HEI Research Report 242

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

Appendix B. Additional Methodological Details Concerning the Satellite Data Analysis

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 B. Additional Methodological Details Concerning the Satellite Data Analysis

Contents

TROPOMI Satellite HCHO Data

TROPOMI formaldehyde data (available 05/2018 to 12/2023) analyses were originally proposed for this study, but were not included for the following reasons:

- Satellite data from different instruments on different platforms is not perfectly comparable. Following other satellite researchers (e.g., Shen et al 2019 using HCHO data from OMI-SAO, SCIAMACHY, GOME-2 from MetOp-A, GOME-2 from MetOp-B, and OMI-BIRA), this could be achieved by defining a conversion factor based on the overlapping dates and times of the satellite retrievals. Between the proposed OMI and TROPOMI, the overlapping dates are 04/2018–06/2022. With just 1.5 years remaining until the end of the study period (12/2023), interpreting the final, short segment of 1.5 years using converted TROPOMI data would be questionable and could potentially lead to skewed/false trend interpretations. More faith would be given to the trend analysis using the OMI data ending 6/2022.
- For the current study objectives of looking at long-term trends, the OMI data are actually the better product since they cover nearly 18 years from 10/2004 to 06/2022. Associated with that are numerous years of studying and using these data, such that satellite data issues are known and can be addressed. In comparison, the TROPOMI HCHO data record is much shorter.
- Without using TROPOMI data, only 1.5 years were "lost" for trend analysis from the end of the current study period, with ~18 years of trend input data from OMI, representing a relatively small loss for the study objective of assessing long-term trends in the Permian Basin.

Thus, within the limited time frame of this 1-year project, it was decided not to include TROPOMI data in order to avoid difficult-to-interpret and potentially biased tendencies of the most recent years.

OMI Data Acquisition and Handling

Aura OMI HCHO data were obtained from the NASA GES-DISC (Goddard Earth Sciences Data and Information Services Center, https://disc.gsfc.nasa.gov/) in NetCDF4 file format. Data were subset by latitude and longitude for the three regions (Permian, Arizona, and Pacific) via the Earthdata subsetting tool. Daily .nc4 files were downloaded using the command line GNU software package 'wget' (Version 1.25.0). NetCDF4 files were converted to .csv data files using the R ncdf4 package (Version 1.23).

Figure B1 shows a typical summer day, a gridded dataset used in this analysis, in which gray pixels represent missing data based on the QA procedures.

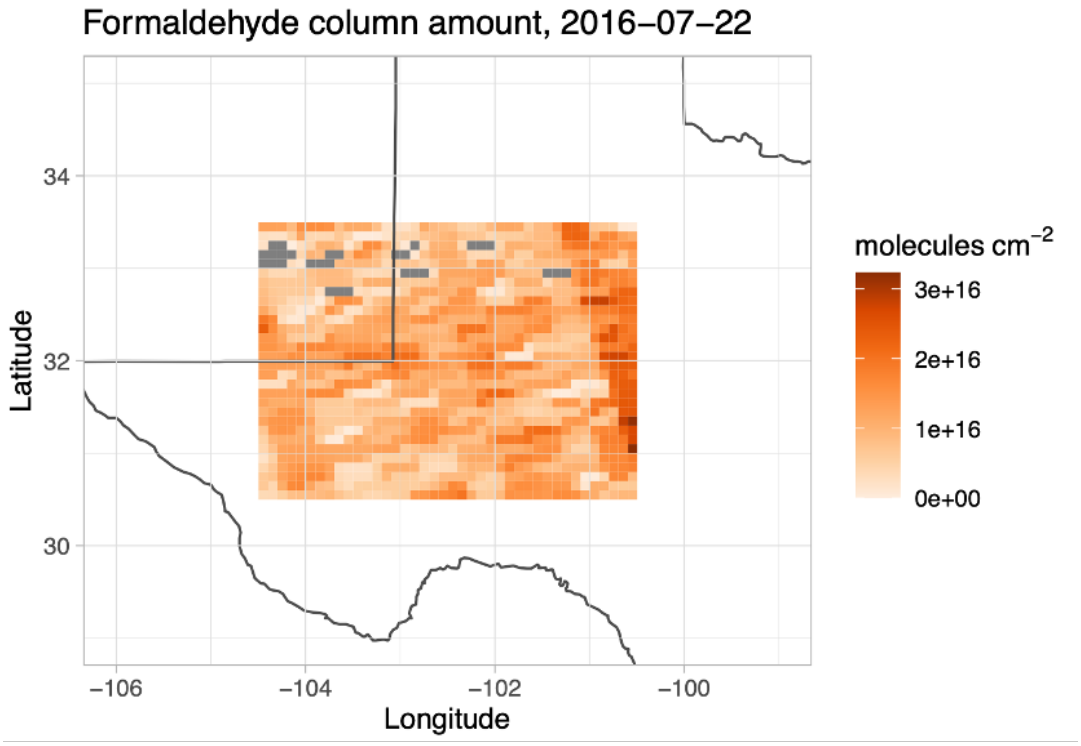

Figure B1. Example daily HCHO gridded data.

Time and Spatial Averaging, and Weighting by Uncertainty

Daily HCHO data were averaged to monthly values, maintaining latitude/longitude grid coordinates (i.e., the daily HCHO 40 × 30 longitude-latitude grid was averaged to a monthly 40 × 30 grid of the study region). Since the NASA data provider and our QA procedures removed several daily grid cell HCHO values (e.g., due to too high cloud cover), monthly averages for a given grid cell were only computed and carried forward if the number of daily values for that grid cell was $N_d \geq 10$ (approximately one-third of possible monthly values, such as to maintain representativeness).

Both temporal averaging and spatial averaging techniques could lead to unintended biases. Therefore, we assessed the potential HCHO VCD uncertainty of using different averaging procedures on any observed trends in the final time series data. The monthly mean was calculated (1) as an unweighted monthly mean, (2) as a monthly mean weighted by the *absolute* data column uncertainty supplied by NASA for each daily grid value (i.e., in molecules/cm^2 units), and (iii) as a monthly mean weighted by the *relative* data column uncertainty (i.e., percent).

The HCHO monthly and spatial means were calculated as unweighted means, means weighted by the *absolute* data column uncertainty (i.e., in molecules/cm^2 units), and means weighted by the *relative* data column uncertainty (i.e., in percent). Monthly mean weighted by the absolute data column uncertainty used the uncertainty supplied by NASA for each daily grid value, and spatial weighting uncertainty used the propagated uncertainty from the monthly mean.

Weights were defined as

$$w_i = \frac{1}{\sigma_i^2} \quad (1)$$

And the weighted mean as

$$\bar{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i} \qquad (2)$$

Higher HCHO values tend to have higher absolute uncertainty. Thus, with absolute uncertainty–weighted means, higher values would be weighted less and dampened out. However, the relative percent uncertainty for higher HCHO tends to be lower. Thus, with relative uncertainty–weighted means, higher values would be weighted more. As expected, the absolute uncertainty–weighted means yielded an absolute lower monthly regional time series, and the relative uncertainty–weighted means led to a higher monthly regional time series (Figure B2). The overall trend, nevertheless — and studying the relationship to O&G production and O&G regulation effects — was equivalent. Since this study focused on long-term trends, the uncertainty weighting of the temporal and spatial means was found not to affect the overall interpretation of this work. However, in studies comparing satellite data with surface measurements or assessing absolute HCHO values, aggregating regional data into periodic spatial means, this weighting would be an important consideration. All results presented in this report show the unweighted means.

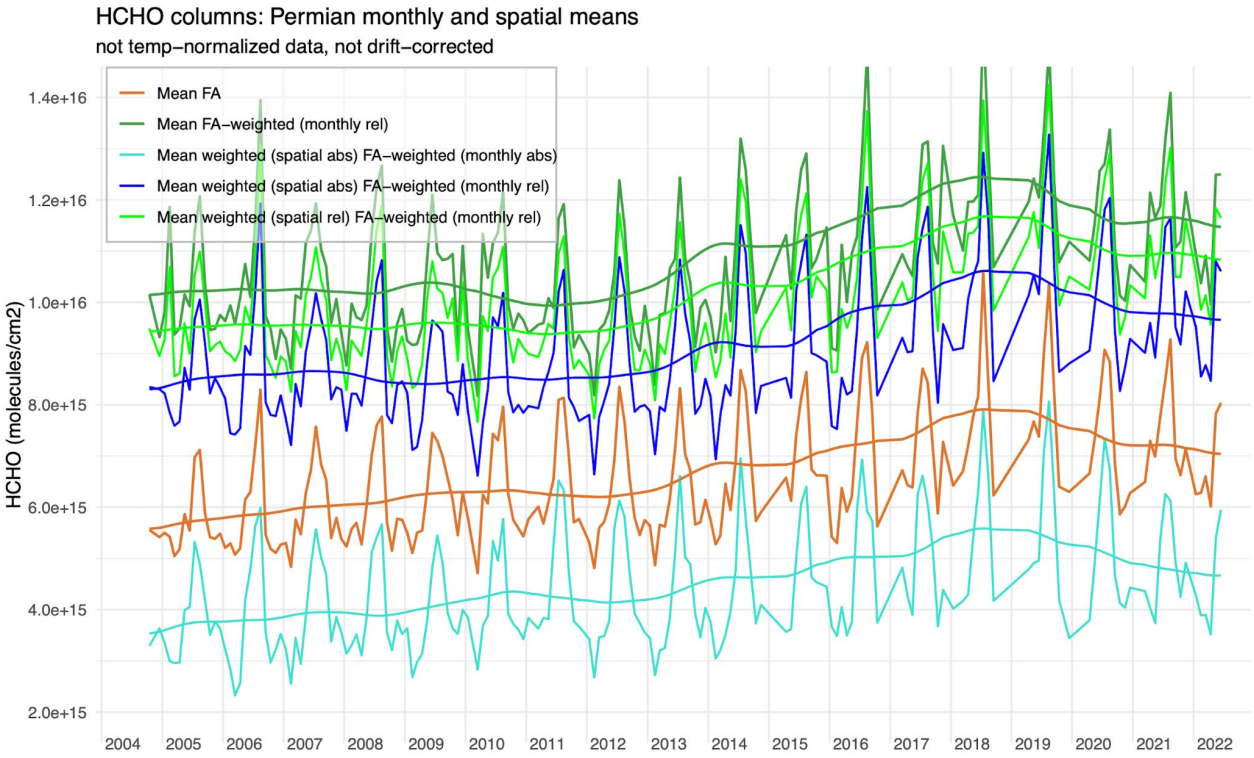

Figure B2. This plot shows different combinations of unweighted ("Mean FA," orange lines), absolute uncertainty–weighted ("Mean weighted (spatial abs) FA–weighted (monthly abs)," turquoise lines: absolute uncertainty for both monthly and spatial averaging), and relative uncertainty–weighted monthly and spatial means ("Mean FA–weighted (monthly rel)," green lines: relative uncertainty for temporal monthly averaging, simple unweighted spatial mean; "Mean weighted (spatial rel) FA–weighted (monthly rel)," bright green lines: relative uncertainty–weighted for both monthly and spatial means; and "Mean weighted (spatial abs) FA–weighted (monthly rel)," blue lines: relative

uncertainty for temporal monthly averaging, absolute uncertainty weighting for spatial averaging) for the entire Permian. Data represent monthly values (jagged lines) and a Gaussian-smoothed time series of the same color.

Daily HCHO values for each grid cell were determined to have no or minimal autocorrelation; therefore, the daily HCHO uncertainty provided with the data were propagated to the monthly uncertainty using

$$\sigma_{mean} = \sqrt{\frac{\sum_i {\sigma_i}^2}{N_d^2}} \quad (3)$$

Monthly latitude-longitude grids were spatially averaged to regional monthly averages for the entire Permian Basin study region (104.5 to 100.5° W, 30.5 to 33.5° N) and also divided into West and East regions at 102.7° W to assess differences in the Delaware (west) and Midland (east) basins. Similar to monthly averaging, for a spatial average to be computed and carried forward for any given month, one-third of possible grid cells were required to be present (entire Permian Basin ≥400 cells out of 1200 total in study region, West Permian ≥180 cells, East Permian ≥220 cells). Also, similar to the temporal averaging, spatial averages were tested as unweighted, as weighted by absolute uncertainty, and as weighted by relative uncertainty. All results presented here are unweighted averages.

OMI Instrument Background Linear Drift Correction and Arizona Background

OMI HCHO data have a known background drift (Shen et al. 2019; Zhu et al. 2017), which was removed from our monthly and spatially averaged Permian data using OMI HCHO VCD data from a remote region over the Pacific at a similar latitude (170 to 167 °W, 30 to 33 °N). Recent work attempting to upgrade the existing OMI HCHO NASA dataset with an improved algorithm seems to have succeeded in eliminating OMI instrument drift and noise and thus found the background HCHO column over the Pacific as constant in time (Ayazpour et al. 2025), supporting the use of the Pacific as a reference for the OMI HCHO drift correction for the current OMI dataset. A linear trend was fit to the Pacific OMI data (quality-assured and averaged in the same manner as the Permian data), and the resulting trend was removed from the Permian Basin data. Under the assumption that HCHO in that region is entirely due to methane oxidation, the observed time development represents the observable background "drift" of the retrieval algorithms' HCHO output. Note that the lowest data points in this time series, at approximately 4×10^{15} molecules/cm^2, have previously been interpreted as the "detection limit" of the instrument (Millet et al. 2008). Figure B3 shows the monthly spatial means and linear fit for the Pacific data, alongside the monthly spatial means and Gaussian smoothed curves for the Permian and Arizona regions before any drift correction. OMI HCHO data for a region in southwest Arizona represents immediate background air entering the Permian Basin and thus was included to better ascertain HCHO changes originating from the Permian region itself.

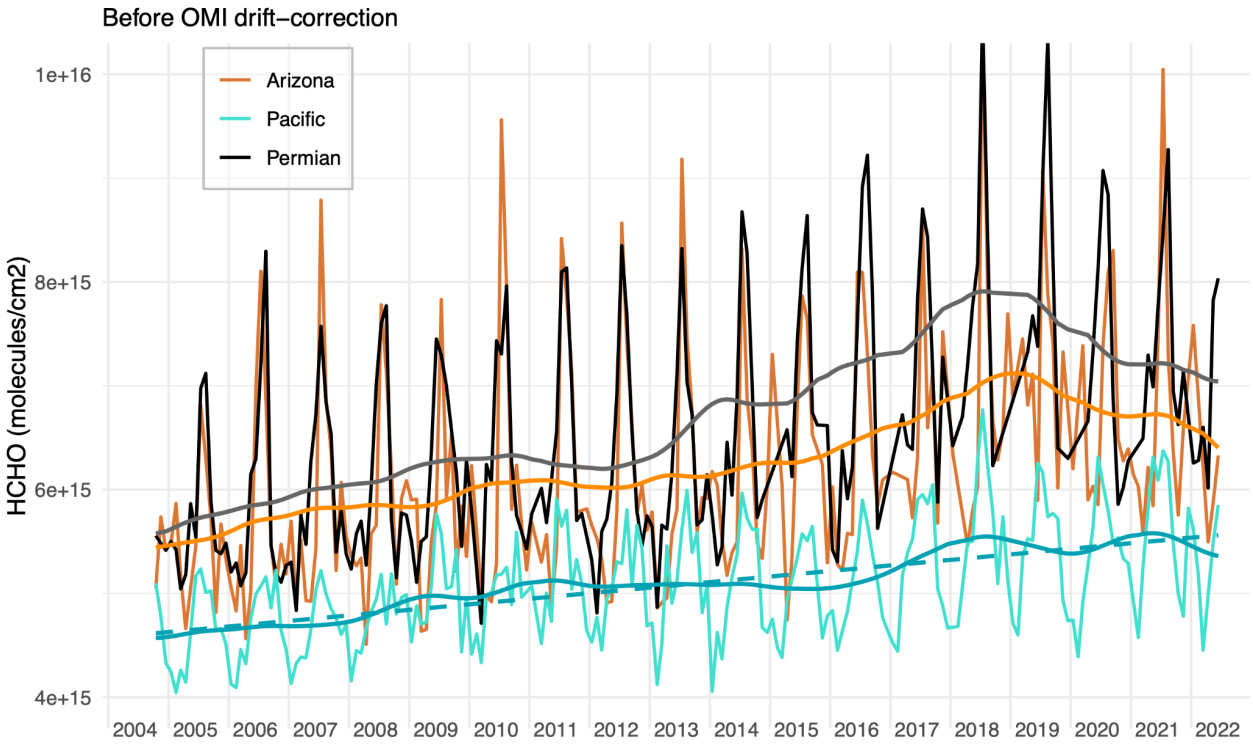

Figure B3. Complete OMI HCHO VCD time series of monthly and regionally averaged data from the Permian Basin (black lines), southern Arizona (orange line), and part of the Pacific reference (cyan line). No trend is expected of the Pacific region, and thus, the dashed blue line represents the drift correction.

For the complete period of OMI HCHO VCD data, we found a drift rate of 5.14×10^{13} molecules/cm^2 per year over the remote Pacific. Shen et al (2019), using a larger region of the Pacific over the shorter timespan 2005–2016, defined a drift rate of 4.3×10^{13} molecules/cm^2 per year; using our defined region of the Pacific for the same 2005–2016 timespan resulted in a compatible 4.6×10^{13} molecules/cm^2 per year.

HCHO Temperature Correction

As tropospheric formaldehyde production from hydrocarbon emissions is dependent on emission rates and atmospheric chemistry, both of which are dependent on surface temperatures, HCHO data were normalized to the mean monthly surface temperature, similar to Zhu et al. (2017). Monthly temperature data from the NASA MERRA-2 model (Modern-Era Retrospective analysis for Research and Applications, Version 2; NASA GES-DISC) at 0.5°×0.625° resolution was matched to the 0.1°×0.1° HCHO gridded data over the Permian Basin. HCHO VCD amounts were then regressed onto the surface air temperature, and the fitted temperature dependency was removed from the data. The effect is predominantly on the seasonality and not on any trends. Figure B4 shows the monthly average HCHO Permian data before and after regressing the monthly mean HCHO columns onto NASA's MERRA-2 model monthly surface temperatures and then normalizing HCHO to the fitted temperature dependency. The effect is predominantly on seasonal HCHO levels rather than longer-term trends.

While background drift and temperature corrections were made to the HCHO data, no other absolute data adjustments were carried out to determine or reflect actual boundary layer formaldehyde concentrations because our interest is less in absolute than in relative changes over time.

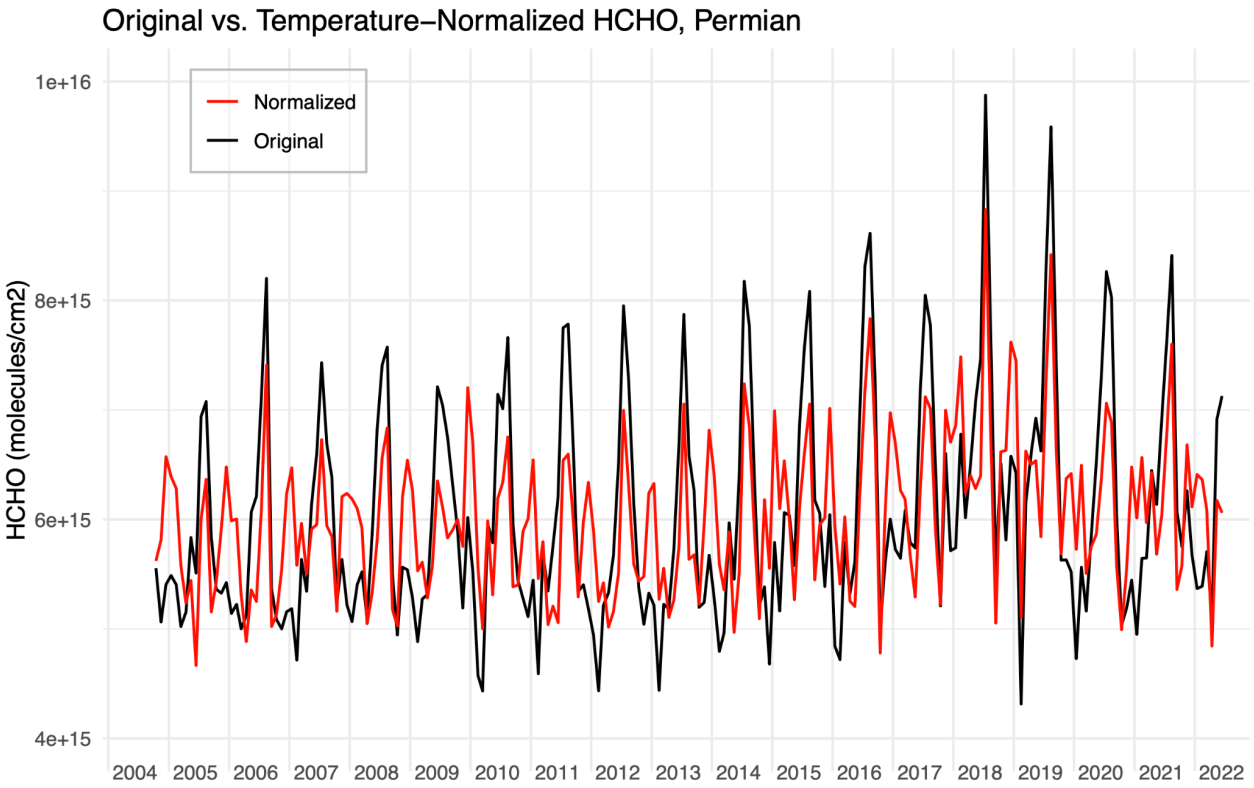

Figure B4. Permian Basin unweighted monthly averages for the OMI satellite HCHO VCD data. In black are the HCHO values before temperature correction, in red are the values after normalization to MERRA-2 monthly mean surface temperatures for that region.

OMI Valid Measurements and Uncertainty

In the monthly and annual HCHO time series of time-averaged and spatially averaged daily gridded HCHO data, the uncertainty is seen to increase with time (Main Report Results section, Figure 43). While the uncertainty associated with individual OMI measurements has remained relatively constant over the timespan of the instrument, there have been fewer valid OMI measurements over time due to a detector *row anomaly* that occurred a few years into operation, thus causing increasingly more data to be flagged; valid data decreased roughly by 30% from 2005 to 2019 (De Smedt et al. 2021). For the Permian data presented here, Figure B5 shows the number of monthly averaged grid cells (N_{cells}) going into the Permian spatial average calculation, with each month providing a maximum of 1200 grid cells (40 0.1° longitude by 30 0.1° latitude cells). Here, the number of months with lower N_{cells} becomes more common in later years, coincident with the step to increase annual average uncertainty (Main Report Results section, Figure 43). The fewer N_{cells} results from fewer valid original daily gridded OMI data also led to fewer grid cells meeting the $N_d \geqq 10$ criteria (number of valid data days per month) to calculate a monthly average. The smaller amount of data feeding into the monthly and annual averages in later years leads to less of a decrease in the averages' propagated uncertainty.

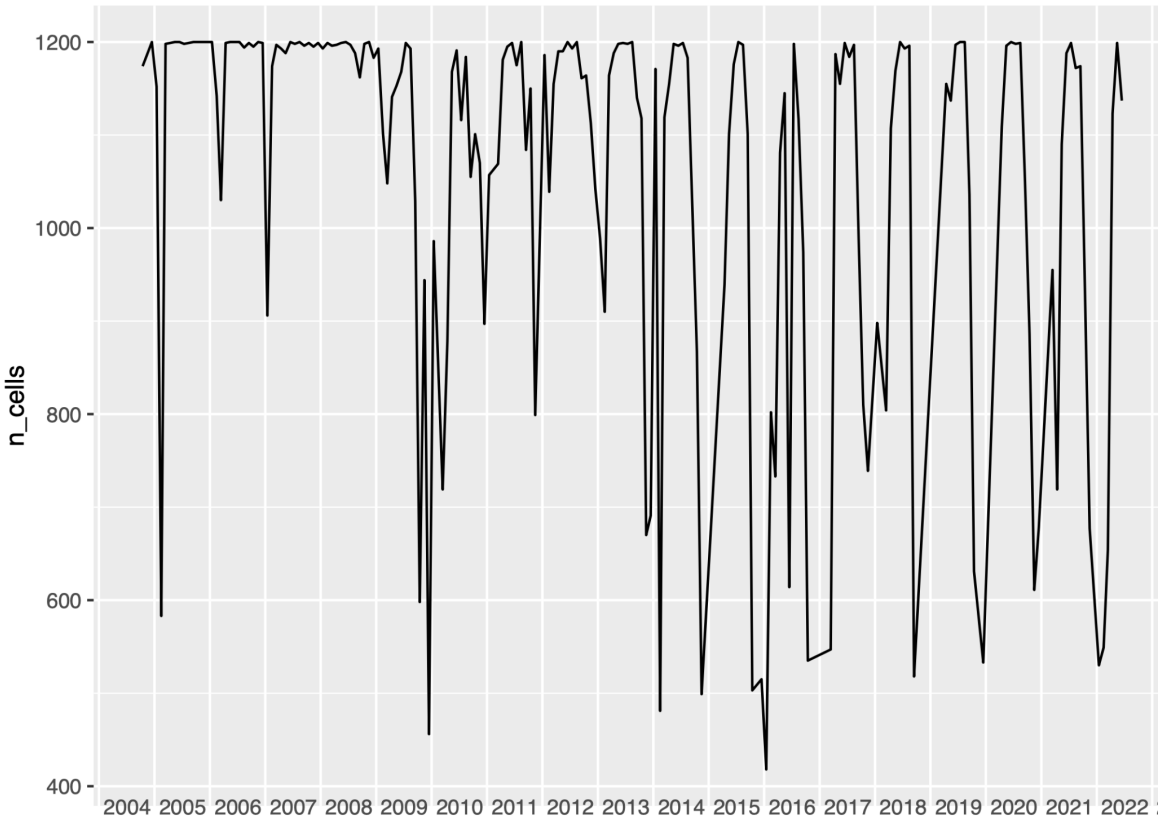

Figure B5. Number of monthly average grid cells for the Permian study region each month (N_{cells}) going into the spatial average calculation. The full study region consists of 1200 0.1°×0.1° cells.

Permian Subregions: Delaware (West) and Midland (East) Basins

The defined Permian study region was divided into west and east subregions, overlying the Delaware and Midland subbasins, respectively, to investigate any contrasts. Figure B6 shows the results as a comparison between the two regions, revealing no significant differences.

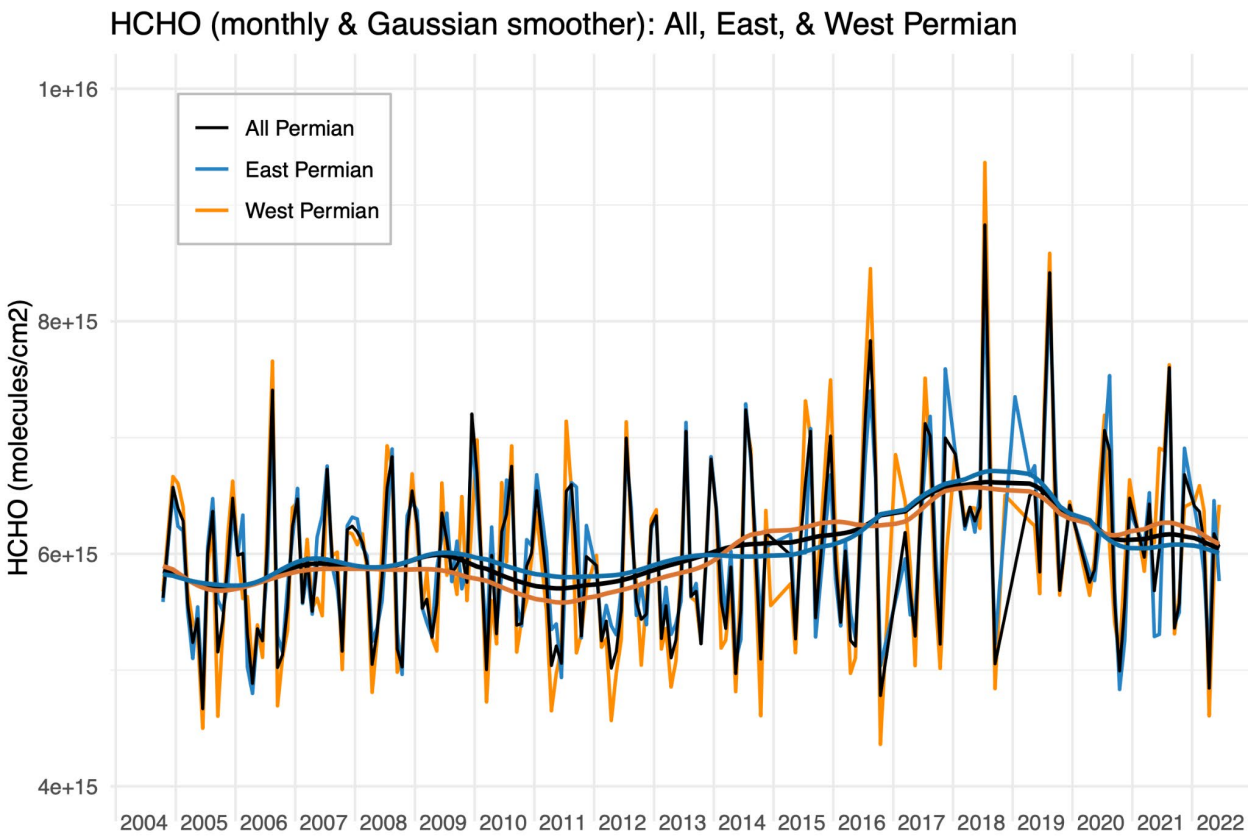

Figure B6. Monthly and spatially averaged time series for the entire Permian study region (black line), and the defined western and eastern subregions divided at 102.7 °W (Permian-Delaware, orange line and trend; Permian-Midland, blue line and trend).