared by cumulative warming and by the linear rate from ordinary least squares regression on decade midpoints. Every region warmed, though by different amounts. From 1975–1984 to 2015–2024 the Northwest gained 1.99 °C (0.51 °C per decade) and the Qinghai–Tibet Plateau gained 1.61 °C (0.40 °C per decade). South China had the smallest rise, 1.09 °C (0.24 °C per decade). The warming was also not steady, as the North–Northeast and South China regions cooled a little in 2005–2014 and then rose again, but the Northwest and the plateau kept warming. Mostly the clusters follow conventional geography, and only three boundary cities change region. Thus the partition learned without a zoning scheme separates warming paths of different size and timing, with faster warming in the north-western interior and on the plateau than on the southern coast.

Keywords: unsupervised learning, spatio-temporal data mining, K-means clustering, ERA5, urban warming

1. Introduction

China has warmed rapidly over the past half-century, by about 1.44 °C during 1961–2013 [1], and the rate varies across regions [2]. Northwest China, in particular, has seen warming alongside wettening, at roughly 0.49 °C per decade [3]. Cities amplify this signal, most visibly in compound hot extremes [4]. Background warming plus the urban heat island (UHI) effect produces uneven local trends: across 92 major cities, urbanization accounted on average for 46.7% of local warming during 1981–2017, especially in minimum temperature, with the largest contribution in the southeast [5]; surface UHI intensity has risen nationwide outside winter, with climate-zone differences [6], and its seasonal pattern can diverge even within a single megacity [7].

Climate reanalysis products such as ERA5 [8] are a typical form of high-dimensional spatio-temporal data, since every grid point carries several variables at hourly steps over many decades.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.37835

Spatio-temporal data mining looks for patterns in such data that are not known in advance, and clustering is one of its main tasks [9]. For regional warming there is no ground-truth label that tells which cities share a warming regime; therefore, supervised models cannot be trained, and unsupervised clustering is a natural choice. K-means is still one of the most widely used clustering algorithms, mainly due to its simplicity and efficiency [10]. Earlier studies have often grouped cities by fixed schemes, such as administrative boundaries or the Köppen–Geiger classification, and clustering forms the groups from the multivariate data itself [2, 11]. ERA5 has been validated over complex terrain in China, including the Qilian Mountains, where it captures observed warming rates well [12], and it provides a consistent long record for this purpose [8].

This paper sets up the analysis as a two-stage mining task with two questions. First, can K-means, applied to standardized feature vectors built from the ERA5 records of 34 cities, find a grouping without labels, and does the same grouping come out as the best of many random starts? Second, what warming patterns appear in the 1975–2024 temperature series of each learned cluster? The main contribution is thus the use of unsupervised learning to mine regional differences in long temperature series, through clusters learned from multivariate ERA5 features. In detail, (1) a simple pipeline turns hourly and monthly ERA5 fields at city grid points into a feature matrix and decadal series; (2) the K-means partition is checked for local minima and compared with conventional zoning; and (3) the mined clusters show warming paths that differ in size and timing, as well as large differences between cities in the same cluster.

2. Data

2.1. Dataset

The analysis uses ERA5, the fifth-generation atmospheric reanalysis from ECMWF [8], available on a 0.25° × 0.25° grid since 1950. Two products were used: (a) hourly single-level data for 2020, from which annual means of the climate variables were computed for clustering; and (b) monthly single-level data from 1975–2024, from which annual and decadal mean 2-meter temperatures were derived.

2.2. Study sample and variables

Thirty-four provincial-level administrative centers are included: 27 provincial capitals, 4 municipalities, the 2 special administrative regions (Hong Kong and Macao), and Taipei. For each city, the nearest ERA5 grid point to the city center was selected. Six ERA5 variables were extracted: 2-meter temperature (t2m), 2-meter dew point (d2m), surface pressure (sp), 10-meter U and V wind components (u10, v10), and total precipitation (tp). From these, five clustering features were formed: temperature, relative humidity (from t2m and d2m), surface pressure, wind speed (from u10 and v10), and precipitation. Temperature was averaged within five consecutive decades: 1975–1984, 1985–1994, 1995–2004, 2005–2014, and 2015–2024. Table 1 lists the decadal temperatures, grouped by the K-means clusters obtained in Section 4.2; Table 2 lists the five climate features for 2020.

DOI: 10.54254/2755-2721/2027.37835

3. Methods

3.1. K-means clustering of ERA5 feature vectors

Each city enters the analysis as a multivariate time series at one ERA5 grid point. For 2020 this series has 8,784 hourly steps for each of six variables, so the 34 cities hold about 1.8 million values. K-means works on vectors of fixed length; therefore, aggregation first removes the time axis, and each of the five features in Section 2.2 is summarized by its annual mean, or by its annual total for precipitation. This step turns the spatio-temporal sample into a data matrix with $n = 34$ rows (cities) and $p = 5$ columns (temperature, relative humidity, surface pressure, wind speed, precipitation), with x_{ij} the value of feature j for city i. City location is not an input; thus any spatial grouping in the result comes from the climate features alone. The features come from one year, and Section 3.2 keeps the partition fixed for the earlier decades.

Variables have different units (°C, %, hPa, m s^{-1}, mm) and scales, so each was z-scored before clustering, as in Equation (1):

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \tag{1}$$

where $\overline{x}_j$ and s_j are the mean and standard deviation of variable j across cities. Standardization prevents high-magnitude variables from dominating Euclidean distance.

K-means partitions the n observations into K clusters by minimizing the within-cluster sum of squares in Equation (2):

$$\mathrm{WCSS}(K) = \sum_{k=1}^{K} \sum_{z_i \in C_k} \|z_i - \mu_k\|^2 \tag{2}$$

where C_k is cluster k and μ_k its centroid. The standard solver, Lloyd's algorithm [13], alternates between assigning each point to its nearest centroid and recomputing centroids until no points move. K was set to 5 so that the learned groups could be compared with five broad regions in common use; the region names in Section 4 came after clustering, from the location of the member cities, and played no part in the algorithm. The procedure was restarted 20 times from random initial centroids, and the run with the lowest WCSS was kept. All clustering was done in Python with scikit-learn. Lloyd's algorithm only reaches a local minimum of Equation (2), so a further check used 1,000 single random starts. The lowest WCSS among them belongs to the same partition reported in Section 4.2; however, only 94 of the 1,000 runs reached it, and this is the reason for using restarts.

3.2. Cluster-level trend estimation

In the second stage, the cluster labels from Section 3.1 index the 1975–2024 temperature series; thus warming is summarized for groups learned from the data and not for administrative units. For each city, annual mean 2-meter temperature was averaged within each decade. Regional decadal means are the average of cities in a cluster. Cumulative warming is the difference between the 2015–2024 and 1975–1984 means. Decadal rate was estimated by ordinary least squares regression of the five regional means on decade midpoint (1979.5, 1989.5, 1999.5, 2009.5, 2019.5), giving a slope in °C per decade. Because only five decadal points enter each regression, the slopes describe long-term trends rather than interannual variability; formal significance tests on annual data are deferred to future work.

DOI: 10.54254/2755-2721/2027.37835

4. Results

4.1. National warming pattern

Every city was warmer in 2015–2024 than in 1975–1984 (Table 1). The rise ranged from 0.59 °C (Beijing) to 2.63 °C (Urumqi). Twenty of the 34 cities cooled temporarily between at least one pair of consecutive decades, 16 of them between 1995–2004 and 2005–2014.

4.2. Clustering result

Applied to the 2020 features in Table 2, K-means assigned the 34 cities to five clusters that roughly track conventional geography (Figure 1). Three boundary cities moved: Taiyuan and Hohhot landed in the Northwest cluster, while Xi'an went to the North–Northeast cluster. Table 1 gives the cluster memberships and regional decadal means.

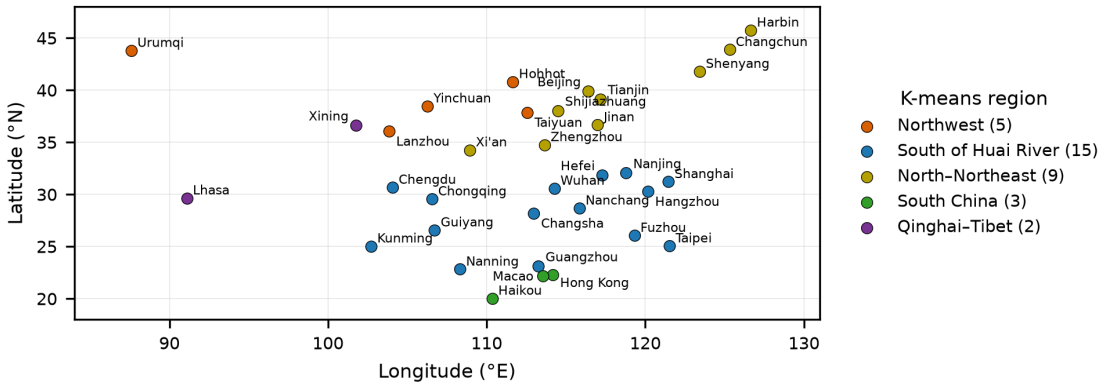


Figure 1. Locations of the 34 cities coloured by K-means cluster (number of cities in parentheses)

4.3. Warming rates by cluster

All five regions warmed between 1975–1984 and 2015–2024 (Table 1; Figure 2a, where each line is labelled with its cumulative increase and OLS rate). The Northwest led at +1.99 °C (0.51 °C per decade), followed by the Qinghai–Tibet Plateau at +1.61 °C (0.40 °C per decade). North–Northeast and South-of-Huai-River gained 1.35 and 1.32 °C (0.32 and 0.33 °C per decade). South China was the least warmed, at +1.09 °C (0.24 °C per decade).

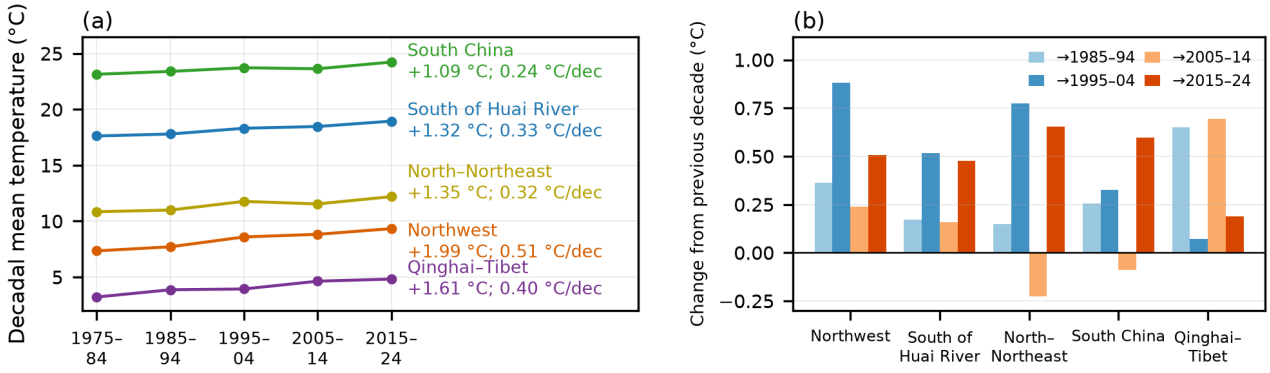


Figure 2. Warming by K-means cluster, 1975–2024: (a) decadal mean temperature; (b) change between consecutive decades

DOI: 10.54254/2755-2721/2027.37835

The decade-to-decade path was uneven (Figure 2b). For most regions, the strongest single-decade rises occurred in the steps into 1995–2004 and into 2015–2024. In 2005–2014, two regions slipped: North–Northeast fell from 11.77 to 11.54 °C, and South China from 23.72 to 23.63 °C. The plateau's regional mean rose every decade, although Lhasa dipped from 4.31 to 3.88 °C in the 1990s.

Table 1. Decadal mean annual temperature (°C) for the 34 cities, 1975–2024, grouped by K-means cluster, with each cluster's mean

Cluster	City	1975–84	1985–94	1995–04	2005–14	2015–24
Northwest	Hohhot	5.02	5.54	6.50	6.92	7.55
	Urumqi	5.14	5.46	6.48	7.32	7.77
	Lanzhou	7.96	8.44	9.15	9.17	9.41
	Yinchuan	8.87	9.23	10.11	10.06	10.59
	Taiyuan	9.72	9.86	10.69	10.65	11.33
	Cluster mean	7.34	7.71	8.59	8.82	9.33
South of Huai River	Shanghai	15.80	16.13	16.94	17.30	17.84
	Chongqing	18.27	18.24	18.66	18.90	19.27
	Hefei	15.91	16.03	16.81	16.77	17.30
	Wuhan	17.09	17.21	17.85	17.86	18.38
	Nanjing	15.42	15.59	16.35	16.44	16.94
	Chengdu	16.46	16.40	17.11	17.23	17.67
	Guiyang	14.68	14.87	15.02	15.02	15.62
	Kunming	15.04	15.13	15.56	16.02	16.08
	Hangzhou	16.07	16.41	17.14	17.52	18.03
	Changsha	17.66	17.76	18.29	18.62	18.89
	Nanchang	18.03	18.23	18.78	18.96	19.57
	Guangzhou	22.60	22.90	23.26	23.14	23.72
	Fuzhou	18.83	18.97	19.46	19.64	20.33
	Taipei	20.49	20.78	21.03	21.31	21.75
	Nanning	21.94	22.23	22.37	22.26	22.76
	Cluster mean	17.62	17.79	18.31	18.47	18.94
North–Northeast	Beijing	12.86	12.66	13.50	13.08	13.45
	Tianjin	12.67	12.84	13.62	13.50	14.11
	Harbin	3.80	4.19	5.09	4.87	5.38
	Changchun	5.34	5.88	6.65	6.21	7.11
	Shenyang	8.14	8.45	9.07	8.83	9.77
	Shijiazhuang	13.44	13.47	14.20	13.99	14.48
	Xi'an	14.20	14.32	15.30	14.85	15.15

DOI: 10.54254/2755-2721/2027.37835

Table 1. (continued)

	Zhengzhou	14.50	14.39	15.08	15.15	16.10
	Jinan	12.64	12.73	13.39	13.37	14.19
	Cluster mean	10.84	10.99	11.77	11.54	12.19
South China	Haikou	24.04	24.26	24.43	24.38	25.07
	Hong Kong	22.68	22.96	23.38	23.28	23.82
	Macao	22.69	22.96	23.35	23.23	23.79
	Cluster mean	23.14	23.39	23.72	23.63	24.23
Qinghai–Tibet	Xining	2.70	3.42	3.99	4.34	4.68
	Lhasa	3.73	4.31	3.88	4.92	4.96
	Cluster mean	3.21	3.86	3.94	4.63	4.82

Table 2. Five climate features for the 34 cities, 2020 (cities in the order of Table 1)

City	T (°C)	RH (%)	SP (hPa)	WS (m s^{-1})	TP (mm)
Hohhot	7.02	46.15	883.41	0.71	336.34
Urumqi	7.77	45.79	896.56	0.91	230.4
Lanzhou	8.65	51.17	803.5	0.79	521.06
Yinchuan	10.28	42.31	887.87	0.46	200.02
Taiyuan	11.02	50.63	907.23	0.65	531.79
Shanghai	17.87	76.17	1016.4	1.7	1639.34
Chongqing	18.74	77.48	975.18	0.49	1350.94
Hefei	17.21	73.34	1013.53	1.1	1490.47
Wuhan	18.06	77.03	1013.35	1.18	2191.0
Nanjing	16.89	74.89	1014.63	1.2	1483.11
Chengdu	17.43	70.89	950.86	0.97	1559.28
Guiyang	14.82	82.0	874.43	1.43	1683.23
Kunming	16.02	68.63	792.44	2.07	1484.31
Hangzhou	18.45	74.65	1013.99	1.04	1794.92
Changsha	18.56	77.48	1007.99	1.52	1905.37
Nanchang	18.95	76.2	1007.72	1.6	1668.89
Guangzhou	23.91	74.01	1011.67	1.89	1526.74
Fuzhou	20.79	74.22	984.52	1.25	1168.06
Taipei	21.99	78.82	976.17	1.66	1593.08
Nanning	22.69	77.97	997.52	1.82	1452.06
Beijing	13.3	49.02	1014.48	0.69	390.02

DOI: 10.54254/2755-2721/2027.37835

Table 2. (continued)

Tianjin	13.95	52.3	1016.21	0.66	501.79
Harbin	5.39	65.53	997.35	1.58	800.02
Changchun	6.34	61.0	991.16	1.82	755.74
Shenyang	9.66	58.71	1008.37	1.29	701.15
Shijiazhuang	14.13	56.89	1005.09	0.46	621.9
Xi'an	14.6	64.24	957.49	0.3	881.8
Zhengzhou	16.04	57.54	995.5	0.49	698.49
Jinan	14.78	58.77	1007.96	0.75	743.49
Haikou	25.35	80.51	1009.19	4.04	979.29
Hong Kong	23.66	80.21	1009.37	3.74	1522.96
Macao	23.87	79.61	1010.71	3.09	1701.65
Xining	4.18	52.95	723.39	0.56	798.18
Lhasa	4.74	43.92	607.58	0.74	839.36

5. Discussion

The clearest pattern is a contrast between the north-western interior and the southern coast. The Northwest rate of 0.51 °C per decade is close to that of Chi et al. [3] for the region and agrees with the usual view that sites at high latitude or elevation warm faster. South China's slower rise likely reflects its warm baseline and maritime buffering; Hong Kong, Macao, and Haikou each warmed by less than 1.2 °C over the period and differ from one another by only 0.11 °C, the narrowest spread of any cluster.

Regional means nonetheless conceal large differences between cities. Within the Northwest, cumulative warming ranged from 1.45 °C in Lanzhou to 2.63 °C in Urumqi; within the South-of-Huai-River cluster, from 0.82 °C in Nanning to 2.04 °C in Shanghai; and within the North–Northeast, from 0.59 °C in Beijing to 1.77 °C in Changchun (Table 1). The spread inside each of these three clusters (1.18–1.22 °C) is larger than the difference in warming between any two cluster means, which is at most 0.90 °C (Northwest versus South China). Shanghai's position at the top of its cluster is consistent with the large urbanization contribution reported for south-eastern cities [5], although a single grid point cannot separate urban from background warming.

The 2005–2014 dip in the North–Northeast and South China regions aligns with the 1998–2012 warming hiatus documented over China, which brought significant cooling to eastern and northwestern China, was absent on the Tibetan Plateau, and has been linked to the negative phase of the Pacific Decadal Oscillation [14]. At the city level the dip was concentrated: 8 of the 9 North–Northeast cities and all 3 South China cities were cooler in 2005–2014 than in 1995–2004, compared with 3 of 15 cities south of the Huai River, 2 of 5 in the Northwest, and neither plateau city. In the decadal means used here, the Northwest and plateau clusters kept rising through the same decades, which suggests the slowdown was geographically uneven rather than uniform across China; the decadal windows, which do not align with 1998–2012, may explain why the northwestern cooling reported for that period does not appear.

DOI: 10.54254/2755-2721/2027.37835

The five clusters mostly match conventional zoning, but the relocations of Taiyuan, Hohhot, and Xi'an show that 2020 climate statistics do not respect provincial boundaries. This is consistent with Duan et al. [11], who found that clustering classified the climates of major Chinese cities better than the Köppen–Geiger scheme. From the data-mining side, this partial agreement is a plausibility check on the partition, since K-means learned it from unlabeled feature vectors with no location input; the three moved cities mark where the data and a fixed scheme disagree. The steps use only gridded data and city coordinates, with no zoning labels; thus they could be applied to other variables, periods or city samples. Since warming differs between and inside the regions, adaptation planning would need to use city-level trends, not only provincial or regional averages, and such planning has been shown to strengthen urban climate resilience [15]. $K = 5$ was chosen to align with conventional divisions; an automatic choice such as the silhouette score would be a natural next step.

This study has three limitations. First, decadal means smooth out year-to-year variability; thus the slopes carry no significance test, and Mann–Kendall or Theil–Sen tests on annual data could add one. Second, the sample covers only provincial capitals, the largest urban sites, and smaller cities may differ. Lastly, an ERA5 grid cell is ~31 km across, and the grid point nearest a city center is a coarse proxy for the urban canopy, mainly in complex terrain such as Lhasa or Chongqing.

6. Conclusion

An unsupervised mining pipeline, K-means clustering of standardized ERA5 features followed by OLS decadal slopes on the clustered temperature series, was applied to 34 Chinese cities over 1975–2024. Based on the five climate features from 2020, clustering gave five regions that mostly follow conventional geography, though Taiyuan, Hohhot, and Xi'an moved across provincial lines. All five regions warmed, though unequally. The Northwest warmed at about twice the rate of South China (0.51 versus 0.24 °C per decade), and the Qinghai–Tibet Plateau at about 1.6 times that rate (0.40 °C per decade). Over time the warming was also uneven, since the North–Northeast and South China regions paused in 2005–2014 and the Northwest and the plateau kept warming. Inside three of the five clusters, city-level warming differed by more than 1.1 °C; therefore, regional averages alone understate what individual cities face. Since the pipeline needs no zoning labels, it can be applied in the same way to other variables, periods or city samples. These estimates rest on decadal means at one grid point per city and on a fixed number of clusters. Trend tests on annual series, an objective choice of K, and a larger sample with smaller cities could improve them.

References

[1] Sun, Y., Zhang, X., Ren, G., Zwiers, F. W., & Hu, T. (2016). Contribution of urbanization to warming in China. *Nature Climate Change, 6*(7), 706–709. https://doi.org/10.1038/nclimate2956

[2] Li, K., Wang, M., & Liu, K. (2022). The study of temperature regionalization in China using complex networks. *International Journal of Climatology, 42*(8), 4445–4459. https://doi.org/10.1002/joc.7478

[3] Chi, H., Wu, Y., Zheng, H., Zhang, B., Sun, Z., Yan, J., Ren, Y., & Guo, L. (2023). Spatial patterns of climate change and associated climate hazards in Northwest China. *Scientific Reports, 13*(1), Article 10418. https://doi.org/10.1038/s41598-023-37349-w

[4] Lin, X., Wang, Y., & Song, L. (2024). Urbanization amplified compound hot extremes over the three major urban agglomerations in China. *Geophysical Research Letters, 51*(2), Article e2023GL106644. https://doi.org/10.1029/2023GL106644

[5] Li, Y., Zhou, D., Yan, Z., & Zhang, L. (2022). Contribution of urbanization to local warming in major cities of China (in Chinese). *Environmental Science, 43*(5), 2822–2830. https://doi.org/10.13227/j.hjkx.202109081

[6] Zhang, Q., Jia, B., Li, T., Yang, Y., & Fang, Y. (2025). Dynamic changes and drives of surface urban heat islands in China. *City and Environment Interactions, 27*, Article 100203. https://doi.org/10.1016/j.cacint.2025.100203

DOI: 10.54254/2755-2721/2027.37835

[7] Yang, N., Lu, C., Ouyang, L., Chen, R., Man, W., Wang, Z., Lin, J., Yu, Q., & Li, Z. (2025). Exploring spatial–temporal evolution patterns of urban heat islands in summer and winter: Evidence from a megacity of China. *Scientific Reports, 15*(1), Article 13592. https://doi.org/10.1038/s41598-025-97284-w

[8] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

[9] Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys, 51*(4), 1–41. https://doi.org/10.1145/3161602

[10] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[11] Duan, H., Li, Q., He, L., Zhang, J., An, H., Ali, R., & Vazifedoust, M. (2024). Climate classification for major cities in China using cluster analysis. *Atmosphere, 15*(7), Article 741. https://doi.org/10.3390/atmos15070741

[12] Zhao, P., & He, Z. (2025). Evaluation of ERA5 reanalysis temperature data over the Qilian Mountains of China. *Journal of Mountain Science, 22*(1), 198–209. https://doi.org/10.1007/s11629-024-9108-8

[13] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[14] Du, Q., Zhang, M., Wang, S., Che, C., Ma, R., & Ma, Z. (2019). Changes in air temperature over China in response to the recent global warming hiatus. *Journal of Geographical Sciences, 29*(4), 496–516. https://doi.org/10.1007/s11442-019-1612-3

[15] Yin, A., & Liu, Z. (2025). How can climate adaptation policy enhance urban climate resilience? Empirical evidence from China. *PLOS ONE, 20*(11), Article e0335736. https://doi.org/10.1371/journal.pone.0335736