Clustering Daily Ozone Cycles as Comparable Regimes Across Stations, Years, and Independent Studies: The Querétaro Metropolitan Area, Mexico

Aldo Ledesma-Durán[1*] and Dara Salcedo[1,2]

[1*]Unidad Multidisciplinaria de Docencia e Investigación (UMDI), Facultad de Ciencias, Universidad Nacional Autónoma de México (UNAM), Blvd. Juriquilla 3001, Santiago de Querétaro, 76230, Querétaro, Mexico.

[2]Instituto de Ciencias de la Atmósfera y Cambio Climático, Universidad Nacional Autónoma de México, Circuito Exterior s/n, Ciudad de México, 04510, CDMX, Mexico.

*Corresponding author(s). E-mail(s): aldo_ledesma@ciencias.unam.mx;

Supplementary Material

Note on numbering. Figures and tables in this Supplementary Material are numbered continuously across all appendices (e.g., Figure A1, A2, followed by B3, B4, ...), rather than restarting at 1 within each appendix, to keep each figure and table identifier unique throughout the manuscript and its supplementary material.

This Supplementary Material is organized in four appendices. Appendix A reports full clustering methodology and validation details (preprocessing, K-Means reproducibility, multi-method cross-validation, clusterability, and the Gap statistic) supporting the choice of $k = 2$ and $k = 5$ discussed in the main text. Appendix B extends the atmospheric characterization of Scenario A at $k = 5$ with the complete set of ancillary variables, phase-lag diagnostics, and shape descriptors. Appendix C provides the corresponding full diagnostics for Scenario B at $k = 5$, including partition-stability tests and station-by-station wind breakdowns. Appendix D addresses cross-scenario and cross-resolution comparability, complementing Section 4 of the main text.

Appendix A Clustering details

Preprocessing and imputation. Missing values in Scenario A were imputed variable-by-variable with a Random Forest regression model, tuned by 5-fold grid search; normalized RMSE ranged from roughly 1% to 7% depending on the variable. Ozone itself had no missing values, so imputation affected only the ancillary variables used for post-hoc characterization, never the clustering input. Gaps in Scenario B, arising from the network's staggered commissioning and routine operational interruptions, were imputed with the station-hour mean, a simpler approach suited to the network's coarser variable set. These gaps were minimal: of the 151,200 station-hour-day cells in the pivoted ozone matrix used for clustering (6,300 day–station records × 24 hours), 14 (0.01%) were missing prior to imputation, confined to a single day–station record at station FEO (station-level range: 0% at CAP, COR, EPG, and JOV to 0.04% at FEO).

K-Means reproducibility. Clustering was performed using K-Means with a fixed random seed (`random_state=42`) throughout MacQueen (1967). Three distinct configurations were used at different stages of the analysis. The final production model, from which all reported cluster assignments and figures are derived, used a single initialization (`n_init='auto'`), rendered deterministic by the fixed random seed, identically for Scenario A and Scenario B. Internal validation of the choice of k (silhouette, Calinski–Harabasz, Davies–Bouldin, and the elbow heuristic, below) used `n_init=50` (`max_iter=500`) identically for both scenarios. The separate multi-method cross-validation at a fixed $k = 5$ (comparing K-Means, Agglomerative-Ward, and SOM by ARI and NMI, below) used `n_init=50` for Scenario A and `n_init=20` for Scenario B, reduced to accommodate its larger sample size (6300 station-days versus 366);

this is the only one of the three configurations that differs between scenarios.

Internal K-Means validation. For Scenario A, the inertia elbow method suggested $k = 5$, while the silhouette coefficient was maximized at $k = 2$ (0.253 versus 0.166 at $k = 5$). For Scenario B, the elbow method also suggested $k = 5$, consistent with Scenario A, while the silhouette, Calinski–Harabasz, and Davies–Bouldin indices all favored $k = 2$. This divergence between the elbow heuristic and the remaining criteria, present in both scenarios, motivated a broader validation strategy.

Multi-method cross-validation. Three independent clustering algorithms — K-Means, Agglomerative clustering (Ward linkage), and Self-Organizing Maps (SOM) — were evaluated using silhouette score, Calinski-Harabasz index, and Davies-Bouldin index Kohonen (1990); Ward (1963); Rousseeuw (1987); Caliński and Harabasz (1974); Davies and Bouldin (1979). For Scenario A, all three methods favored $k = 2$ across all three metrics (Table A1; K-Means: silhouette 0.253, CH 162.95, DB 1.436; SOM: silhouette 0.326; Agglomerative: silhouette and CH optimal at $k = 2$, DB marginally better at $k = 3$). At a fixed $k = 5$, no pair of methods reached strong agreement (Table A2; ARI < 0.6 for all six method pairs, including K-Means against its own production reference, ARI = 0.451). For Scenario B, this convergence was even more unanimous: all three methods and all three metrics favored $k = 2$ without exception (K-Means silhouette 0.307, Agglomerative 0.279, SOM 0.374); cophenetic correlation at $k = 5$ was low (0.499), and inter-method agreement was weak (ARI 0.29–0.58), except between K-Means and its own production reference (ARI = 0.765) and between SOM and that same reference (ARI = 0.675), both of which exceeded the informal strong-agreement threshold of 0.6.

Despite this difference in initializations between scenarios (`n_init=50` vs. 20, above), silhouette, Calinski–Harabasz, and Davies–Bouldin values were stable across it to three decimal places. The markedly higher production-versus-validation ARI in Scenario B (0.765) relative to Scenario A (0.451) indicates that the $k = 5$ solution is comparatively more sensitive to reinitialization in Scenario A, and comparatively more stable in Scenario B, despite the smaller number of initializations used for the latter (Fig. A1).

Clusterability. The Hopkins statistic confirmed a genuine clustering tendency in both scenarios, above the level expected from uniform random data ($H = 0.739 \pm 0.010$ for Scenario A, $H = 0.8396 \pm 0.0019$ for Scenario B; 100 Monte Carlo runs) Hopkins and Skellam (1954).

Gap statistic. Using the standard (unrotated) null reference, the Gap statistic did not converge within $k = 1$–11 for either scenario; correcting the null by PCA rotation Tibshirani et al. (2001) converged to $k = 2$ for Scenario A and, more conservatively, to $k = 1$ for Scenario B, with the underlying Gap values at $k = 1$ and $k = 2$ close (1.8255 versus 1.8171) – a narrow margin rather than a strong preference for no clustering structure.

See main text for full discussion of the $k = 2/k = 5$ trade-off for the choice of the number of clusters.

Robustness to normalization and initialization. As a further check beyond the multi-method comparison above, the production clustering (ozone cycle only, min-max scaled) was compared against 100 independent re-clusterings of the same ozone cycle under z-score scaling and random seeds (Table A3). ARI favors the coarser $k = 2$ partition (A: 0.644; B: 0.648, vs. $k = 5$'s 0.539/0.564) Hubert and Arabie (1985), while NMI favors $k = 5$ (A: 0.632; B: 0.611, vs. $k = 2$'s 0.596/0.529) Strehl and Ghosh (2002) – the same ARI/NMI trade-off already seen above, now under a change in scaling rather than in algorithm.

Summarizing ancillary variables by cluster for interpretability. As described in the main text (Section 3.1.1), differences between clusters for each variable were tested using one-way ANOVA with Benjamini–Hochberg (BH) correction. For each ancillary variable, five candidate daily statistics (minimum, maximum, mean, range, and standard deviation) were computed and compared across clusters. The statistic yielding the highest discriminating F-value was retained for that variable in the resulting cluster-summary tables. This post-hoc selection involves five correlated tests per variable and is not covered by the BH correction, which addresses the family of variables. Accordingly, the cluster mean is used uniformly as the default summary statistic for cluster characterization in the main text and Supplementary Material, with one exception: wind speed is additionally reported using its daily minimum in some cases because of its direct mechanistic interpretation as a floor on sustained ventilation; both statistics yield the same cluster ranking for this variable.

Sensitivity of a priori design choices. Two design choices were fixed a priori rather than tuned: the 25-hour representation of the daily cycle, and the percentile thresholds of the Scenario A operational criterion (Section 3.1.1). Neither materially changes the results: an alternative 24-hour representation left the robust $k = 2$ partition essentially unchanged (ARI ≥ 0.965 against production, both scenarios; $k = 5$ more sensitive, ARI 0.434–0.585, consistent with its weaker support throughout), and two alternative percentile-threshold pairs for the operational criterion (0.55% at 80th/20th, 3.01% at 70th/30th, versus 1.64% at the original 75th/25th) still placed every qualifying day exclusively in the high-ozone cluster.

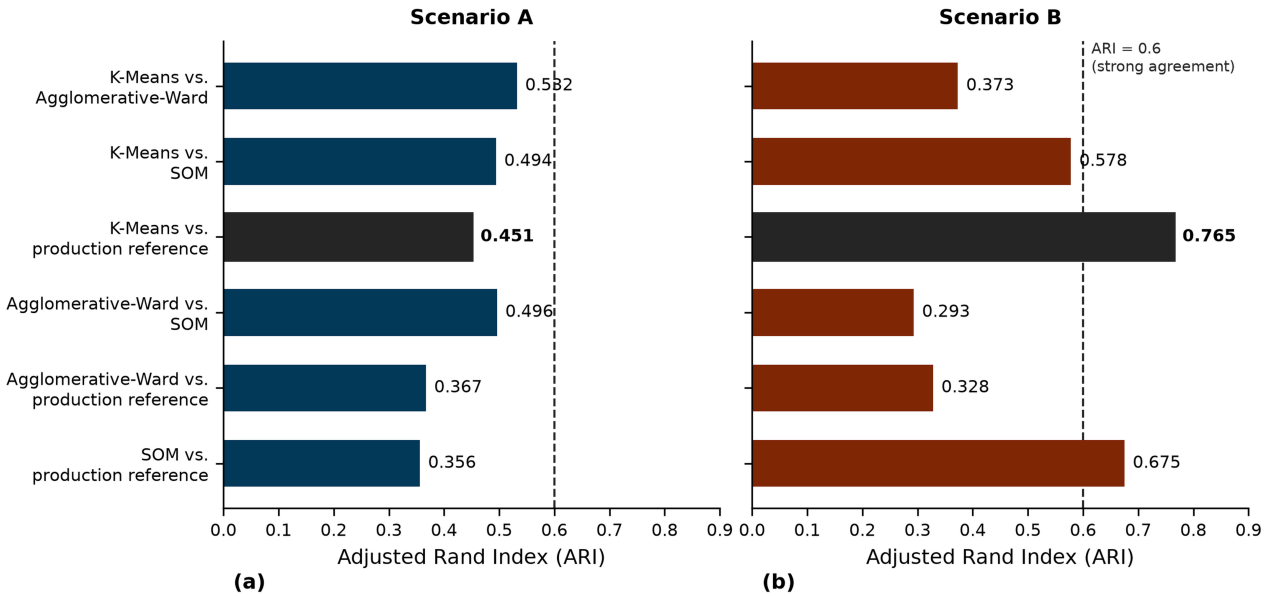


Fig. A1: Pairwise Adjusted Rand Index (ARI) at $k = 5$ across clustering methods, for Scenario A and Scenario B. The K-Means production-versus-validation pair is highlighted; the dashed line marks the informal strong-agreement threshold (ARI = 0.6).

	Scenario A						Scenario B					
	$k = 2$			$k = 5$			$k = 2$			$k = 5$		
Method	Sil.	CH	DB	Sil.	CH	DB	Sil.	CH	DB	Sil.	CH	DB
K-Means	0.253	162.95	1.436	0.166	98.00	1.569	0.307	3380.46	1.273	0.176	1961.81	1.533
Agglomerative-Ward	0.207	122.55	1.651	0.147	85.94	1.627	0.279	3105.21	1.345	0.121	1585.91	1.679
SOM	0.326	66.91	1.156	0.242	46.51	1.206	0.374	322.44	1.035	0.252	212.16	1.204

Table A1: Internal clustering validation indices (silhouette coefficient, Calinski–Harabasz index, Davies–Bouldin index) at $k = 2$ and $k = 5$, for the three clustering algorithms evaluated (K-Means, Agglomerative with Ward linkage, Self-Organizing Maps) and both monitoring-network scenarios. Higher is better for silhouette and Calinski–Harabasz; lower is better for Davies–Bouldin. Across all three methods and both scenarios, every index favors $k = 2$ over $k = 5$ without exception.

Appendix B Scenario A ($k = 5$) – full atmospheric characterization and phase structure

Ozone correlations, primary partition ($k = 2$). Exact statistics for the four variables discussed in Section 3.1.1 (main text): ozone was negatively correlated with NO ($r = -0.377$, $p_{\text{adj}} = 2.0 \times 10^{-12}$) and NO_2 ($r = -0.223$, $p_{\text{adj}} = 1.9 \times 10^{-4}$); NO_3 and wind speed showed no significant correlation with ozone after BH correction ($p_{\text{adj}} = 0.244$ and 0.669, respectively).

Section 3.1.2 reports a compact, 9-variable summary of the atmospheric conditions distinguishing the five Scenario A clusters. The full set of 21 ancillary variables (all available chemical species, meteorological variables, and wind direction) is shown here for completeness (Figures B2 and B3).

As a complement to the normalized cluster-mean summaries above, Figure B4 shows the full mean diurnal cycle (0–23h, not just a single normalized value) for each of the 21 ancillary variables plus ozone, by cluster. This makes visible, for example, whether two clusters that share a similar mean value for a given variable differ in the timing of that variable's daily cycle – information the normalized bar summaries above do not show.

The lag between ozone and each ancillary variable was computed per cluster, on the cluster's mean (smoothed) 25-hour diurnal cycle: a full cross-correlation between the two mean cycles, restricted to $|\text{lag}| \leq 24$ h, with the reported lag being the hour of maximum correlation. This identifies relative phase between average cycles, not a causal or day-by-day relationship. The resulting hour-of-peak-cross-correlation lag between ozone and each ancillary variable (Figures B5 and B6) is compatible with the pattern already described in the main text (Section

Method pair	Scenario A		Scenario B	
	ARI	NMI	ARI	NMI
K-Means vs. Agglomerative-Ward	0.532	0.604	0.373	0.501
K-Means vs. SOM	0.494	0.582	0.578	0.619
K-Means vs. production reference	0.451	0.585	0.765	0.783
Agglomerative-Ward vs. SOM	0.496	0.602	0.293	0.454
Agglomerative-Ward vs. production reference	0.367	0.485	0.328	0.482
SOM vs. production reference	0.356	0.488	0.675	0.663

Table A2: Pairwise agreement (Adjusted Rand Index and Normalized Mutual Information) between the three clustering methods at a fixed $k = 5$, for both scenarios. "Production reference" denotes the K-Means labels used throughout the manuscript for all reported cluster assignments and figures (a single deterministic initialization, `n_init='auto'`, `random_state=42`); the "K-Means" row/column instead refers to the independently re-fit K-Means solution computed inside the validation notebook for direct comparison against Agglomerative-Ward and SOM (`n_init=50` for Scenario A, `n_init=20` for Scenario B). ARI ≥ 0.6 (the informal threshold used to flag "strong" agreement elsewhere in this analysis) is reached for only two of six pairs, both in Scenario B (K-Means vs. production reference, and SOM vs. production reference).

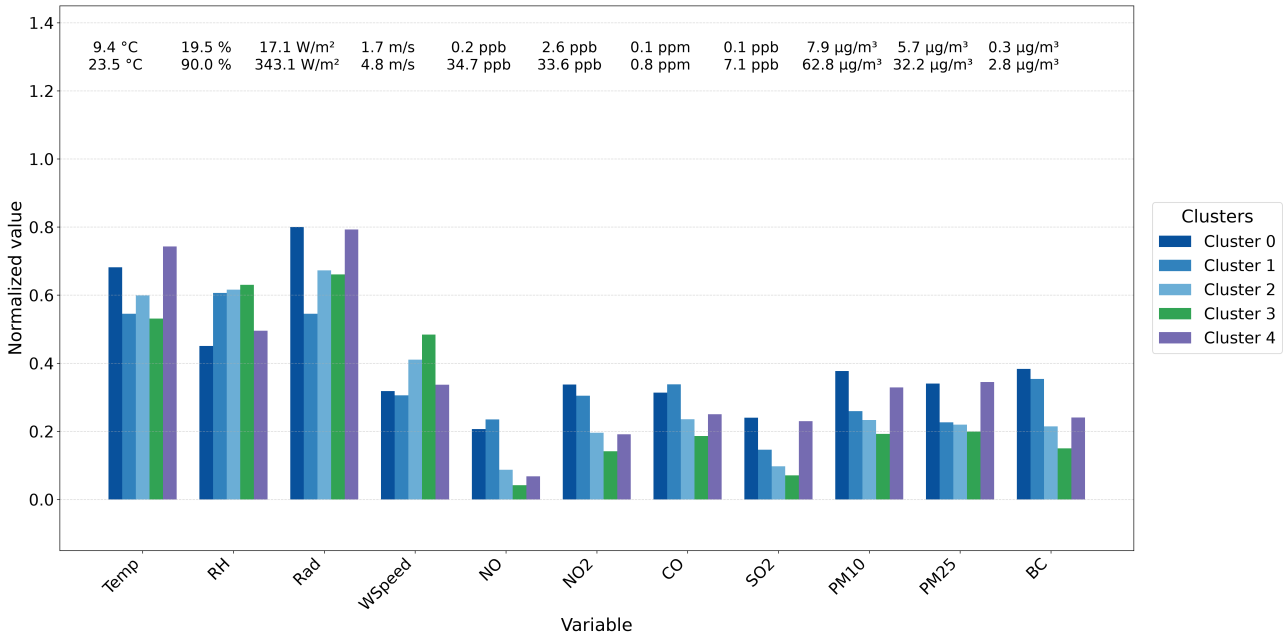


Fig. B2: Normalized cluster-mean values of the 11 atmospheric variables, Scenario A ($k = 5$). The absolute minimum and maximum values of each variable are specified at the top.

Table A3: Structural robustness of clustering to normalization and initialization (ozone-only re-clustering under z-score scaling vs. the min-max-scaled production labels, 100 random seeds). Values are ARI and NMI (mean ± std).

k	Scenario	ARI	NMI
2	A	0.644 ± 0.000	0.596 ± 0.000
2	B	0.648 ± 0.0001	0.529 ± 0.0001
5	A	0.539 ± 0.012	0.632 ± 0.007
5	B	0.564 ± 0.001	0.611 ± 0.001

3.1.2): NO, CO, and black carbon lead ozone by several hours (exact per-cluster values in Figure B5; NO and CO lead by 7–10 hours), all three peaking during the morning traffic rush hour, consistent with a shared combustion source. PM10 lags by 6–9 hours, and temperature and radiation are essentially synchronous with ozone (lag 0–2h) in all five clusters. Because this ordering is shared across clusters, it is not used as a cluster-discriminating feature.

Table B4 reports BH-adjusted Tukey pairwise comparisons for NO, solar radiation, and ozone across all cluster pairs with statistically equivalent NO. Three of the four such pairs also show significant differences in both radiation and ozone (clusters 0/1, 2/4, and 3/4); the fourth (clusters 2 and 3) shows no significant radiation difference either ($p_{\text{adj}} = 0.999$), and is therefore not part of the radiation-mediated pattern discussed in the main text.

Beyond these chemical findings, the five clusters also differ systematically in the shape and level of the diurnal ozone cycle itself, illustrating how a finer

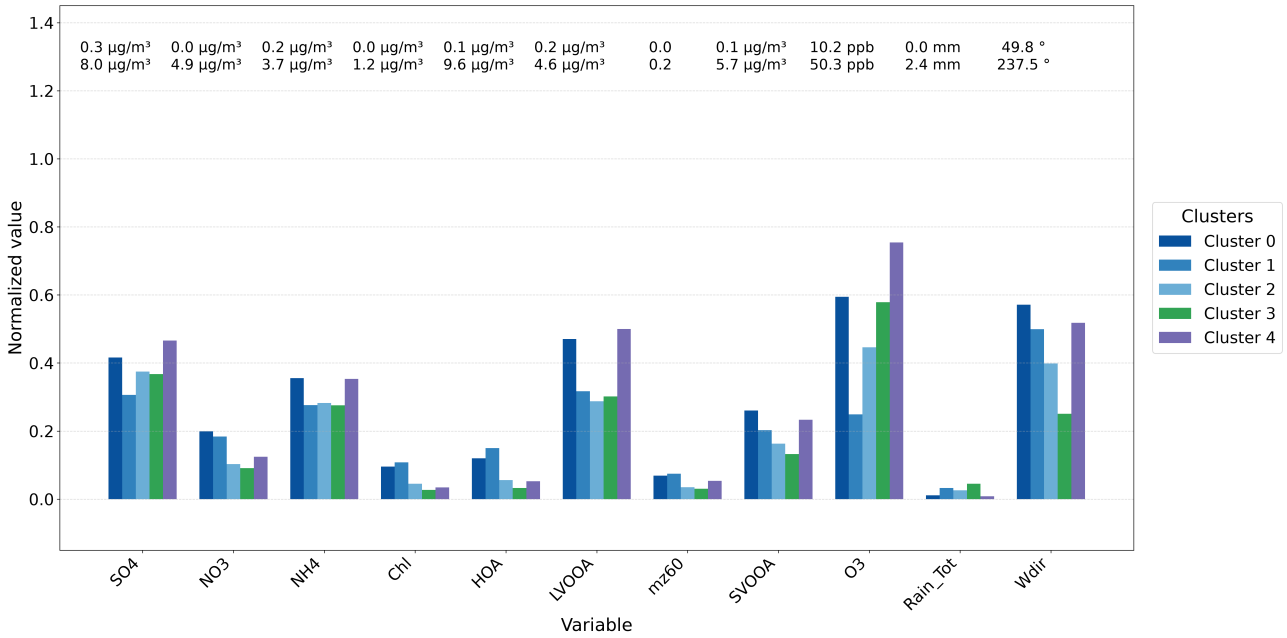


Fig. B3: Normalized cluster-mean values of the remaining 10 atmospheric variables, Scenario A ($k = 5$). The absolute minimum and maximum values of each variable are specified at the top.

Cluster pair	NO p_{adj}	Rad p_{adj} (diff.)	O_3 p_{adj} (diff.)
0 vs 1	0.79	< 0.001 (-83.1 W/m^2)	< 0.001 (-13.9 ppb)
2 vs 4	0.87	< 0.001 ($+39.2$ W/m^2)	< 0.001 ($+12.4$ ppb)
3 vs 4	0.86	< 0.001 ($+43.1$ W/m^2)	< 0.001 ($+7.0$ ppb)
2 vs 3	0.23	0.999 (n.s.)	< 0.001 ($+5.3$ ppb)

Table B4: BH-adjusted Tukey pairwise comparisons for NO, solar radiation, and ozone, restricted to the four cluster pairs with statistically equivalent NO, Scenario A ($k = 5$). The first three pairs show the radiation-mediated pattern discussed in the main text; the fourth (2 vs 3) does not, since radiation is also statistically equivalent in that pair.

partition characterizes what kind of day each record represents (Figure B7). Cluster 0 ($n = 67$) shows a sharp, late-peaking cycle: above-average ozone (mean 34.0 ppb) reached through the fastest morning rise of the five clusters, peaking around 15:00–16:00h (e.g., 2015-09-08). Cluster 1 ($n = 78$) has the lowest mean and maximum ozone of the five clusters (20.1 ppb, 44.0 ppb) but, proportionally, the widest relative amplitude, indicating a cycle that is low in absolute terms yet strongly oscillating (e.g., 2016-01-11). Cluster 2 ($n = 101$, the largest) shows no extreme value on any of the five descriptors and represents the closest approximation to a typical day in the campaign (e.g., 2015-08-01). Cluster 3 ($n = 60$) is characterized by the flattest, most compressed cycle of the five – low maximum, high minimum, slowest morning rise, earliest peak (∼13:40h) – and corresponds to the same structurally atypical cluster whose wind-ventilation regime is examined in Section C.1; no additional mechanism is proposed here beyond what is outlined there. Cluster 4 ($n = 60$) shows the highest mean and maximum ozone of the five clusters (40.4 ppb, 66.7 ppb) together with a fast rise and pronounced afternoon decline, consistent with the campaign's highest photochemical-loading days (e.g., 2015-04-27, 2015-08-09).

Seasonal concentration was examined for the five clusters (Figure B8). Two of the five showed a clear pattern: cluster 1, the least photochemically active regime (lowest ozone, temperature, and radiation; highest NO and NO_2), concentrated in November–December (50% of its 78 days), consistent with the winter concentration already identified for the corresponding low-ozone regime at $k = 2$ (Section 3.1.1) and reinforcing that finding across independently derived partitions. Cluster 4, the highest-ozone, highest-radiation regime, concentrated in April (28% of its 60 days). Cluster 0 showed a milder tendency toward April–May (32.8% of its 67 days, versus ∼17% expected by chance) that did not reach the concentration threshold used for clusters 1 and 4. The remaining two clusters (2, 3) showed no comparable seasonal concentration, distributed across the annual cycle.

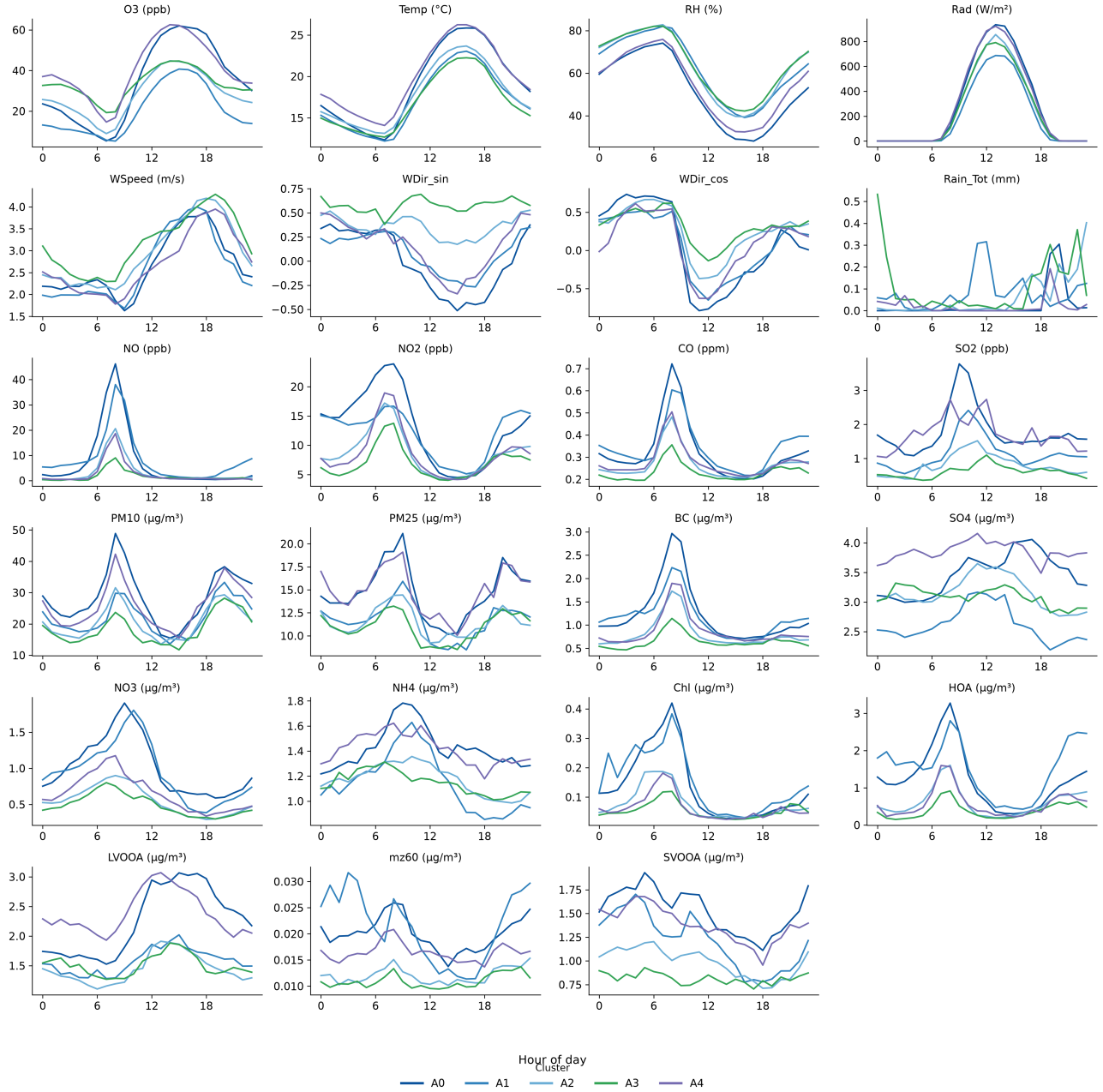


Fig. B4: Mean diurnal cycle (0–23h) of ozone and all 21 ancillary variables, by cluster, Scenario A ($k = 5$).

Detailed contrasts: Scenario A cluster 3 wind speed. All four BH-corrected pairwise contrasts of cluster 3 against the remaining Scenario A clusters were significant (cluster 3 vs. 0, $p_{\text{adj}} = 1.1\times10^{-8}$; vs. 1, $p_{\text{adj}} = 1.1\times10^{-8}$; vs. 2, $p_{\text{adj}} = 0.030$; vs. 4, $p_{\text{adj}} = 6.0\times10^{-5}$). One pair (cluster 3 vs. 2) is more sensitive to the correction procedure than the other three: it remains significant under the targeted, four-contrast BH correction reported above but not under a full Tukey-Kramer correction across all ten pairwise cluster comparisons ($p_{\text{adj}} = 0.097$), reflecting the larger family-wise penalty of testing all ten pairs rather than the four contrasts of direct interest; the other three contrasts are significant under both procedures.

Appendix C Scenario B, $k = 5$ – full diagnostics

This section expands the exploratory $k = 5$ diagnostics for Scenario B summarized in Section 3.2.2, with the full atmospheric characterization and partition-stability tests referenced there.

Only three of five clusters nest cleanly (>90%) within a single $k = 2$ cluster; the remaining two split across both $k = 2$ groups (cluster 2, 69.8%/30.2%, and cluster 3, 8.6%/91.4%). The 70/30 split (cluster 2) was examined for an explanatory mechanism: the two subgroups differ in seasonal occurrence (one concentrated in spring, the other in summer), ozone,

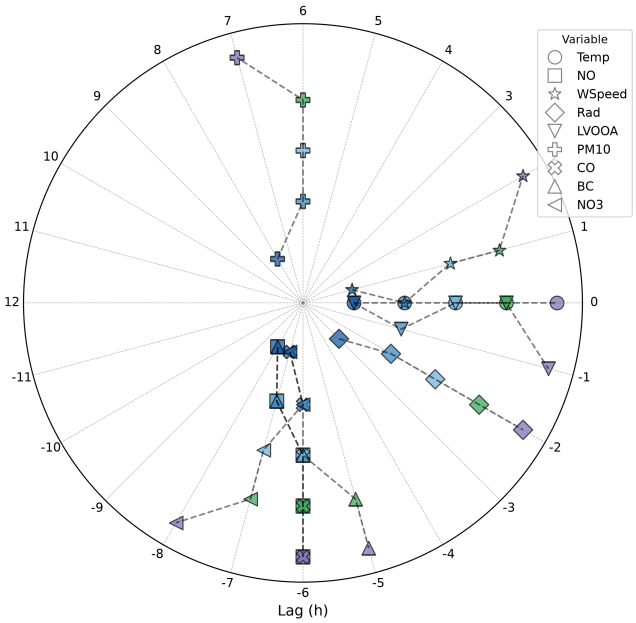


Fig. B5: Peak cross-correlation lag (hours, relative to ozone) by cluster, Scenario A ($k = 5$), for Temp, NO, WSpeed, Rad, LVOOA, PM10, CO, BC, and NO_3.

temperature, and relative humidity, but not in monitoring station, year, or NO_2. Wind speed also did not explain the split (Welch $t = 1.477$, $p = 0.140$; Mann-Whitney $p = 0.206$); solar radiation was not available for testing in this network. No confirmed mechanism has been identified; it is reported as descriptive evidence of partition instability at $k = 5$ rather than an independent physical finding.

At $k = 5$, wind direction does not split cleanly into consistent and inconsistent clusters: cluster 2 is the only cluster whose network-pooled dominant sector (60–90°, 25.4%) is not shared by any other, while clusters 0 and 3 share a dominant sector (210–240°, 14.8% and 15.1% of their hours, respectively) and clusters 1 and 4 share another (90–120°, 19.3% and 23.8%). No cluster shows full agreement across all five stations. Cluster 2, the closest to a shared pattern, is in fact a station-block split (CAP, COR, and FEO favoring 90–120°; EPG and JOV favoring 60–90°); cluster 4 shows an analogous partial split (CAP, COR, and FEO favoring 90–120°; EPG favoring 60–90°; JOV favoring 210–240°). Clusters 0, 1, and 3 show no dominant sector shared by more than two of the five stations (see the station-by-cluster wind distribution in Figure C9). The NO_2–ozone lag did not show a stable two-group split across partitions: at $k = 2$, one cluster showed a positive lag (+8h) and the other a negative lag (−8h); at $k = 5$, the split was 3-against-2, a reminder that cluster indices are not stable identities across different values of k. Table 3 reports the complete station-by-cluster distribution.

Seasonal structure at $k = 5$ showed one cluster concentrated narrowly in May, one more frequent in August (with a broader distribution and higher weekend frequency), one predominant in winter (October–March), one bimodal (July and December), and cluster 3 concentrated in spring (March–May, 46.6% of its days). The complete station-by-cluster distribution, already introduced in the main text (Table 3), is not repeated here.

Table 4 summarizes the key atmospheric variables distinguishing the five Scenario B clusters; the complete set of 11 ancillary variables (plus ozone itself, shown for reference alongside them using the same convention as Scenario A) is shown in Figures C10 and C11, the mean diurnal ozone cycle in Figure 7, and the station-by-cluster wind distribution in Figure C9.

As a complement to the normalized cluster-mean summaries above, Figure C12 shows the full mean diurnal cycle (0–23h) for each of the 11 ancillary variables plus ozone, by cluster.

A concise cluster-by-cluster summary is given in the main text (Section 3.2.2). Exact per-station wind-direction breakdowns are shown in Figure C9.

A note on circular wind direction.

The circular mean wind direction reported in Figure C11 can diverge from the dominant directional bin reported in Table 4. The clearest case is cluster 2: its circular mean (88.0°) sits at the edge of its dominant bin (60–90°), but this apparent agreement is not evidence of station-level consistency – the underlying pattern is a station-block split, with three stations (CAP, COR, FEO) favoring the adjacent 90–120° sector and two (EPG, JOV) favoring 60–90°, rather than a shared direction at every station. No cluster in the corrected network shows full agreement across all five stations; station-level heterogeneity is the norm across the five-cluster partition rather than the exception. For this reason, both the dominant-bin values in Table 4 and the circular means in Figure C11 should be read as network-level summaries rather than indicators of per-station consistency.

Scenario B cluster 2: internal split. Wind speed was tested and rejected as an explanation for the internal split within Scenario B cluster 2 (Welch $t = 1.48$, $p = 0.140$). The minority subgroup is concentrated in spring and shows higher ozone and temperature with lower relative humidity than the majority subgroup.

C.1 Atypical clusters

Both scenarios contain a cluster that stands apart on structural, rather than purely chemical, grounds. In Scenario A this signature converges on a single cluster and is largely resolved below; in Scenario B it scatters across three clusters and stays genuinely unresolved – the two outcomes together are why this subsection is titled as it is. Figure C14 shows the wind

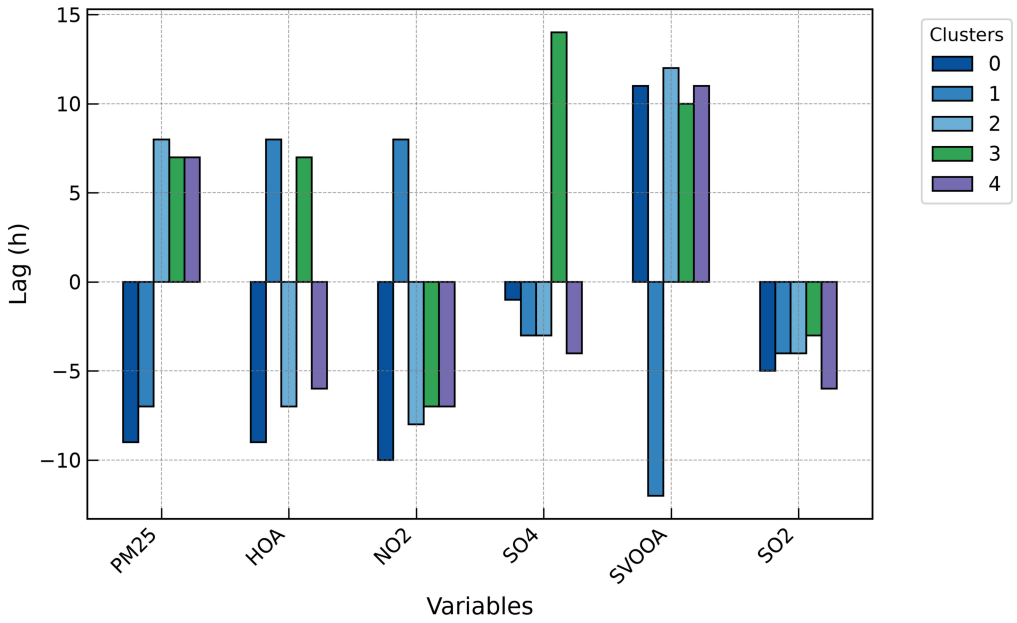


Fig. B6: Peak cross-correlation lag (hours, relative to ozone) by cluster, Scenario A ($k = 5$), for PM2.5, HOA, NO_2, SO_4, SVOOA, and SO_2.

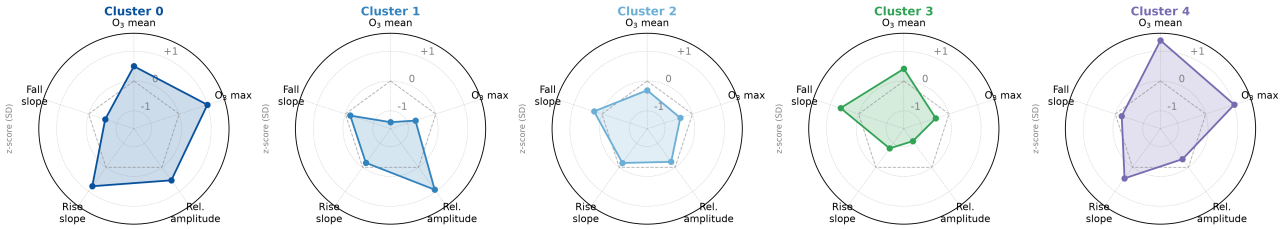


Fig. B7: Z-score profile of the five Scenario A clusters ($k = 5$) across the two level descriptors (daily mean and maximum ozone) and three shape descriptors (relative amplitude, morning rise rate, afternoon fall rate).

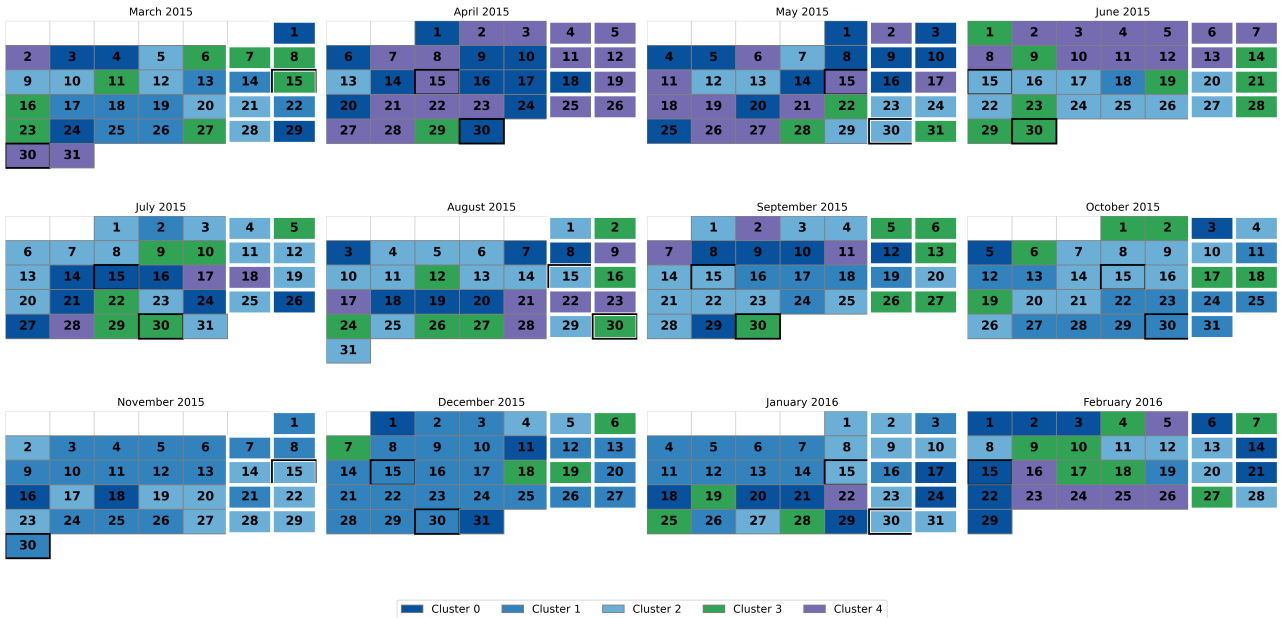


Fig. B8: Annual calendar of daily cluster assignment, Scenario A ($k = 5$). Each cell is the cluster of that calendar day; unlike the Scenario B calendar (Figure 7), Scenario A has a single station, so no cross-station agreement statistic applies here.

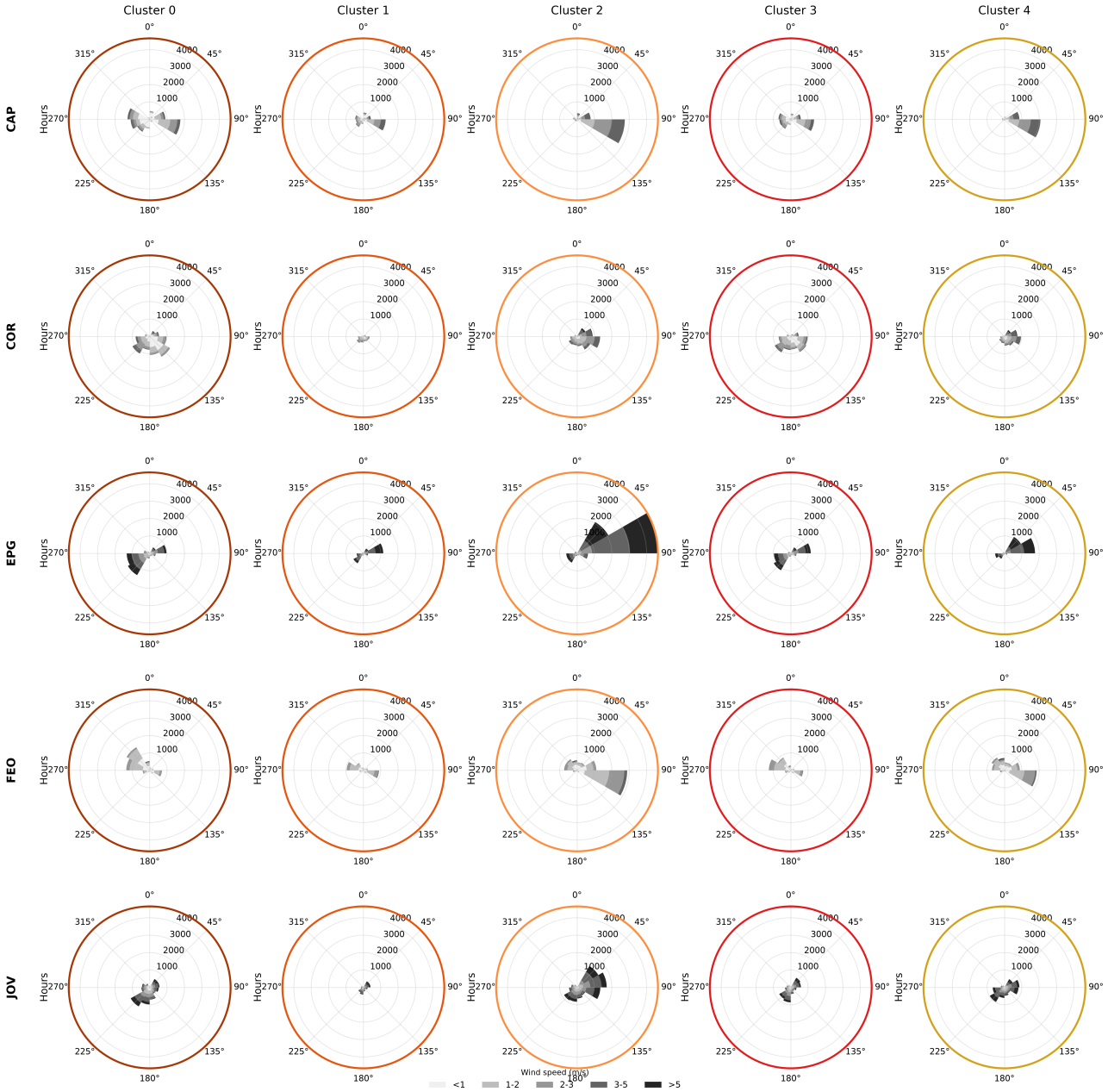


Fig. C9: Wind-direction and wind-speed distribution for each of the five Scenario B clusters ($k = 5$), broken down by monitoring station (rows: CAP, COR, EPG, FEO, JOV; radial scale uniform across panels).

direction and diurnal ozone signature of Scenario A's atypical cluster alongside the diurnal cycles of the two Scenario B clusters discussed below.

In Scenario A, cluster 3 ($n = 60$) combines three distinguishing features in a single group: a mean wind direction markedly different from the other four (96.9° against a 124.6–157.1° range; Figure 5), the flattest diurnal ozone amplitude of the five clusters (Figure B7), and the only partition that does not nest cleanly into $k = 2$ (55.0%/45.0% split). Three plausible causal explanations were tested. A distinct rainfall regime and altered nighttime NO titration were both rejected (not significant, or significant against only one of the other four clusters). Wind speed, by contrast, discriminates cluster 3 from every other cluster ($F = 15.11$, $p_{\text{adj}} = 2.5 \times 10^{-11}$; full pairwise contrasts, including one that is more sensitive to the correction procedure used, in Supplementary Material B). Cluster 3's wind rose is dominated by wind speeds above 5 m/s from a narrow, near-unidirectional sector, together with the lowest concentration of every measured precursor and particulate species (NO, NO_2, CO, PM10, PM2.5, black carbon) among the five clusters. This combination – elevated wind speed, a persistent unidirectional flow, and depressed precursor and particulate loading – is consistent with a mechanically ventilated regime that suppresses local accumulation of both

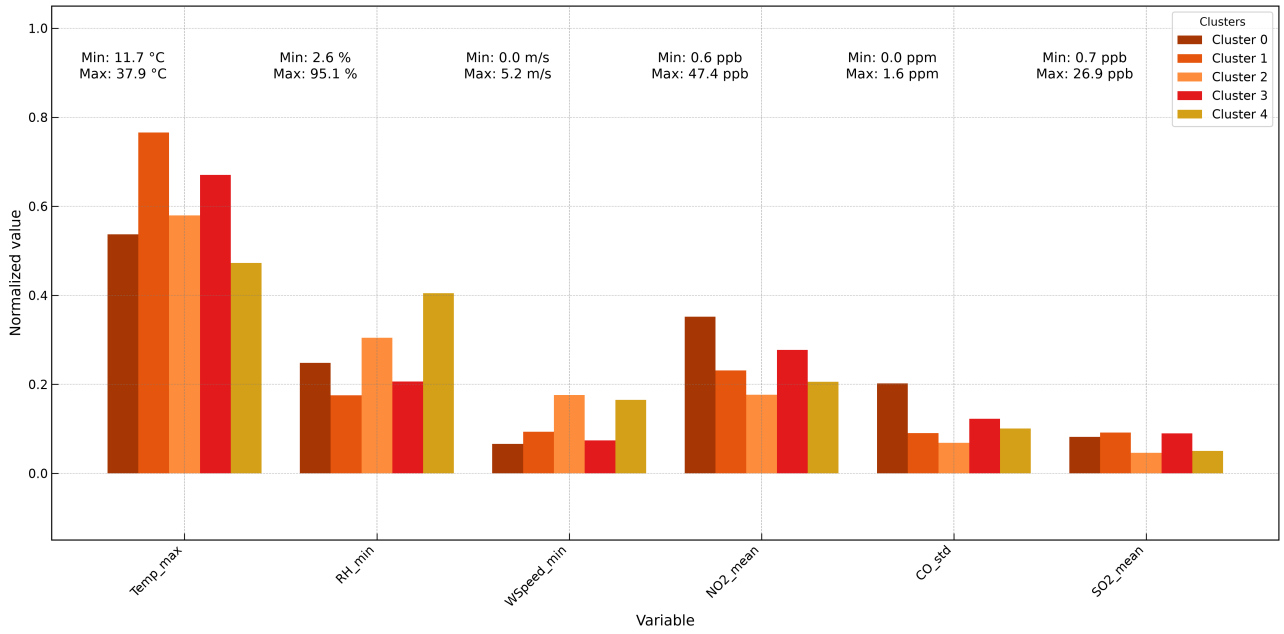


Fig. C10: Normalized cluster-mean values for six ancillary variables, Scenario B ($k = 5$): Temp, RH, WSpeed, NO_2, CO, SO_2. All variables show the cluster mean except WSpeed (minimum, as in Table 4).

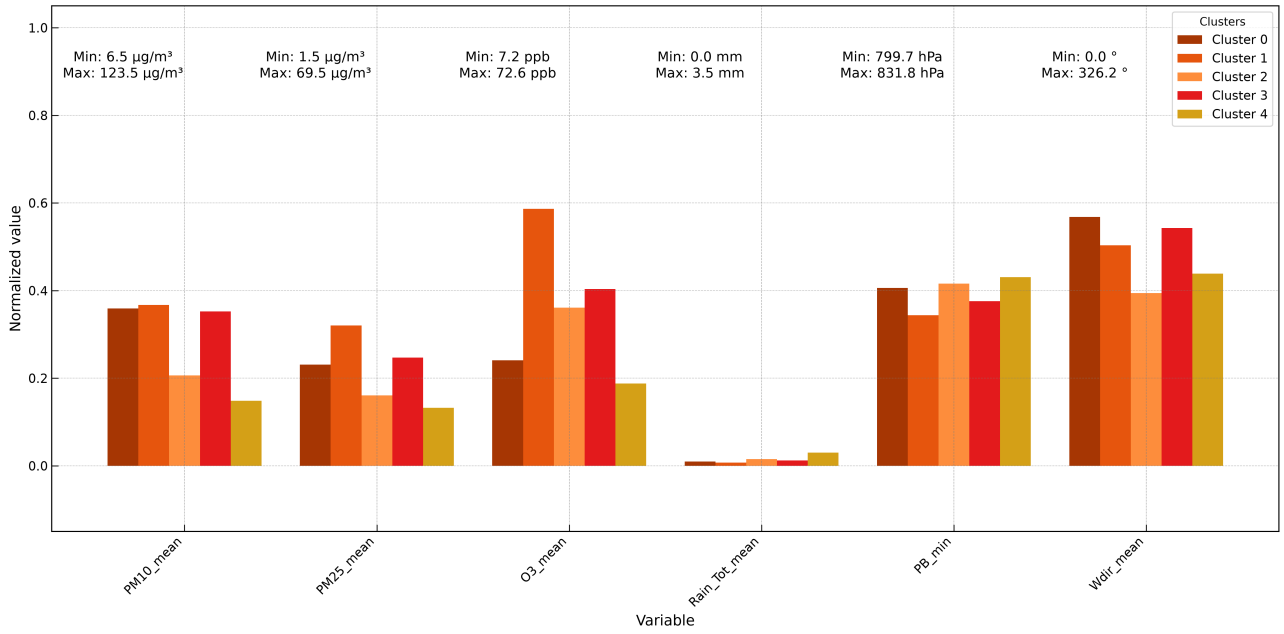


Fig. C11: Normalized cluster-mean values for the remaining ancillary variables, Scenario B ($k = 5$): PM_{10}, $PM_{2.5}$, precipitation, barometric pressure (PB), and circular mean wind direction (Wdir), together with ozone itself (O_3) shown for reference alongside the ancillary variables, not as one of them (Section 2.2 for its role as the clustering input). PM_{10}, $PM_{2.5}$, and precipitation reflect station-respecting aggregation; Wdir uses the circular mean, consistent with Scenario A.

precursors and ozone: a plausible physical mechanism for the flat diurnal amplitude that distinguishes this cluster, supported by convergent statistical evidence across multiple independent variables, though not independently replicated or tested against alternative explanations such as boundary-layer or synoptic effects.

In Scenario B, two of the three markers of atypicality now converge on the same cluster: cluster 2 ($n = 1630$) shows both the least clean nesting into $k = 2$ of any Scenario B cluster (69.8%/30.2%, versus $\geq$91% for every other cluster) and a dominant wind sector (60–90°) not shared by any other cluster (Figure C9). Only the flattest diurnal cycle overall

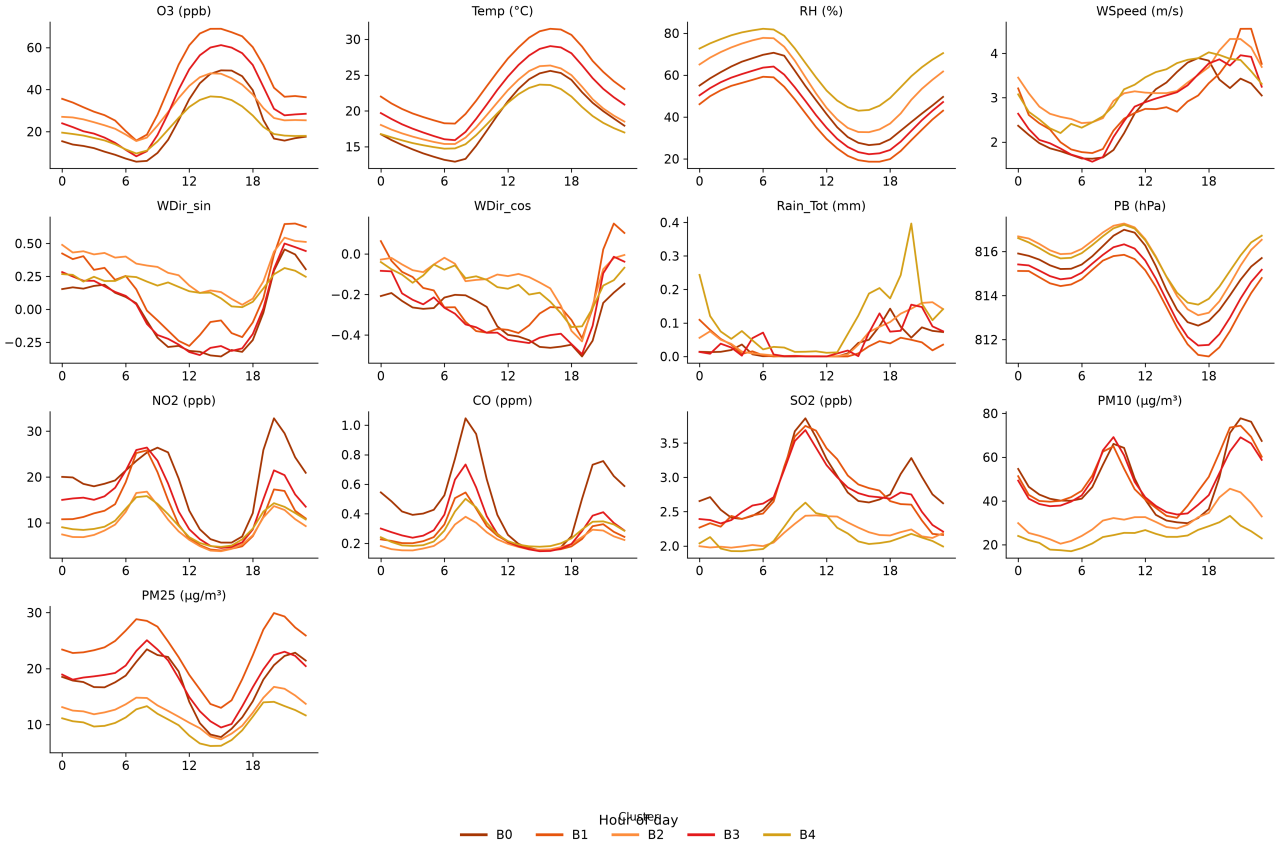


Fig. C12: Mean diurnal cycle (0–23h) of ozone and all 11 ancillary variables, by cluster, Scenario B ($k = 5$).

belongs to a different cluster, cluster 4, which nests essentially perfectly into $k = 2$ (100%/0%). This partial convergence is closer to Scenario A's pattern, where all three markers align on a single cluster, than to three fully independent markers. For cluster 2's internal split, wind speed has already been tested and rejected as an explanation (Supplementary Material C); the minority subgroup (the 30.2% share noted above) is concentrated in spring and shows higher ozone and temperature with lower relative humidity, but — unlike Scenario A's cluster 3 — rainfall and nighttime titration have not yet been tested against it, leaving this an open question in both scenarios, for different reasons.

A replicated ventilation mechanism, and two further atypical signatures.

The mechanical-ventilation mechanism supported for Scenario A cluster 3 (Section C.1; ANOVA $F = 15.11$, $p_{\text{adj}} = 2.5 \times 10^{-11}$) finds an independent counterpart in Scenario B cluster 2.

A one-way ANOVA on wind speed mean across the five Scenario B clusters was likewise highly significant ($F = 158.68$, $p = 8.0 \times 10^{-129}$; restricted to $n = 5598$ day–station records with valid hourly wind data). Tukey HSD post-hoc contrasts confirm that cluster B2 (mean wind speed 2.85 m/s) differs significantly from all four other clusters (Figure C9), including B4 (2.61 m/s; $p_{\text{adj}} < 0.0001$), combining the highest wind speed of the network with the lowest precursor and particulate load among the five clusters (Section 3.2.2) – the same signature, elevated ventilation clearing precursors and flattening the diurnal cycle, found independently in a distinct scenario, dataset, and time period.

B2 and B4 are not statistically indistinguishable in wind speed, and the two clusters do not support a single shared mechanism expressed differently. B2's ventilation signature is self-contained: the strongest wind of the network, the lowest precursor load, and the flattest relative amplitude in the shape-descriptor radar (Figure 9). B4 shares essentially no meteorological driver with B2 – the two clusters remain statistically indistinguishable only in SO_2 and barometric pressure, neither a plausible ventilation driver – and instead stands out through elevated peak-timing instability in the diurnal ozone cycle (67 of 106 atypical cases across the network; Section 4.1), for which no single meteorological driver among the variables available here has been confirmed. Scenario B at $k = 5$ therefore presents two atypical clusters rather than one mechanism with two expressions: B2, which combines the ventilation-driven flattening above with the convergent nesting and wind-direction anomaly reported in Section C.1, and B4's unexplained peak-timing instability, left as

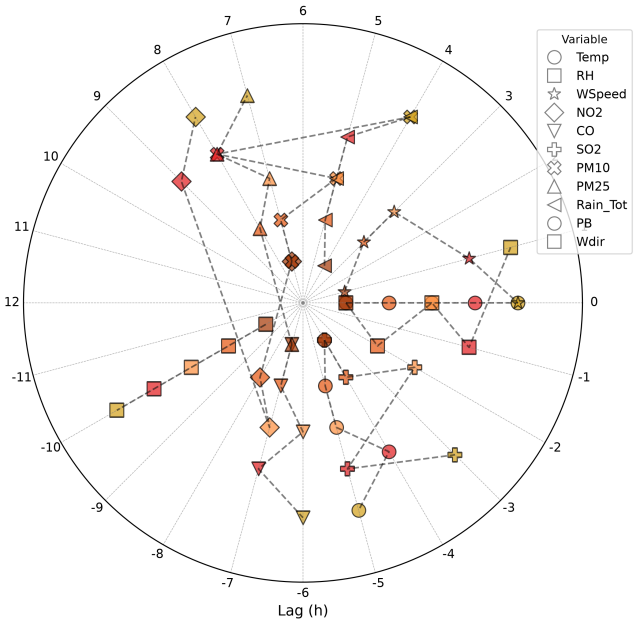


Fig. C13: Peak cross-correlation lag (hours, relative to ozone) by cluster, Scenario B ($k = 5$), for Temp, RH, WSpeed, NO_2, CO, SO_2, PM_{10}, $PM_{2.5}$, precipitation, barometric pressure (PB), and wind direction (Wdir).

an open question for future work with variables not available in the current network (e.g., boundary-layer height or upstream emissions proxies).

Appendix D On comparability

Standardized residuals, cross-station agreement (Section 4.2, main text). For the highest- and lowest-agreement station pairs (COR–JOV and COR–FEO), the large positive standardized residuals that fell on the contingency-table diagonal ranged 14.6–20.3 for COR–JOV and 10.6–15.7 for COR–FEO.

Shape versus level, additional diagnostics. The multinomial logistic classifier reported in Section 4.1 achieved balanced accuracy of 0.851 ± 0.006 (level descriptors alone), 0.754 ± 0.005 (shape descriptors alone), and 0.912 ± 0.007 (both combined; 5-fold cross-validation, mean $\pm$ SD across folds). The complementary ANCOVA (cluster membership as factor, daily mean ozone as covariate) found substantial partial effect sizes for all three retained shape descriptors (η_p^2: relative amplitude 0.477, afternoon fall rate 0.284, morning rise rate 0.264).

The exclusion of cycle asymmetry as a fourth shape descriptor (Section 4.1) reflects that it is computed relative to the same peak-hour reference as peak timing, and is therefore collinear with it ($r = 0.85$) by construction; retaining both would have double-counted a single geometric feature rather than adding independent information.

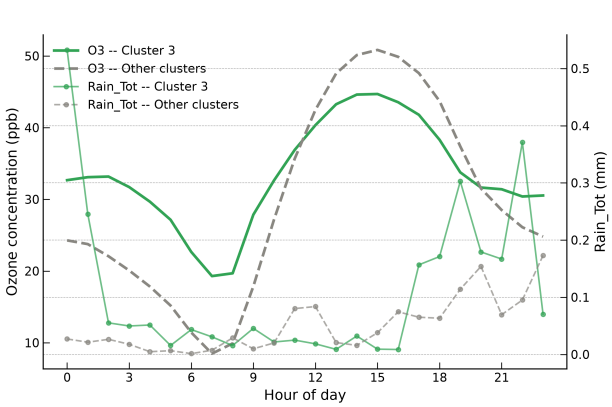
The peak-timing instability noted in the main text (106 of 6300 records, 1.7%, concentrated in cluster 4) was checked against two candidate data-quality explanations, both rejected: imputation (0% of the 106 cases involved an imputed hour) and early-commissioning artifacts at newly installed stations (4.7% of atypical cases fell within the first 60 days of station operation, indistinguishable from the 4.8% base rate). Cluster 4's atypical-peak rate (5.1% internally) is the highest in the network, followed by cluster 2 (2.3%), with the remaining three clusters at $\leq 0.14\%$; because a cycle with little diurnal contrast has, by construction, a poorly defined peak, this is an expected consequence of the flat structure already reported for that cluster (Section C.1), not an independent data-quality concern.

Cross-scenario correspondence at $k = 2$. Figure D15 shows the $k = 2$ analogue of Figure 13 (main text), in the same joint PCA space and using identical axis limits for direct visual comparison. At this coarser resolution, both cluster pairs (A0↔B1, A1↔B0) are fully reciprocal mutual nearest neighbors, with no non-reciprocal correspondence to display.

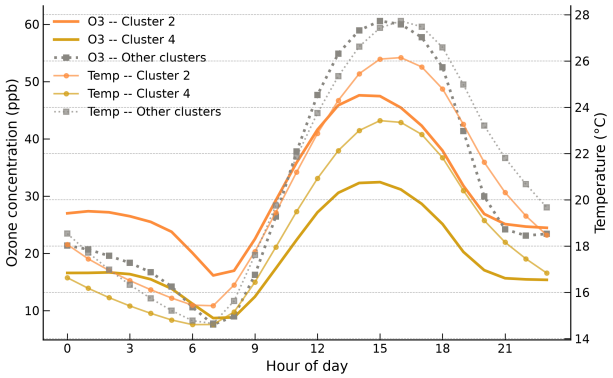
Persistence. Regime persistence, measured as the mean length of consecutive days remaining within the same cluster, was computed on Scenario A's single daily series directly and, for Scenario B, on the regional daily label (the modal cluster across the five stations for each date) rather than per station; it differed between scenarios and depended on the partition examined. At $k = 5$, Scenario A regimes were short-lived (mean run length 1.48–2.53 days), while Scenario B regimes persisted longer (mean 1.79–2.47 days, with maxima up to 21 days). At $k = 2$, both scenarios showed longer persistence than at $k = 5$ (Scenario A: 5.5–6.2 days; Scenario B: 4.1–7.0 days), an expected consequence of coarser partitioning rather than an independent dynamical finding. Persistence and the geometric correspondence assessed next are two independent diagnostics of how the two independently trained solutions relate to each other – one about how long a regime lasts once it starts, the other about how similar two regimes' underlying ozone cycles are.

Sensitivity of comparability metrics to clustering resolution To evaluate whether the comparability results reported in Section 4 depend on the choice of k, the cross-station and cross-period comparability analyses were repeated using the statistically robust $k = 2$ partition (Supplementary Material A) and compared directly against the corresponding $k = 5$ results.

Cross-station regime agreement (Cohen's κ, 10 station pairs) preserved its rank order strongly across resolutions (Spearman $\rho = 0.81$, Pearson $r = 0.86$): the most concordant pair (COR–JOV) and the least concordant pair (COR–FEO) were identical at both



(a) Diurnal cycle, Scenario A, cluster 3 vs. rest of clusters clusters



(b) Diurnal cycle, Scenario B, clusters 2 and 4 vs. rest of

Fig. C14: Wind direction and diurnal cycle of the atypical cluster in Scenario A (cluster 3), alongside the diurnal cycle of the two clusters in Scenario B that show weak nesting into $k = 2$ (cluster 2) or the flattest overall amplitude (cluster 4). Unlike Scenario A, where wind direction, amplitude, and nesting converge on a single cluster, in Scenario B these markers point to different clusters (Section C.1), so no single wind-rose panel for Scenario B is shown here. Scenario A panel shows O3 and precipitation; Scenario B panel shows O3 and temperature (used as the closest available proxy given structural gaps in precipitation coverage across stations).

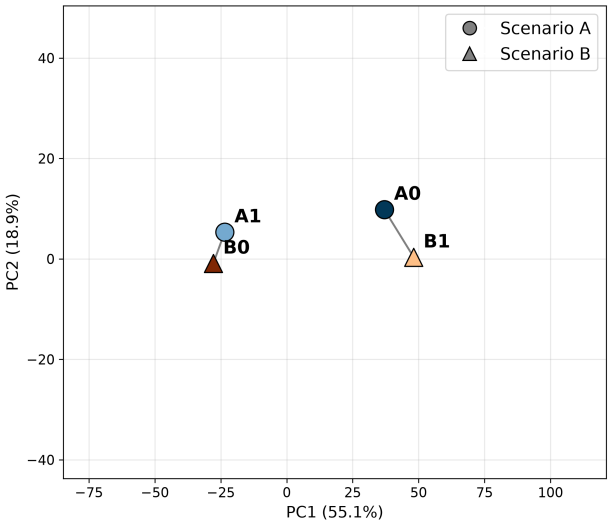


Fig. D15: Cross-scenario correspondence of cluster centroids ($k = 2$) in the same joint PCA space as Figure 13 (main text; identical axis limits for comparability). Circles: Scenario A clusters (A0–A1); triangles: Scenario B clusters (B0–B1). Both pairs (A0↔B1, A1↔B0) are mutual nearest neighbors.

$k = 2$ and $k = 5$. All 10 pairs remained significant after Benjamini-Hochberg correction at both resolutions. Kappa values were systematically higher at $k = 2$ than at $k = 5$ for every pair, an expected consequence of comparing fewer categories rather than evidence of stronger agreement in absolute terms.

Cross-period regime divergence (Jensen–Shannon distance, 15 station-period pairs) preserved its magnitude ordering across resolutions less strongly than cross-station agreement did (Spearman $\rho = 0.80$, Pearson $r = 0.92$), with values systematically smaller at $k = 2$. Statistical power to detect regime shifts between periods dropped sharply at the coarser resolution: of the 10 consecutive-period comparisons (Period 1–2, Period 2–3, across five stations), 6 remained significant after BH correction at $k = 5$ (matching Section 4.3), versus only 2 at $k = 2$ (CAP and FEO, both Period 2–3). The clearest example of this collapse is EPG between Period 1 and Period 2, where the JS distance fell from 0.115 at $k = 5$ to 0.007 at $k = 2$ – a shift that is clearly detectable with five regimes all but disappears when collapsed to two, consistent with reduced statistical power and coarser categories in a lower-dimensionality contingency table rather than a qualitative disagreement between resolutions.

Taken together, these results indicate that spatial (cross-station) comparability is robust to clustering resolution, while cross-period sensitivity specifically requires the higher resolution of $k = 5$ to detect interannual regime instability – an even sharper contrast under the period-based analysis (6/10 vs. 2/10 significant consecutive-period comparisons) than under a calendar-year comparison – one of the central findings of this study. Figure D16 summarizes both comparisons (panel a: cross-station; panel b: cross-period).

Hypothesis 1: convergent validation of Step-1 geometric correspondence via auxiliary variables. Section 4.4 reports that the Step-1 geometric correspondence between Scenario A and Scenario B clusters also holds for auxiliary variables the joint PCA never used. Full method and detail follow.

Variables and preprocessing. Ten cluster fingerprints (5 Scenario A, 5 Scenario B) were compared on the variables measured comparably by both networks:

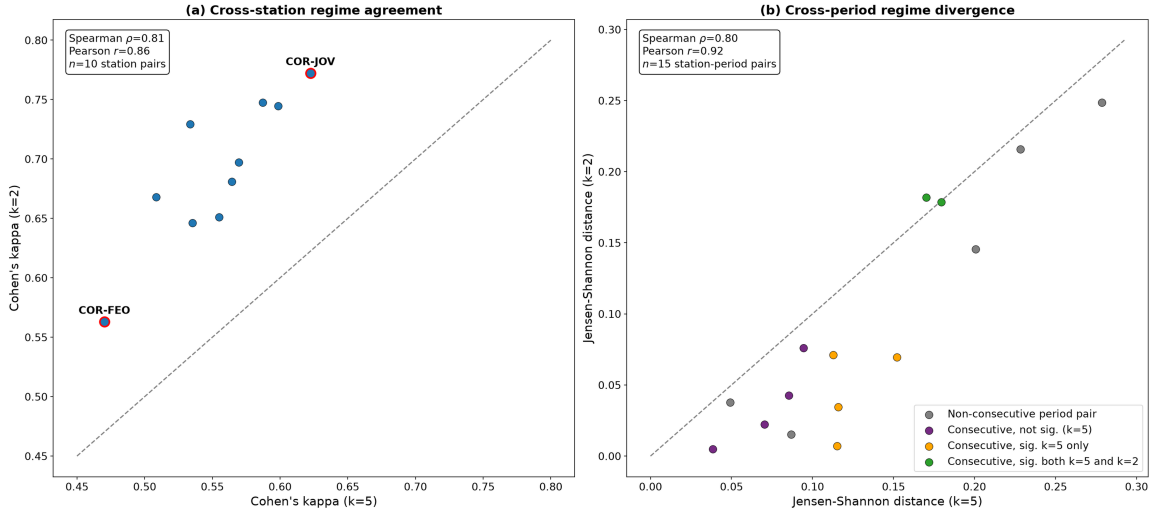


Fig. D16: Sensitivity of comparability metrics to clustering resolution ($k = 2$ vs. $k = 5$), Scenario B. (a) Cross-station Cohen's κ (10 station pairs). (b) Cross-period Jensen-Shannon distance (15 station-period pairs, Periods 1–3; Section 4.3). See text for statistics.

temperature, relative humidity, wind speed, NO_2, CO, SO_2, PM_{10}, $PM_{2.5}$, and precipitation (linear variables, z-scored across the pooled ten fingerprints), plus wind direction, decomposed into sine/cosine components before z-scoring to respect its circular scale. Ozone itself was excluded from the primary test, since it is the variable the clustering was built on; a secondary version including ozone is reported below for context only, and gives a similar result ($r = 0.48$, exact $p = 0.017$, identical to the primary test to the precision reported here).

Test. The 5×5 Euclidean distance between Scenario A and Scenario B fingerprints in this standardized auxiliary-variable space (D_{aux}, Table D5) was correlated (Pearson r on the 25 flattened entries) against the Step-1 geometric distance matrix (D_{geo}). Because there are only $5! = 120$ possible relabelings of the Scenario B side, the null distribution was enumerated exactly rather than Monte Carlo sampled: $r = 0.39$ against a null of mean -0.01, s.d. 0.24; $z = 1.70$; exact $p = 0.017$. Spearman rank agreement gives $\rho = 0.41$ ($p = 0.044$, parametric).

	B0	B1	B2	B3	B4
A0	4.04	4.69	3.28	3.52	3.65
A1	5.47	6.84	3.49	5.49	2.15
A2	7.19	7.14	2.33	6.42	2.57
A3	9.13	8.83	3.95	8.28	3.98
A4	5.57	4.77	2.10	4.30	3.40

Table D5: Auxiliary-variable distance matrix D_{aux} (standardized Euclidean distance, temperature/RH/wind/precursors/precipitation, ozone excluded), Scenario A rows against Scenario B columns.

B3 replicated ambiguity, in full. B3's closest auxiliary match (A0) and second-closest (A4) reproduce the same two clusters implicated in its Step-1 geometric ambiguity (Figure D17), in a data source entirely independent of the ozone cycles used in Step 1.

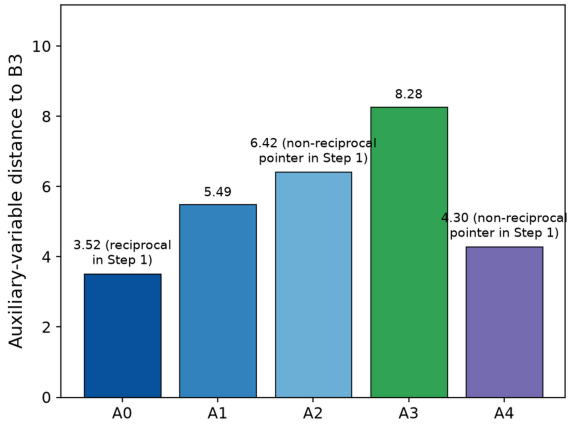


Fig. D17: Auxiliary-variable distance from cluster B3 to each Scenario A cluster. A0 (B3's Step-1 reciprocal partner) and A4 (the cluster that pointed to B3 in Step 1 without reciprocation) are both closer than A1, A2, or A3 – the same ambiguity found geometrically in Step 1 reappears in an independent set of variables.

Imputed chemistry. As an illustrative, purely exploratory corollary (not a confirmatory test, given $n = 5$), each Scenario B cluster's Step-1 nearest-neighbor Scenario A donor would predict an AMS-level chemical fingerprint unavailable in the operational network; where Scenario B does measure a rough proxy, its own fingerprint is directionally,

though not significantly, consistent with that prediction ($PM_{2.5}$ $\rho = 0.82$, $p = 0.089$; SO_2–SO_4 $\rho = 0.67$, $p = 0.219$).

Cross-model sequence-comparison toolkit: validation protocol. Section 4.4 reports three statistics (distance-weighted kappa, banded DTW, transition-matrix comparison) applied to Scenario A against each complete year at FEO. Before trusting those results on the real question, the three statistics were validated against two cases chosen to have a known-recurrence answer – one testing the statistical machinery alone, one testing the full cross-model mapping.

Scope of the validation. The two cases below were chosen to have a known-recurrence answer under two different sources of uncertainty: one tests the statistical machinery alone, the other tests the full cross-model mapping. A case with no expected correspondence is deliberately not included: Scenario A and Scenario B measure the same regional atmosphere, so there is no principled reason to expect *no* correspondence between them, only a weaker one than between two periods of the *same* station – and Section 4.3 already shows that a single station's own regime composition drifts across periods, which can make a same-station, different-period comparison a harder test than comparing against an external, fixed reference. The comparison between Scenario A and FEO is accordingly reported directly as the main-text application (Table 7, all three available FEO years) rather than as a validation case with a predetermined expected answer.

Cross-model distance matrix. The cross-model distance matrix used in Step 1 and throughout this validation is built from each scenario's already-published canonical cluster labels (the 'k-means' column of `02_kmeans_clusters.csv` for Scenario A, and `cluster_labels_k5.csv` for Scenario B), not from an independent re-fit of either scenario's raw cycles: even with an identical algorithm and random seed, such a re-fit does not reproduce either scenario's canonical partition, so the cluster identities entering the distance matrix would not match the labels used everywhere else in this manuscript. All numbers reported here and in the main text therefore use the distance matrix built from each scenario's canonical labels.

Validation cases. **POSC** (cheap positive control): FEO Period 1 against FEO Period 3 (definitions as in Section 4.3), using the *same* already-trained Scenario B model on both sides, so the cross-model weight matrix reduces to the trivial 0/1 mismatch matrix. Validates the permutation-null machinery in isolation. **POSR** (real positive control): FEO Period 2 against FEO Period 3, using two *independently* trained k=5 models – one fit only on Period 2's data, one only on Period 3's, following the same pipeline as the main Scenario B model (per-day 24-hour ozone cycle, global min-max normalization within the training window, `KMeans(n_clusters=5, random_state=42, n_init='auto')`), with its own freshly computed distance matrix (joint PCA on the pooled raw ppb cycles, 3 components; centroid = mean PCA-space position of each model's own assigned days; Figure D19). This case is unaffected by the labeling correction above, since it never touches either scenario's canonical labels.

Statistics, in full. For two label sequences of equal length aligned index-by-index, with cross-model weight matrix w_{ij} (normalized centroid distance, Step 1):

$$\kappa_w = 1 - \frac{\sum_{ij} w_{ij} O_{ij}}{\sum_{ij} w_{ij} E_{ij}} \tag{D1}$$

where O_{ij} is the observed joint proportion of index positions with series-A label i and series-B label j, and $E_{ij} = O_{i\cdot} O_{\cdot j}$ is the joint proportion expected under independence given each series' own marginal frequency (the standard weighted-kappa generalization of Cohen, 1968).

Banded DTW (restricted to $|t_A - t_B| \leq 14$ days) provides a more conservative, corroborating check: it aligns the two label sequences using the same substitution cost, reporting the total alignment cost and the mean day-offset along the optimal path.

The transition-matrix comparison estimates each series' first-order Markov transition matrix, $T^{(A)}_{ij} = P(\text{label}_{t+1}{=}j \mid \text{label}_t{=}i)$, translates $T^{(A)}$ into series B's label space via a soft, row-stochastic map $M_{ij} = \exp(-w_{ij}/\sigma)/\sum_k \exp(-w_{ik}/\sigma)$ with bandwidth $\sigma = \overline{w}$, and compares $\tilde{T}^{(A)} = M^\top T^{(A)} M$ against $T^{(B)}$ entrywise (Pearson r on the 25 flattened transition probabilities).

Nulls. Kappa's null permutes series B's labels (2000 draws). DTW's null circularly shifts series B by a random offset (2000 draws). The transition-matrix null holds each row's diagonal (self-persistence) fixed at its observed value and randomly permutes only which of the other four labels receives the remaining probability mass. A fully-memoryless null (shuffling series B completely before recomputing $T^{(B)}$) would instead produce spuriously significant results even where no correspondence is expected, because almost any daily regime sequence has a diagonal-heavy transition matrix purely from day-to-day weather continuity; holding the diagonal fixed isolates whether the specific switching *pattern*, not that generic persistence, matches.

Reading the validation. Table D6 summarizes both cases. Kappa and the transition-matrix comparison detect the known recurrence in both POSC and POSR, including the harder case where the cross-model mapping itself is genuinely at stake. DTW detects POSC clearly but loses significance in POSR,

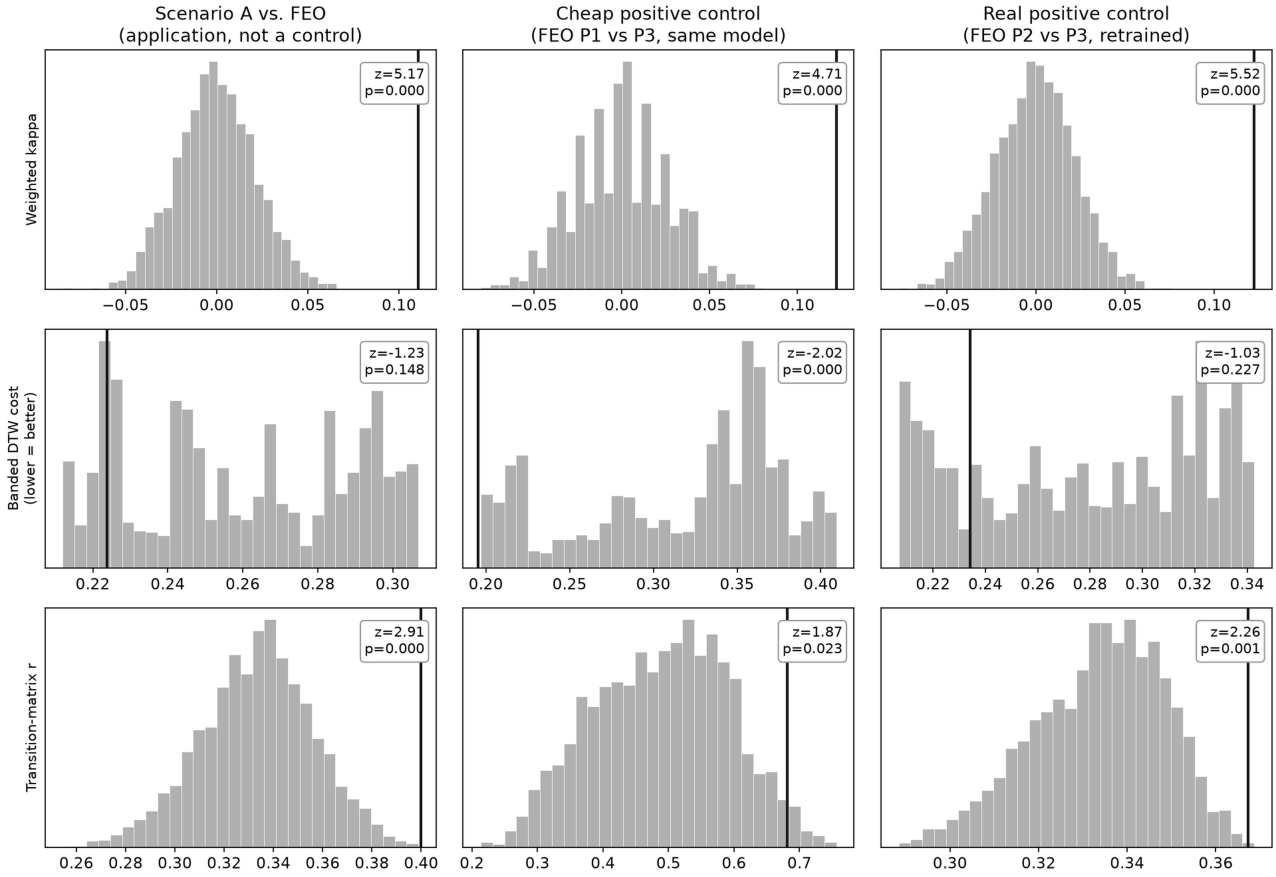


Fig. D18: Observed statistic (red line) against its permutation null (gray histogram), for all three statistics (rows) and both validation cases, together with the Scenario A vs. FEO comparison shown for reference (columns).

	Kappa	DTW	Trans. r
POSC	0.122 (4.71, ¡0.001)	−2.02 (¡0.001, 6.2 d)	0.681 (1.87, 0.024)
POSR	0.123 (5.52, ¡0.001)	−1.03 (0.227, 6.5 d)	0.367 (2.26, 0.001)

Table D6: Validation results, both validation cases. DTW z: negative means the true alignment cost beat its null. Full null distributions in Figure D18.

consistent with it being the least reliable of the three once real cross-model mapping uncertainty is introduced – which is why it is reported as corroborating, not primary, throughout Section 4.4

References

Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974) https://doi.org/10.1080/03610927408827101

Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979) https://doi.org/10.1109/TPAMI.1979.4766909

Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193–218 (1985) https://doi.org/10.1007/BF01908075

Hopkins, B., Skellam, J.G.: A new method for determining the type of distribution of plant individuals. Annals of Botany **18**(2), 213–227 (1954) https://doi.org/10.1093/oxfordjournals.aob.a083391

Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990) https://doi.org/10.1109/5.58325

MacQueen, J.: Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1**, 281–297 (1967)

Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.

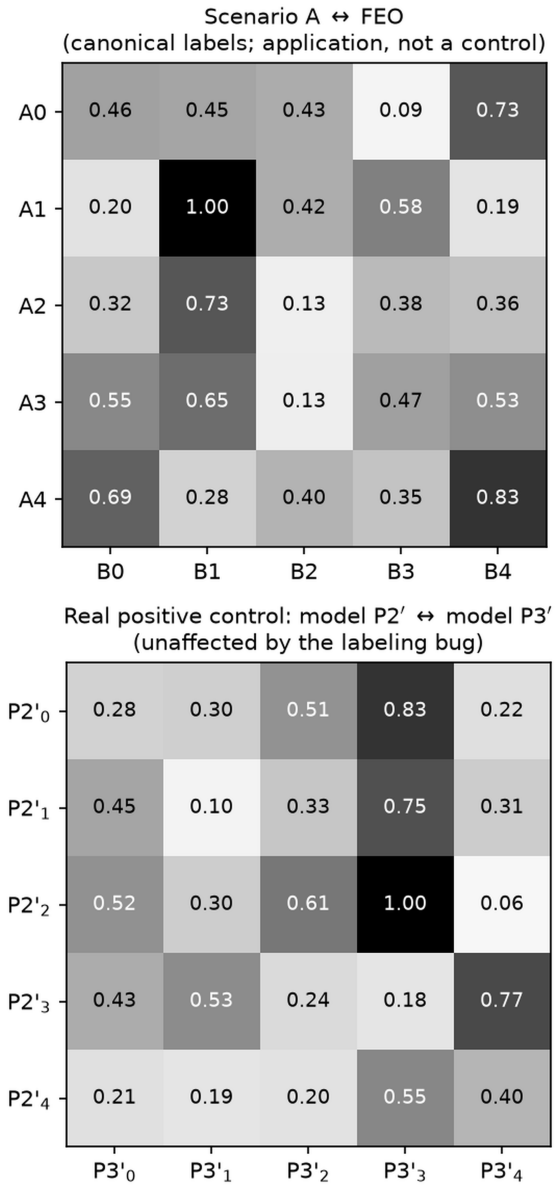


Fig. D19: The two genuine cross-model distance matrices used as substitution cost: Scenario A ↔ FEO (canonical labels, Step 1) and the freshly retrained POSR pair.

Journal of Computational and Applied Mathematics **20**, 53–65 (1987) https://doi.org/10.1016/0377-0427(87)90125-7

Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research **3**, 583–617 (2002) https://doi.org/10.1162/153244303321897735

Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(2), 411–423 (2001) https://doi.org/10.1111/1467-9868.00293

Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (1963) https://doi.org/10.1080/01621459.1963.10500845