Contract No. 582-23-45978
Work Order No. 16
Tracking No. 2025-19
Task 8.2

Prepared for:

Texas Commission on Environmental Quality
12100 Park 35 Circle MC 164
Austin, TX 78753

Prepared by:

Ramboll
7250 Redwood Blvd., Suite 105
Novato, California 94945

June 27, 2025

Biogenic Model Evaluation Using Satellite Formaldehyde

Final Report

PREPARED UNDER A CONTRACT FROM THE
TEXAS COMMISSION ON ENVIRONMENTAL QUALITY

The preparation of this document was financed through a contract from the State of Texas through the Texas Commission on Environmental Quality.
The content, findings, opinions and conclusions are the work of the author(s) and do not necessarily represent findings, opinions or conclusions of the TCEQ.

Biogenic Model Evaluation Using Satellite Formaldehyde
Final Report

Ramboll
7250 Redwood Boulevard
Suite 105
Novato, CA 94945
USA

T +1 415 899 0700
https://ramboll.com

Contents

Table of Figures

Table of Tables

List of Acronyms and Abbreviations

AGU	American Geophysical Union
AMF	Air Mass Factor
BEIS	Biogenic Emission Inventory System
BELD	Biogenic Emissions Landuse Database
BVOC	Biogenic Volatile Organic Compound
CAMx	Comprehensive Air Quality with Extensions
CMAQ	Community Multiscale Air Quality
CO	Carbon Monoxide
CONUS	Continental United States
EPA	Environmental Protection Agency
ESA	European Space Agency
GES DISC	Goddard Earth Sciences Data and Information Services Center
HCHO	Formaldehyde
MB	Mean Bias
MEGAN	Model of Emissions of Gases and Aerosols from Nature
MPE	Model Performance Evaluation
NASA	National Aeronautics and Space Administration
NetCDF	Network Common Data Form
NMB	Normalized Mean Bias
NME	Normalized Mean Error
NO2	Nitrogen Dioxide
OFFL	Offline
qa_flag	Quality Assurance Flag
QF	Quality Flag
RPRO	Reprocessed
S5P	Sentinel-5 Precursor
SCD	Slant Column Density
SIP	State Implementation Plan
TCEQ	Texas Commission on Environmental Quality
TEMPO	Tropospheric Emissions: Monitoring of Pollution
TM5	Tracer Model 5
TROPOMI	TROPOspheric Monitoring Instrument
U.S.	United States
VCDs	Vertical Column Densities
VMR	Volume Mixing Ratio
VOC	Volatile Organic Compound

Project Summary

The goal of this project was to evaluate biogenic emission inventories by comparing formaldehyde concentrations estimated by CAMx using two different biogenic inventories—BEIS and MEGAN—with satellite observations from instruments such as TROPOMI. To ensure reliable comparisons, Ramboll developed a process to align the satellite and model data in time and space, using Python tools and adjustments for differences in vertical profiles. Overall, the model tended to overestimate HCHO levels, with BEIS showing better alignment with satellite data than MEGAN, especially in regions with high natural emissions.

Executive Summary

The Texas Commission on Environmental Quality (TCEQ) uses the Comprehensive Air Quality Model with Extensions (CAMx) to simulate ozone and particulate matter ($PM_{2.5}$) incorporating biogenic emissions from either the Biogenic Emission Inventory System (BEIS) or the Model of Emissions of Gases and Aerosols from Nature (MEGAN). These two biogenic models can produce widely varying emission estimates, making it challenging to determine which is more appropriate for air quality modeling in Texas and surrounding regions.

Biogenic volatile organic compounds (BVOCs), particularly isoprene, are major precursors of formaldehyde (HCHO) through atmospheric oxidation, especially during the growing season in regions with high biomass. Satellite instruments such as the TROPOspheric Monitoring Instrument (TROPOMI) and Tropospheric Emissions: Monitoring of POllution (TEMPO) provide high-resolution data of atmospheric HCHO, which can be used to evaluate biogenic emission inventories.

For this project, Ramboll developed a suite of Python-based tools to prepare and analyze tropospheric HCHO vertical column density (VCD) data from satellites and CAMx simulations. These tools mapped Level 2 satellite retrievals to CAMx model domains (at 4 km and 12 km resolution), aligned satellite observation times with CAMx output times, and applied necessary vertical corrections using CAMx-simulated HCHO profiles. This workflow established the spatial, temporal, and vertical consistency needed for comparison between satellite data and model outputs.

The evaluation of CAMx simulated HCHO VCDs against TROPOMI produced several key findings:

- CAMx consistently overestimated HCHO VCDs relative to TROPOMI, with the overestimation less pronounced in simulations using BEIS.
- Although the CAMx simulations using both biogenic models overestimated HCHO, those using BEIS showed closer agreement with TROPOMI than simulations using MEGAN, particularly in regions with elevated biogenic emissions.
- Strong spatial correlations were observed in the Southeastern U.S. Pine Forests, the Ozarks, and the South Texas Agricultural Region, suggesting that CAMx captured regional HCHO patterns despite the positive bias.
- Variations in fire emission inventories contributed to differences in modeled HCHO, especially in areas and years affected by wildfire activity.

The consistent overestimation of HCHO by CAMx underscores the need for further investigation into biogenic emissions inputs and chemical mechanisms. Ramboll recommends the following next steps to better understand the observed discrepancies:

1. Compare with other modeling studies, such as Hoque et al. (2024) and Goldberg et al. (2022), to assess differences in emission inventories, chemical mechanisms, and satellite data processing.
2. Investigate the role of isoprene emissions and oxidation chemistry by analyzing the correlation between CAMx-modeled isoprene and HCHO VCD differences relative to TROPOMI.
3. Evaluate additional VOC precursors, including anthropogenic species like acetylene and highly reactive VOCs, particularly in urban and industrial areas.

Together, these efforts will support TCEQ’s State Implementation Plan (SIP) modeling by informing the selection of biogenic emission models, which play a key role in ozone modeling.

1.0 Introduction

Atmospheric oxidation of biogenic volatile organic compound (BVOC) emissions is an important source of formaldehyde (HCHO) and the dominant source over regions with high biomass during the growing season (Fortems-Cheiney et al., 2012). Anthropogenic volatile organic compound (VOC) emissions are also important HCHO precursors and may dominate when anthropogenic emissions are high and/or when biogenic emissions are low (Hoque et al., 2024). Wildfires and other biomass burning sources emit VOCs and therefore contribute to HCHO. Additionally, the slow oxidation of methane provides a global HCHO background or "floor" (Levy, 1972).

Satellite instruments such as TROPOspheric Monitoring Instrument (TROPOMI) and Tropospheric Emissions: Monitoring of POllution (TEMPO) can detect atmospheric column HCHO at a resolution of 3.5 km x 5.5 km for TROPOMI and 2.0 km x 4.5 km for TEMPO. These high-resolution measurements align well with Texas Commission on Environmental Quality (TCEQ) modeled emissions, which have a resolution as fine as 4 km. While TROPOMI has been operational since 2017 and provides consistent coverage for TCEQ's modeling years of 2019 and 2023, TEMPO is relatively new, with quality-assured retrievals available only since August 2023. Therefore, TEMPO data were used for training purposes only.

Isoprene emissions often dominate BVOCs and have a substantial impact on air quality in Texas and other regions. However, the two widely used biogenic emission models, the Biogenic Emission Inventory System (BEIS) and the Model of Emissions of Gases and Aerosols from Nature (MEGAN), produce widely varying emission estimates. This makes it difficult to determine the most appropriate biogenic model for air quality modeling. Satellite-detected HCHO columns provide a valuable proxy for evaluating these biogenic emissions.

This project assessed biogenic emission inventories by comparing TROPOMI tropospheric HCHO vertical column densities (VCDs) with HCHO VCDs derived from four simulations using the Comprehensive Air quality Model with extensions (CAMx; Ramboll (2024)). Each simulation used a different combination of meteorological data (2019 and 2023) and biogenic emissions. For both years, one simulation used BEIS version 3.7 with the Biogenic Emissions Landuse Database (BELD) version 5, while the other used MEGAN version 3.2. This work will support TCEQ's State Implementation Plan (SIP) modeling efforts by evaluating the biogenic emission models, which play a key role in ozone and $PM_{2.5}$ chemistry.

2.0 TROPOMI HCHO Processing

TROPOMI is a satellite instrument on board the Copernicus Sentinel-5 Precursor (S5P), which was launched by the European Space Agency (ESA) on October 13, 2017. It monitors air pollution by measuring gases like nitrogen dioxide (NO_2), HCHO, and carbon monoxide (CO) using light from different parts of the electromagnetic spectrum — ultraviolet, visible, and infrared. The satellite follows a sun-synchronous orbit with an equator crossing time near 13:30 local time, providing consistent daily coverage for each location. ESA provides formaldehyde data as part of its standard data products, specifically focusing on how much is present in the lower atmosphere, known as "tropospheric vertical column densities".

Ramboll used two sets of TROPOMI data: the reprocessed (RPRO) Level 2 data for 2019 and offline (OFFL) Level 2 data for 2023, both from version 02-04-01. These datasets were downloaded from the National Aeronautics and Space Administration's (NASA) Goddard Earth Sciences Data and Information Services Center (GES DISC) website[1]. Access to this data requires a free NASA EarthData account. Level 2 satellite data provide geolocated retrievals of atmospheric traces gases and aerosol derived from raw satellite measurements, mapped onto a swath-based grid. Level 3 data are processed from Level 2 data by mapping them onto a uniform spatial and temporal grid, which involves averaging over space and time.[2]

For comparison with CAMx, the TROPOMI Level 2 data were mapped to the CAMx model grid (4 km or 12 km resolution) using a nearest-neighbor method to create a Level 3 product on a common grid. We only included good-quality satellite data, specifically pixels with a quality assurance (qa_flag) value above 0.5. This threshold ensures that the data meet several conditions: no error flags, a cloud radiance fraction at 340 nm below 0.5, a Solar Zenith Angle (SZA) of 70° or less, surface albedo of 0.2 or lower, no snow or ice warnings, and an air mass factor above 0.1. These filters helped exclude poor-quality measurements affected by cloud cover, high surface reflectivity, or instrument issues. When multiple good-quality satellite measurements fell within the same grid cell on the same day, we treated each one as its own data point and averaged them to get a single, daily value for that grid cell.

Even though our Level 3 data are on the same horizontal grid as the CAMx model, there is still an important difference: the satellite retrievals rely on "a priori" assumptions about the vertical distribution of HCHO in the atmosphere, while CAMx uses its own simulated vertical profiles. These differences can introduce artificial mismatches when comparing the two, so we applied some additional adjustments to help account for these vertical differences.

Satellite-based measurements of tropospheric HCHO are sensitive to both its vertical distribution and the air mass factor (AMF) used in retrieval. To limit artificial differences when comparing to CAMx outputs, we adjusted the model vertical profiles to better align with the satellite retrievals, following the method described by Boersma et al. (2004) and Lamsal et al. (2010), which incorporates both the averaging kernel and the AMF into the analysis. Specifically, we applied TROPOMI's averaging kernels—both the total (A_k) and tropospheric (A_k^{trop}) versions—at each vertical level (k), using HCHO vertical profiles (X_k^{camx}) from the CAMx simulations. This helps correct for biases introduced by the satellite's built-in assumptions, which are based on the global Tracer Model 5 (TM5; Huijnen et al., 2010) at a coarser spatial resolution of 1°×1°.

1 https://tropomi.gesdisc.eosdis.nasa.gov/data/S5P_TROPOMI_Level2/S5P_L2__HCHO___HiR.2/

2 https://www.earthdata.nasa.gov/about/esdis/esco/standards-practices/data-processing-level-definitions

The final adjusted satellite-based estimate of tropospheric HCHO, referred to as Ω_{camx}, was calculated using the following equations (Lamsal et al., 2010; Palmer et al., 2001):

$$\Omega_{camx} = \frac{M^{trop}}{M^{trop\_camx}} \Omega_{TROPOMI} \quad (2\text{-}1)$$

$$M^{trop\_camx} = \frac{M^{trop} \sum_k A_k^{trop} X_k^{camx}}{\sum_k X_k^{camx}} \quad (2\text{-}2)$$

$$A_k^{trop} = \frac{M}{M^{trop}} A_k \quad (2\text{-}3)$$

where, trop refers to the tropopause, $\Omega_{TROPOMI}$ is the gridded satellite tropospheric HCHO VCD, M and M^{trop} are the total and tropospheric AMFs derived from the a priori TM5 model, and M^{trop_camx} is a recalculated tropospheric AMF that uses the vertical HCHO profile from CAMx, X_k^{camx}. The CAMx vertical HCHO profile was interpolated to TROPOMI's pressure levels using linear interpolation to align with the satellite retrieval layers. Tropopause pressure, needed to separate tropospheric and stratospheric columns, was obtained from TROPOMI data, which is based on TM5 model output.

To align with the TROPOMI overpass time (~13:30 local time), we used CAMx model outputs from 13:00 Central Time for the Texas region. A ±30-minute window was applied to select the closest time match. Since CAMx provides hourly averaged outputs, this approach helps makes sure the satellite data and model results are as closely aligned in time as possible. To enable this comparison, TROPOMI observation times, originally reported in UTC, were converted to Central Time to match the local time used in the CAMx output.

2.1 Overview of Python Tool

A Python-based tool was developed to grid Level 2 TROPOMI HCHO data and apply averaging kernel revisions using the CAMx HCHO profile. The tool consists of two Python scripts executed using a C-shell wrapper script for batch processing.

2.1.1 Gridding TROPOMI Data

The first script, "grid_tropomi_L2_HCHO.dom_daily.py", takes TROPOMI Level 2 swath data and maps those onto a target modeling domain using a nearest-neighbor method. It is designed to work with both 4 km and 12 km grid resolutions to match the CAMx model grid resolution. To run the script, users need to provide a domain-specific latitude-longitude NetCDF file.

The script processes several important variables, including tropospheric HCHO VCDs, the averaging kernel, the satellite a priori profile, the tropopause layer index, and vertical pressure coordinates. All these variables are then saved to daily NetCDF files tailored for the user-defined domain.

The vertical pressure levels used in the scripts are calculated using hybrid coefficients from the TM5 model, following this formula[3]:

$$p(n,k,j,i) = a_p(k) + b(k) \cdot p_s(n,j,i) \qquad (2\text{-}4)$$

[3] https://sentiwiki.copernicus.eu/__attachments/1673595/S5P-L2-DLR-PUM-400F%20-%20Sentinel-5P%20Level%202%20Product%20User%20Manual%20Formaldehyde%20HCHO%202022%20-%202.4.pdf?inst-v=87ef0ca0-8091-4ed6-bc9f-05ea3a6bc632

where $p(n,k,j,i)$ is the pressure at each grid point (time n, vertical level k, latitude j, and longitude i); $a_p(k)$ and $b(k)$ are the hybrid coordinate coefficients for level k, and $p_s(n,j,i)$ is the surface pressure at the corresponding location and time.

Only retrievals with a qa_value above 0.5 are retained to ensure data quality. For efficient nearest-neighbor matching, the script uses the *KD-tree* method from "scipy.spatial". Input and output of netCDF files are handled with the netCDF4 package, and numerical calculations are performed using "numpy". Figure 2-1 displays TROPOMI tropospheric HCHO VCDs for a single day (September 17, 2023), showing both the original Level 2 data and the gridded Level 3 data over the 12 km Continental U.S. (CONUS) modeling domain.

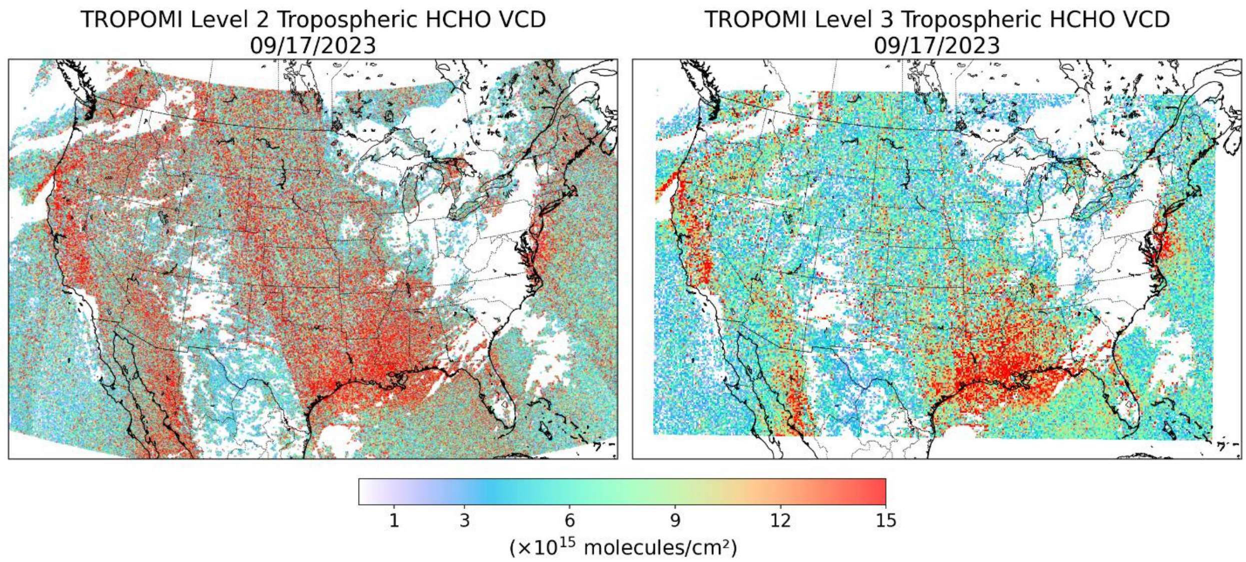

Figure 2-1. TROPOMI tropospheric HCHO VCDs from Level 2 and gridded Level 3 data for a single day (September 17, 2023) over the 12 km CONUS modeling domain.

2.1.2 Averaging Kernel Correction

The second script, "hcho_vcd.camx.tropomi.dom_daily.py", applies averaging kernel corrections to the gridded TROPOMI HCHO data using vertical HCHO profiles from the CAMx model to make it more accurate. Specifically, it corrects for biases caused by the built-in assumptions (from the TM5 global model) used in the satellite retrievals. To align with the TROPOMI overpass time, which occurs between 12:00 and 15:00 Central Time (CT) across the continental U.S., the script extracts CAMx model output within that same time window. The corrected data are saved in daily netCDF files specific to each modeling domain. These files are set up with 24 hourly time steps, but only the 12:00–15:00 period contains actual data; the other time slots are just placeholders to keep the file structure consistent.

3.0 TEMPO HCHO Processing

The TEMPO instrument is a spectrometer onboard the Intelsat 40e geostationary satellite. It provides daytime measurements of HCHO across North America, with both high temporal and spatial resolution. Ramboll used Level 2 HCHO data (version 3) from NASA's EarthData portal, which allows users to select specific regions and timeframes. After making a selection, the portal generates a script to download the data; note that a NASA EarthData login is required.

The Level 2 data include several key variables, such as slant column density (SCD), AMF, scattering weights, temperature profile, and terrain height. We created Level 3 daily averages by mapping each TEMPO observation to the CAMx model grid (at either 4 km or 12 km resolution) using a nearest-neighbor approach. We also converted the satellite observation times from UTC to U.S. Central Time to match CAMx output and applied a ±30-minute window to align the timing as closely as possible.

To make sure we used only high-quality satellite data, we kept only those pixels where the main data quality flag was 0 (meaning the data passed all quality checks) and the effective cloud fraction was below 0.2. This filtering helped minimize issues related to clouds or retrieval errors. If multiple good-quality observations fell within the same model grid cell during a single hour, we averaged them to create a single hourly value.

To make the comparison between TEMPO and CAMx more accurate, we recalculated the AMF using TEMPO's scattering weights and the vertical profiles of HCHO from CAMx, following the approach outlined by Palmer et al. (2001). We first interpolated the CAMx vertical pressure levels to match those used by TEMPO. Then, to separate the tropospheric and stratospheric portions of the HCHO column, we calculated the tropopause pressure using the World Meteorological Organization's (WMO, 1957) definition. This step relied on the temperature profile and terrain height provided in the TEMPO Level 2 data.

Using the resulting tropospheric HCHO profile, we calculated the tropospheric AMF (AMF_{CAMx}) with the following equation:

$$AMF_{CAMx} = \sum_k \left(scattering\ weight_k \cdot \frac{\Delta\Omega_k}{\Omega_{trop}} \right) \quad (3\text{-}1)$$

where $\Delta\Omega_k$ is the partial HCHO column in layer k (in molecules/cm^2) and Ω_{trop} is the total tropospheric HCHO column from CAMx (also in molecules/cm^2).

We then used this recalculated AMF to adjust the TEMPO SCD and derive the tropospheric VCD:

$$VCD_{TEMPO} = \frac{SCD_{TEMPO}}{AMF_{CAMx}} \quad (3\text{-}2)$$

This approach ensures that the satellite retrievals and model output are compared on a consistent basis, accounting for differences in vertical sensitivity and atmospheric structure within the troposphere.

3.1 Overview of Python Tool

A Python-based tool was developed to process and grid Level 2 TEMPO column HCHO SCD data and apply CAMx-derived AMFs to calculate the tropospheric HCHO VCDs. The tool includes two main Python scripts and a C-shell wrapper for batch processing.

3.1.1 Gridding TEMPO Data

The first script, "grid_tempo_L2_HCHO.dom_daily.py", maps TEMPO Level 2 swath data to a user-defined CAMx modeling domain using a nearest-neighbor approach. It grids several key variables, including column HCHO SCDs, scattering weights, AMF, terrain height, temperature, and vertical pressure levels. The script outputs daily netCDF files containing hourly gridded data for the entire domain.

Vertical pressure levels are calculated using GEOS-CF model hybrid coefficients, following the method described in Equation (2-4). To ensure data quality, only retrievals with main_data_quality_flag =0 and eff_cloud_fraction < 0.2 are included. Multiple TEMPO pixels may fall within a single CAMx grid cell; each pixel is treated individually, and those that pass the quality check are averaged to represent the grid cell. The gridding method mirrors that used for TROPOMI data to maintain consistency.

Figure 3-1 displays TEMPO tropospheric HCHO SCDs for a single day (September 17, 2023) at 9:00 AM CT, showing both the original Level 2 data and the gridded Level 3 data over the 12 km CONUS modeling domain.

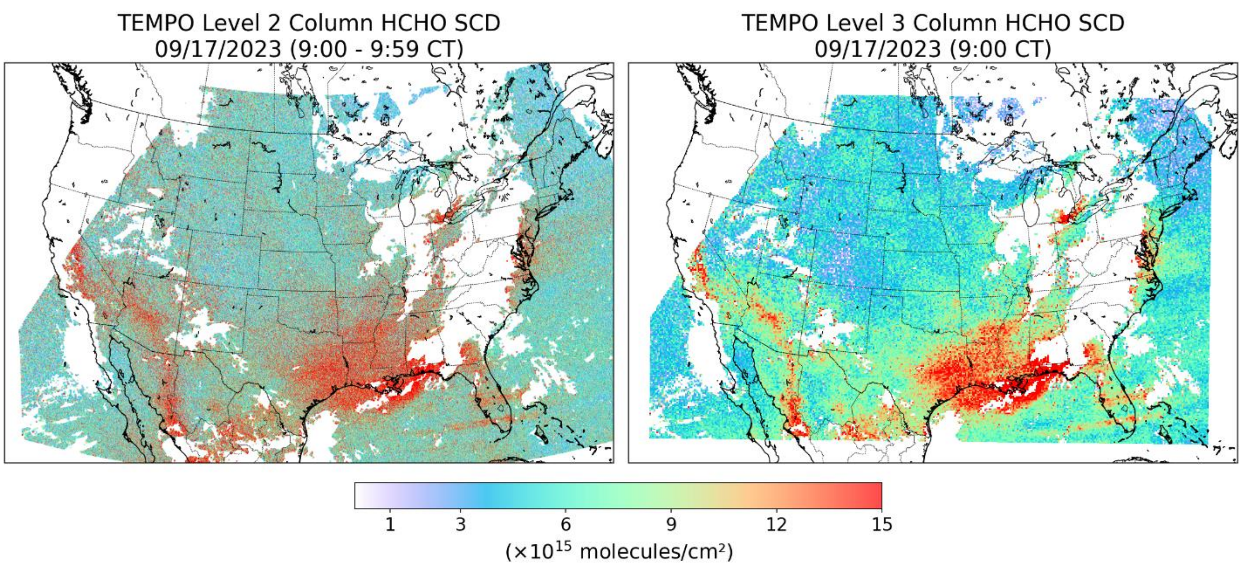

Figure 3-1. TEMPO tropospheric HCHO VCDs from Level 2 and gridded Level 3 data for a single day (September 17, 2023) at 9:00 AM CT over the 12 km CONUS modeling domain.

3.1.2 Air Mass Factor Correction

The second script, "hcho_vcd.camx.tempo.dom_daily.py", recalculates the tropospheric AMFs using CAMx vertical HCHO profiles and TEMPO scattering weights to improve consistency between the satellite retrievals and the model data.

TEMPO timestamps (originally in UTC) are converted to U.S. Central Time to match the CAMx output. The script interpolates CAMx pressure levels to match those used in the TEMPO data, identifies the tropopause using WMO criteria, and isolates the tropospheric portion of the HCHO column. The recalculated AMF is then applied to convert TEMPO SCDs to tropospheric VCDs.

The resulting daily netCDF files are structured with 24 hourly time steps in Central Time, though valid data are only populated for the specific hours when TEMPO observations are available and passed quality screening. The remaining time steps are included in the file structure but contain no data.

4.0 CAMx HCHO VCD Processing

To calculate HCHO VCDs from CAMx model outputs, we started by interpolating the CAMx model HCHO volume mixing ratios (VMRs), reported in parts per million by volume (ppmv), to match the pressure levels used in TROPOMI and TEMPO retrievals (Lamsal et al., 2010; Nawaz et al., 2024; Palmer et al., 2001). This vertical interpolation was performed for each grid column in the CONUS 12 km and Texas 4 km domains, and then converted into partial columns (in molecule/cm^2). The interpolation used pressure (P) data from the WRFCAMx output.

TROPOMI and TEMPO L2 data were gridded separately to match the CAMx resolutions as described in Sections 2.0 and 3.0, respectively. To ensure reliable comparisons[4,5], only pixels with valid quality flags (QF) from each satellite dataset were included. As a result, CAMx vertical profile interpolation was performed only in areas where valid satellite data were available, allowing consistent and reliable comparisons between modeled and satellite-derived HCHO VCDs.

To maintain consistency in defining the tropospheric column, we identified the tropopause level separately for the gridded TROPOMI and TEMPO datasets. This ensures that both CAMx and satellite data use the same upper pressure level when calculating VCDs. Once the tropopause was defined, we summed the HCHO partial columns from all model layers below it to obtain the final vertically integrated tropospheric HCHO VCD.

The final CAMx tropospheric HCHO VCDs were saved to daily netCDF files for each modeling domain and for both biogenic emissions scenarios (BEIS and MEGAN) for the years 2019 and 2023. Each file includes 24 hourly time steps in Central Time, consistent with TCEQ’s CAMx output format. However, valid satellite comparisons are generally limited to the 12:00–15:00 CT window, which aligns with the TROPOMI overpass period and covers all available TEMPO observation hours over the CONUS.

A flowchart illustrating this process is shown in Figure 4-1.

[4] https://sentiwiki.copernicus.eu/__attachments/1673595/S5P-MPC-BIRA-PRF-HCHO%20-%20Sentinel-5P%20Formaldehyde%20Level%202%20Readme%202023%20-%202.7.pdf?inst-v=4318c067-be91-4544-be2e-16af66246c9f

[5] https://asdc.larc.nasa.gov/documents/tempo/ATBD_TEMPO_CH2O.pdf

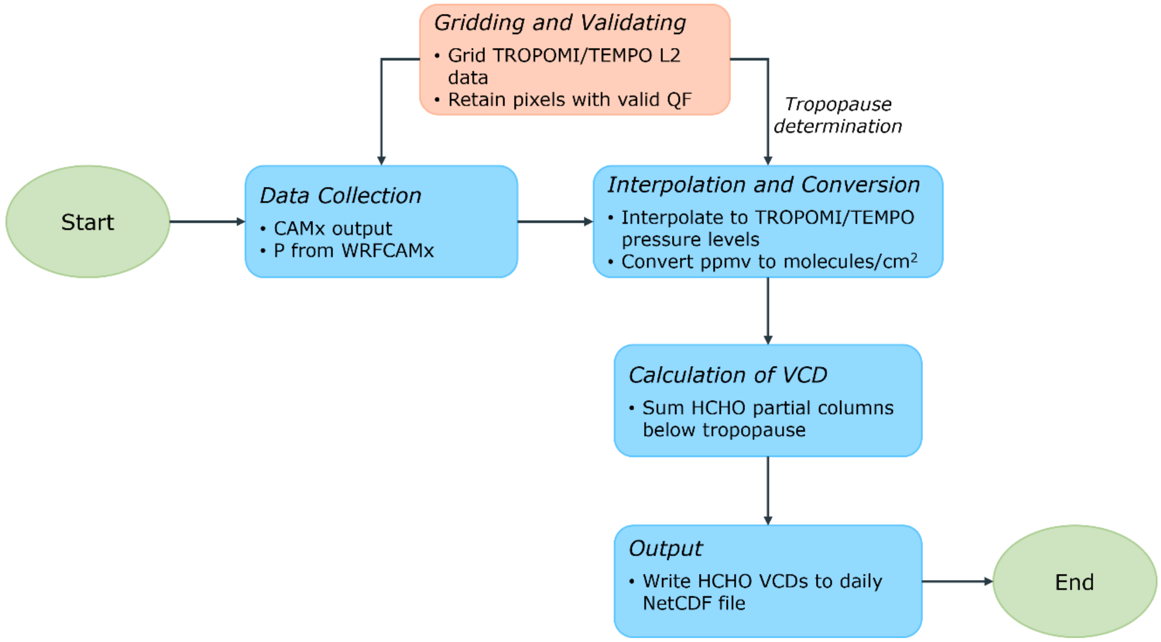

Figure 4-1. Flowchart showing the process used to calculate HCHO VCDs from CAMx outputs. The steps include interpolating to match gridded TROPOMI/TEMPO pressure levels, converting VMRs to partial columns, identifying the tropopause, integrating partial columns to compute VCDs, and exporting the results to netCDF format. Only TROPOMI/TEMPO retrievals with valid quality flags are used in this process.

4.1 Overview of Python Tool

A Python-based tool was developed to calculate CAMx HCHO VCDs by integrating several key components:

- Gridded satellite pressure levels for the CONUS 12 km and Texas 4 km domains, used to interpolate CAMx VMRs to satellite retrieval levels and convert them to partial columns (in molecules/cm²).
- Tropopause level information, used to define the upper boundary of the tropospheric column for integration.

The same process is applied separately for TROPOMI and TEMPO using two dedicated scripts:

- hcho_vcd.camx.tropomi.dom_daily.py (for TROPOMI)
- hcho_vcd.camx.tempo.dom_daily.py (for TEMPO)

These scripts handle the full CAMx VCD calculation workflow and are typically run via a C-shell wrapper script for batch processing. The script performs the following steps:

1. Vertical Interpolation to Satellite Data Pressure Levels

To allow for a direct comparison between CAMx and satellite data, CAMx VMRs are linearly interpolated from the model's native pressure levels to the gridded satellite pressure levels. For TROPOMI, this interpolation is performed for each grid column using CAMx output aligned with the satellite overpass time of 13:30 local time. For TEMPO, which provides hourly measurements throughout the day, the interpolation is performed for each available hourly data using the corresponding hourly CAMx output. The interpolation is carried out using the "interp1d()" function from the "scipy.interpolate" library with the *kind='linear'* option to apply standard linear interpolation between pressure levels.

Since both CAMx and satellite datasets use pressure-based vertical coordinates, pressure levels are first sorted in descending order (from the surface to the top of the atmosphere) prior to interpolation. In some cases, the lowest TROPOMI or TEMPO pressure level may lie above the CAMx surface pressure, meaning it corresponds to a higher altitude. When this happens, interpolation can produce missing values near the surface. To avoid underestimating near-surface contributions to the total column, we fill these gaps using the CAMx surface value.

2. Conversion of CAMx HCHO VMR (ppmv) to Partial Columns (molecules/cm2)

HCHO partial column in each CAMx layer is calculated as:

$$Partial\ Column\ (molecules/cm^2)\ = VMR_{ppmv} \cdot 10^{-6} \cdot \Delta P \cdot \frac{N_A}{(g \cdot M_{air})} \cdot \frac{1}{10^4} \qquad (4\text{-}1)$$

where:

- VMR_{ppmv} = Volume mixing ration of HCHO from CAMx (ppmv), interpolated to TROPOMI/TEMPO pressure levels
- ΔP = Pressure difference between levels (Pa)
- N_A = Avogadro's constant (6.022×10^{23} molecules/mol)
- M_{air} = Molar mass of dry air (28.97 g/mol)
- g = Gravitational acceleration (9.80665 m/s^2)

3. Tropopause Filtering

To isolate the tropospheric portion of the column, we mask out CAMx partial columns above the tropopause. For TROPOMI, the tropopause level is identified using the accompanying gridded data, which provides a layer index indicating the vertical level of the tropopause for each grid column. For TEMPO, we calculate the tropopause pressure using the WMO definition, as described in Section 3.0. Based on this information, all CAMx layers above the tropopause are excluded before integrating the column, ensuring only tropospheric contributions are included in the final VCD calculation.

4. Vertical Integration

The final tropospheric HCHO VCDs are calculated by summing the interpolated partial columns vertically, from the surface up to the tropopause level:

$$HCHO\ VCD\ = \sum_{k=1}^{k_{trop}} Partial\ Column_k \qquad (4\text{-}2)$$

where, k_{trop} represents the tropopause index derived from the gridded TROPOMI or TEMPO data for each grid column. This integration step ensures that only tropospheric portion of the column is included in the final HCHO VCD values.

5. Output to NetCDF

Daily HCHO VCDs are calculated and saved in netCDF format at hourly resolution in Central Time, using the "Dataset" class from the *netCDF4* library. Each netCDF file corresponds to a specific modeling domain and emission scenario (2019 or 2023, using either BEIS or MEGAN). Valid data are stored between 12:00 and 15:00 CT, corresponding to the TROPOMI overpass window across the CONUS and covers all available hourly observations from TEMPO.

5.0 Spatial Masking for Six Ecoregions

To evaluate CAMx-simulated HCHO VCDs against TROPOMI observations, we conducted a comparative analysis across six ecologically distinct regions, as shown in Figure 5-1: East Texas Forests, agricultural areas of Central Texas, Ozark oak forests, Southeastern U.S. Pine Forests, South Texas Agricultural Region, and the Boreal Forests of Ontario, Canada. These ecoregions were selected to represent areas with an abundance of high-emitting trees and other vegetation, both within and outside of Texas.

For each of the six ecoregions, we processed HCHO VCDs from both TROPOMI and CAMx at two spatial resolutions: Texas 4 km for regions within Texas and CONUS 12 km for regions outside Texas. The processing workflow first selects the appropriate domain based on the ecoregion and reads the corresponding netCDF file containing HCHO VCDs from both datasets. For the Southeastern U.S. Pine Forest region, CAMx HCHO columns were averaged over two-time steps, 12:00 and 13:00 CT, to account for the satellite overpass occurring around 13:30 local time, which spans both the Central and Eastern time zones. While 13:00 CT aligns well with the Central portion, including the 12:00 CT time step ensures coverage of the Eastern portion. For all other regions, only the 13:00 CT time step was used.

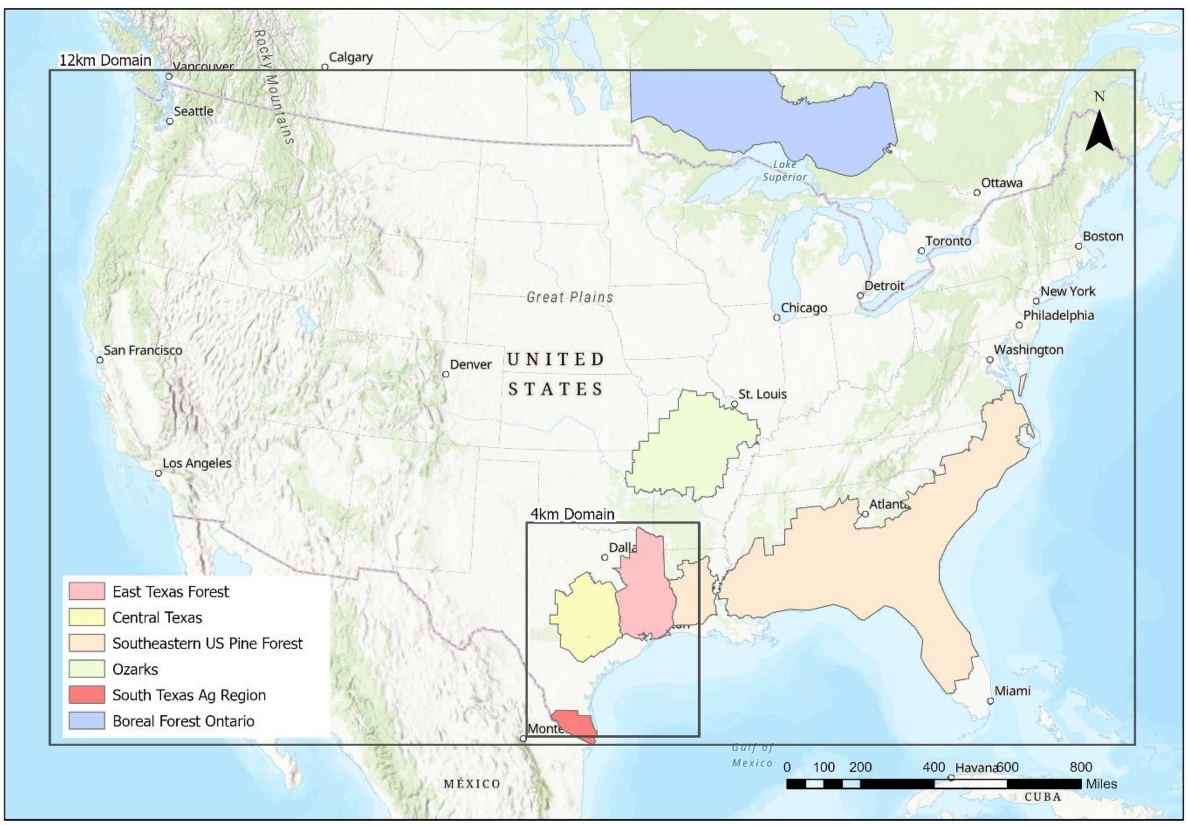

Figure 5-1. Map showing the six ecoregions for spatial analysis: the East Texas forests, agricultural areas of Central Texas, Ozark oak forests, southeastern U.S. pine forests, the agricultural region of Mexico near South Texas, and the boreal forests of Ontario. The map also displays the boundaries of the 4 km and 12 km CAMx modeling domains.

The Python script, "hcho_vcd.ecoregions.daily.py", along with a C-shell wrapper, automates the extraction of TROPOMI and CAMx HCHO data for the six ecoregions and generates a daily netCDF file for each region. A shapefile for each ecoregion was used in the processing workflow to enable spatial data extraction. The script uses the "pyproj" library to convert geographic coordinates (latitude and longitude) into a projected coordinate system (Lambert Conformal Conic) to facilitate spatial analysis. The "Shapely" library is used to construct grid cell polygons and compute their intersections with ecoregion boundaries, while "GeoPandas" handles reading, reprojecting, and performing spatial operations on the shapefiles. These tools work together to isolate grid cells that intersect each ecoregion and extract the corresponding CAMx and TROPOMI values. The final output contains data limited to the ecoregions of interest.

6.0 Python Library and Module Requirements

The scripts are designed to work with Python 3.x and depend on several scientific Python libraries to handle data processing, interpolation, file input/output, and time handling. Table 6-1 summarizes the key libraries and modules used in the script along with installation guidance. Core libraries include "numpy" and "xarray" for working with arrays and labeled multi-dimensional datasets, "netCDF4" for reading and writing netCDF files, and "scipy" for linear interpolation routines. The "datetime" and "os" modules support time-based operations and file system interactions. All required packages should be installed in the Python environment before running the script. The table below provides a brief description of each library/module along with the recommended installation command:

Table 6-1. Python libraries and modules required for running the tool.

Library/Module	Description	Installation Command
numpy	Provides support for numerical computations and array operations.	pip install numpy
pandas	Provides data structures and analysis tools for tabular data.	pip install pandas
xarray	Handles multi-dimensional labeled datasets, ideal for climate/model data.	pip install xarray
netCDF4	Enables reading and writing netCDF files using the Dataset interface.	pip install netCDF4
scipy	Used for interpolation and scientific computing (interp1d function).	pip install scipy
datetime	Standard library module for manipulating dates and times.	(Standard library)
pytz	Supports timezone conversions.	pip install pytz
os	Standard library module for interacting with the operating system.	(Standard library)
sys	Standard module to access system-specific parameters and functions.	(Standard library)
Collections	Provides specialized container datatypes like defaultdict.	(Standard library)
geopandas	Extends pandas to work with geospatial data	pip install geopandas
shapely	Performs geometric operations (e.g., intersections, buffers) on spatial objects	pip install shapely
pyproj	Performs cartographic projections and coordinate transformations	pip install pyproj
matplotlib	Used for creating static, animated, and interactive visualizations	pip install matplotlib
cartopy	Provides cartographic tools for geographic plotting with Matplotlib	pip install cartopy

7.0 Comparison of CAMx Simulated and Satellite-Derived HCHO Columns

A comparative analysis was performed for the six ecoregions described above during the ozone season (April through September) in 2019 and 2023 using two biogenic emission models: MEGAN and BEIS. The evaluation included several key diagnostic plots and metrics.

First, mean spatial distributions of HCHO VCDs were generated for both CAMx and TROPOMI, broken out by region, year, and biogenic emissions models. These maps highlight regional differences between modeled and observed HCHO levels across the study domain. In addition, bias maps (CAMx − TROPOMI) were created to show the spatial patterns of model differences and to help identify areas of systematic over- or underestimations.

Scatter density plots comparing CAMx and TROPOMI HCHO VCDs were used to assess overall agreement between the model and observations. Each plot includes key statistical metrics, such as normalized mean bias (NMB), normalized mean error (NME), the coefficient of determination (R^2), and the linear regression fit. Together, these metrics provide a quantitative assessment of model accuracy and the consistency of spatial patterns between CAMx and TROPOMI.

Finally, box and whisker plots summarize the distribution of daily HCHO values from CAMx and TROPOMI across all six ecoregions. These plots provide insight into how well CAMx captures the variability observed by TROPOMI, including the spread, central tendency, and presence of outliers.

7.1 Seasonal Mean and Bias Maps

Figure 7-1 and Figure 7-2 show the mean HCHO VCDs averaged over April-September 2019 for the CONUS 12 km and Texas 4 km domains, respectively. CAMx generally overestimates HCHO VCDs, particularly in the southeastern U.S., although simulations using BEIS biogenic emissions produce lower values compared to those using MEGAN. The simulated HCHO VCDs appear smoother than the satellite retrievals, which is expected given the inherent noise in HCHO satellite data. The noise contributes to the pixelated appearance of the TROPOMI VCDs, even after multi-month averaging. Despite these differences, CAMx captures key spatial features observed in satellite data, such as elevated HCHO over the southeastern U.S., localized enhancements in parts of California, and high concentrations over urban areas.

Satellite-derived tropospheric HCHO columns are strongly influenced by the vertical HCHO profiles, and the AMF used in the retrieval process. An averaging kernel describes how much the retrieved value at each atmospheric level reflects the true state versus the prior assumptions used in the retrieval process. To enable a consistent comparison between modeled and observed HCHO column densities, both the averaging kernel and the AMF must be accounted for in the analysis (Boersma et al., 2004; Lamsal et al., 2010). The impact of applying the averaging kernel correction to TROPOMI data is evident in subtle changes to the spatial patterns of TROPOMI fields in Figure 7-1 and Figure 7-2. The spatial patterns of the corrected TROPOMI columns differ depending on which biogenic emissions inventory is used in the CAMx simulations. These differences highlight the influence of biogenic VOC emissions on the vertical distribution of HCHO, which in turn affects how the averaging kernel correction modifies the observed TROPOMI columns.

At 4 km resolution, CAMx shows more distinct pockets of elevated HCHO column densities across Texas compared to the coarser 12 km grid. The finer resolution captures localized spatial variability, improving the representation of regional patterns. Mean maps for 2023, along with summary statistics and bias plots for all six ecoregions are provided in Appendix A.

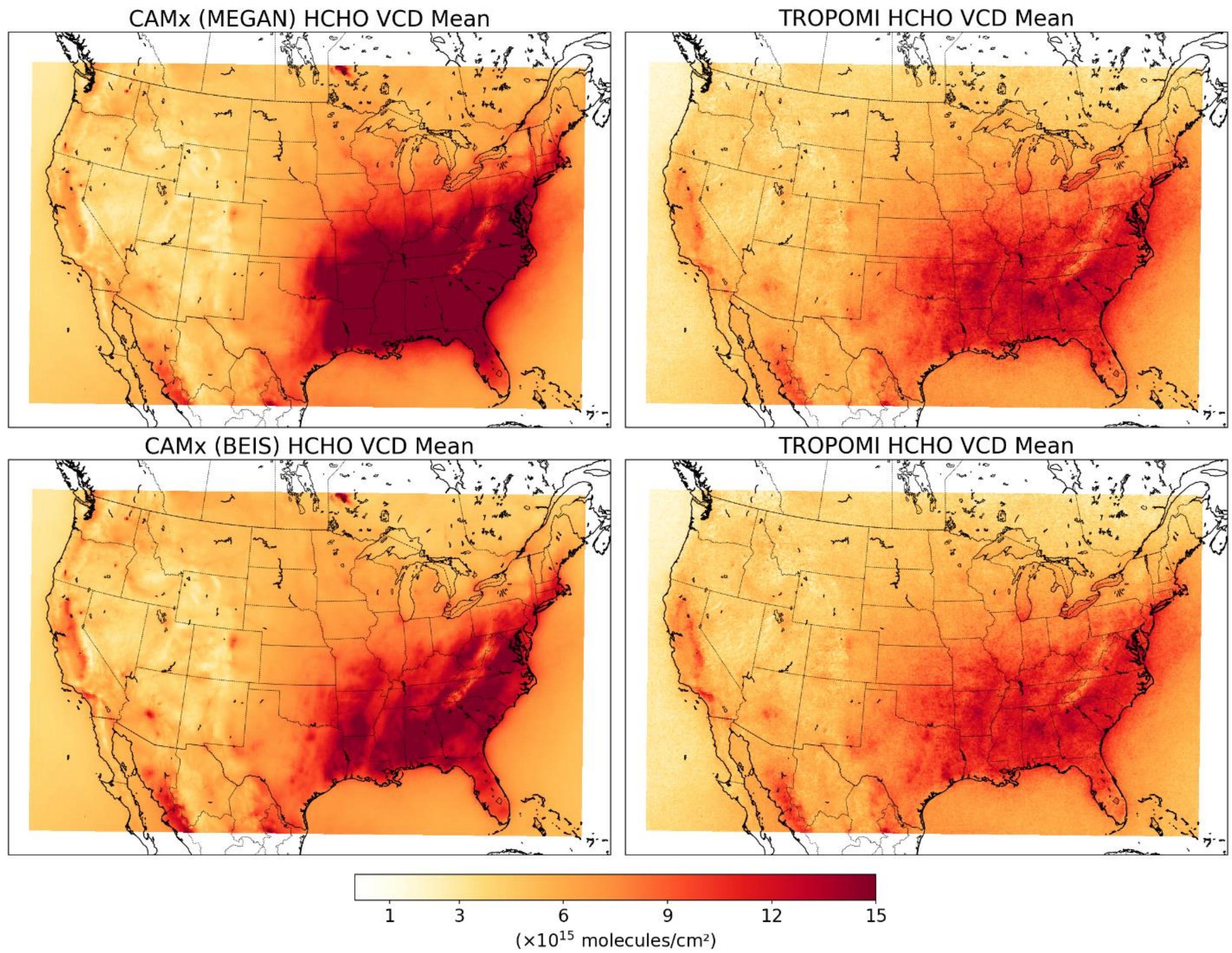

Figure 7-1. Mean HCHO VCDs for April–September 2019 from CAMx simulations and TROPOMI observations over the CONUS 12 km domain. Results are shown for CAMx simulations using two biogenic emissions models: MEGAN and BEIS. Differences in TROPOMI fields reflect the impacts from applying model-based averaging kernel corrections.

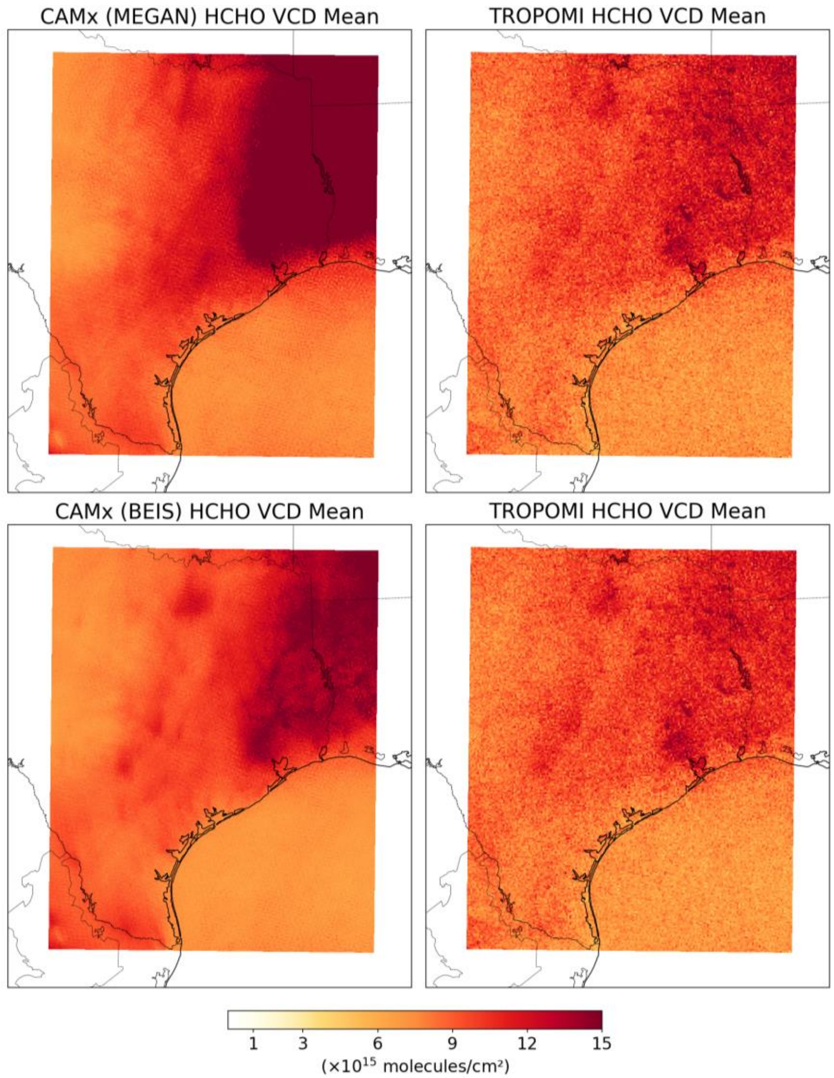

Figure 7-2. Mean HCHO VCDs for April–September 2019 from CAMx simulations and TROPOMI observations over the Texas 4 km domain. Results are shown for CAMx simulations using two biogenic emissions models: MEGAN and BEIS. Differences in TROPOMI fields reflect the impacts from applying model-based averaging kernel corrections.

7.2 Scatter Plots

Scatter plots comparing CAMx and TROPOMI HCHO VCDs were developed to evaluate model performance across the six ecoregions for April–September 2019 and 2023. Each scatter plot includes the linear fit equation, coefficient of determination (R^2), normalized mean bias (NMB), normalized mean error (NME), and mean bias (MB). Summary statistics for each ecoregion and year are provided in Table 7-1, with the corresponding scatter plots included in the Appendix B.

CAMx simulations using BEIS show a varied pattern of underestimations and overestimations of HCHO VCDs relative to TROPOMI, as reflected by the range of negative and positive NMB and MB values across ecoregions. In contrast, simulations with MEGAN consistently produced positive NMB values, pointing to a systematic overestimation. NMB values range from -5% to 24% for BEIS and from 6% to 56% for MEGAN, suggesting that BEIS-based simulations more closely align with satellite observations. Among the ecoregions, the Ozarks, Southeastern U.S. Pine Forest, and South Texas Agricultural Region show the highest R^2 values (~0.57–0.78), indicating relatively strong spatial correlation between CAMx and TROPOMI. In contrast, Central Texas, East Texas Forest, and the Boreal Forest (Ontario) show the lowest R^2 values (~0.2 or less in 2019), reflecting weaker agreement, though slight improvements are seen in 2023. Across the board, R^2 values tend to be slightly higher in simulations using BEIS compared to those using MEGAN.

Table 7-1. Statistical summary comparing CAMx and TROPOMI HCHO VCDs across six ecoregions for April–September 2019 and 2023. Metrics include NMB, NME, MB (×10^{15} molecules/cm^2), and R^2. Results are shown for CAMx simulations using MEGAN and BEIS biogenic emission inventories.

Ecoregions	NMB (%) MEGAN	NMB (%) BEIS	NME (%) MEGAN	NME (%) BEIS	MB (×10^{15} molec/cm^2) MEGAN	MB (×10^{15} molec/cm^2) BEIS	R^2 MEGAN	R^2 BEIS
2019								
Central Texas	7.5	-1.5	14.1	8.9	0.7	-0.1	0.131	0.169
East Texas Forest	48.5	16.6	48.6	18.5	5.6	1.8	0.181	0.188
Ozarks	36.8	5.8	36.8	7.5	4.2	0.6	0.666	0.684
Southeastern U.S. Pine Forest	55.9	19.3	55.9	19.8	6.6	2.2	0.572	0.568
South Texas Ag Region	20.9	22.1	21.0	22.1	1.6	1.7	0.579	0.646
Boreal Forest (Ontario)	38.2	24.3	38.4	25.5	1.6	1.0	0.006	0.010
2023								
Central Texas	21.8	-4.8	22.0	9.2	2.4	-0.5	0.272	0.291
East Texas Forest	31.8	4.1	32.1	9.7	4.2	0.5	0.278	0.290
Ozarks	17.5	-1.0	17.6	4.6	2.0	-0.1	0.709	0.763
Southeastern U.S. Pine Forest	33.1	17.9	33.4	18.4	3.7	2.0	0.515	0.474
South Texas Ag Region	29.3	23.4	29.3	23.4	2.2	1.7	0.779	0.644
Boreal Forest (Ontario)	5.5	13.1	10.6	14.6	0.3	0.7	0.098	0.124

The linear fit equations shown in the figures in Appendix B provide additional insight into CAMx performance. Slopes greater than 1 for some regions and close to or below 1 in others indicate that CAMx can overestimate and underestimate HCHO VCDs relative to TROPOMI, depending on the biogenic model used. Simulations using MEGAN generally produce steeper slopes than those with BEIS, indicating higher isoprene emission estimates in MEGAN. Regions with higher R^2 values tend to have slopes closer to 1 for CAMx BEIS simulations, suggesting better agreement with TROPOMI in both the variability and magnitude of HCHO VCDs. However, regions with weaker R^2 values (Central Texas, East Texas Forest, and Boreal Forest) display slopes from 0.50 to 0.91. For example, the East Texas Forest region shows MEGAN slopes between 0.78 and 0.91, and the Boreal Forest shows a notable improvement in slope, increasing from ~0.2 in 2019 to 0.6 in 2023.

These findings for the Southeastern U.S. Pine Forest are consistent with recent work by Hoque et al. (2024), who reported a correlation coefficient (r) of 0.92 between TROPOMI HCHO and CHASER model simulations across the eastern U.S. (32°-43° N, 71°-95° W), averaged over 2019 and 2020. While the spatial agreement in our study for the Southeastern U.S. Pine Forest is somewhat lower (R^2 = 0.5-0.6, or r ~ 0.7), the relatively higher R^2 and improved slope observed with BEIS biogenic emissions suggest that CAMx similarly captures key spatial patterns in regions influenced by biogenic emissions.

Goldberg et al. (2022) evaluated TROPOMI tropospheric HCHO columns using the HCHO TROPOMI products v1.1, which is known to have a low bias of approximately 25% (De Smedt et al., 2021). To address this, Goldberg et al. (2022) applied a bias correction factor of 1.25. Both the original and bias-corrected TROPOMI products are compared with CAMx HCHO columns using the MEGAN biogenic emissions model, with and without applying the averaging kernel. Their analysis found that CAMx underestimated HCHO in Central Texas by 25% relative to the bias-corrected TROPOMI product. Our study shows that CAMx simulations using BEIS generally have lower biases overall. However, despite these improvements, the low R^2 and slope values suggest that CAMx still struggles to capture the spatial variability of HCHO across the region.

7.3 Box and Whisker Plots

Figure 7-3 shows box and whisker plots of HCHO VCDs across the six ecoregions for April-September 2019 and 2023. Results from CAMx simulations using the MEGAN and BEIS biogenic emission models are compared against TROPOMI satellite observations, providing insights into model performance and variability.

Across most ecoregions, CAMx using BEIS shows median and mean values comparable to TROPOMI, except in the Southeastern U.S. Pine Forest and South Texas Agricultural Region, where CAMx means are about 30% higher. CAMx with MEGAN consistently simulates higher HCHO VCDs, but in 2023, the differences and spread are smaller than in 2019 across most regions, except in Central Texas, where the differences actually increased.

The interquartile range (IQR) of CAMx with MEGAN is broader than that of TROPOMI, indicating greater spatial variability in the CAMx simulations. The whiskers, which extend to the 5^{th} and 95^{th} percentiles, provide insights into the extremes of the data distribution. While both CAMx and TROPOMI display noticeable whiskers, those for CAMx with MEGAN are generally longer, reflecting a wider spread in the lower and upper percentiles, especially in regions like Central Texas and Southeastern U.S. Pine Forest. Outliers (points beyond the whiskers) are present in both datasets for 2019 and 2023. In 2019, CAMx shows sporadic extreme values that are not seen in the TROPOMI data in the Boreal Forest region. This discrepancy may be related to differences in fire emissions inventories: CAMx used the Fire Inventory from NCAR (FINN) for 2019, whereas the 2023 simulations used the Fire Emission Inventory (FEI) developed by Ramboll based on GFAS (Ramboll, 2024). An investigation into fire-related discrepancies in 2019 between CAMx and TROPOMI in the Boreal Forest region of

Ontario (Figure A8) revealed that although both datasets captured fire events, CAMx produced more exaggerated HCHO VCDs than TROPOMI.

Both CAMx and TROPOMI show higher HCHO VCDs with MEGAN compared to BEIS, consistent with MEGAN's higher isoprene emission estimates (Carlton & Baker, 2011; Sorochkina & Shah, 2024). Overall, there is no dramatic shift between 2019 and 2023 in most regions, though some (notably Central Texas and East Texas Forest) show slightly higher TROPOMI values in 2023.

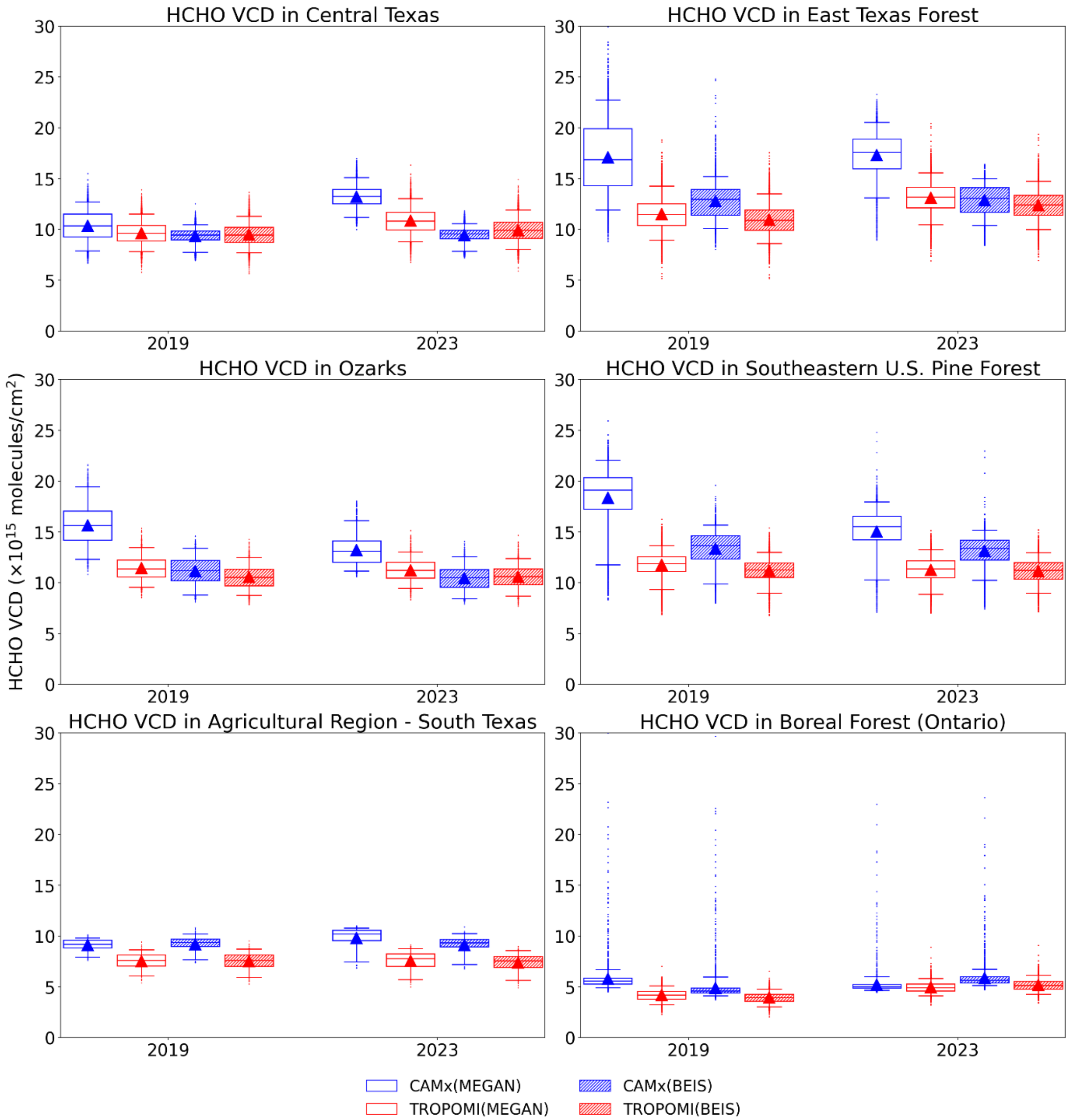

Figure 7-3. Box and whisker plots of HCHO VCDs across six ecoregions for April–September 2019 and 2023. CAMx results are shown in blue, TROPOMI in red, with BEIS-based simulations indicated by hatching. Boxes represent the IQR, the horizontal line indicates the median, and the triangle marks the mean. Whiskers extend to the 5th and 95th percentiles, with outliers shown as individual points.

8.0 Review of Recent Studies on Biogenic Emissions Evaluation Using Satellite HCHO

Several recent studies have used satellite HCHO data to evaluate biogenic emissions, demonstrating their value in assessing emission inventories and the performance of air quality models.

- Harkey et al. (2021) compared simulations of Environmental Protection Agency's (EPA) Community Multiscale Air Quality (CMAQ) air quality model with HCHO measurements from ground-based monitors and two different satellite instruments, OMI-SAO and OMI-QA4ECV, for the summers of 2011 and 2016 across the U.S. The CMAQ configurations differ in their biogenic emission inputs (BEIS 3.14, BEIS 3.61 or MEGAN 2.3) and chemical mechanism (CB05 or CB6r3). HCHO evaluation focused on the summer months (June–August) because biogenic precursors, and thus HCHO amounts, are greatest over the U.S. during these warm months. They found that while CMAQ captured general spatial patterns, it underestimated total column HCHO compared to satellite observations, particularly in the western U.S., and biogenic emissions played a significant role in model agreement with satellite data. However, model bias results differed between comparisons against the OMI-SAO data product (from NASA) or the OMI-QA4ECV data product (from ESA). The simulation using MEGAN produced larger HCHO columns than the simulations that used BEIS. All CMAQ configurations showed low daily correlations with OMI HCHO (ranging from $r = 0.13$ to 0.38). However, they demonstrated significantly higher monthly correlations (ranging from $r = 0.35$ to 0.73).

- De Smedt et al. (2021) conducted a comprehensive assessment of satellite-based HCHO retrievals by comparing data from the TROPOMI and OMI instruments. The study assessed the performance of TROPOMI, a newer instrument with finer spatial resolution, against its predecessor OMI. They analyzed differences in HCHO column retrievals between the two satellites and validated both datasets using observations from a global network of ground-based MAX-DOAS instruments. The authors found good agreement between OMI and TROPOMI HCHO data, especially when considering variations related to cloud correction methods, and highlight the improved precision of TROPOMI, particularly for low HCHO levels and at shorter timescales.

  The study highlighted key spatial and seasonal patterns across the U.S., including the strong summertime HCHO enhancements in the southeastern U.S. driven by biogenic emissions, as well as prominent HCHO signals from wildfire events in the western U.S. (e.g., summers of 2018 and 2020). Additionally, urban areas such as Houston, Dallas, and Los Angeles exhibited clear HCHO signals linked to anthropogenic VOC emissions. Overall, the improved resolution of TROPOMI enabled finer-scale detection of both natural and anthropogenic HCHO sources across the U.S.

- Goldberg et al. (2022) compared TROPOMI data for NO2 and HCHO columns with CAMx model simulations for April-September of 2019. Biogenic emissions were estimated using MEGAN version 3.1. The authors adjusted TROPOMI v1.1 HCHO columns upward by 25% to correct for a bias reported by De Smedt et al. (2021). Compared to these adjusted HCHO columns, CAMx showed an underestimate of HCHO in central and western Texas. In eastern Texas, the CAMx model's magnitude and spatial patterns for HCHO matched the bias-corrected TROPOMI product better. The model bias relative to the bias-corrected TROPOMI HCHO product was -7.9% in eastern Texas and -25.0% in central Texas.

- Hoque et al. (2024) evaluated global HCHO simulations from the CHASER V4.0 chemical transport model at 2.8° × 2.8° spatial resolution. The model's outputs for HCHO were compared against several independent observational datasets, including satellite data from TROPOMI (v1.1) and OMI, aircraft measurements from the Atmospheric Tomography (Atom) mission (Thompson et al., 2022), and ground-based MAX-DOAS observations. They analyzed the model's ability to reproduce the spatial distribution, seasonal variability, and vertical profiles of HCHO across different global regions. They also investigated the impact of uncertainties in anthropogenic and biogenic emissions on the model's performance and estimated the relative contributions of different emission sources to regional HCHO levels.

The CHASER model performance over North America, specifically the Eastern United States (E-USA) and Western United States (W-USA), was evaluated primarily using satellite observations from TROPOMI and OMI, and to some extent, airborne measurements from the ATom mission. Overall, CHASER showed strong agreement with TROPOMI HCHO columns across the U.S., with high spatial correlation and simulated magnitudes that closely matched observations. The overall bias for the US as a whole was reported as only 2%. However, the authors acknowledged that a 25% adjustment to the v1.1 TROPOMI HCHO columns may be needed to correct for known negative biases relative to ground-based observations, as noted by De Smedt et al. (2021).

For the E-USA, the simulated seasonality in spatial mean HCHO columns correlated strongly with observations (r=0.97) and the simulated amplitude of the seasonal modulation (74%) matched the observed amplitude (74%). In summer, the peak HCHO variability (around 1.2×10^{16} molec. cm^{-2}) coincided with the peak in model isoprene concentrations, indicating a strong biogenic contribution during summer. In winter, when anthropogenic VOC emissions are significant drivers of HCHO variability, the model showed strong agreement with HCHO-columns, suggesting the simulated contribution from these sources was reasonable.

For the W-USA, the simulated seasonal amplitude in spatial mean HCHO-columns (62%) was close to the observed (65%). However, model-satellite discrepancies were prominent in W-USA during summer and autumn. This suggests a potential model underestimation of biogenic HCHO levels in W-USA during these seasons, possibly linked to uncertainties in the biogenic emission inventory and the isoprene chemical mechanism.

Overall, their analysis highlighted that biogenic emissions are the dominant contributor to HCHO in the eastern U.S. during summer, aligning with peak isoprene emissions. These findings emphasize the critical importance of improving biogenic emissions and isoprene chemistry in models for accurately simulating HCHO, consistent with the recommendations from this study.

This study complements these efforts by providing a detailed regional analysis using CAMx with both MEGAN and BEIS biogenic emissions, evaluated against TROPOMI data. It is worth noting that the studies referenced above used TROPOMI v1.1, whereas this analysis used TROPOMI version 2.4, which includes improved retrieval algorithms. Our analysis confirmed that CAMx simulations using MEGAN overestimate HCHO VCDs relative to TROPOMI, while BEIS produced better agreement—consistent with findings from Harkey et al. (2021). Strong spatial correlations in the southeastern U.S., Ozarks, and southern Texas further corroborate the importance of biogenic precursors in driving HCHO spatial patterns, as highlighted by De Smedt et al. (2021) and Hoque et al. (2024).

9.0 Conclusion

9.1 Summary

Atmospheric oxidation of BVOCs is an important source of HCHO and the dominant source over regions with high biomass during the growing season (Fortems-Cheiney et al., 2012). Anthropogenic VOC emissions also are important HCHO precursors and may dominate when anthropogenic emissions are high and/or when biogenic emissions are low (Hoque et al., 2024). Wildfires and other biomass burning sources emit VOCs and therefore contribute to HCHO. Additionally, the slow oxidation of methane provides a global HCHO background or "floor" (Levy, 1972).

Satellite instruments such as TROPOMI and TEMPO can detect atmospheric column HCHO at a resolution of 3.5 km x 5.5 km for TROPOMI and 2.0 km x 4.5 km for TEMPO. These high-resolution data align well with TCEQ's air quality model grids, which have a resolution as fine as 4 km. While TROPOMI has been operational since 2017 and provides consistent coverage for TCEQ's modeling years of 2019 and 2023, TEMPO is relatively new, with quality-assured retrievals available only since August 2023. Therefore, TEMPO data was used for training purposes only.

Isoprene is often the dominant BVOC and has a substantial impact on air quality in Texas and other regions. However, the two widely used biogenic emission models, BEIS and MEGAN, produce widely varying emission estimates. This makes it difficult to determine the most appropriate biogenic model for air quality modeling. Satellite-detected HCHO columns provide a valuable proxy for evaluating these biogenic emission inventories.

This project evaluated biogenic emissions by comparing TROPOMI tropospheric HCHO VCDs with CAMx simulations that incorporated different meteorological data (2019 and 2023) and biogenic emissions inventories. For each year, one simulation used BEIS version 3.7 with the Biogenic Emissions Landuse Database (BELD) version 5, while the other used MEGAN version 3.2.

Ramboll developed Python tools to process satellite data and align those with CAMx outputs, ensuring spatial, temporal, and vertical consistency. The methodology included gridding Level 2 satellite data to CAMx domains, aligning timestamps, applying averaging kernel corrections, recalculating air mass factors, and interpolating CAMx vertical profiles to satellite pressure levels.

9.2 Key Findings

- CAMx tended to overestimate HCHO VCDs across most ecoregions in both years and with both biogenic emission inventories, though the overestimation was less pronounced when using BEIS.
- CAMx simulations using the BEIS biogenic emissions inventory generally agreed better with TROPOMI than those using MEGAN, particularly in high biogenic emission regions.
- Despite bias, CAMx showed strong spatial correlation (high R^2 values) in several regions, with slope indicating the model captures the magnitude of HCHO VCDs reasonably well.
- Using a finer 4 km spatial resolution within Texas improved the simulation of localized HCHO patterns.
- CAMx and TROPOMI showed the closest agreement in the Southeastern U.S. Pine Forest, the Ozarks, and the Southern Texas Agricultural region. In Central Texas, East Texas Forest, and the Boreal Forests of Ontario, the model tended to overestimate HCHO and correlations were weaker.

- Variations in fire emissions inventories also influenced modeled HCHO outputs, particularly in years with substantial wildfire activity.

9.3 Recommendations

The HCHO VCDs comparisons between CAMx simulations and TROPOMI satellite observations show that CAMx consistently overestimates HCHO VCDs, particularly in the regions dominated by biogenic emissions. To address these discrepancies and improve model performance, Ramboll recommends the following:

1. Compare the findings of this study to other published studies: Review findings alongside other research, such as Hoque et al. (2024) and Goldberg et al. (2022), which found smaller biases between their CHASER global model and TROPOMI HCHO columns over the Eastern U.S. It would be useful to understand similarities and differences in emission inventories, model chemistry and TROPOMI data version.
2. Analyze Isoprene Influence: Investigate the relationship between daily differences in CAMx and TROPOMI HCHO VCDs with CAMx isoprene (ISPD) outputs to assess the role of biogenic emissions and isoprene oxidation chemistry.
    a. A strong positive correlation would suggest that underpredicted isoprene emissions, incomplete oxidation pathways, or underestimated HCHO yields contribute to model underestimation.
    b. A weak or absent correlation may indicate that other factors—such as anthropogenic emissions, fire contributions, or chemical mechanism limitations—are more influential in driving the discrepancies.
3. Evaluate Contributions from other VOCs: Consider additional VOC precursors, such as acetylene and highly reactive VOCs, particularly in urban or industrial regions. While isoprene is the primary biogenic HCHO precursor, anthropogenic VOCs can significantly influence HCHO formation. Including these in the analysis may help understand CAMx–TROPOMI differences in areas with lower biogenic and higher urban influence.

10.0 References

Boersma, K. F., Eskes, H. J., & Brinksma, E. J. (2004). Error analysis for tropospheric NO2 retrieval from space. *Journal of Geophysical Research: Atmospheres*, *109*(4). https://doi.org/10.1029/2003jd003962

Carlton, A. G., & Baker, K. R. (2011). Photochemical Modeling of the Ozark Isoprene Volcano: MEGAN, BEIS, and Their Impacts on Air Quality Predictions. *Environmental Science & Technology*, *45*(10), 4438–4445. https://doi.org/10.1021/es200050x

De Smedt, I., Pinardi, G., Vigouroux, C., Compernolle, S., Bais, A., Benavent, N., et al. (2021). Comparative assessment of TROPOMI and OMI formaldehyde observations and validation against MAX-DOAS network column measurements. Atmospheric Chemistry and Physics, 21(16), 12561–12593. https://doi.org/10.5194/acp-21-12561-2021

Fortems-Cheiney, A., Chevallier, F., Pison, I., Bousquet, P., Saunois, M., Szopa, S., et al. (2012). The formaldehyde budget as seen by a global-scale multi-constraint and multi-species inversion system. *Atmospheric Chemistry and Physics*, *12*(15), 6699–6721. https://doi.org/10.5194/acp-12-6699-2012

Goldberg, D. L., Harkey, M., de Foy, B., Judd, L., Johnson, J., Yarwood, G., & Holloway, T. (2022). Evaluating NOx emissions and their effect on O3 production in Texas using TROPOMI NO2 and HCHO. *Atmospheric Chemistry and Physics*, *22*(16), 10875–10900. https://doi.org/10.5194/acp-22-10875-2022

Harkey, M., Holloway, T., Kim, E. J., Baker, K. R., & Henderson, B. (2021). Satellite Formaldehyde to Support Model Evaluation. Journal of Geophysical Research: Atmospheres, 126(4), 1–18. https://doi.org/10.1029/2020JD032881

Henderson, B., Szykman, J., Valin, L., Simon, H., Liu, X., Chance, K., et al. (2024). Validation and Early Applications of the Tropospheric Emission: Monitoring of POllution Nitrogen dioxide and Formaldehyde using Pandora and TROPOMI (pp. 1–20). Retrieved from https://gaftp.epa.gov/air/aqmg/bhenders/presentations/AGU/Henderson_AGU2024_TEMPO.pdf

Hoque, H. M. S., Sudo, K., Irie, H., He, Y., & Khan, M. F. (2024). Evaluating CHASER V4.0 global formaldehyde (HCHO) simulations using satellite, aircraft, and ground-based remote-sensing observations. *Geoscientific Model Development*, *17*(14), 5545–5571. https://doi.org/10.5194/gmd-17-5545-2024

Huijnen, V., Williams, J., Van Weele, M., Van Noije, T., Krol, M., Dentener, F., et al. (2010). The global chemistry transport model TM5: Description and evaluation of the tropospheric chemistry version 3.0. *Geoscientific Model Development*, *3*(2), 445–473. https://doi.org/10.5194/gmd-3-445-2010

Lamsal, L. N., Martin, R. V., Van Donkelaar, A., Celarier, E. A., Bucsela, E. J., Boersma, K. F., et al. (2010). Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: Insight into the seasonal variation of nitrogen oxides at northern midlatitudes. *Journal of Geophysical Research Atmospheres*, *115*(5), 1–15. https://doi.org/10.1029/2009JD013351

Levy, H. (1972). Photochemistry of the lower troposphere. *Planetary and Space Science*, *20*(6), 919–935. https://doi.org/https://doi.org/10.1016/0032-0633(72)90177-8

Nawaz, M. O., Johnson, J., Yarwood, G., De Foy, B., Judd, L., & Goldberg, D. L. (2024). An intercomparison of satellite, airborne, and ground-level observations with WRF-CAMx simulations of NO2columns over Houston, Texas, during the September 2021 TRACER-AQ campaign. *Atmospheric Chemistry and Physics*, *24*(11), 6719–6741. https://doi.org/10.5194/acp-24-6719-2024

Palmer, P. I., Jacob, D. J., Chance, K., Martin, R. V, Spurr, R. J. D., Kurosu, T. P., et al. (2001). Air mass factor formulation for spectroscopic measurements from satellites: Application to formaldehyde retrievals from the Global Ozone Monitoring Experiment. *Journal of Geophysical*

Research: Atmospheres, *106*(D13), 14539–14550. https://doi.org/https://doi.org/10.1029/2000JD900772

Ramboll. (2024). CAMx v7.3 User's Guide. Retrieved from https://www.camx.com/Files/CAMxUsersGuide_v7.30.pdf

Sorochkina, K., & Shah, T. (2024). Comparison of Biogenic Emission Estimates from BEIS and MEGAN Models for Texas and the Continental U.S.A. In *23 rd Annual CMAS Conference*. Retrieved from https://www.cmascenter.org/conference/2024/abstracts/CMAS2024_extabstract_KiraSorochkina_2024Nov19.pdf

World Meteorological Organization (WMO). (1957). Meteorology – A Three-Dimensional Science: Second Session of the Commission for Aerology (pp. 134–138). WMO Bulletin.

APPENDIX A

2023 Continental U.S. Comparison and Spatial Plots by Ecoregion for 2019 and 2023

Appendix A 2023 Continental U.S. Comparison and Spatial Plots by Ecoregion for 2019 and 2023.

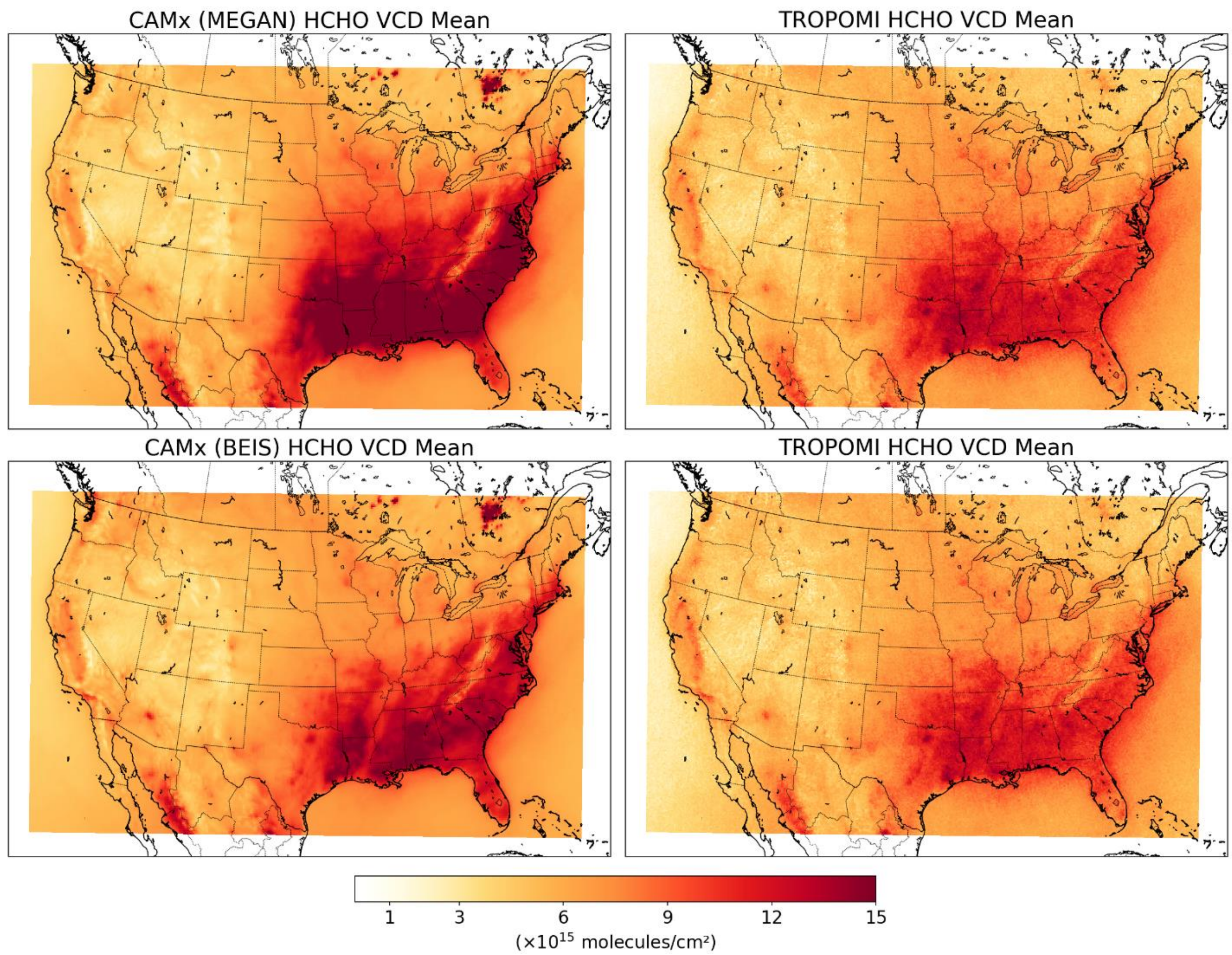

Figure A1. Mean HCHO VCDs for April–September 2023 from CAMx simulations and TROPOMI observations over the U.S. 12 km domain. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

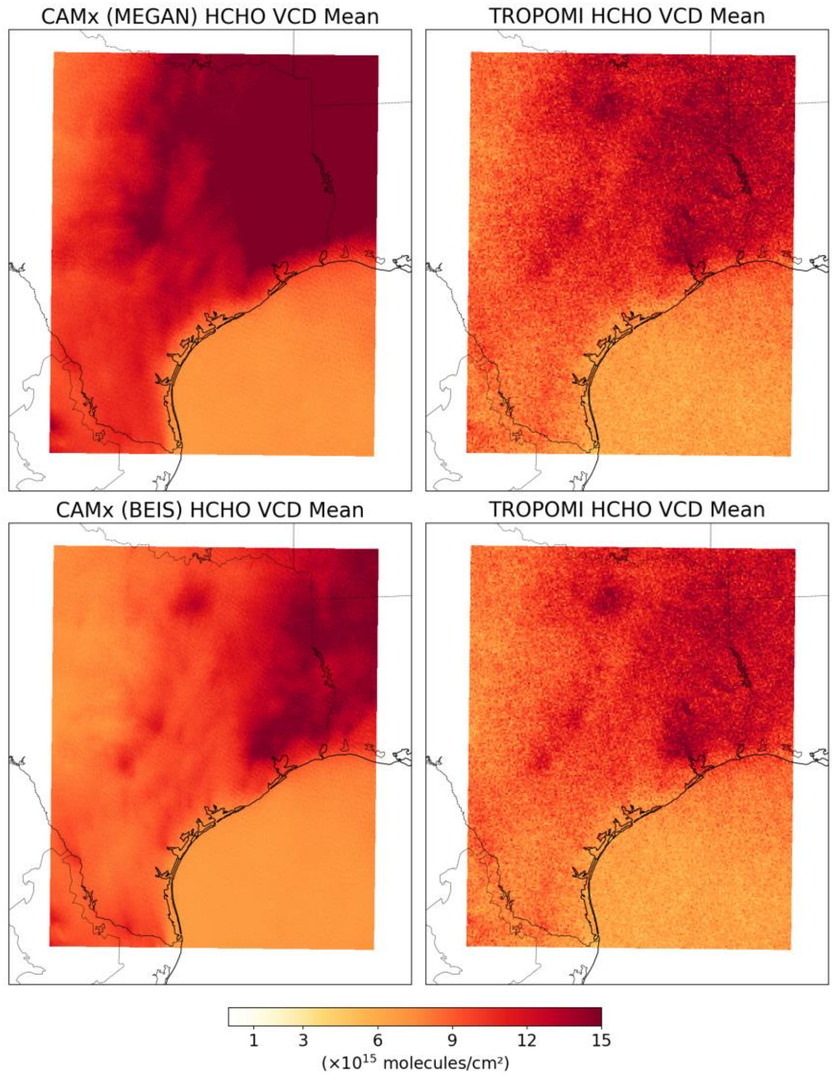

Figure A2. Mean HCHO VCDs for April–September 2023 from CAMx simulations and TROPOMI observations over the Texas 4 km domain. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

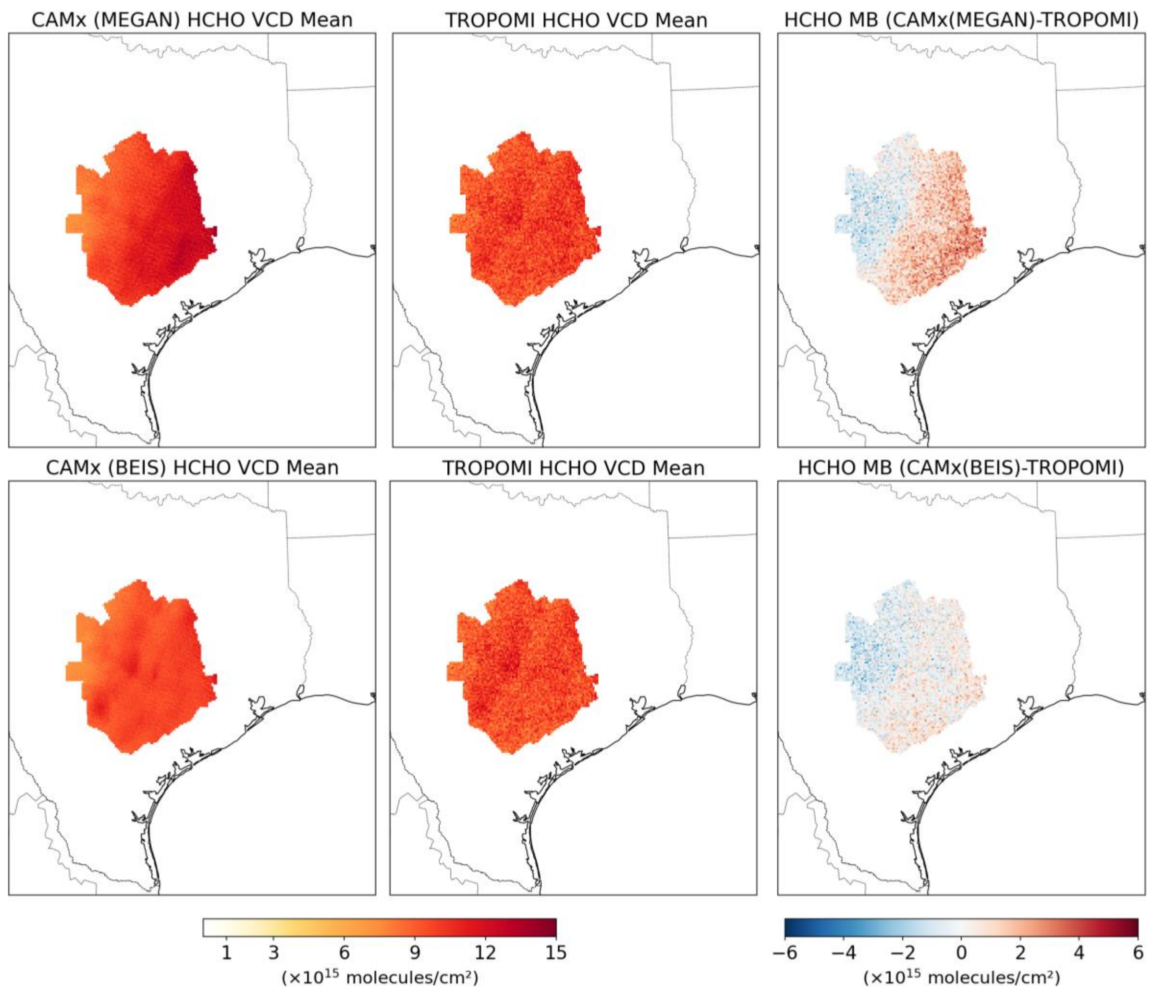

Figure A3. Mean HCHO VCDs and bias for April–September 2019 from CAMx simulations compared with TROPOMI observations in the Central Texas region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

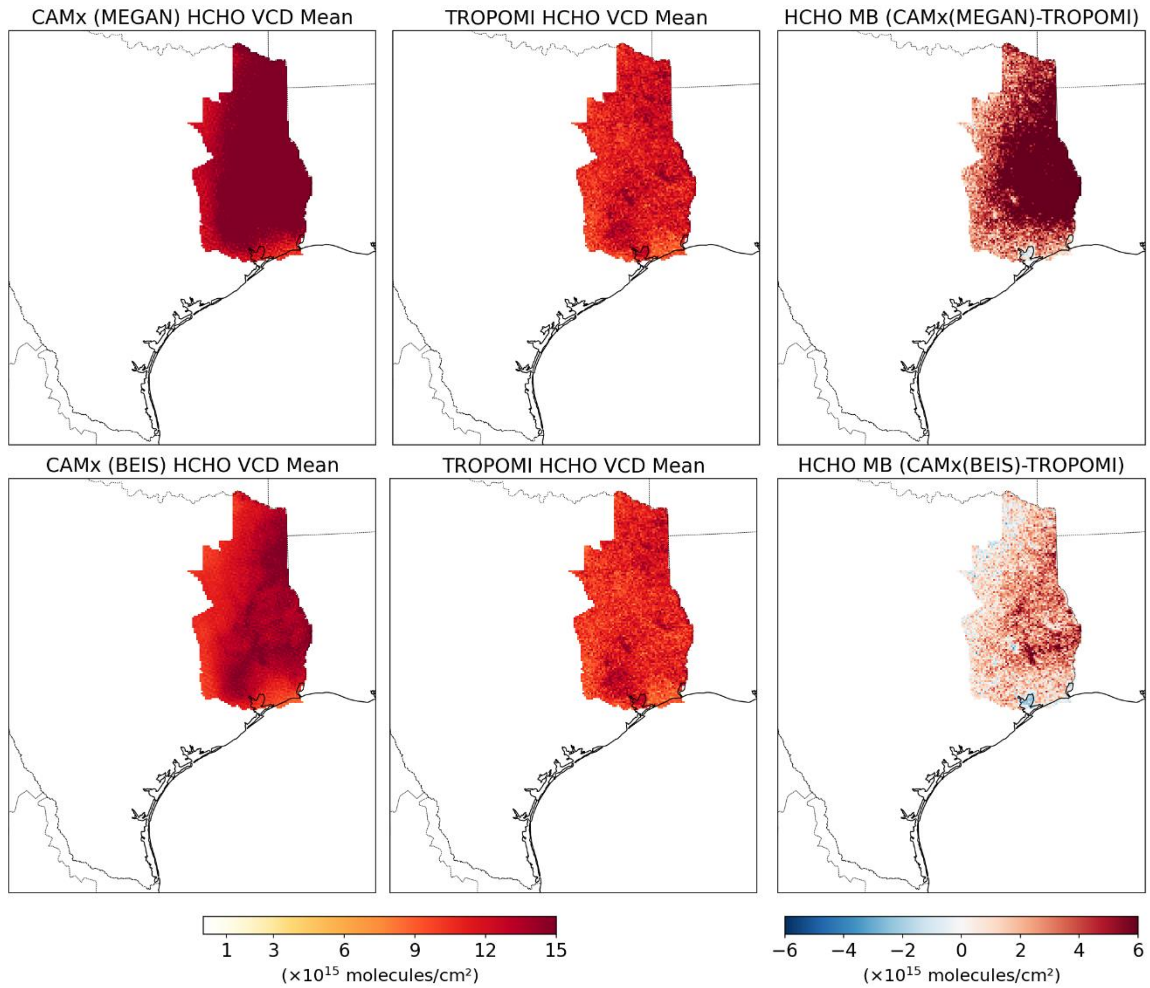

Figure A4. Mean HCHO VCDs and bias for April–September 2019 from CAMx simulations compared with TROPOMI observations in the East Texas Forest region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

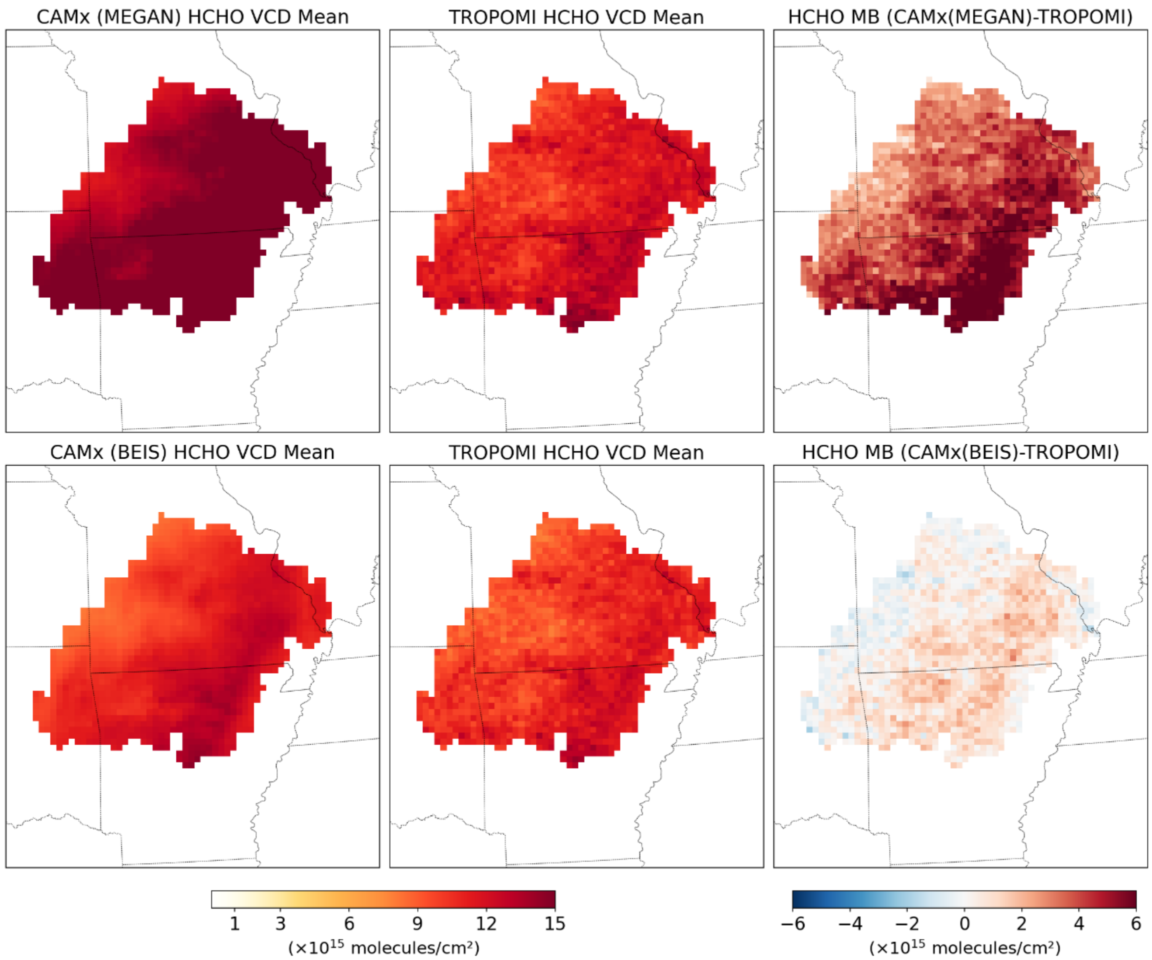

Figure A5. Mean HCHO VCDs and bias for April–September 2019 from CAMx simulations compared with TROPOMI observations in the Ozarks region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

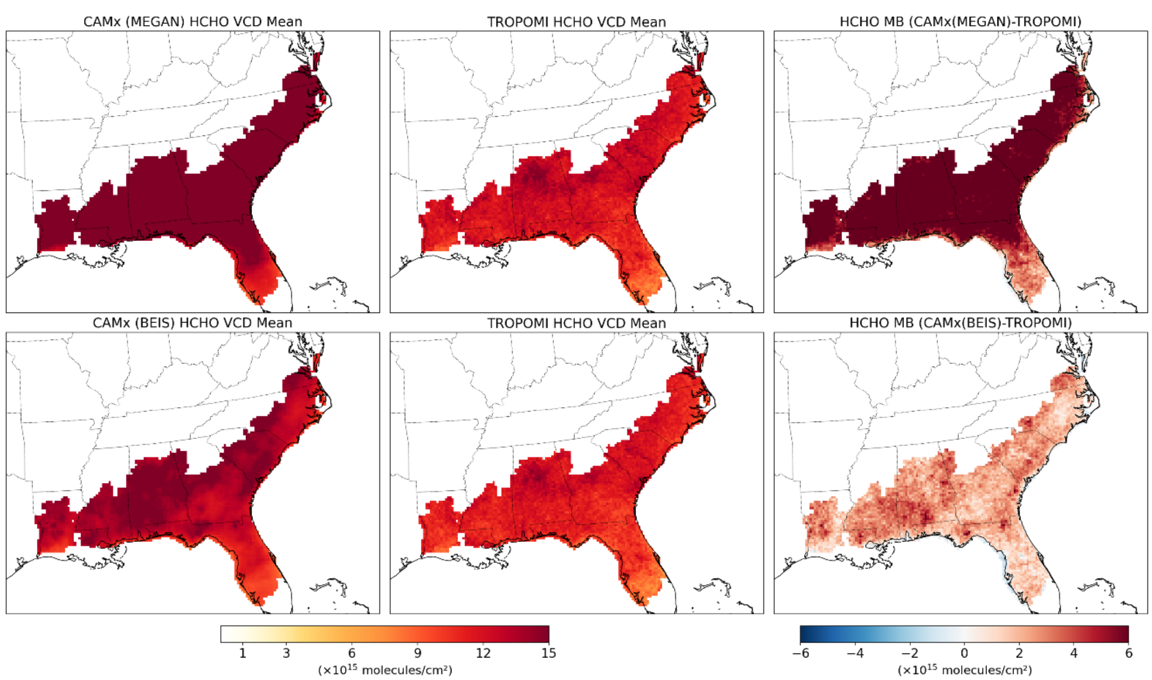

Figure A6. Mean HCHO VCDs and bias for April–September 2019 from CAMx simulations compared with TROPOMI observations in the Southeastern U.S. Pine Forest region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

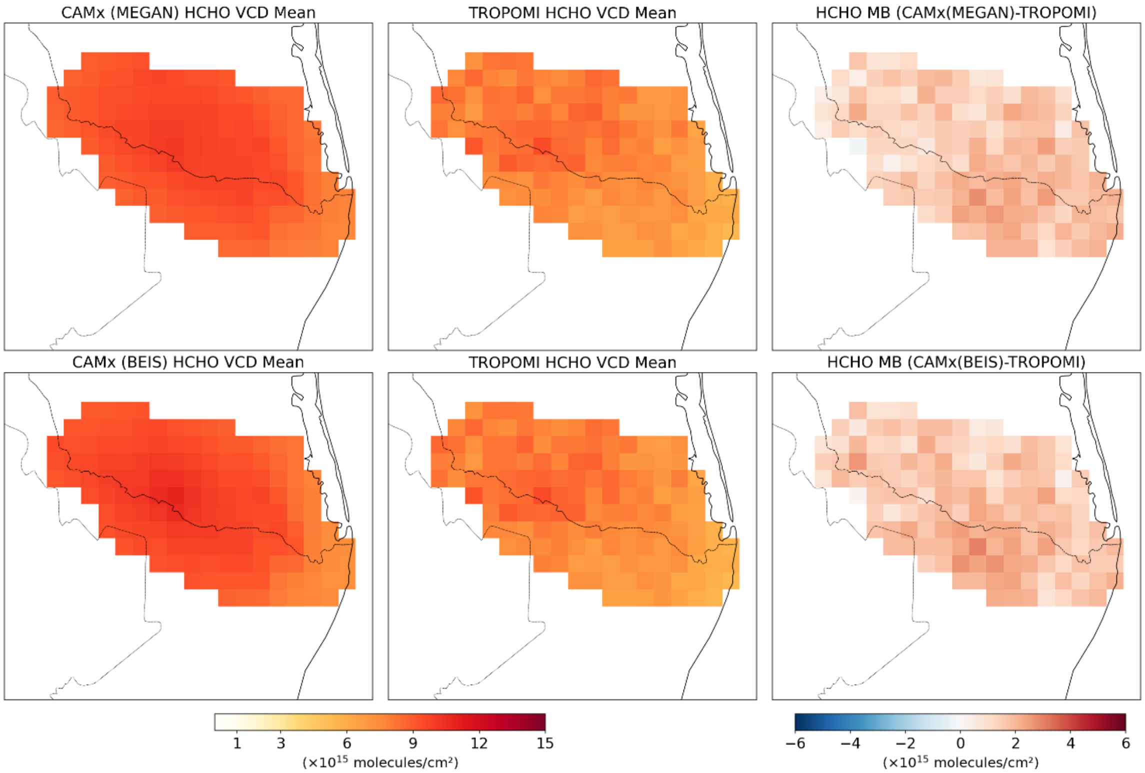

Figure A7. Mean HCHO VCDs and bias for April–September 2019 from CAMx simulations compared with TROPOMI observations in the South Texas Agricultural region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

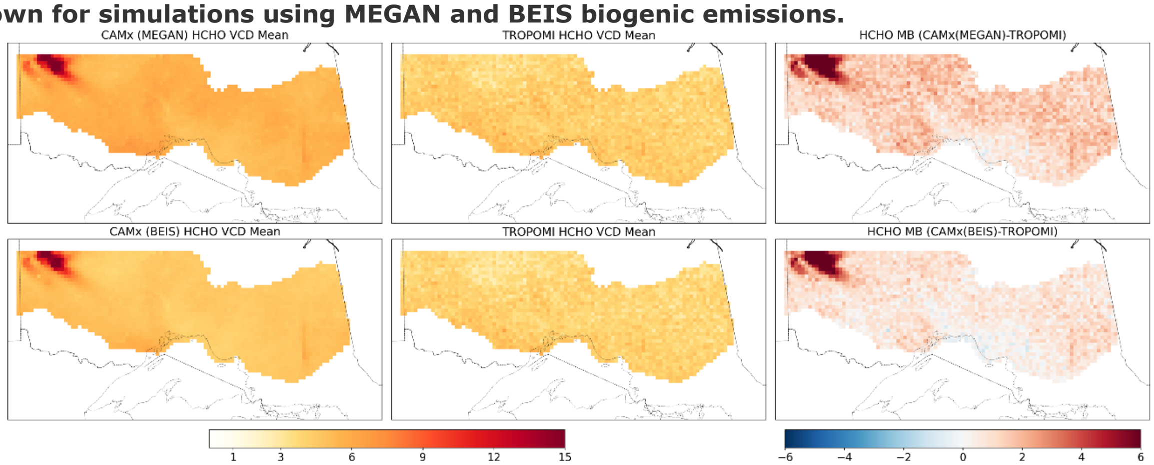

Figure A8. Mean HCHO VCDs and bias for April–September 2019 from CAMx simulations compared with TROPOMI observations in the Boreal Forest (Ontario) region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

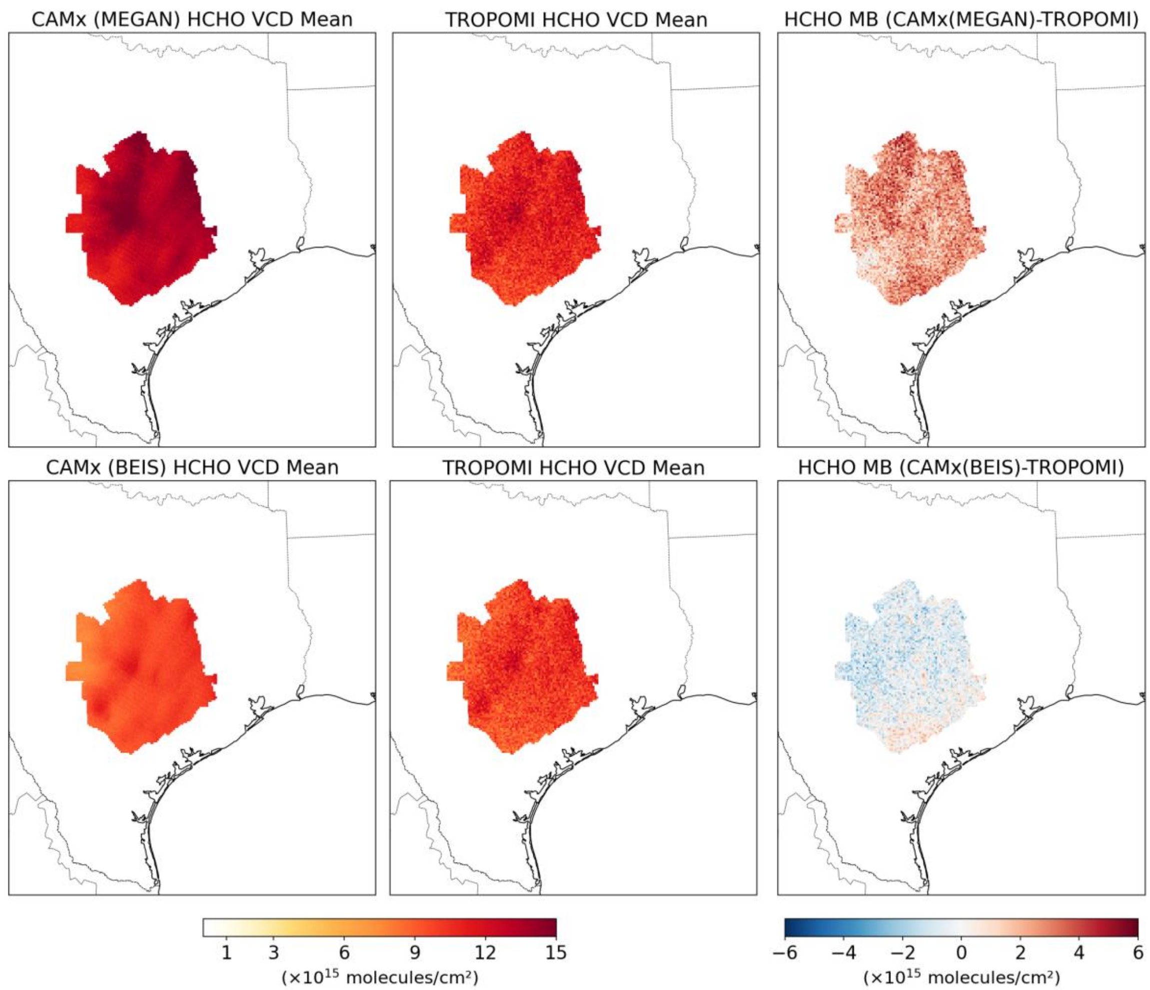

Figure A9. Mean HCHO VCDs and bias for April–September 2023 from CAMx simulations compared with TROPOMI observations in the Central Texas region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

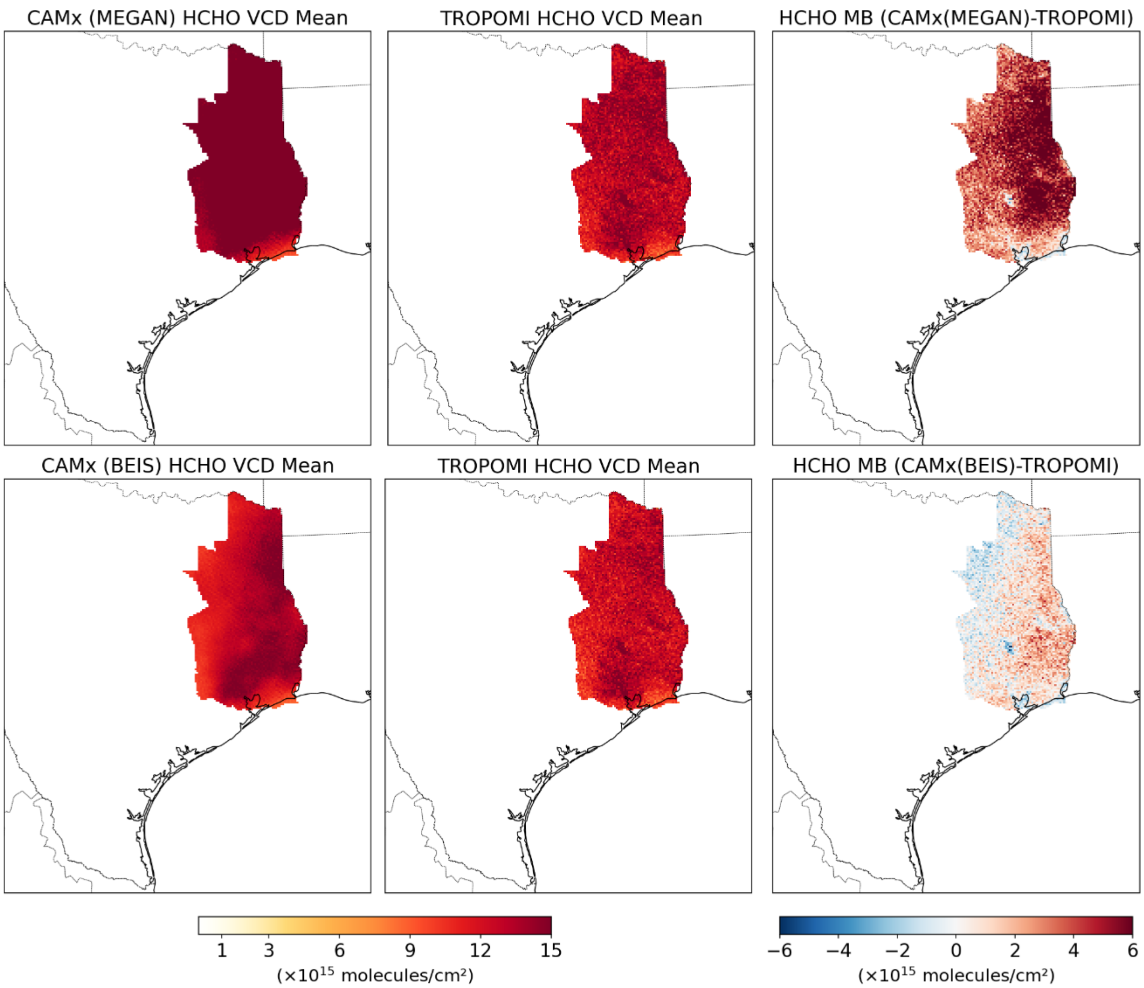

Figure A10. Mean HCHO VCDs and bias for April–September 2023 from CAMx simulations compared with TROPOMI observations in the East Texas Forest region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

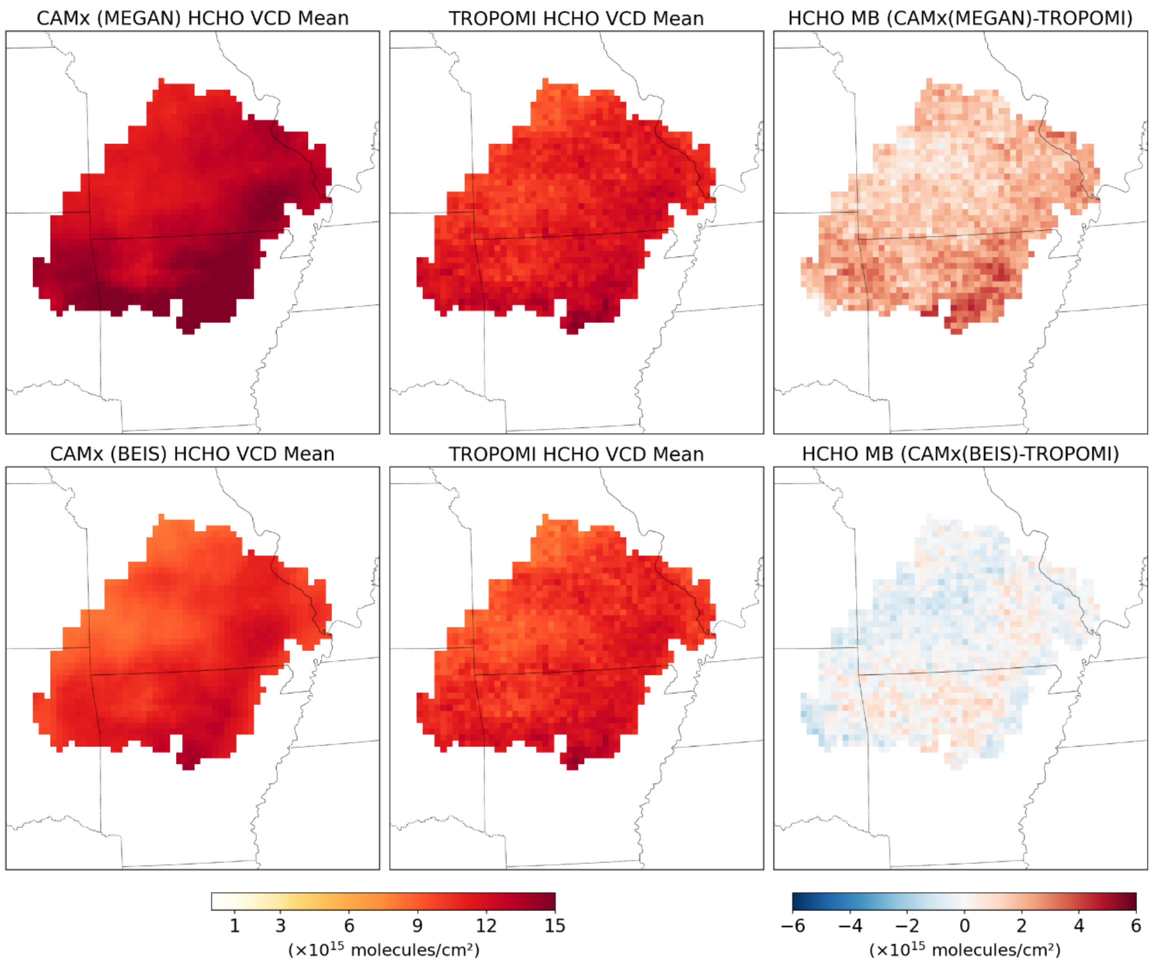

Figure A11. Mean HCHO VCDs and bias for April–September 2023 from CAMx simulations compared with TROPOMI observations in the Ozarks region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

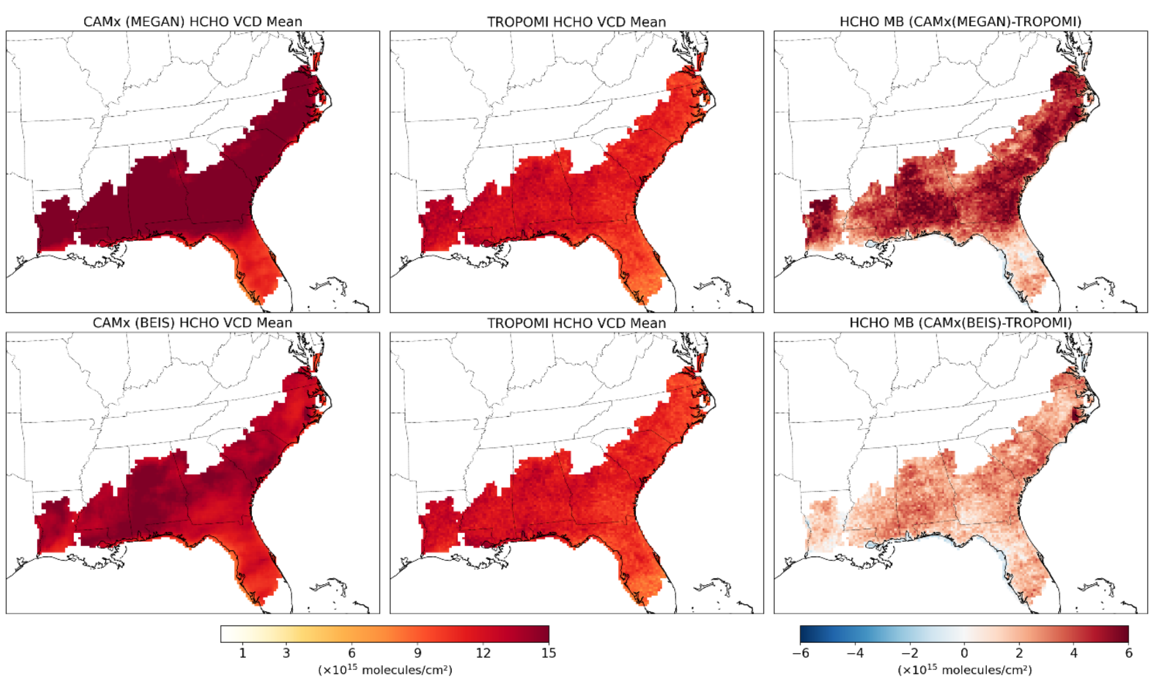

Figure A12. Mean HCHO VCDs and bias for April–September 2023 from CAMx simulations compared with TROPOMI observations in the Southeastern U.S. Pine Forest region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

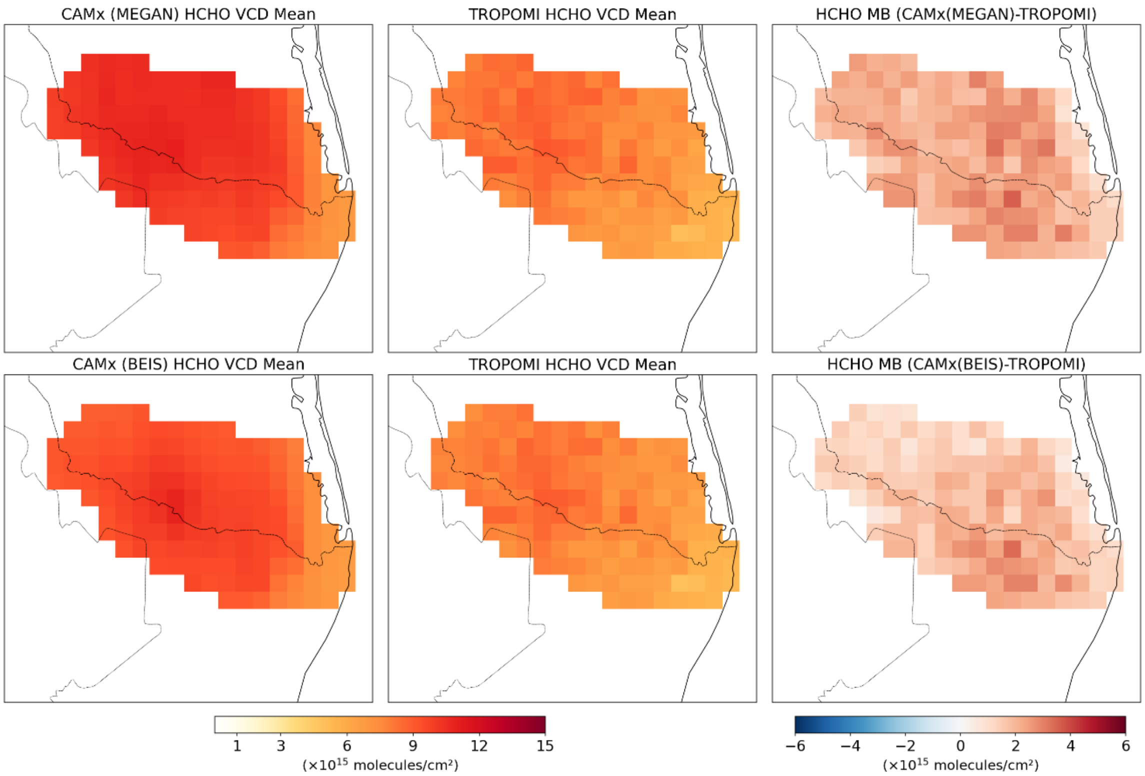

Figure A13. Mean HCHO VCDs and bias for April–September 2023 from CAMx simulations compared with TROPOMI observations in the South Texas Agricultural region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

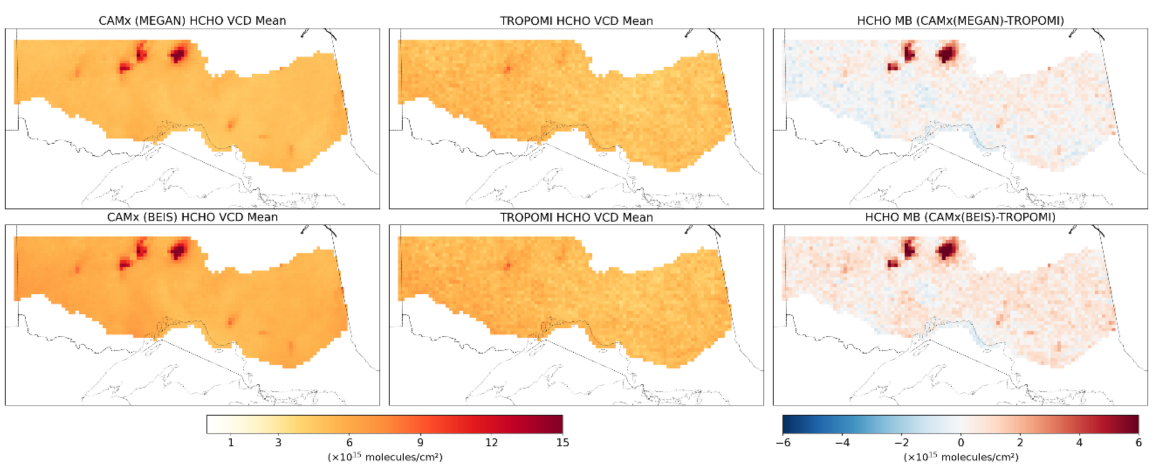

Figure A14. Mean HCHO VCDs and bias for April–September 2023 from CAMx simulations compared with TROPOMI observations in the Boreal Forest (Ontario) region. Results are shown for simulations using MEGAN and BEIS biogenic emissions.

APPENDIX B

Scatter Density Plots for Each Ecoregion Comparing CAMx and TROPOMI HCHO VCDs for 2019 and 2023

Appendix B Scatter Density Plots for Each Ecoregion Comparing CAMx and TROPOMI HCHO VCDs for 2019 and 2023

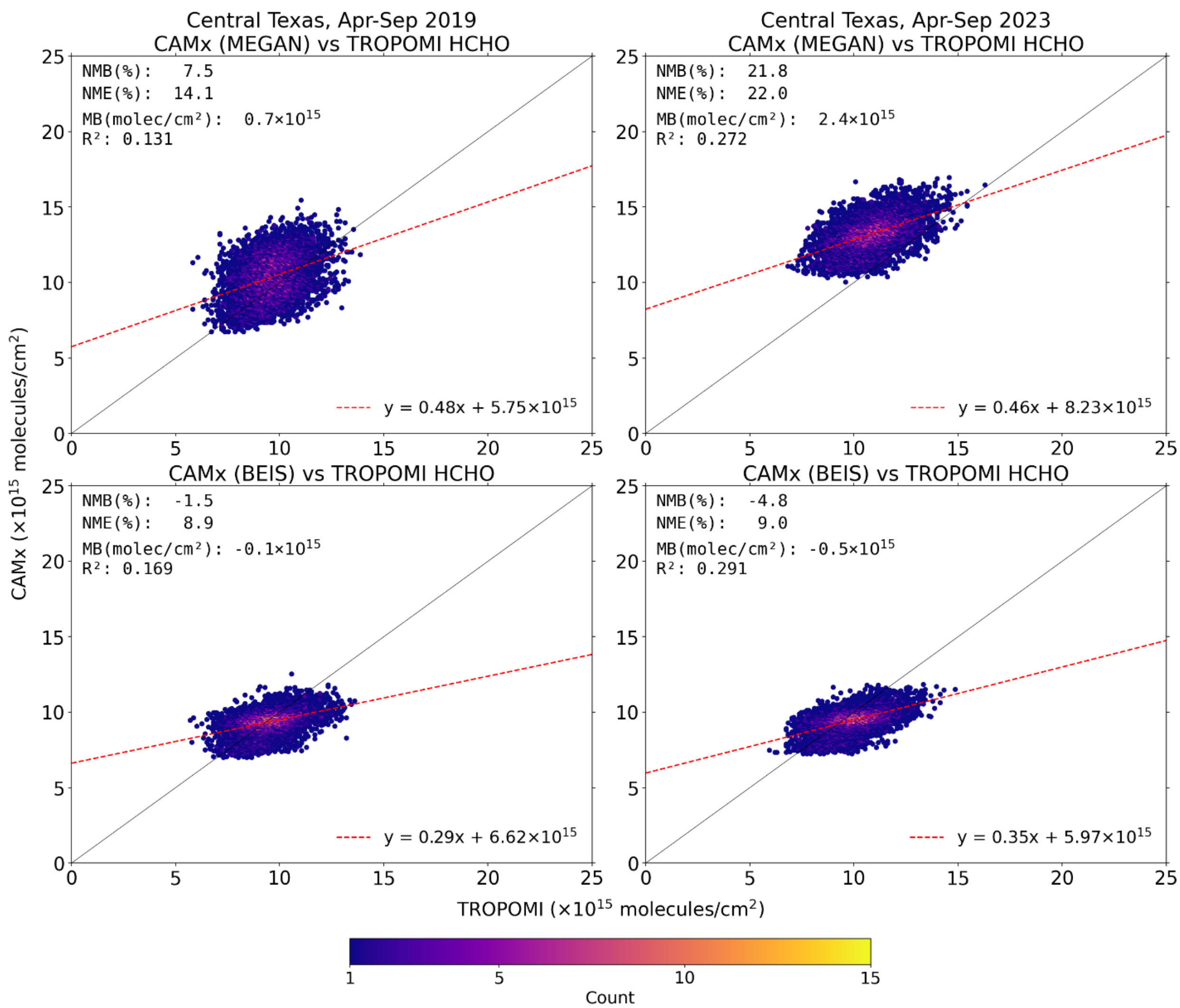

Figure B1. Scatter density plots comparing CAMx and TROPOMI HCHO VCDs for April–September in 2019 and 2023 in Central Texas. Results are shown separately for simulations using MEGAN and BEIS biogenic emissions. Each plot includes the linear fit equation, coefficient of determination (R^2), normalized mean bias (NMB), normalized mean error (NME), and mean bias (MB). The color shading represents point density, with warmer colors indicating areas of higher density. The 1:1 reference line is shown for reference.

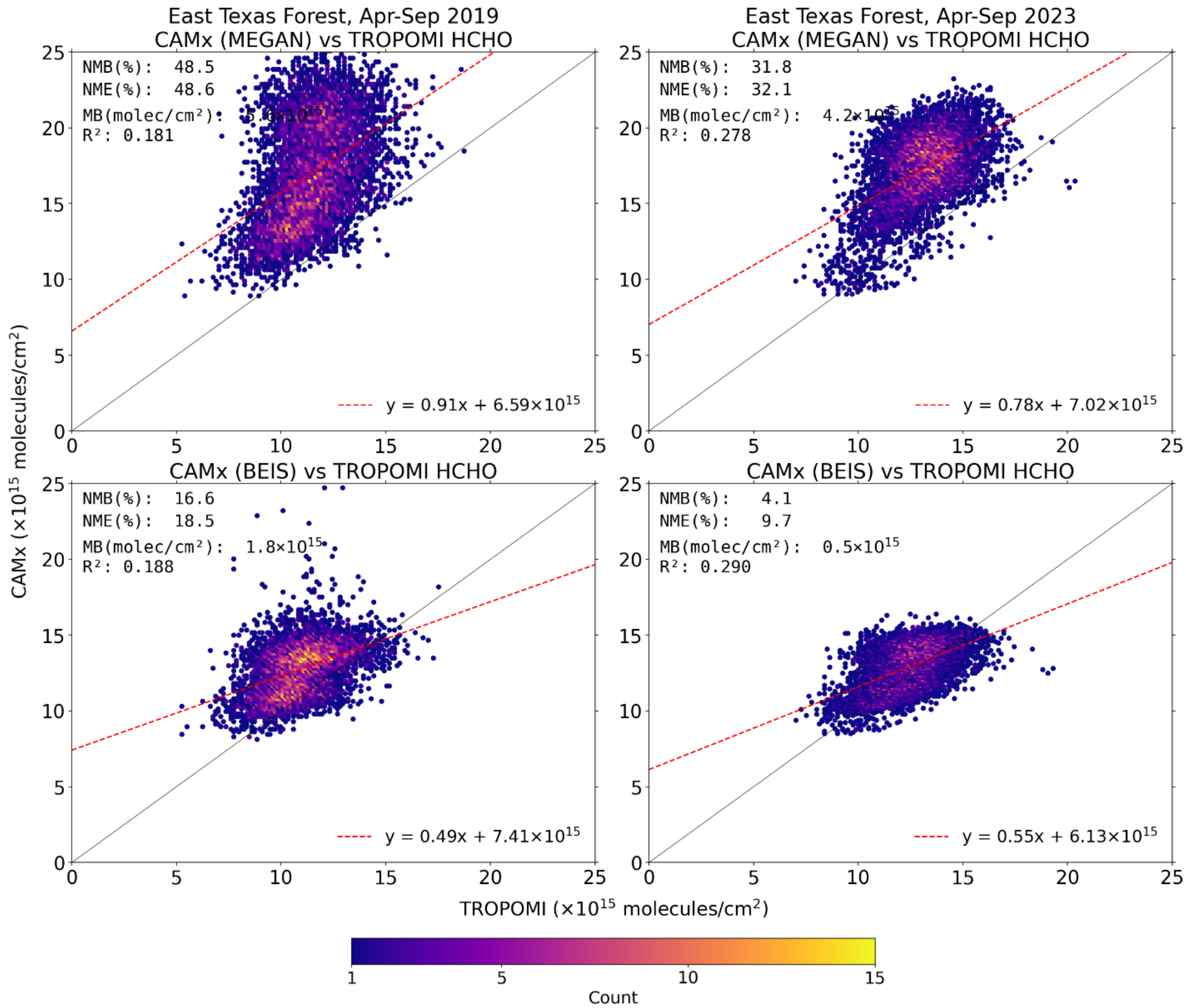

Figure B2. Scatter density plots comparing CAMx and TROPOMI HCHO VCDs for April–September in 2019 and 2023 in East Texas Forest. Results are shown separately for simulations using MEGAN and BEIS biogenic emissions. Each plot includes the linear fit equation, coefficient of determination (R^2), normalized mean bias (NMB), normalized mean error (NME), and mean bias (MB). The color shading represents point density, with warmer colors indicating areas of higher density. The 1:1 reference line is shown for reference.

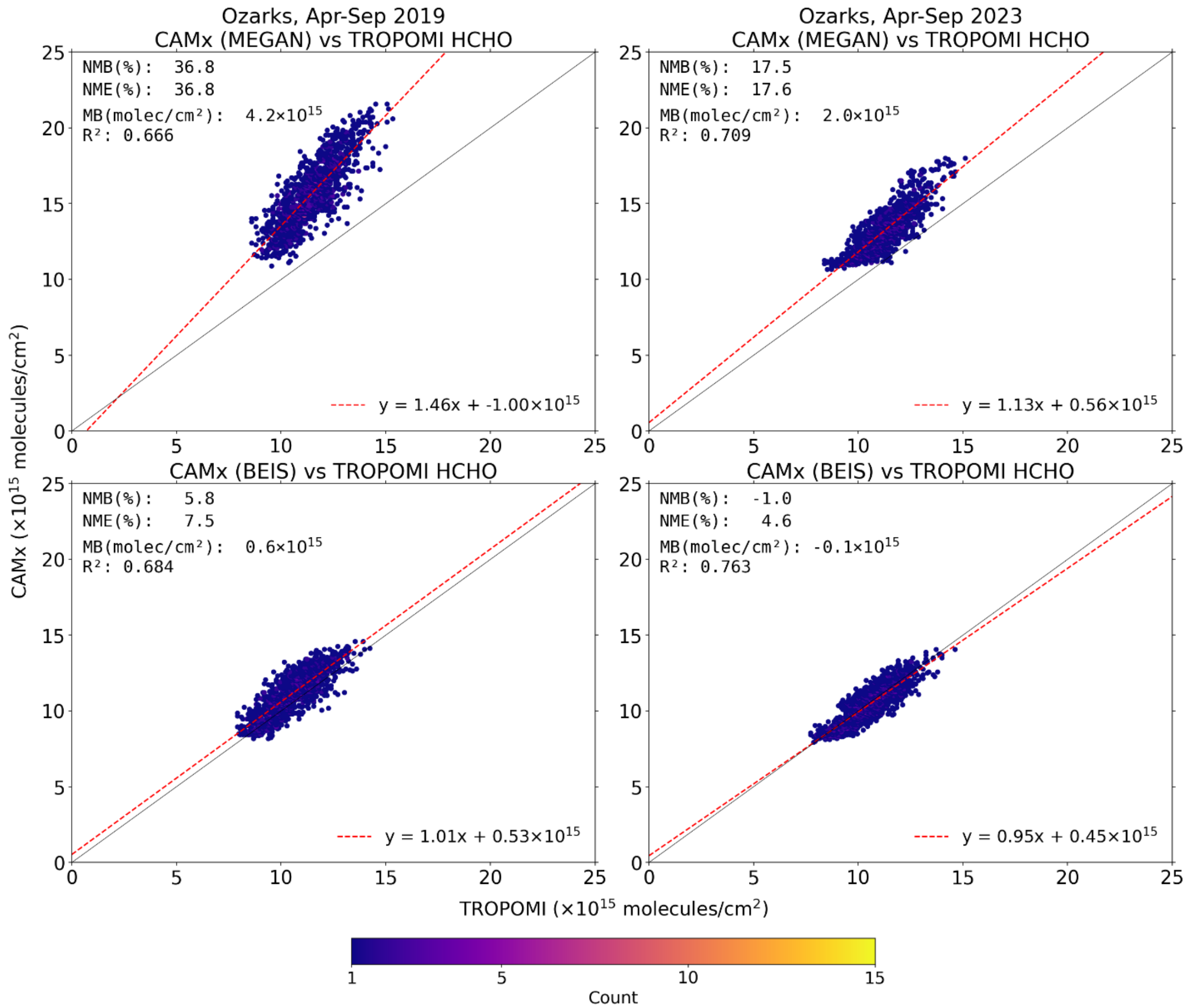

Figure B3. Scatter density plots comparing CAMx and TROPOMI HCHO VCDs for April–September in 2019 and 2023 in Ozarks. Results are shown separately for simulations using MEGAN and BEIS biogenic emissions. Each plot includes the linear fit equation, coefficient of determination (R^2), normalized mean bias (NMB), normalized mean error (NME), and mean bias (MB). The color shading represents point density, with warmer colors indicating areas of higher density. The 1:1 reference line is shown for reference.

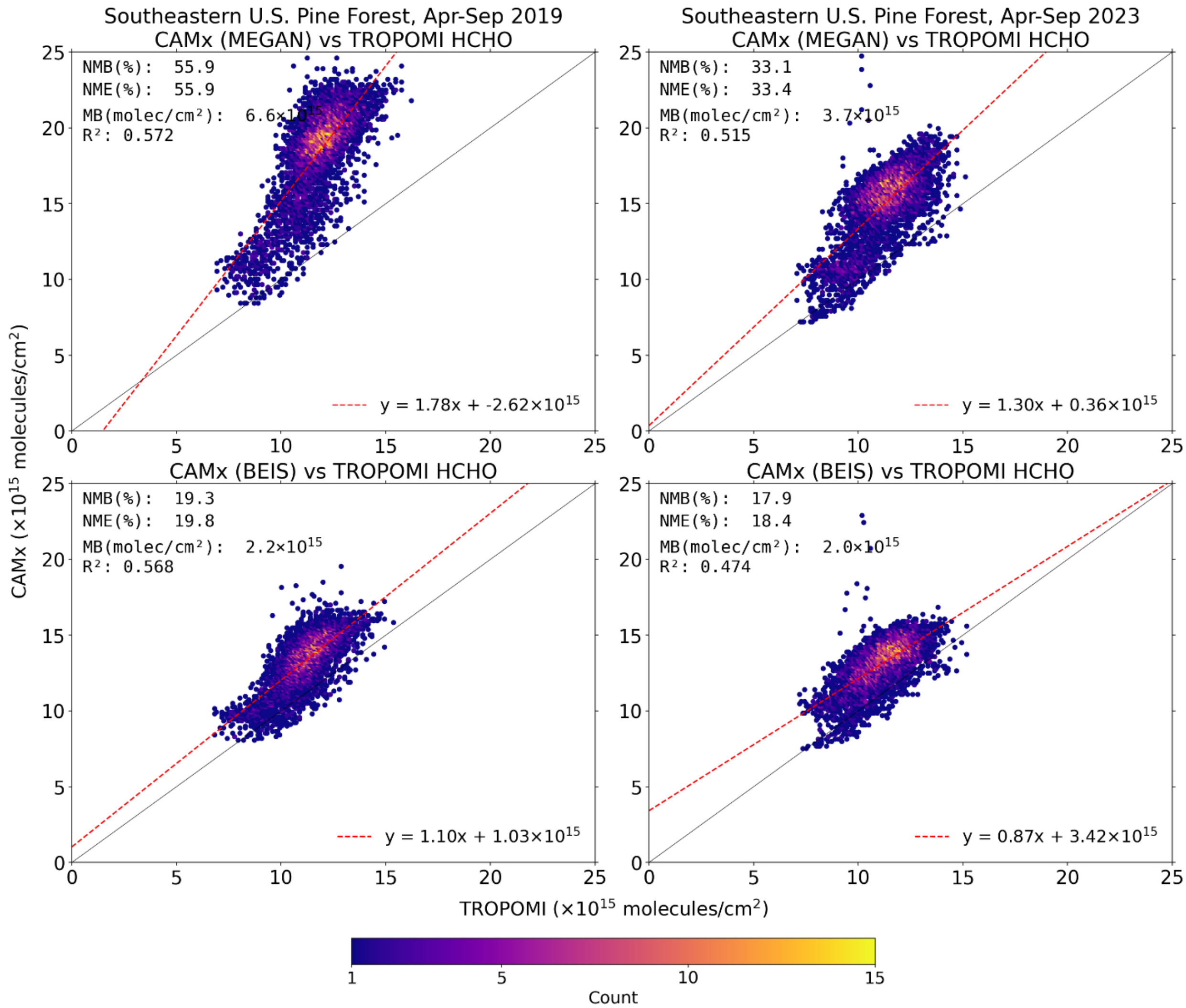

Figure B4. Scatter density plots comparing CAMx and TROPOMI HCHO VCDs for April–September in 2019 and 2023 in Southeastern U.S. Pine Forest. Results are shown separately for simulations using MEGAN and BEIS biogenic emissions. Each plot includes the linear fit equation, coefficient of determination (R^2), normalized mean bias (NMB), normalized mean error (NME), and mean bias (MB). The color shading represents point density, with warmer colors indicating areas of higher density. The 1:1 reference line is shown for reference.

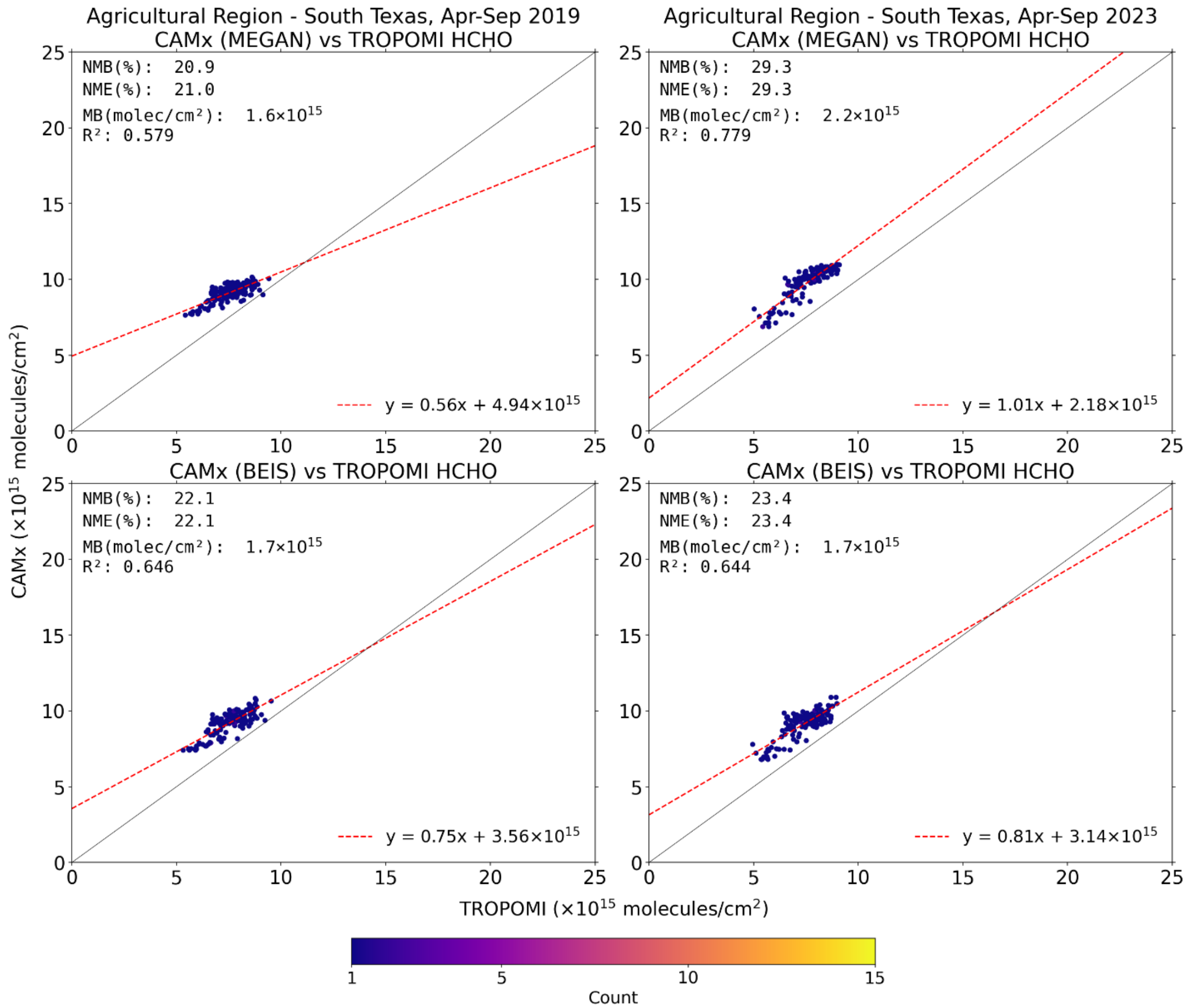

Figure B5. Scatter density plots comparing CAMx and TROPOMI HCHO VCDs for April–September in 2019 and 2023 in South Texas Agricultural Region. Results are shown separately for simulations using MEGAN and BEIS biogenic emissions. Each plot includes the linear fit equation, coefficient of determination (R^2), normalized mean bias (NMB), normalized mean error (NME), and mean bias (MB). The color shading represents point density, with warmer colors indicating areas of higher density. The 1:1 reference line is shown for reference.

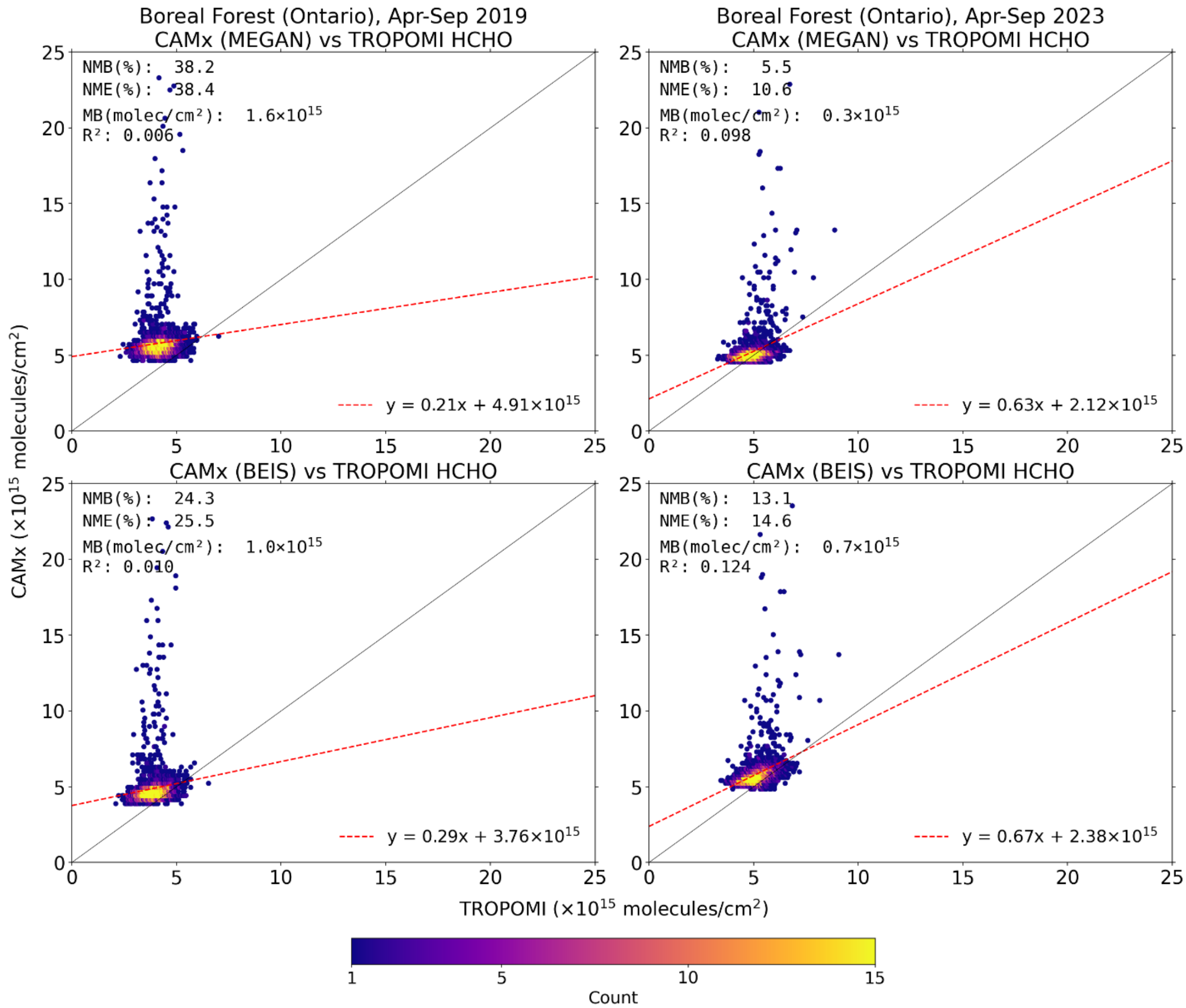

Figure B6. Scatter density plots comparing CAMx and TROPOMI HCHO VCDs for April–September in 2019 and 2023 in Boreal Forest (Ontario). Results are shown separately for simulations using MEGAN and BEIS biogenic emissions. Each plot includes the linear fit equation, coefficient of determination (R^2), normalized mean bias (NMB), normalized mean error (NME), and mean bias (MB). The color shading represents point density, with warmer colors indicating areas of higher density. The 1:1 reference line is shown for reference.