HEI Research Report 242

Air Quality Trends in Texas and Colorado Associated with Unconventional Oil and Gas Development

Appendix A. Detailed Method Descriptions

Gunnar W. Schade and Detlev Helmig et al.

Correspondence may be addressed to Dr. Gunnar W. Schade, Department of Atmospheric Sciences, Texas A&M University, 3150 TAMU, College Station, TX 77843-3150; email: gws@geos.tamu.edu.

Although this report was produced with partial funding by the United States Environmental Protection Agency under Contract No. 68HERC19D0010 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and may not reflect the views of the Agency; thus, no official endorsement by the Agency should be inferred. This report also has not been reviewed by private party institutions, including those that support HEI Energy, and may not reflect the views or policies of these parties; thus, no endorsement by them should be inferred.

Appendix A. Detailed Method Descriptions

Contents

A1. Detailed Descriptions of Methods Used to Analyze AQM Station Data from Texas

Texas Ground-Based Measurement Data

NMHC, NO_x, and meteorological data for the 11 TCEQ monitoring stations that were considered in this study were obtained via the TCEQ Texas Air Monitoring Information System (TAMIS) web interface (https://www17.tceq.texas.gov/tamis/index.cfm?fuseaction=report.main), along with associated qualifier and minimum detection limit metadata. Data were obtained from the earliest starting date available for each station and compound, through 12/31/2023. For NMHC and NO_x data, the starting dates range from 1997 to 2015. All TCEQ datasets were stored on a PC, on a Texas A&M-hosted OneDrive in cloud space, and on an external hard drive.

Splines Trend Variations: Log-Scale Fit, Quantiles Fits, Deseasonalized Data Fit

The Texas NMHC data, like most environmental data, do not have a normal distribution; therefore, the spline trend fits were also assessed on a log scale and for different data quantiles. For the logarithmic fit, the data were first natural log transformed, then the natural cubic spline trend was fit, followed by the fit results transformed back. The resulting trend line is offset to the lower *geometric mean*, but the long-term trend tendency, which is the concern of this study, was found to be nearly identical. This finding held for both the hourly and sparser canister-sample NMHC data. Figure C1-2 in Appendix C shows representative example outcomes for 1-hour NMHC data from Eagle Mountain Lake and Floresville, and 24-hour NMHC canister data every 6 days from Mineral Wells.

The natural cubic spline trend fitting routine was applied to the 98th, 95th, 90th, and 20th quantiles of the data. Figures in Appendix C show examples of this quantile fitting for Floresville ethane, Mineral Wells benzene, and Eagle Mountain Lake benzene data, respectively. Higher quantile fits show higher amplitude variations and lower quantile fits show lower amplitudes, but all quantiles maintained the same general trend tendencies. This implies that much of the trend variation is driven by the variation in higher NMHC levels, while the background levels are more stable, which supports the more local/regional influence that is the focus of this study.

The seasonal cycle can also potentially influence the trend estimation, particularly at the start and end points of the data; therefore, the spline long-term trend fit was tested after first deseasonalizing the data. To remove the seasonal cycle, a repeating periodic cubic B-spline was fit to the data. After removing the periodic B-spline fit, the same natural cubic spline routine was used to fit a trend line. Representative example outcomes are shown in Figure 15 in the main report and in the figures in Appendix C. The trend calculation with the seasonal cycle removed is seen to have negligible effects on the trend estimation.

A2. Detailed Descriptions of Methods Used to Analyze AQM Station Data from Colorado

Remote Reference Data Stations

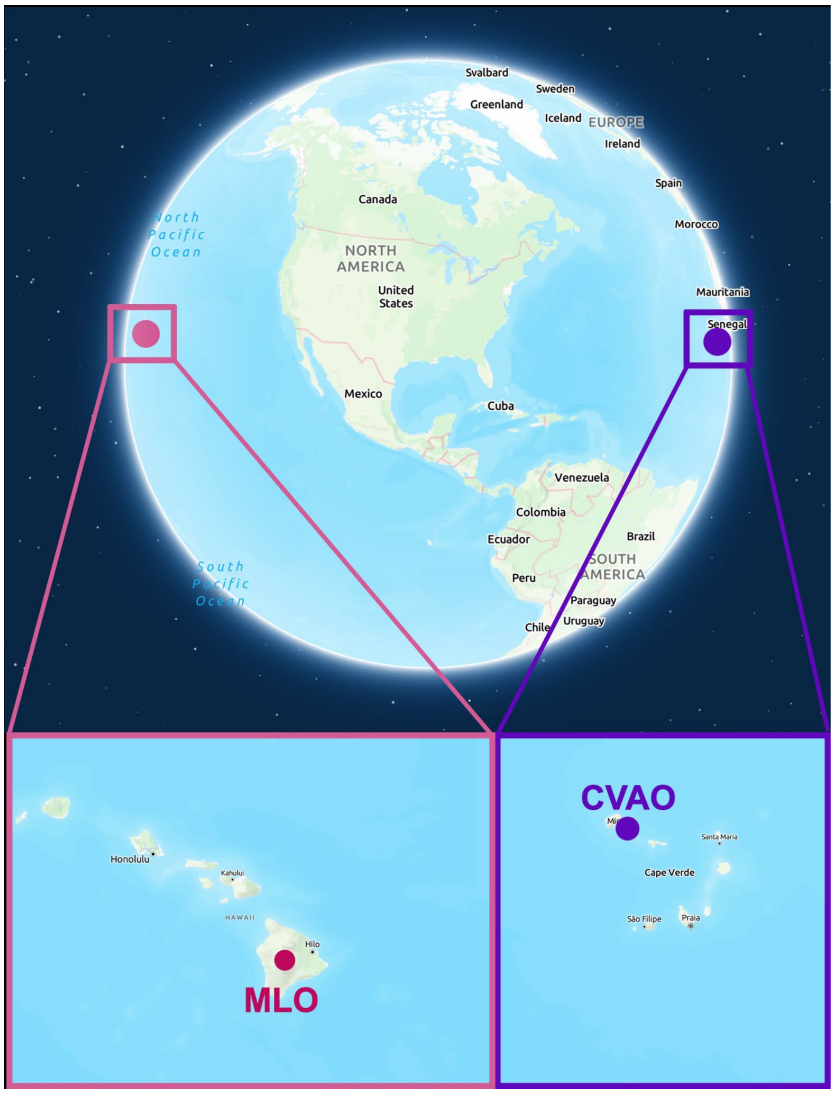


Figure A2-1. *Map of reference station locations.*

Colorado Ground-Based Measurement Data

Sector Choice Algorithms

Maximum Normalized Difference in Median Monthly Ethane

Given that ethane is a well-documented hydrocarbon emission from oil and gas activities and has very few other significant sources, the first algorithm focused on maximizing the normalized difference between median monthly ethane values in the background versus oil and gas (O&G) sectors, which we

would expect to be significant if chosen correctly. This approach allowed for an objective assessment of sector definitions based on the highest-ranked combinations of background and O&G sectors.

Methodology:

- Sector Window Size: A loop was implemented to evaluate sector sizes ranging from 80 to 140 degrees in 10-degree increments.
- Wind Speed Threshold: The dataset was filtered to include only wind speeds exceeding 3 m/s.
- Combination Evaluation: All possible background versus O&G sector combinations were tested. For each pair, the algorithm calculated the average difference between the monthly median ethane mole fraction divided by its standard deviation of the O&G sector and the background sector.
- Ranking: The top 15–30 sector combinations with the highest normalized differences were identified, and two to eight options were selected for further testing in the next algorithm, **Table A1.**

Table A1. Example Sector Combinations Selected for the LUR

Average Scaled Difference Rank	**Background Sector**	**O&G Sector**	**Average Scaled Difference**
Rank 1	(180°, 270°)	(350°, 80°)	0.371
Rank 2	(180°, 270°)	(350°, 110°)	0.371
Rank 3	(170°, 270°)	(350°, 80°)	0.371
Rank 60	(160°, 280°)	(350°, 100°)	0.366

Highest Coverage i-/n-Pentane Ratio-Based Wind Direction Sampling

To further refine sector definitions, a second Monte Carlo-style algorithm evaluated the selected sector combinations based on the i-/n-pentane ratio, a known indicator of O&G emissions.

Methodology:

1. Input Sector Choice: The top-ranked sector combinations from the previous algorithm served as inputs.
2. Random Sampling: For each potential O&G sector:
    - The dataset was randomly subsampled (60% of data above the 3 m/s threshold) 200 times.
    - For each sample, wind directions (wdr) were collected where the i-/n-pentane ratio fell within the predetermined “O&G” range (0.8–1.2 noninclusive).
3. Sector Size Optimization: A simple search algorithm iteratively tested sector sizes (site-dependent, typically 90°–140°) to maximize the percentage of collected wdr points within the sector.
    - A minimum threshold percentage was set (e.g., 95%), with larger sector sizes tested until the maximum size was reached. If no size matched the threshold percentage, the minimum percentage threshold was incrementally decreased by 5%, and the process was repeated.
4. Winning Combination:

- The first combination achieving the maximum percentage of i-/n-pentane ratio matches at the smallest sector size was identified.
- Thresholds for input sector choice inclusion in the winning combination (e.g., 80%, 90%) were recorded.
- The frequency of input sector choice agreement with simulated i-/n-pentane ratio-based sectors was calculated, and the best-performing sectors were selected, prioritizing higher agreement or larger wdr sector sizes if differences were negligible.

Table A2 summarizes the results of this procedure.

Table A2. Wind Sector Choice Algorithm Results for the LUR Site

Sector Type	Input Sector Choice	80% Coverage Simulation Count	90% Coverage Simulation Count	95% Coverage Simulation Count
Oil and gas	(350°, 70)	200	200	200
Oil and gas	(350°, 80)	200	200	200
Oil and gas	**(350°, 90°)***	200	194	193
Oil and gas	(350°, 100°)	200	190	78
Oil and gas	(350°, 110°)	196	35	9
	***selected**	Out of 200 Simulations		

The highest coverage sampling method was repeated for the "clean" western (background) sector choices with an i/n-pentane ratio range of ≤0.8 or ≥1.2. Modifications by site were primarily limited to adjusting maximum allowed background and O&G sector wind direction window sizes, expecting certain sites, for instance, BSE, to have a larger O&G sector. However, final sector definitions remained within the standard 90°–140° allowance, as shown in Table A3. Sector decisions were made using data up to summer 2024, the most recent quality-assured data at the time.

Table A3. Final Wind Sector Selections for the Northern Colorado Front Range Sites

Site	Western	Oil and Gas
BSE	190°–290°	330°–110°
BRZ	200°–320°	350°–90°
ECC	200°–290°	350°–90°
LUR	190°–280°	350°–90°
BNP	170°–270°	300°–60°

HYSPLIT Back-Trajectory Analysis

We performed an analysis to compare locally sourced meteorological measurements (with 1-minute time resolution) for use as a wind sector filter, with NOAA HYSPLIT (version win64U 5.0.0b) trajectories, a widely established model in the meteorology community. The highest resolution trajectories possible were selected. Back trajectories from the 3 km High-Resolution Rapid Refresh (HRRR) model were obtained from NOAA's Air Resources Laboratory (ARL) data archive and computed locally using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model. This model estimates the path an air parcel travels back in time, using 15-minute radar data to reconstruct the trajectory at an hourly time resolution (https://gsl.noaa.gov/focus-areas/regional-models?utm_source=chatgpt.com). Each of the colored lines in Figures A2-1 and A2-2 represents the predicted path the air took over the course of an hour using the HYSPLIT model. The error in HYSPLIT trajectory estimates is reported to be 15%–30% of the total travel distance, arising from a combination of physical, computational, measurement, and forecast uncertainties (https://www.arl.noaa.gov/documents/workshop/NAQC2007/HTML_Docs/trajerro.html).

To compare local meteorological data with HYSPLIT outputs, aggregated wind measurements were paired with their corresponding trajectories using a starting location and time as close as possible. Because HYSPLIT predictions are instantaneous and limited in spatial resolution, pairing real-world wind observations with modeled trajectories introduces additional uncertainty. To address this issue, numerous temporal aggregations of the local wind data were tested. The proportion of HYSPLIT trajectories falling within the expected wind sectors (see Table A4) was calculated after applying directional and speed filters (>3 m/s) using the site's wind measurements. Hourly-averaged wind measurements were found to produce the best agreement with the HYSPLIT trajectories when compared with other aggregation intervals (e.g., 10-minute, 30-minute). There was less agreement between the trajectories and our wind data from the western sector as compared to the O&G sector, likely due to more consistent wind behavior over the plains under easterlies when compared to air flow across the mountains to the west. However, we note that many western sector mismatches were only off by small degrees. No changes to final wind sector selections were made based on this analysis.

Figures A2-1 and A2-2 provide visualizations of the HYSPLIT 1-hour back-trajectory analysis, with the LUR site at the center (receptor site), and the back-trajectory origination points indicating the presumably correct air mass 1-hour origin. An approximately centered polar coordinate system is overlaid, with colored bars outlined in black indicating the start and end degrees of the selected wind sectors.

Table A4. Agreement Between Local Wind Sector Assignments at the LUR site and Air Mass Origin Sectors

	Predicted O&G	**Predicted Western**
Incorrect	103	118
Correct	295	119

As predicted by 1-hour HYSPLIT back trajectories and presented in a confusion matrix format. LUR O&G sector: (350°–90°), LUR western sector: (190°–280°).

"Predicted O&G" and "Predicted Western" refer to the sectors determined by the HYSPLIT model based on the air parcel's origin 1 hour prior. "Correct" indicates the number of instances where the locally measured wind sector matched the HYSPLIT-predicted sector, while "Incorrect" indicates mismatches. The O&G sector results were correct two-thirds of the time, while half the western sector produced mismatches. Although mismatches are more frequent for the boundaries of the western sector, Figure

A1-2 confirms that the air mass origins were overwhelmingly from the west, as expected, and did not overlap with the O&G sector. Figures A2-2 and A2-3 provide visualizations of the HYSPLIT 1-hour back-trajectory analysis, with the LUR site at the center (receptor site), and the back-trajectory origination points indicating the presumed correct air mass 1-h origin. An approximately centered polar coordinate system is overlaid, with colored bars outlined in black indicating the start and end degrees of the selected wind sectors.

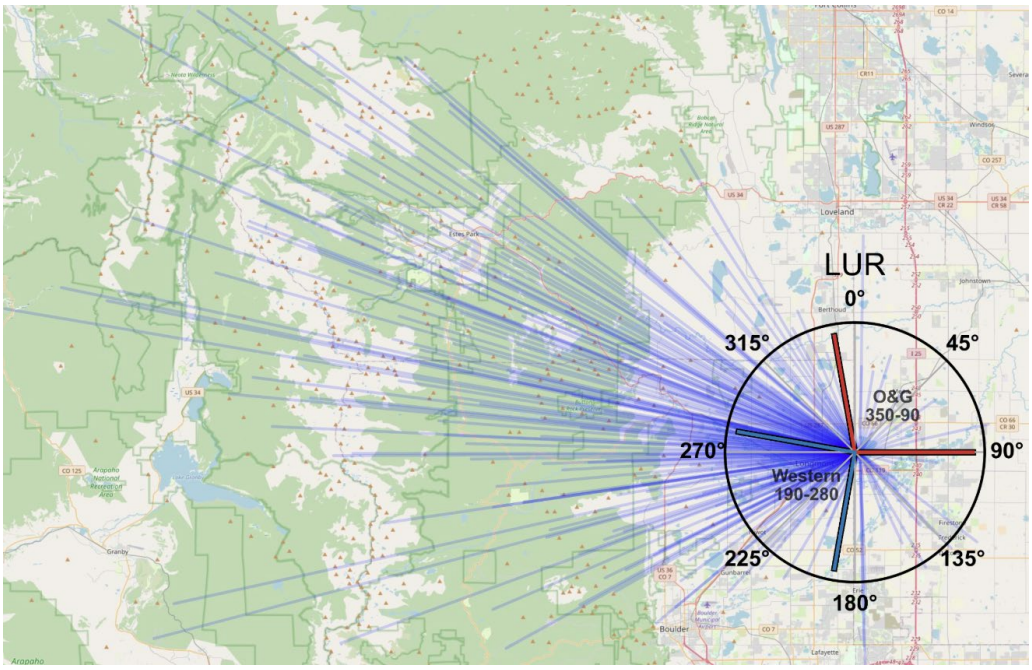


Figure A2-2. *One-hour HYSPLIT back trajectories matched to the LUR western "clean" sector, 190°–280°.*

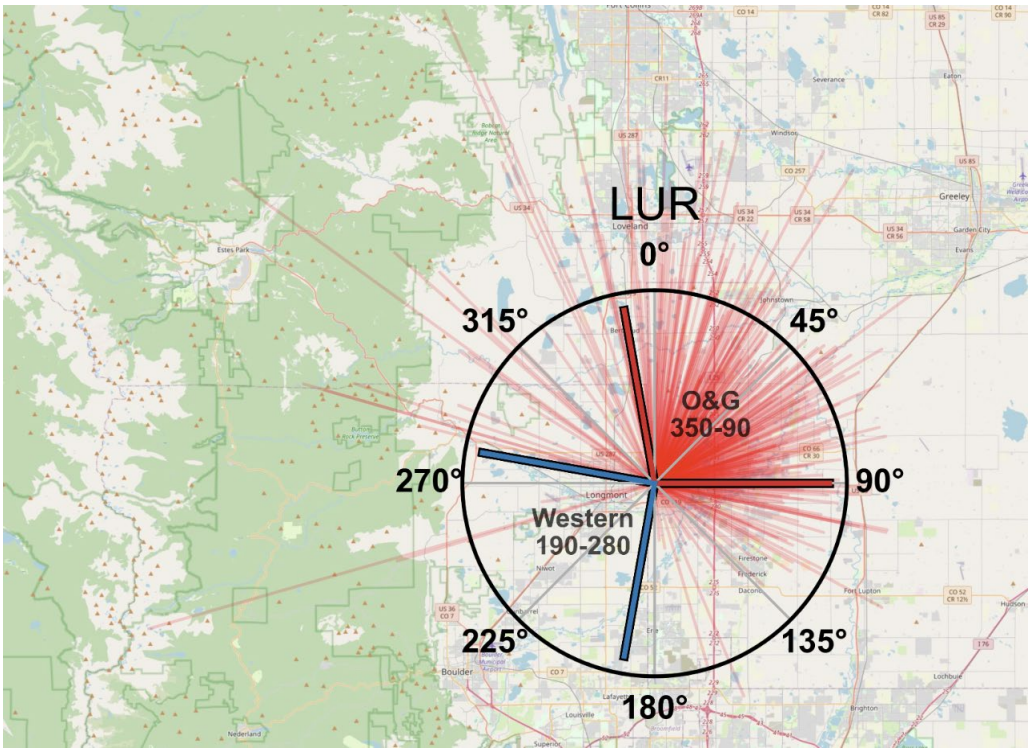


Figure A2-3. *One-hour HYSPLIT back trajectories matched to the LUR O&G sector direction wind data, 350°–90°.*

Further Wind Filtering for Wind Sectors

Sector-specific input data underwent further filtering before trend analysis to isolate sector origin data by requiring neighboring wind direction consistency. For the trend analysis, we chose only NMHC data points whose wind direction was in the same sector for at least 1 hour prior. For methane data, only data points surrounded by hours or centered on hours with at least two-thirds of the data points matching the same sector were chosen. Although this method further reduced the dataset size, it helped exclude periods in which wind conditions were more variable and, therefore, a weaker connection to the true air mass origin. A limitation of this approach is that the removal of data points may occur inconsistently across different time periods, disproportionately affecting some periods more than others. This uneven data loss may have exaggerated existing phenomena during linear interpolation in the NOAA tool before the fit function application — for instance, periods naturally prone to greater wind variability could exhibit amplified spikes. If these spikes aligned with expected seasonal maxima, they could artificially increase or decrease the trend line slope.

Colorado Analysis Tools and Uncertainty Information

All analyses except bivariate polar plots were produced using Python (Version 3.12.0) scripts. Standard data analysis packages, including *Pandas* (V2.2.2), *Numpy* (V1.26.4), and *Matplotlib* (V3.9.0), were used. NOAA trend tool scripts, CCGCRV, were downloaded, and a Python script was constructed to run the tool using command line prompts, save outputs, and load them into Python for analysis. The driver program ccgcrv.py was prompted using arguments detailed in ccgcrv.pdf. All NOAA script files used were obtained from https://gml.noaa.gov/aftp/user/thoning/ccgcrv

NOAA Trend and OLS Regression Uncertainty

The NOAA trend curve output uncertainty bounds (orange CI bands) were computed using the long-term trend's standard deviation output, defined by NOAA as the combination of polynomial fit variance and the long-term cutoff residual filter (FFT) variance. This output is reported as one average for the entire fitting routine. The NOAA website specifies the residual chi-square statistic as a multiplicative factor in the trend error calculation; however, it was excluded here because the chi-square statistic was often highly inflated, leading to unrealistically large uncertainty estimates. Total trend error was computed at the 95% confidence level following

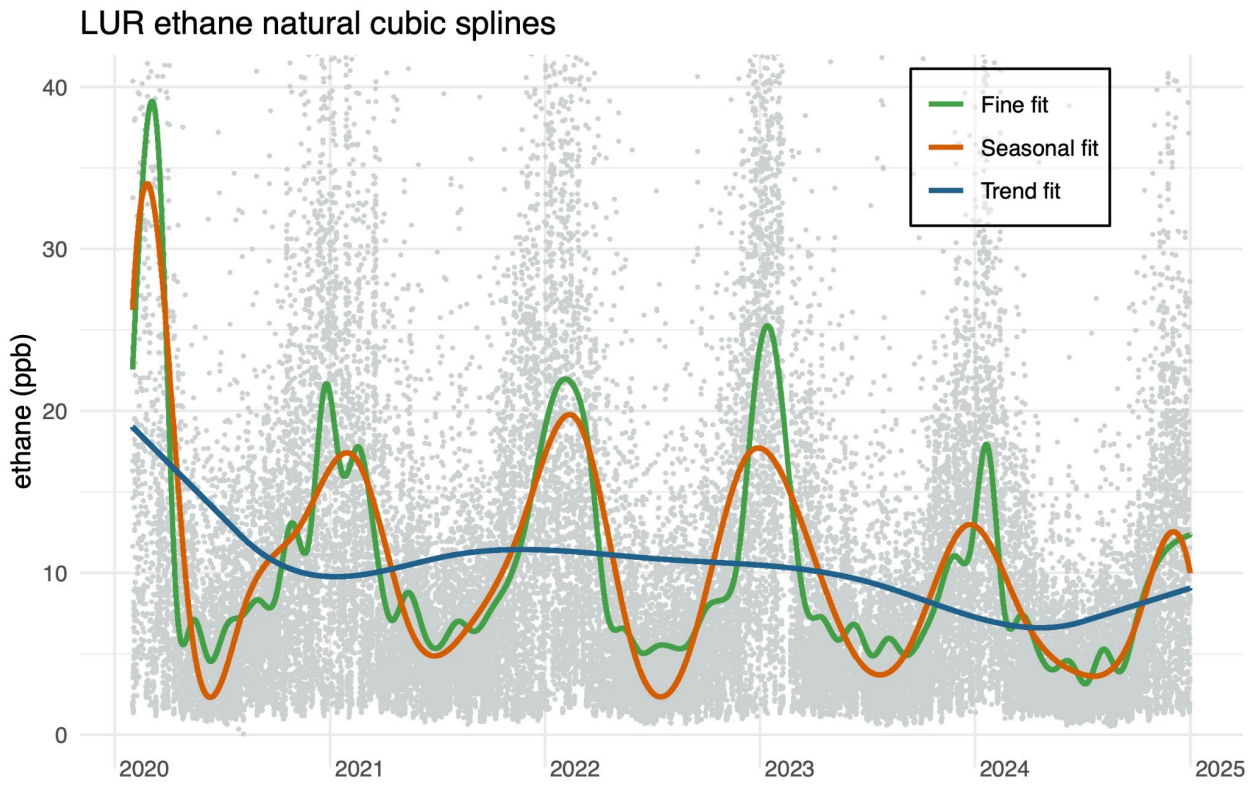
$$\text{trend error} = 1.96 \times \text{trend standard deviation} \quad (1)$$

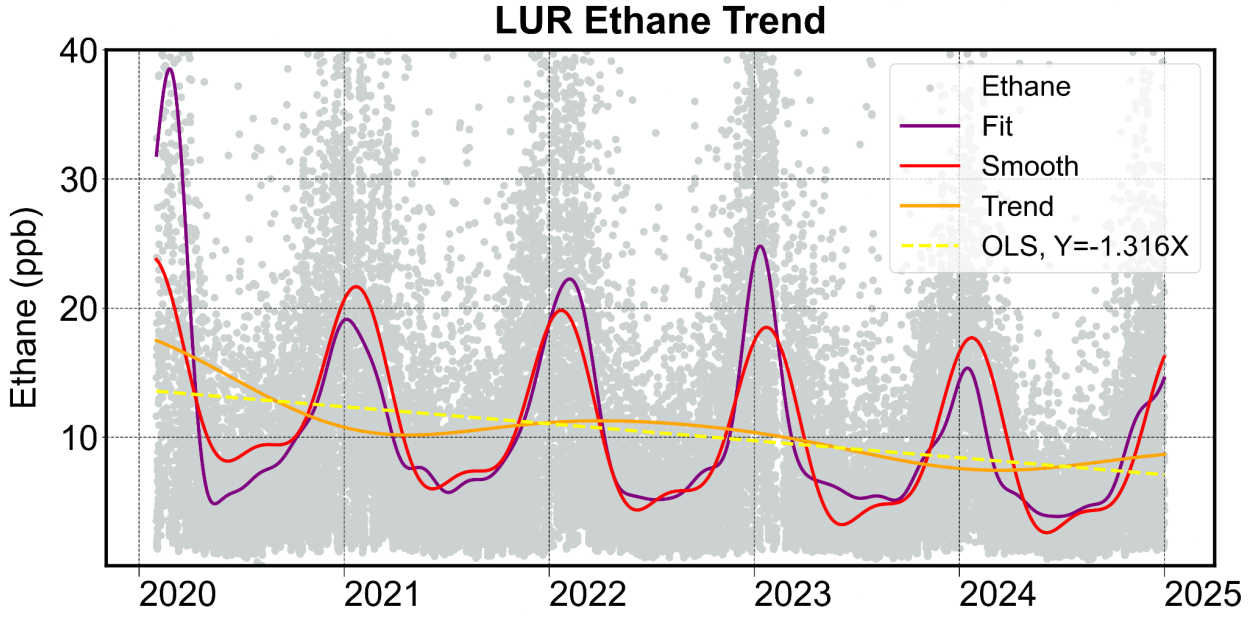
The OLS regression line on the trend curve, which we used to compare linear long-term trends among sites, was computed using the *Statsmodels* Python package (V0.14.5). For computing the OLS linear regression slope confidence intervals, the Moving Block Bootstrap method was used based on Künsch (1989) to account for autocorrelation in the NOAA trend values. As a result of the resampling process, the resulting confidence bounds can be asymmetrical. The block length was automatically chosen using the Python *ARCH* bootstrap package function optimal_block_length (Sheppard et al., 2024). Many trends, regression slopes, and seasonal behaviors of ethane, propane, and benzene indicate a decrease in molar mixing ratios over time, particularly for the longer record sites BRZ and LUR. This shared downward movement is reflected in the upper percentiles of the percentile plots (see Appendix C). However, in many cases, the NOAA trend confidence intervals are so large that they encompass the start and end values. In comparison, the linear regression fit on top of the NOAA trend is technically considered statistically significant in most O&G sector cases, as the 95% confidence interval bounds do not include a zero-slope value. This finding makes our statistical significance interpretation somewhat ambiguous.

Multisite comparison graphs were created using NOAA tool output columns, including the trend and smooth outputs, and OLS regression lines fitted on the NOAA trend curves.

A3. Trend Method Comparison

Method comparison between splines and NOAA trend tool for Colorado hourly ethane data, LUR site.





Number of harmonics: 2 Number of polynomials: 3

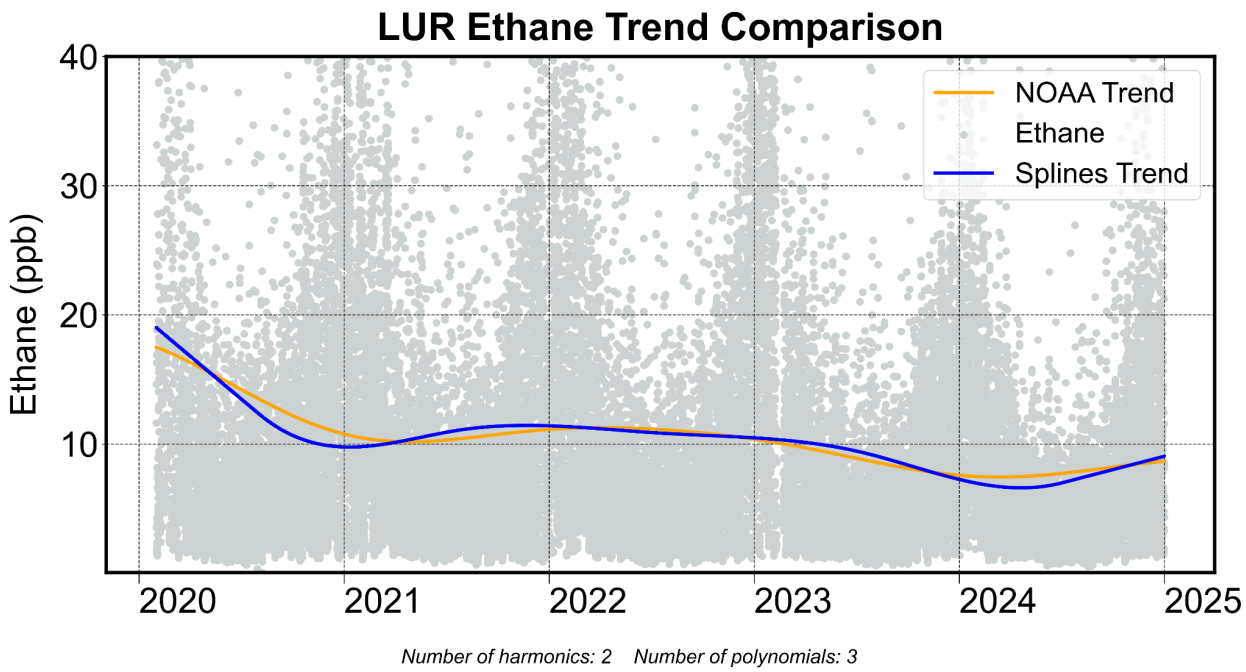


Figure A3-1. *Comparative analysis of trends between the splines and NOAA trend tool methodologies, similar to Figure 10 in the main report, but for the Colorado LUR site. Top panel: splines analysis output; middle panel: NOAA trend tool output; bottom panel: direct comparison of calculated long-term trend.*

A4. NMF Analysis Details

Data Selection and Reduction

Raw data were obtained as described in Section A1. Each raw data time series was inspected in Excel to determine which volatile organic compounds (VOCs) listed showed consistent abundances in the data, which compounds were listed but contained no data, and which compounds were listed but were rarely observed. The data set was then reduced by eliminating columns that had no data, had overwhelmingly missing data, and any non-data columns except the date column and the origin site's lat-lon data. Lastly, a set of seven chlorinated compounds was selected to remain in the data set; compounds that were expected to represent ubiquitous species of low but steady abundance and not expected to be emitted from the oil and gas industry or other regional sources. The retained set of compounds is listed in Table A5.

Table A5. VOC Compound Selections for Site-Specific NMF Analyses

Alkanes			
ethane	propane	n-butane	iso-butane
n-pentane	isopentane	n-hexane	2-methylpentane
3-methylpentane	2,2-dimethylbutane	2,3-dimethylbutane	n-heptane
2-methylhexane	3-methylhexane	2,3-dimethylpentane	2,4-dimethylpentane
n-octane	2-methylheptane	3-methylheptane	2,2,4-trimethylpentane
2,3,4-trimethylpentane	n-nonane	n-decane	n-undecane
Naphthenes			
cyclopentane	cyclohexane	methyl-cyclopentane	methyl-cyclohexane
Aromatic VOCs			
benzene	toluene	ethyl-benzene	m/p-xylene
o-xylene	n-propyl-benzene	isopropyl-benzene	styrene
1,2,3-trimethyl-benzene	1,2,4-trimethyl-benzene	1,3,5-trimethyl-benzene	o-ethyl-toluene
m-ethyl-toluene	p-ethyl-toluene	p-diethyl-benzene	
Alkenes			
ethene (ethylene)	propene	1-butene	cis-2-butene
trans-2-butene	1,3-butadiene	1-pentene	2-methyl-butene
cis-2-pentene			
Other VOCs			
acetylene (ethyne)			
Chlorinated VOCs			
chloro-methane	dichloro-methane	chloroform	methyl-chloroform
dichlorodifluoromethane	trichlorodifluoromethane	carbon-tetrachloride	

Among the compounds listed in Table A5, cis-2-pentene was later dropped from the raw dataset as well. Thus, a time series of a total of 60 compounds was initially read into R for the NMF calculations.

To further scrutinize the input data, all compound time series were plotted in R to visualize quantitative data density. For several compounds — shaded light gray in Table A5 — for which quantitative data density was typically low (below detection limit, LOD, and listed as "0"), we combined abundances into a sum of isomers, such as "trimethyl-benzenes." This created fewer compounds but overall reduced entries <LOD. Furthermore, several compounds were completely dropped from a site's data set after inspection if their data abundance >LOD was less than a few percent of the total available or large periods without data existed (shaded dark gray in Table A5). This further reduced the number of compounds entering the NMF analysis, leading to a range of a minimum of 47 (OHE site) to a maximum of 60 (OLN site) compounds used.

Lastly, the time series of all compounds was scrutinized to identify individual extreme values in a compound's time series that could skew the NMF outcome. Such values were of two kinds:

- Missing detection and quantification of a compound that is ordinarily quantified in the vast majority of samples, manifesting as a "drop to zero" in the time series (low outlier); and
- Quantification of a compound at more than ten times its average abundance, manifesting as a "spike" in the time series (high outlier).

In the former case, the respective zeros in the timeline were replaced by the data median; in the latter case, the spike was reduced to half its size through a division by two.

Next, in several data sets, a small number of missing data were replaced with time series medians in lieu of discarding the whole sample from the data set, despite the fact that most compounds were measured. Lastly, the uncertainty for replaced data in the error matrix was raised to a high value, thereby downgrading the impact these data may have on the uncertainty-weighed NMF calculations.

NMF Analysis Interpretation

After data preparation, all compounds were ordered by compound class, similar to the order presented in Table A5, then normalized to a [0,1] range as described in Section 3 in the main report, and lastly rounded to six significant digits. The NMF was executed in R version 4.2.2, and the results were plotted to visualize the outcome. Figure A4-1 shows an example image plot of the NMF's basis and coefficient matrices. The basis matrix plot illustrates changes over time, and in this example, shows the diminishing contributions of compounds associated with diesel engine emissions. The coefficient plot illustrates each factor's VOC composition, which, due to the (here: bottom-to-top) ordering of compounds by class (Table A5), helps with interpreting the potential source(s) contributing to or dominating each factor.

Factors were tentatively assigned to an individual air pollutant source based on prior experience of using NMF analyses in similar environments for source apportionment. The guiding principles for this exercise, based on known compositions of typical air pollutant source categories (mobile and stationary), were as follows:

1. Road/car traffic emissions, dominated by internal combustion engine emissions, are characterized by medium chain-length alkanes, such as pentanes, short-chained alkenes, particularly ethene and acetylene associated with tailpipe exhaust, and simple aromatic compounds, such as toluene; VOC ratios associated with such emissions (dominant in most urban areas) are an isotopic pentane ratio of 2:1 (i-C5/n-C5) or higher, and a toluene to benzene ratio of 2:1 or higher.
2. Diesel engine emissions are characterized by longer chain alkanes and alkenes, and particularly larger ("heavy") aromatic compounds such as trimethylbenzenes (C8- and C9-aromats).
3. a) O&G production-related emissions are characterized mostly by saturated hydrocarbons, particularly short-chain alkanes such as ethane, propane, and butanes, but also by naphthenes

such as methyl-cyclohexane, and medium-chain-length, straight-chain alkanes such as n-pentane and n-hexane; a VOC ratio typically used to identify such sources is a higher-than-urban isotopic pentane ratio (i-C5/n-C5) of one or lower.
b) O&G production-related gas flaring emissions may be characterized by short-chain alkanes and alkenes, such as ethane and ethene, alongside benzene, with these unsaturated hydrocarbons representing the inefficient oxidation by the external combustion source.

4. Other stationary or area sources may be characterized by a more distinct VOC mix, which may be closely related to the dominant VOC(s) in the calculated NMF factor, such as in the case of the gas-fired power plant in the example of the NMF output for the LRB site.

In all cases investigated, the imputed chlorinated compounds were presumed to represent a mostly invariant "background," interpreted to reflect a pollutant contribution present in the absence of the more distinct pollution source factors we intended to identify. These compounds, and the factor resulting from their inclusion into the NMF, were envisioned to function as an independent measure of the fidelity of the NMF, presuming that a high fidelity output (i) positively identifies an independent factor for this group of compounds, and (ii) shows that neither do these compounds score significantly on other factors, nor do other VOCs score significantly on this "background factor." While condition (i) was satisfied by all NMF outputs obtained, condition (ii) was broken to different degrees, suggesting that the NMF in these cases was of lower quality. However, note that in the case provided in Figure A4-1, the VOCs scoring on "background factor" two included ethane and propane, the abundances of which do include a significant background concentration in air far away from their sources. Therefore, not all infringements of presumption (ii) should be interpreted as undermining the NMF output's fidelity.

The second piece of information informing about the NMF output's fidelity was a calculation of how well the calculated matrix **V** represented the original matrix **X**. This was done via a dataset variability comparison of the total sum of squares of V (tss) with the residual sum of squares between V and X (rss), expressed as a percentage ($1 - \text{rss}/\text{tss}$). If a weighted NMF had been implemented, the respective weighted tss and rss were used. A leave-one-out procedure was used to determine which individual contribution each factor made to explaining the dataset's total variability, by reconstructing the data matrix without the factor in question. While this procedure is generally more appropriate than reconstructing the data matrix using each factor by itself, we found that the results between these two calculations were similar.

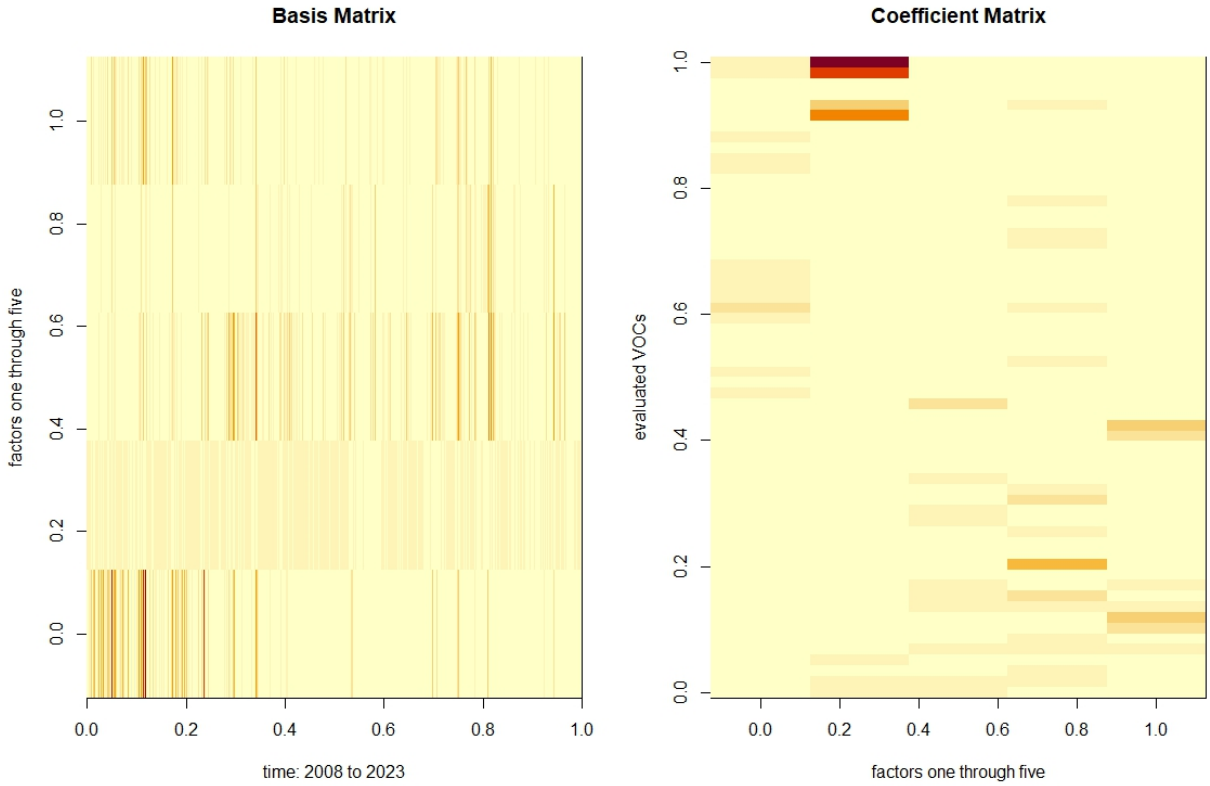


Figure A4-1. *Generic image plots of the basis and coefficient matrix of the OLN site data NMF output. The factors on the left plot illustrate changes over time, such as the diminishing impact of factor one – tentatively identified as diesel emissions – and the lack of change in factor two – the "background factor." The coefficient matrix on the right illustrates the factor's VOC composition, identifying factor two as the "background factor" due to its loadings on the chlorinated compounds, and factor three (center) as the likely O&G factor, due to its loadings on short-to-median chain length alkanes and methyl-cyclohexane.*

References

Künsch HR. 1989. The jackknife and the bootstrap for general stationary observations. Ann Stats 17:1217–1241, https://doi.org/10.1214/aos/1176347265.

Sheppard K, Khrapov S, Lipták G, van Hattem R, Hammudoglu J, Rob Capellini R, et al. 2024. bashtage/arch: Release 7.2 (v7.2.0). Zenodo; https://doi.org/10.5281/zenodo.14035889.