GIScience & Remote Sensing

ISSN: 1548-1603 (Print) 1943-7226 (Online) Journal homepage: www.tandfonline.com/journals/tgrs20

Reconstructing long-term (2003–2019) global high-resolution XCO_2: bridging observational gaps with machine learning

Soomin Hwang, Hyunyoung Choi, Yoojin Kang & Jungho Im

To cite this article: Soomin Hwang, Hyunyoung Choi, Yoojin Kang & Jungho Im (2026) Reconstructing long-term (2003–2019) global high-resolution XCO_2: bridging observational gaps with machine learning, GIScience & Remote Sensing, 63:1, 2627042, DOI: 10.1080/15481603.2026.2627042

To link to this article: https://doi.org/10.1080/15481603.2026.2627042

GISCIENCE & REMOTE SENSING
2026, VOL. 63, NO. 1, 2627042
https://doi.org/10.1080/15481603.2026.2627042

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Reconstructing long-term (2003–2019) global high-resolution XCO_2: bridging observational gaps with machine learning

Soomin Hwang[a,1], Hyunyoung Choi[a,1], Yoojin Kang[b] and Jungho Im[a,c,d]

[a]Department of Civil, Urban, Earth, and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea; [b]Department of Forestry, Environment and Systems, Kookmin University, Seoul, Republic of Korea; [c]Graduate School of Carbon Neutrality, UNIST, Ulsan, Republic of Korea; [d]Graduate School of Artificial Intelligence, UNIST, Ulsan, Republic of Korea

ABSTRACT

Atmospheric carbon dioxide (CO_2), a long-lived and well-mixed greenhouse gas, is a key driver of global warming. Accurate, long-term monitoring of its spatiotemporal variability is essential for understanding carbon dynamics. While the Orbiting Carbon Observatory-2 (OCO-2) satellite provides one of the most precise column-averaged CO_2 (XCO_2) measurements, its limited spatial coverage and short record since 2014 constrain long-term global analysis. Many studies thus highly rely on chemical transport models (e.g. Copernicus Atmosphere Monitoring Service (CAMS) and CarbonTracker) when applying machine learning (ML) approaches. However, their coarse resolutions often lead to spatial smoothing. In this context, we present a novel ML-based framework based on residual learning with the Light Gradient Boosting Machine (LGBM) to reconstruct global, gap-free XCO_2 at 0.1° resolution for the period 2003–2019. By explicitly modeling the residuals between high precision OCO-2 observations and the coarse resolution CAMS-EGG4 reanalysis, the proposed framework mitigates spatial smoothing effects and enables the extension of XCO_2 estimates beyond the temporal coverage of the OCO-2 mission. The resulting product was strictly validated through internal cross-validation (random, spatial, and temporal) and external in situ validation, showing strong agreement with OCO-2 satellite observations ($R^2 = 0.93$–0.96, RMSE = 0.80–1.11 ppm) and ground-based measurements ($R^2 = 0.98$, RMSE = 1.17 ppm), respectively. Compared to CAMS-EGG4, the LGBM-based XCO_2 product also outperforms by offering higher accuracy and resolving the spatial smoothing limitations caused by its coarse resolution. By bridging gaps in satellite data across space and time, this high-resolution XCO_2 product enhances applications in climate research, emission source attribution, and greenhouse gas policy assessment.

ARTICLE HISTORY
Received 27 November 2025
Accepted 1 February 2026

KEYWORDS
Carbon dioxide; OCO-2; machine learning; high resolution; global

1. Introduction

Recently, there has been a significant increase in the public's awareness of greenhouse gases (GHGs) due to climate change, which has led many countries to implement regulatory policies aimed at reducing emissions (Lee et al. 2023). Atmospheric carbon dioxide (CO_2) is one of the most important GHGs in the Earth's atmosphere (Hong et al. 2023). For decades, industrial development and the combustion of fossil fuels have been the primary drivers of anthropogenic CO_2 emissions, resulting in a long-term annual increase in atmospheric CO_2 concentrations (Friedlingstein et al. 2022). According to the latest annual report released by the National Oceanic and Atmospheric Administration (NOAA), the global average atmospheric CO_2 concentration reached a new record high of 422.8 parts per million (ppm) in 2024 (https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide, last accessed: 30 October 2025). This implies that atmospheric CO_2 is now 50% higher than it was before the Industrial Revolution, at about 280 ppm (He et al. 2023). Additionally, as a long-lived GHG, CO_2 intensifies the greenhouse effect to become more pronounced and contributes to the increasing frequency of extreme climate events (Jin et al. 2022).

CONTACT Jungho Im ersgis@unist.ac.kr
[1]The first two authors equally contributed to the paper.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15481603.2026.2627042.

Recognizing that terrestrial ecosystems are major sources of anthropogenic CO_2 emissions while oceans act as crucial carbon sinks (Jung et al. 2024; Zhang et al. 2025), comprehensive monitoring of both land and ocean is essential to capture the full dynamics of the global carbon cycle (Zhu et al. 2024). Therefore, sustained and accurate monitoring of atmospheric CO_2 across space and time is crucial for understanding its environmental impacts at the global scale.

XCO_2, the column-averaged dry-air mole fraction of CO_2, is the primary variable used for atmospheric CO_2 monitoring. Monitoring XCO_2 can be achieved through multiple approaches, including ground-based measurements, aircraft-based gas collection, chemical transport models (CTMs), and satellite observations. Ground-based monitoring involves direct measurements of XCO_2 concentrations at specific locations (Sha et al. 2019). Various organizations around the globe maintain ground networks that provide highly precise GHG measurements, with the Total Carbon Column Observing Network (TCCON) as a representative example. In addition, aircraft-based gas collection provides vertical profiles of atmospheric CO_2 and is widely used for inversion analyses (Pitt et al. 2022). However, both ground- and aircraft-based measurements are constrained by limited spatial distribution and irregular temporal intervals, making global monitoring challenging. CTM offers spatially and temporally continuous products by combining various observational inputs with atmospheric transport and chemistry models. The Copernicus Atmosphere Monitoring Service (CAMS) generates the EGG4 global XCO_2 reanalysis product by incorporating satellite retrievals into a chemical transport model (Inness et al. 2019). While offering global coverage, their coarse spatial resolution and time-dependent biases arising from the change in the assimilated satellite instrument constrain the detection of accurate XCO_2 variations (Agustí-Panareda et al. 2023). Lastly, satellite remote sensing enables high-resolution global monitoring of XCO_2 concentrations with broad spatial coverage and regular temporal frequency (Zhu et al. 2025). Satellite missions such as the Scanning Imaging Absorption SpectroMeter for Atmospheric ChartographY (SCIAMACHY), the Greenhouse Gases Observing Satellite (GOSAT) series, and the Orbiting Carbon Observatory (OCO) satellite series retrieve XCO_2 by measuring the absorption of sunlight at specific wavelengths reflected from the Earth's surface (Heymann et al. 2015). This technique provides consistent, large-scale observations, overcoming the limitations of ground-based stations and reanalysis data, and is broadly adopted in recent studies (Wang et al. 2023).

The OCO-2 XCO_2 product is widely known for its high accuracy when validated with ground-based TCCON data, outperforming earlier satellites such as GOSAT and SCIAMACHY and even its successor OCO-3 (Jin et al. 2022; Yang et al. 2025). Despite the highest accuracy, a common limitation of current satellite-based XCO_2 monitoring is the large observation gaps in both space and time. These gaps make it difficult to comprehensively observe the spatiotemporal variation of XCO_2, especially when analyzing long-term and large-scale events (Sheng et al. 2021). For instance, Cusworth et al. (2023) attempted to quantify annual XCO_2 emissions from power-plant emissions using OCO-series data, but the narrow spatial coverage and short observation period led to high-uncertainty results. Likewise, Guan et al. (2024b) tried to figure out the oceanic contribution to interannual XCO_2 variability with OCO-2, but it was challenging owing to satellite-based products' small magnitude of the ocean imprint. To address this issue, considerable research has been conducted to fill these gaps using geostatistical modeling, multi-source ensemble, and machine learning (ML) techniques (Jin et al. 2022; Zhang and Liu 2023). In particular, ML has recently proven its efficiency and feasibility in producing gap-filled outputs by leveraging the seamless characteristics of CTM-based data (He et al. 2022; Zhang and Liu 2023; Park, Lee, and Park 2025).

While ML-based efforts have contributed to filling the observation gaps in satellite-based XCO_2, several challenges remain. These include limited applicability due to a focus on regional-scale implementations, insufficient spatiotemporal validation despite the goal of reconstructing substantial missing values, and the reliance on computationally intensive data processing. Although global, seamless XCO_2 products are now available, a common and persistent limitation remains. In particular, most existing approaches are based on direct XCO_2 estimation frameworks in which coarse resolution CTM products are used as dominant input features (Zhang et al. 2023; Guan et al. 2024a). Under such conditions, the derived estimates tend to inherit the spatially smoothed characteristics of the CTMs, thereby limiting their suitability for fine-scale regional analyses.

Therefore, in this study, we aim to (1) reconstruct a long-term global XCO_2 product with OCO-2 level accuracy using a residual-based ML approach, (2) ensure strict validation methods across diverse spatial and temporal domains, (3) demonstrate improved accuracy and spatial resolution over existing data, and (4) enable long-term (2003–2019), and high-resolution (0.1°) analysis of XCO_2 trends at regional to global scales comprehensively.

2. Study area and data

2.1. Study area

The study area covers the global domain (70°N–70°S, 180°W–180°E), including both land and ocean (Figure 1). This broad coverage can support a comprehensive assessment of XCO_2 variability across space and time, enabling integrated monitoring of the global carbon cycle. Polar regions above 70° north and south latitude were excluded due to the low quality of satellite products in these regions (Trishchenko, Garand, and Trichtchenko 2019). In terms of the temporal domain, we focus on generating long-term monthly global seamless XCO_2 from January 2003 to December 2019, which extends about eleven years prior to the launch of OCO-2 satellite. This period was also selected considering the availability of CAMS-EGG4 reanalysis that will be used for the ML-based XCO_2 reconstruction framework.

2.2. Data

Table 1 provides an overview of datasets, including their sources, variables, abbreviations, and spatio-temporal resolution. In the following sections, the abbreviations listed here will be used to describe the data.

2.2.1. OCO-2 satellite data

The OCO-2 satellite, launched in July 2014, is the National Aeronautics and Space Administration's (NASA) first Earth remote sensing satellite. Its primary objective is to provide precise and accurate measurements to improve our understanding of the carbon cycle. OCO-2 operates in a sun-synchronous orbit and measures XCO_2 using high-resolution imaging grating spectrometers in the near-infrared (NIR) and shortwave

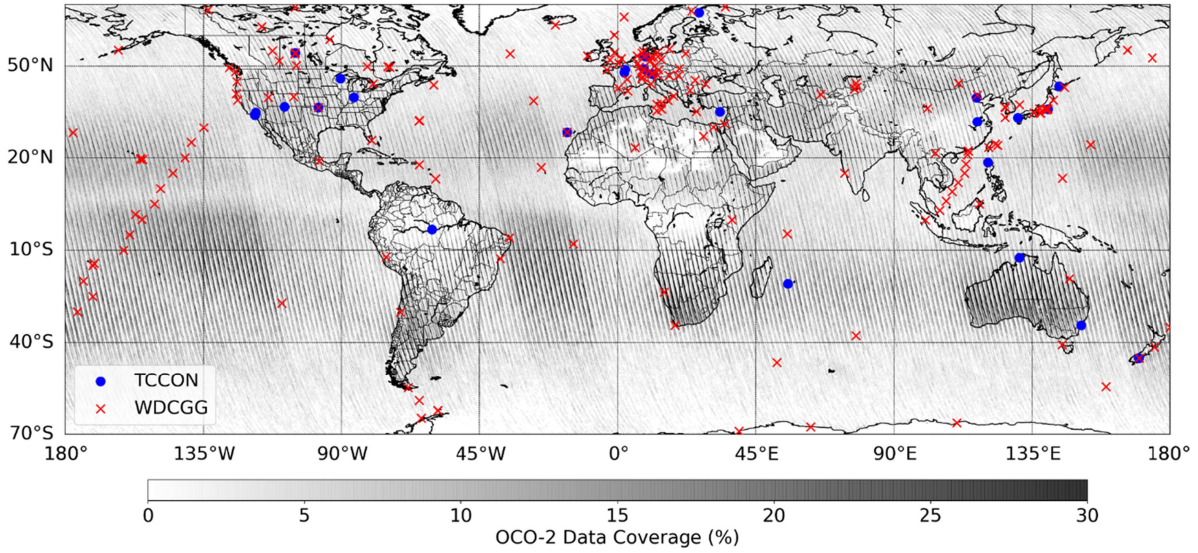

Figure 1. Study area with total carbon column observing network (TCCON) stations (blue circles) and world data centre for greenhouse gases (WDCGG) stations (red crosses). The background image represents the monthly pixel-wise Orbiting Carbon Observatory-2 (OCO-2) data coverage (%) after quality control, accumulated over its operational period (September 2014 to December 2019) within the overall study period. Darker shades indicate regions with frequent satellite observations, while white areas correspond to regions with persistent data gaps.

Table 1. Summary of data used in this study. For ECMWF variables, superscripts a, b, and c denote the data source and spatial resolution: a indicates ERA5-Land (0.1°), b represents ERA5 (0.25°), and c refers to the use of ERA5-Land over land and ERA5 over ocean.

Source	Variables	Abbreviation	Spatial resolution	Temporal resolution
OCO-2	XCO_2	OCO-2 XCO_2	2.25 km x 1.29 km	Daily
CAMS-EGG4	XCO_2	CAMS-EGG4 XCO_2	0.75° x 0.75°	3-hourly
ECMWF ERA5	Total evaporation[a]	TE	0.25° x 0.25°/0.1° x 0.1°	Monthly
	Sea surface temperature[b]	SST		
	Boundary layer height[b]	BLH		
	Mean sea level pressure[b]	MSL		
	Total column water[b]	TCW		
	Wind speed[c]	WS		
	2 m dewpoint temperature[c]	D2M		
	2 m temperature[c]	T2M		
	Surface pressure[c]	SP		
Auxiliary	Fossil Fuel Emission	FFE	1 km	Monthly
	Road density	RoadDens	5 arc min (~8 × 8 km)	–
	Population density	PopDens	Vector data 2.5 min (~5 km)	5 years
	Land-ocean ratio mask	LOratio	–	–
	Sinusoidal seasonal cycle	mmsine	–	–
TCCON	XCO_2	TCCON XCO_2	Point	Various
WDCGG	Surface CO_2	WDCGG CO_2	Point	Various

infrared (SWIR) spectral bands, employing nadir and glint observation modes to optimize data quality over land and ocean. In these operating modes, the provided XCO_2 data is recorded in eight adjacent cross-track footprints, each approximately 2.25 km (along-track) × 1.29 km (cross-track), resulting in a total swath width of about 10 km (Kira and Sun 2020). The satellite revisits the same location every 16 days, with measurements collected during the ascending orbit phase at approximately 13:30 local transit time. For this study, we utilized the OCO-2 version 10 Level 2 Full Physics (OCO2_L2_Lite_FP_10r) products from September 2014 to December 2019 during the accessible service period. Note that there are missing values in the OCO-2 data from July 31, 2017, to September 18, 2017, due to a temporary instrument anomaly (He, Wang, and Wang 2024).

2.2.2. CAMS-EGG4 reanalysis

The CAMS reanalysis is the latest global reanalysis dataset of atmospheric composition, such as pollution gases and GHGs, supported by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Inness et al. 2019). CAMS global GHG reanalysis (CAMS-EGG4) focuses on GHGs, including CO_2 and methane (CH_4), which generally have longer lifetimes compared to other air pollutants. It applies the 4D-Var assimilation method to integrate the forecasts from the Integrated Forecasting System with multiple satellite products (i.e., Envisat, Metop-A/B, and GOSAT) (Agustí-Panareda et al. 2023). Notably, OCO-2 XCO_2 observations are not assimilated in the CAMS-EGG4 reanalysis system, allowing CAMS-EGG4 and OCO-2 to be treated as independent information sources for column CO_2. CAMS-EGG4 can produce long-term, gridded, seamless data from 2003 to 2020 with spatial and temporal resolutions of 0.75° × 0.75° and 3 hours, respectively. The CO_2 column-mean molar fraction (unit: ppm) variable from CAMS-EGG4 single-level chemical vertical integrals was downloaded from the CAMS Atmosphere Data Store (ADS) (https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-ghg-reanalysis-egg4?tab=overview, last accessed: 30 October 2025). However, CAMS-EGG4 data have several well-known issues. The main problem is that the anthropogenic emissions used were not adjusted for any Coronavirus Disease 2019 (COVID-19) lockdowns in 2020. Therefore, 2020 was excluded from the study period due to its inherent uncertainty.

2.2.3. ECMWF ERA5

The meteorological variables were obtained from the ECMWF Reanalysis 5th Generation (ERA5) data (https://cds.climate.copernicus.eu/, last accessed: 30 October 2025). ERA5 is a global atmospheric reanalysis product that integrates model outputs with diverse observations worldwide to produce a physically consistent and spatially complete dataset (Hersbach et al. 2023). To facilitate various climate applications, monthly mean averages have been pre-calculated and provided by ECMWF. We utilized the monthly single-level data following the target temporal resolution. The horizontal resolution of ERA5 is 0.25° × 0.25°.

However, for land areas, the ERA5-Land data with $0.1° \times 0.1°$ resolution was utilized to provide more detailed meteorological conditions (Muñoz-Sabater et al. 2021). While all other meteorological variables were used as provided, wind speed (WS) was calculated using the formula provided by ECMWF (Equation (1)).

$$WS = \sqrt{U^2 + V^2} \tag{1}$$

where U is a 10 m u-component of wind (m/s) and V is a 10 m v-component of wind (m/s).

2.2.4. Auxiliary data

The selection of auxiliary data is based on the understanding that human-related activities are the main source of CO_2 emissions across the globe. To reflect this, we utilized data on fossil fuel emissions, road density, and population density. The fossil fuel emission data were obtained from the Open-Data Inventory for Anthropogenic Carbon dioxide (ODIAC, version 2023) (https://www.cger.nies.go.jp/en/, last accessed: 30 October 2025), which provides a high spatial resolution of 1 km for global CO_2 emissions from fossil fuel combustion. This dataset is widely used in various research applications, such as urban emission estimation, CO_2 flux inversion, and observing system design experiments (Oda and Maksyutov 2011). For road density data, we utilized the GRIP4 vector datasets provided by the Global Roads Inventory Project (GRIP), available on the GLObal BIOdiversity model for policy support (GLOBIO) website (https://www.globio.info/, last accessed: 30 October 2025). These datasets include detailed global road networks, which are useful for analyzing transportation-related emissions and understanding their impact on CO_2 emission levels. In terms of population data, we used the GPW v4 (Gridded Population of the World, Version 4) datasets, which were available from the Socioeconomic Data and Applications Centre (SEDAC; https://sedac.ciesin.columbia.edu/, last accessed: 27 May 2025). These population density rasters offer estimated human populations at five-year intervals beginning in the year 2000, consistent with national censuses and population registers.

Moreover, to enhance the representation of XCO_2 dynamics, additional variables were incorporated. Recognizing that XCO_2 variations are generally less pronounced over oceans compared to land, a land-ocean ratio mask (LOratio) was introduced to distinguish whether a pixel predominantly represents land or ocean. This accounts for the greater regional variability observed over terrestrial areas compared to the relatively stable conditions over oceans. Furthermore, to reflect the seasonal cycle of XCO_2, with peaks in April–May and troughs in August–September, a sinusoidal seasonal cycle variable (mmsine) was introduced by transforming the 12 months into a sine function ranging from −1 to 1. This formulation provides a continuous and cyclic representation of the annual seasonal pattern, rather than treating months as independent categorical variables, consistent with the smoothly varying seasonal cycle observed in long-term XCO_2 records, such as the Keeling Curve (Keeling et al. 1976).

2.2.5. TCCON measurements

The TCCON is a global network of ground-based Fourier transform spectrometers that record direct solar spectra in the NIR spectral region. The objectives of the network are to provide precise column-averaged dry-air mole fractions of atmospheric constituents (e.g. CO_2, CH_4, N_2O, HF, CO, H_2O, and HDO) and enhance our comprehension of the carbon cycle (Wunch et al. 2011). TCCON has become an essential validation source for a wide range of satellites (e.g. GOSAT, OCO-2, and TANSAT) and model simulation data (e.g. CAMS and CarbonTracker) due to its outstanding precision (Agustí-Panareda et al. 2023). In this study, XCO_2 measurements of the latest version (GGG2020; Laughner et al. 2024) from a total of 30 stations were utilized as external validation data. The spatial distribution of TCCON stations used in this study is shown in Figure 1, and note that the period of data availability for each station varies depending on their operation schedule. Station details are summarized in Table S1. The TCCON data were obtained from the TCCON Data Archive hosted by CaltechDATA at https://tccondata.org (last accessed: 30 October 2025).

2.2.6. WDCGG measurements

The World Data Centre for Greenhouse Gases (WDCGG) is a global data archive established under the Global Atmosphere Watch (GAW) programme by the World Meteorological Organization (WMO). WDCGG focuses on surface-based measurements, providing long-term and reliable records essential for understanding regional

carbon dynamics. These measurements are collected using diverse sampling methods, such as in situ monitoring, flask sampling, mobile platforms, and ship-based observations. A total of 178 stations were utilized in this study. Among the available hourly, daily, and monthly products, monthly averaged products were chosen to align with the temporal resolution of our analysis. Unlike TCCON, WDCGG offers surface CO_2 data, making it an indirect validation source for satellite- and model-derived XCO_2 estimates. In this study, WDCGG surface CO_2 measurements were used to cross-check reconstructed XCO_2 product. The data were obtained from the WDCGG archive hosted at https://gaw.kishou.go.jp (last accessed: 30 October 2025).

2.2.7. Emission inventories: EDGAR and GFED

Two global emission datasets were additionally used to support the analyses in Sections 4.4 and 4.5: the Emissions Database for Global Atmospheric Research (EDGAR) and the Global Fire Emissions Database (GFED). The latest version of EDGAR_2024_GHG, developed by the European Commission's Joint Research Centre, provides global anthropogenic greenhouse gas emissions at a $0.1^\circ \times 0.1^\circ$ spatial resolution. It includes both monthly and annual gridded products with detailed sector-specific inventories, covering 1970–2023 for annual data and 2000–2023 for monthly data. Large-scale biomass burning, including savanna and forest fires, as well as land-use, land-use change, and forestry (LULUCF) sources and sinks, are excluded from the emission totals (Crippa et al. 2023). In this study, the monthly sector-specific grid maps for the Transport and Power Industry sectors were used in Section 4.4. Fossil-fuel CO_2 emissions from the International Energy Agency (IEA)-EDGAR CO_2 dataset included in the same release were employed in Section 4.5. The data were obtained from the EDGAR repository (https://edgar.jrc.ec.europa.eu; last accessed: 30 October 2025).

The GFED, jointly developed by NASA and the Vrije Universiteit Amsterdam, provides global gridded estimates of fire emissions and burned areas by integrating satellite observations of fire activity and vegetation productivity. It quantifies CO_2 emissions from natural fire events such as savanna burning and forest fires. The current version, GFED5 Beta, offers 0.25° gridded products for 2002–2023. In this study, the CO_2 variable from the GFED5_Beta_monthly_emissions product was used in Section 4.5. The data were obtained from the official GFED repository (https://www.globalfiredata.org; last accessed: 30 October 2025).

3. Methods

Figure 2 illustrates the overall workflow of this study. The methodology comprises four main steps: (1) data preprocessing, (2) modeling, (3) model evaluation, and (4) analysis. Initially, multi-source datasets were integrated and resampled onto a unified spatiotemporal grid. A Light Gradient Boosting Machine (LGBM) model was then trained to retrieve high-resolution monthly XCO_2 for the OCO-2 operational period (hereafter, P1) and subsequently applied to reconstruct XCO_2 prior to the OCO-2 operational period (hereafter, P2). The reconstructed long-term XCO_2 product was analyzed to evaluate its accuracy, spatial enhancement, persistent emission hotspots, and global growth trends. A detailed description is provided in the following sections.

3.1. Data preprocessing

Given the varying spatiotemporal resolutions of the datasets listed in Table 1, all datasets were resampled to a uniform $0.1^\circ \times 0.1^\circ$ monthly grid using bilinear interpolation. This resampling was performed to ensure consistency across the integrated multi-source datasets; however, it may introduce interpolation-related uncertainties, such as spatial representativeness errors and subgrid-scale smoothing effects (Pogson and Smith 2015).

For OCO-2, only measurements with "xco_2_quality_flag = 0" were used to ensure data quality, following the guidelines outlined in the OCO-2 Algorithm Theoretical Basis Document. Daily XCO_2 data were aggregated by averaging all valid OCO-2 observations within each target grid cell. To reduce the influence of uneven or sparse orbital sampling, monthly means were computed only for grid cells containing at least 30 individual observations per month. This threshold was selected as a practical minimum for aggregating track-based OCO-2 observations from their native footprint (1.29 km × 2.25 km) to the target grid (0.1°

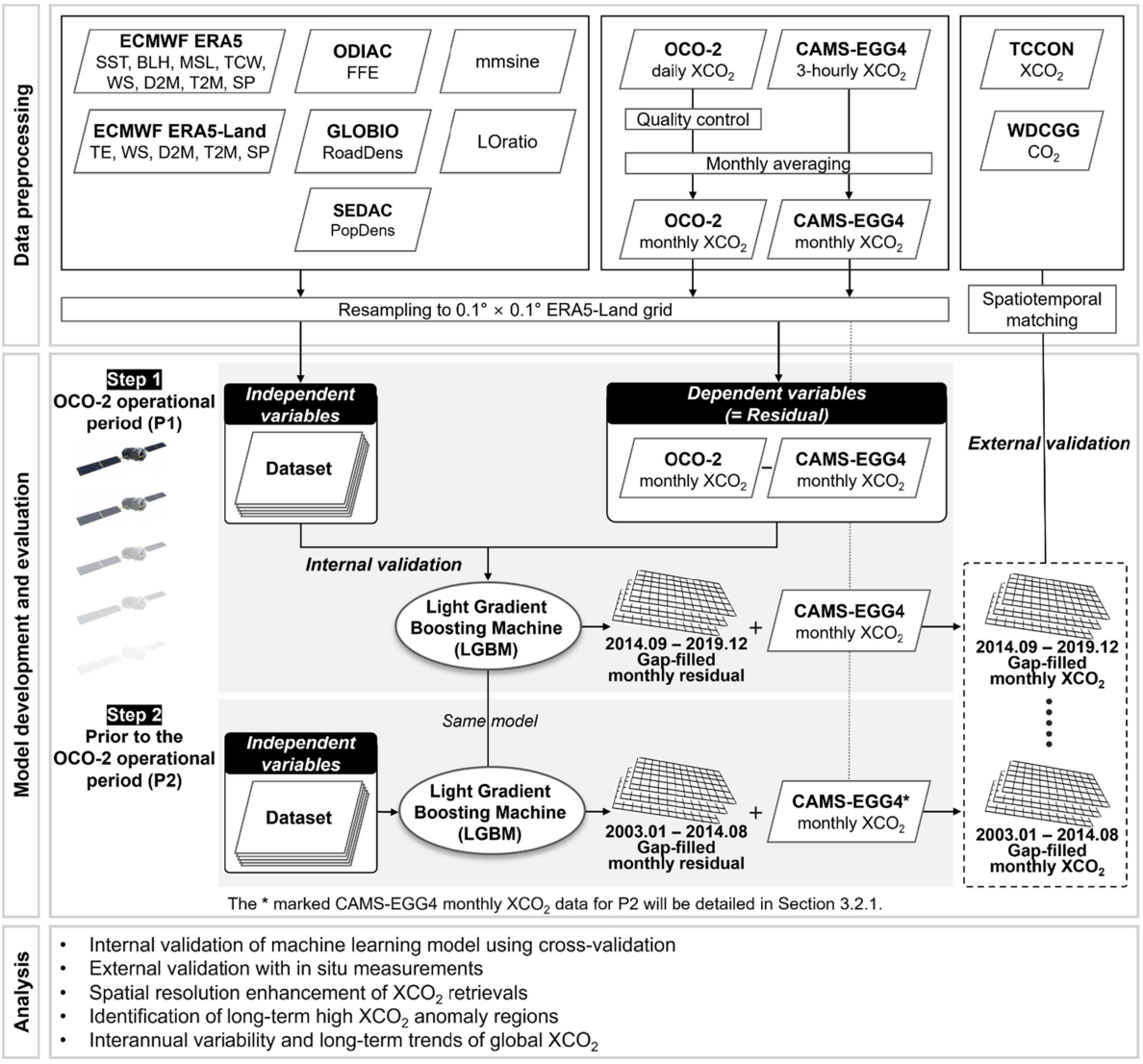

Figure 2. Schematic overview of the proposed framework for reconstructing a global, long-term, high-resolution XCO_2 product. The framework consists of four processes: data preprocessing, modeling, model evaluation, and analysis.

resolution). Sensitivity analyses using alternative thresholds (10–40 observations per month) indicate that model performance is largely insensitive to the threshold choice, whereas spatial coverage decreases rapidly with higher thresholds (Figures S1 and S2).

For the CAMS-EGG4 XCO_2, the 3-hourly values were temporally interpolated and converted from UTC to local time to match the OCO-2 local overpass time (i.e., 13:30). This conversion was performed using navigational time zones derived from longitude; statutory time deviations were not considered (Figure S3). Although this simplified approach may introduce uncertainties related to solar geometry, its potential influence is expected to be minimal due to the exclusion of high-latitude regions (above 70°) that experience strong seasonal variations in solar angle and day length. Regarding the auxiliary variables, those that are static or updated at five-year intervals were temporally mapped to the monthly target resolution using a stepwise assignment. For LOratio variable, a 50 km radius circular mask was used to compute the ratio of land to ocean within each grid cell. Ground-based observations from the TCCON, which have varying temporal resolutions across different stations, were aggregated to monthly averages to ensure consistency with the temporal resolution of the reconstructed XCO_2 product. For each TCCON station, LGBM-derived XCO_2 values were averaged within a circular area of approximately 100 km radius centred on the station (He et al. 2022; Yang et al. 2025).

3.2. Modeling

3.2.1. Residual learning

In this study, we define the residual as the difference between OCO-2 XCO_2 and CAMS-EGG4 XCO_2 at each $0.1^\circ \times 0.1^\circ$ pixel in the global monthly map. This reflects the discrepancy between satellite observations and reanalysis products, primarily arising from the coarse resolution and assimilation-related uncertainties in the CAMS-EGG4 system (Custódio, Borrego, and Relvas 2022; Agustí-Panareda et al. 2023). Rather than directly predicting OCO-2 XCO_2 from CTM-based XCO_2, as in previous studies (Mustafa and Xu 2025; Guan et al. 2024a), we employ a residual learning framework. This design alleviates extrapolation errors arising from the secular increase of atmospheric CO_2 and lessens reliance on coarse reanalysis inputs, thereby reducing the spatial biases inherent to CAMS-EGG4. Mathematically, the residual at each grid cell (i, j) is computed as Equation (2):

$$Residual_{(i,j)} = OCO\text{-}2_{XCO_{2(i,j)}} - CAMSEGG4_{XCO_{2(i,j)}} \tag{2}$$

The final $Reconstructed_{XCO_2}$ is obtained by adding the predicted residual to the original CAMS-EGG4 XCO_2. Furthermore, to address the known time-dependent biases of CAMS-EGG4 reported in the official validation report (Ramonet et al. 2021; Mustafa and Xu 2025), an additional bias correction term, δ_{bias}, was applied (Equation (3)). As detailed in Equation (4), the mean bias for each period (P1 and P2) was calculated as the average difference between CAMS-EGG4 and TCCON values across all available stations. Here, $\mu_{bias,P1}$ and $\mu_{bias,P2}$ denote the mean biases for P1 and P2, respectively. Accordingly, δ_{bias} is set to zero for P1 because residual learning against OCO-2 observations already accounts for bias adjustment, whereas for P2, $\mu_{bias,P2}$ is aligned with the $\mu_{bias,P1}$ baseline to compensate for a time-dependent drift in CAMS-EGG4 during P2.

$$Reconstructed_{XCO_{2(i,j)}} = CAMS\ EGG4_{XCO_{2(i,j)}} + Residual_{(i,j)} + \delta_{bias} \tag{3}$$

$$\delta_{bias} = \begin{cases} 0, & if\ t \in P1 \\ \mu_{bias,P1} - \mu_{bias,P2}, & if\ t \in P2 \end{cases} \tag{4}$$

3.2.2. Model development

The overall modeling process is divided into two key steps based on the availability of OCO-2 satellite products: (1) estimation during the OCO-2 operational period within our study domain, from September 2014 to December 2019 (P1), and (2) reconstruction for the period lacking OCO-2 satellite data, from January 2003 to August 2014 (P2) (Figure 2). In Step 1, an algorithm is trained using multi-source input features to fill spatial gaps in the global monthly OCO-2 XCO_2 product. Step 2 applies the trained model developed in Step 1 to reconstruct XCO_2 concentrations without direct satellite measurements for P2. This two-step approach allows the model to learn from high-quality reference data from P1 and apply the learned patterns to extend the product backward in time.

To implement this framework, the LGBM model was chosen for its computational efficiency and high predictive accuracy with large-scale datasets (Kang et al. 2024). It employs a leaf-wise gradient boosting strategy that enables efficient split optimization, making it well-suited for capturing complex spatial and temporal patterns in XCO_2 concentrations. LGBM also incorporates Gradient-based One-Side Sampling (GOSS), which retains all samples with large gradients while randomly sampling those with smaller gradients, thereby accelerating training without sacrificing accuracy. Previous studies have demonstrated the strong performance of the LGBM model in atmospheric applications, including the estimation of XCO_2 as well as air pollutants, owing to its ability to model nonlinear relationships and leverage multi-source data (He et al. 2022; Choi et al. 2023; Meng et al. 2024). The LGBM parameters were optimized using the Python package named "*hyperopt*". This Bayesian optimization employs a probabilistic surrogate model and an acquisition function to efficiently search the hyperparameter space (Snoek, Larochelle, and Adams 2012). Each parameter was explored within the ranges listed in Table S2 and finalized with the following settings: boosting type '*goss*', learning rate 0.09, maximum depth 9, minimum child weight 2, number of estimators 800, number of leaves 300, feature fraction by tree 0.80, and subsample 0.513. These settings ensure

computational efficiency, accurate representation of carbon dynamics, and robust performance, addressing the challenges of global-scale data modeling.

3.3. Model evaluation

In this study, we evaluated the predictive performance of the proposed LGBM model using two validation approaches: internal *N*-fold cross-validation (CV) using a training dataset and additional external validation of reconstructed XCO_2 results using TCCON and WDCGG observations. Firstly, we applied various CV methods—random (RDCV), spatial ($SPCV_{lon}$ and $SPCV_{lat}$), and temporal ($TPCV_{yr}$ and $TPCV_{mon}$)—to evaluate how accurate and robust the LGBM model trained by targeting XCO_2 residual values over the period for which OCO-2 data is available. *N*-fold CV is a technique for evaluating the performance of a model by dividing the given data into *N* subsets, using one subset for testing and the remaining *N*-1 subsets for training across *N* iterations. RDCV randomly splits into 10 subsets based on the sample without any other considerations. SPCV and TPCV were applied to evaluate the reliability and robustness of the models across spatial and temporal dimensions, respectively. $SPCV_{lon}$ and $SPCV_{lat}$ split the whole sample into 12 and 14 folds, respectively, by dividing the global study area into 30° intervals of longitude and 10° intervals of latitude. $TPCV_{yr}$ and $TPCV_{mon}$ split the sample by year and month, respectively, to identify if the model appropriately represented the seasonality and annual increase in XCO_2. Secondly, in situ XCO_2 observations (i.e., TCCON and WDCGG), which were not used in the training of the LGBM model, were employed exclusively as independent data sources for external validation to evaluate the reliability of the long-term XCO_2 reconstruction.

The performance of the proposed model was evaluated using five widely known accuracy metrics: slope [unitless], square of the Pearson correlation coefficient [R^2; Equation (5)], mean absolute error [MAE, unit: ppm; Equation (6)], root mean square error [RMSE, unit: ppm; Equation (7)], and relative RMSE [rRMSE, unit: %; Equation (8)]. The rRMSE was included as a complementary metric to express errors relative to the mean XCO_2 concentration, thereby facilitating comparisons across different regions. The expressions for each metric are as follows:

$$R^2 = \frac{\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\left[(\sum_i x_i - \bar{x})^2 (\sum_i y_i - \bar{y})^2 \right]^{1/2}} \tag{5}$$

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\mathrm{n}} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} \tag{7}$$

$$rRMSE = 100 \times \frac{RMSE}{\bar{y}} \tag{8}$$

where x_i and y_i are predicted and actual values of the i^{th} sample respectively, $\bar{x}$ and $\bar{y}$ is the average of the predicted and actual data and n is the total number of samples. Although the LGBM model was trained to predict the residual between OCO-2 and CAMS-EGG4 XCO_2, all evaluation metrics were calculated from the reconstructed XCO_2—that is, the sum of the predicted residuals and the CAMS-EGG4 baseline—to directly assess the accuracy of the final product (Figure 2). The slope and R^2 represent the degree of linear agreement between predicted and observed values, where the slope reflects the proportional relationship and R^2 quantifies the overall goodness of fit. The MAE measures the average absolute error, while the RMSE captures the overall error magnitude with greater sensitivity to larger deviations. The rRMSE is calculated by normalizing RMSE with the mean of observed values, yielding a unitless metric that facilitates comparison across different scales and mitigates misinterpretation caused by magnitude differences.

3.4. XCO_2 anomaly analysis

To examine patterns beyond the year-to-year increase in atmospheric XCO_2, we computed two types of anomalies from the reconstructed 17-year LGBM-based XCO_2 product, each tailored to specific analytical objectives. For the long-term spatial distribution of high-emission regions (Section 4.4), the anomaly at each grid cell was calculated

by subtracting the global median XCO_2 from the 17-year mean value at that cell, highlighting persistently elevated concentrations while reducing the impact of outliers (Hakkarainen et al. 2019). For interannual variability linked to El Niño–Southern Oscillation (ENSO) (Section 4.5), a two-step method was used at every grid-cell *(x, y)*, as formulated in Equation S1. We first removed long-term nonlinear trends over time *(t)* by fitting a third-degree polynomial to the monthly XCO_2 time series at each grid cell (Figure S4b). Second, the monthly climatological mean, which is a function of calendar month *(m)* was subtracted to isolate deviations from the seasonal cycle (Figure S4c) (Thoning, Tans, and Komhyr 1989; Keppel-Aleks et al. 2014). These approaches provide a consistent basis for hotspot detection and ENSO-related variability analysis, respectively.

4. Results and discussion

4.1. Internal validation of machine learning model using cross-validation

Table 2 and Figure S5 summarize the overall performance of the LGBM-based XCO_2 reconstruction model, evaluated using different CV methods. The results demonstrated that the model consistently reconstructed global XCO_2 concentrations on a monthly scale with high accuracy ($R^2 = 0.93–0.96$) and low errors (RMSE = 0.80–1.11 ppm). Among the various validation strategies, RDCV yielded the highest accuracy across all error metrics (e.g. $R^2 = 0.96$, MAE = 0.60 ppm, and RMSE = 0.80 ppm), while also exhibiting a negligible standard deviation (SD) across the folds, ensuring reliable performance. This evidence indicates that the data-driven model achieves the most stable and reliable performance when training samples are randomly and evenly distributed without spatiotemporal constraints, which is consistent with the findings of previous studies (Li, Wu, and Wang 2023; Wu et al. 2024). However, as the primary objective of this study is to fill spatial gaps in OCO-2 observations and further reconstruct historical XCO_2 values, rigorous spatiotemporal validation remains essential. The results from spatial (i.e., $SPCV_{lon}$ and $SPCV_{lat}$) and temporal (i.e., $TPCV_{yr}$ and $TPCV_{mon}$) validation exhibited relatively larger SDs across the folds compared to RDCV. This was particularly evident in $SPCV_{lat}$ and $TPCV_{yr}$, likely reflecting the uneven distribution of observations across latitude and year, which can amplify variability in fold-wise performance. Nevertheless, despite the larger fold-wise variability, the overall model accuracy remained robust and comparable across different validation schemes, with stable performance observed at both pixel- and month-wise levels as well (Figures S6 and S7). Furthermore, as shown in Figure S8, the residuals used as the learning targets exhibited errors that were small relative to the observed residual range (−13.97 ppm to +7.35 ppm), with MAE values of 0.60–0.84 ppm and RMSE values of 0.80–1.11 ppm. These results indicate that the model not only reproduces the long-term growth trend of XCO_2 but also effectively resolves and corrects the spatio-temporal variability of the residuals through the integrated use of multi-source data as supported by the feature importance analysis (Figure S9). Collectively, these findings highlight the robustness of the residual learning framework across unseen spatiotemporal domains and underscore its potential for high-fidelity, long-term XCO_2 monitoring.

Table 2. Internal model validation results for P1, evaluated using different cross-validation methods. RDCV, $SPCV_{lon}$, $SPCV_{lat}$, $TPCV_{yr}$, and $TPCV_{mon}$ represent random, spatial (longitude- and latitude-based), and temporal (yearly and monthly) cross-validation, respectively. Values in parentheses indicate the standard deviation of accuracy across individual folds. The highest accuracy for each metric is highlighted in bold, while the largest standard deviation across folds is underlined.

Validation method	Slope	R^2	MAE (ppm)	RMSE (ppm)	rRMSE (%)
RDCV	**0.95** (0.0000)	**0.96** (0.0002)	**0.60** (0.0008)	**0.80** (0.0016)	**0.20** (0.0004)
$SPCV_{lon}$	**0.95** (0.0100)	0.94 (0.0095)	0.75 (0.0862)	0.99 (0.1162)	0.24 (0.0290)
$SPCV_{lat}$	0.94 (0.0325)	0.94 (0.0284)	0.75 (0.2410)	0.99 (0.3156)	0.24 (0.0781)
$TPCV_{yr}$	0.94 (0.0509)	0.93 (0.0925)	0.83 (0.0903)	1.08 (0.1141)	0.27 (0.0259)
$TPCV_{mon}$	0.94 (0.0510)	0.93 (0.0264)	0.84 (0.1282)	1.11 (0.1883)	0.28 (0.0469)

Figure 3 illustrates the spatial and temporal CV results of the XCO_2 reconstruction model, highlighting fold-specific accuracy metrics. This comprehensive evaluation offers insights into the model's performance and variability under different spatial and temporal validation scenarios. In the spatial CV results based on the longitudinal segmentation (Figure 3a and c), notable differences in model accuracy were observed across regions with varying land-to-ocean proportions. The highest accuracy, characterized by the largest R^2 values and lowest rRMSE, was observed over the Pacific Ocean—where the land fraction is minimal. In contrast, regions with a higher land fraction, particularly within certain longitudinal bands, exhibited larger errors. This may reflect the increased complexity of terrestrial ecosystems and anthropogenic influences, which possibly contributed to higher uncertainty in XCO_2 retrievals over land (Guan et al. 2024a). In the $SPCV_{lat}$ results (Figure 3b and d), the model errors were also larger in the Northern Hemisphere, likely due to higher emissions and greater variability in XCO_2 concentrations. Accuracy declined most notably in high-latitude regions of the Northern Hemisphere (i.e., 60–70°N), where sparse OCO-2 coverage and land-dominated surfaces result in greater spatial heterogeneity. By contrast, the ocean-covered Southern Hemisphere (i.e., 60–70°S) exhibits more uniform conditions and is less affected by limited observational coverage. This relationship is reflected in Figure S10, which shows that regions with greater land fractions generally correspond to lower R^2 and higher rRMSE across spatial CV folds. Complementing the spatial patterns, the continent-level evaluation (Figure S11) shows that the model achieved R^2 values of 0.91–0.96 and rRMSE values of 0.20–0.35%, indicating stable performance and the absence of systematic bias even in regions with sparse training data. The $TPCV_{yr}$ results exhibited relatively large fold-to-fold variations in accuracy (Table 2 and Figure 3e), with noticeably lower performance in the earliest and latest years of the training period. This pattern may reflect the upward trend in annual XCO_2 concentrations due to industrialization, leading to out-of-range issues. In contrast, the $TPCV_{mon}$ results showed relatively small variations, with the highest accuracy occurring in April, when XCO_2 concentrations typically peak in the Northern Hemisphere due to springtime accumulation at higher latitudes. Conversely, the lowest accuracy was observed in August, a period characterized by minimum concentrations associated with strong photosynthetic uptake during the growing season.

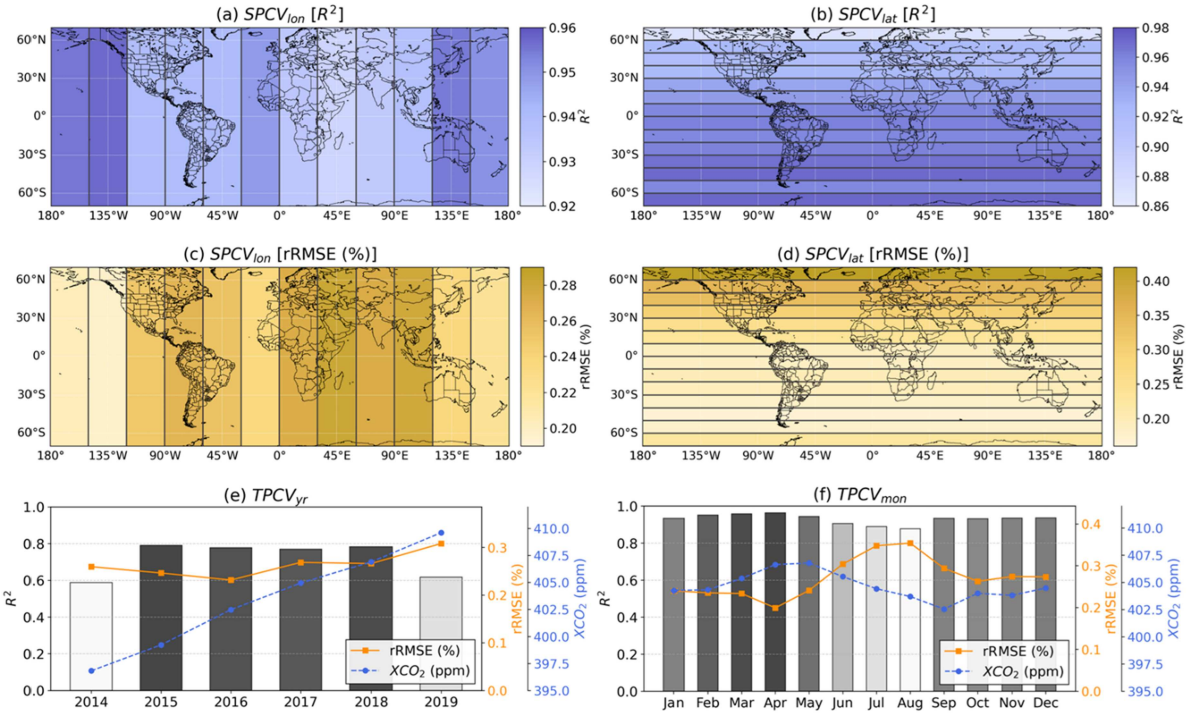

Figure 3. Spatial and temporal cross-validation results of the XCO_2 reconstruction model. (a) and (b) show spatial validation results for $SPCV_{lon}$ and $SPCV_{lat}$, respectively, using R^2, while (c) and (d) display corresponding rRMSE (%) values. (e) and (f) present temporal validation results for $TPCV_{yr}$ and $TPCV_{mon}$, with R^2 represented by bar graph, rRMSE (%) by orange solid lines, and XCO_2 (ppm) by blue dashed lines. The bars are shaded using a grayscale gradient, where darker colours indicate higher R^2 values.

4.2. External validation with in situ measurements

Figure 4 presents the external validation results, comparing XCO_2 estimates from the LGBM model and the CAMS-EGG4 reanalysis against in situ observations from TCCON and WDCGG. These ground-based observations enabled the evaluation of reconstructed XCO_2 in regions and periods not covered by

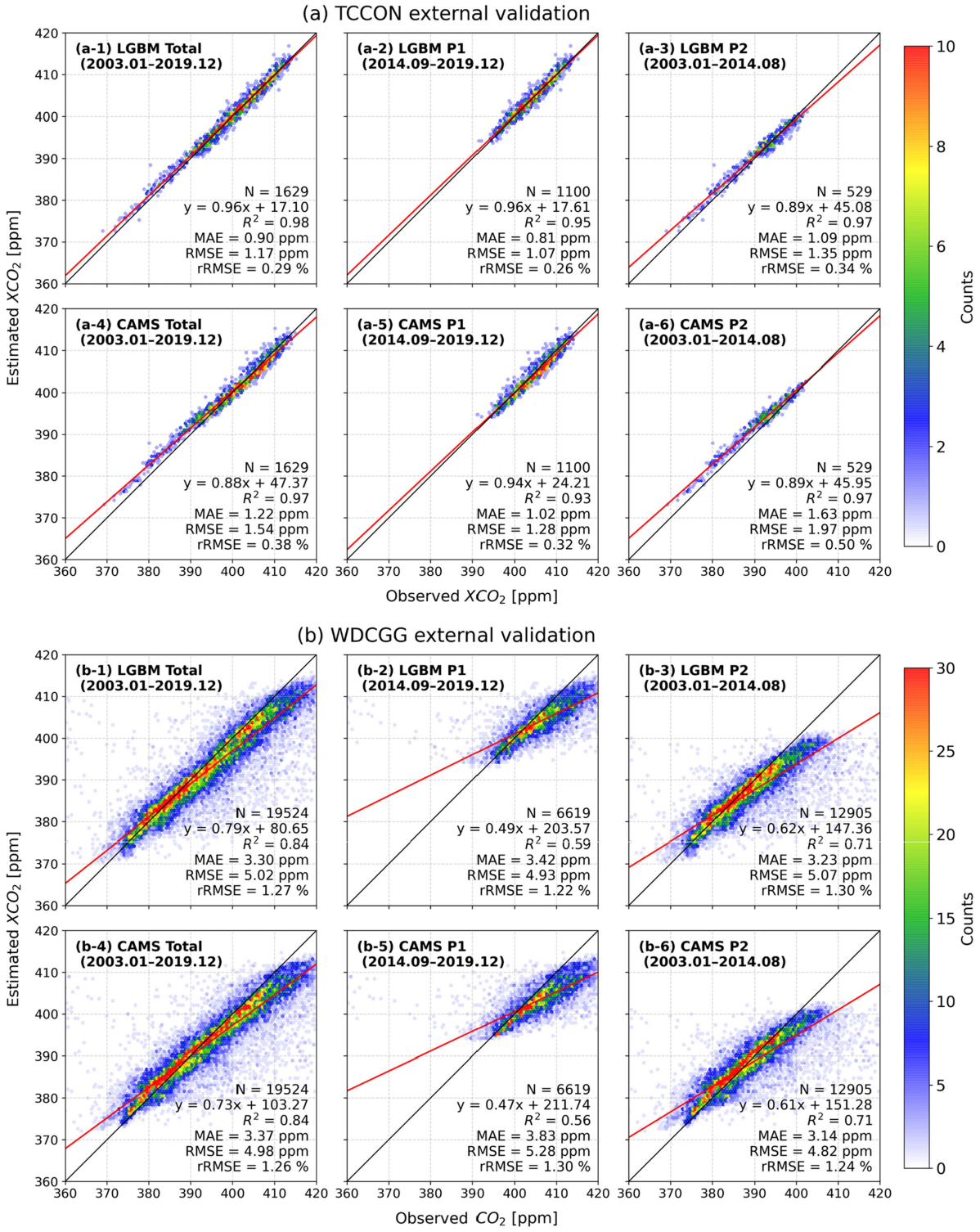

Figure 4. Scatter plots comparing observed in situ values (x-axis) with estimated XCO_2 values (y-axis) for validation against (a) TCCON and (b) WDCGG stations, respectively. For each dataset, results from the Light Gradient Boosting Machine (LGBM) model are shown in subplots 1–3 and from the Copernicus Atmosphere Monitoring Service Global Greenhouse Gas (CAMS-EGG4) reanalysis in subplots 4–6.

OCO-2. To evaluate the advantages of the high-resolution (0.1°) LGBM-based XCO_2 product, the coarser CAMS-EGG4 XCO_2 data were resampled to the same spatial resolution using the nearest-neighbor method, enabling comparison on a consistent spatial scale. As shown in Figure 4a, the LGBM model consistently exhibited smaller errors than CAMS-EGG4 reanalysis when compared with TCCON observations, reducing RMSE by approximately 0.37 ppm (24%) over the entire period (2003–2019). Notably, this improvement persisted across both P1 and P2, with the LGBM model substantially mitigating the positive bias observed in CAMS-EGG4 reanalysis during P2, which corresponds to the satellite data gap period. This result indicates that the residual learning approach effectively mitigates the out-of-range overestimation observed during P2 when directly predicting OCO-2 XCO_2 (Figure S12), which stems from the mismatch in concentration ranges between P1 and P2. These findings demonstrate the ability of the LGBM model to generate a high-resolution XCO_2 products that align closely with ground-based observations across both recent and historical timeframes. To complement the sparse coverage of TCCON, WDCGG surface CO_2 data were used as an additional indirect validation source. As shown in Figure 4b, both the LGBM model and CAMS-EGG4 reanalysis show comparable agreement with WDCGG observations in terms of large-scale variability and temporal consistency. Although this constituted an indirect validation, the LGBM model reconstructed XCO_2 at a finer resolution of 0.1°, while exhibiting performance comparable to that of CAMS-EGG4, which operates at a coarser resolution of 0.75°. This enhanced spatial resolution enables more precise characterization of regional variations and emission patterns, thereby strengthening the utility of the product for high-resolution XCO_2 monitoring and localized analysis.

To further evaluate the robustness of the validation beyond aggregated performance metrics, comparisons with TCCON were conducted at both the station and seasonal levels (Figure 5). Station-wise evaluation over the full analysis period shows that the LGBM model achieves high agreement with TCCON across most sites, while exhibiting smaller and more stable mean biases than CAMS-EGG4, with no evidence of systematic degradation at specific locations (Figure 5a). Notably during P2, the application of the additional δ_{bias} term leads to a systematic reduction in mean bias relative to CAMS-EGG4, as demonstrated by the results in Figure S13. This behavior is also supported by the station-averaged comparison (Figure S14), in which LGBM-based XCO_2 estimates remain more closely aligned with TCCON than CAMS-EGG4 even after aggregation at the station level.

Seasonal analyses indicate that during P1, both products exhibit relatively limited seasonal biases, whereas during P2, CAMS-EGG4 shows pronounced positive seasonal biases that are substantially reduced in the bias-corrected LGBM estimates (Figure 5b). Overall, these station- and season-resolved results demonstrate that the improved performance of the LGBM model is consistently maintained across sites and seasons, thereby reinforcing the reliability of the high-resolution XCO_2 product beyond overall correlation and RMSE metrics.

Figure 6 presents a comprehensive evaluation of monthly XCO_2 biases between the model products and TCCON observations from 2003 to 2019. By evaluating performance at the station and monthly levels, this analysis offers a more detailed perspective on how model accuracy varies across space and time, complementing the aggregated assessment shown in Figure 4a. The mosaic plots reveal that the LGBM model (Figure 6a) exhibits generally lower and more stable biases than the CAMS-EGG4 reanalysis (Figure 6b), effectively mitigating the temporal drift previously observed in CAMS-EGG4 reanalysis—namely, overestimation in earlier years and underestimation in more recent years. Figure S15 highlights the statistical improvement achieved by the LGBM-based product over CAMS-EGG4 reanalysis. Our reconstructed XCO_2 product reduced the RMSE from 1.56 ppm to 1.16 ppm, the MAE from 1.25 to 0.90 ppm, and the standard deviation from 1.54 to 1.14 ppm. The mean bias also improved slightly, from 0.27 ppm to 0.22 ppm, indicating a closer agreement between the LGBM model and TCCON observations. The narrower and more centered distribution of the reconstructed XCO_2 product biases around zero further confirmed a substantial reduction in both error magnitude and variability. These overall findings demonstrate that our proposed framework offers a clear performance advantage over existing reanalysis products by effectively reducing systematic bias and error variability across space and time. This highlights its value as a reliable and high-resolution alternative to existing XCO_2 reanalysis datasets.

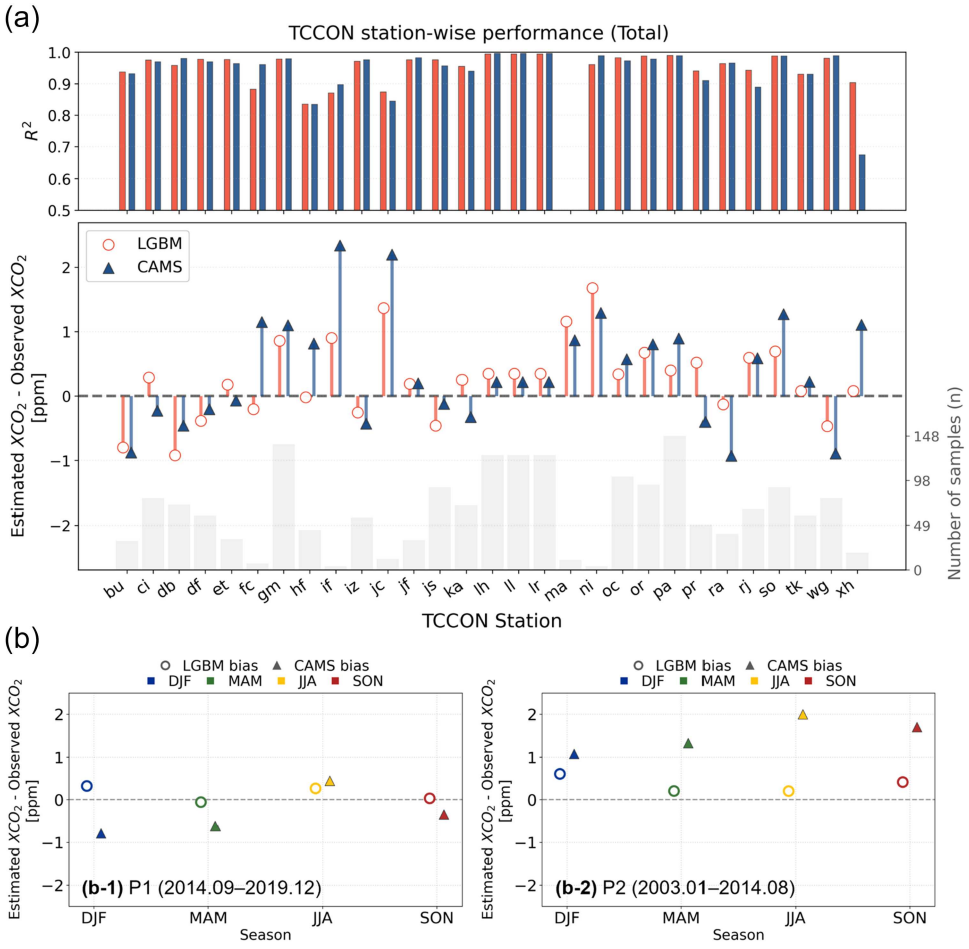

Figure 5. Station- and season-resolved validation of estimated XCO_2 against TCCON observations. (a) Station-wise R^2 and mean bias (estimated XCO_2 − observed XCO_2) for the proposed LGBM model and CAMS-EGG4 across the full analysis period. Grey bars denote the number of samples at each station. (b) Seasonal mean biases for the OCO-2 period (P1) and the pre–OCO-2 period (P2).

4.3. Spatial resolution enhancement of XCO_2 retrievals

To evaluate the spatial resolution improvement of the reconstructed XCO_2 product relative to existing products, regional distributions were analyzed over two representative regions—the United States (July 2019) and East Asia (May 2019). These regions were selected because their relatively high XCO_2 concentration requires accurate monitoring, and they have a dense network of TCCON stations (Figure S16). The selected periods reflected high OCO-2 coverage over the mid-latitudes (Zheng, Zhang, and Zhang 2023) and increased TCCON station availability in recent years, allowing for more comprehensive validation. Figure 7 compares the spatial XCO_2 patterns derived from the three sources. While OCO-2 provided high-precision XCO_2 measurements, it suffered from substantial data gaps caused by cloud contamination and unfavorable surface conditions (He et al. 2022; Li, Wu, and Wang 2023) (Figure 7a and d). On average, at our target monthly resolution, the mean global coverage of OCO-2 was 0.55%, with a maximum of 0.75% and a minimum of 0.17%. Although CAMS-EGG4 reanalysis could complement these gaps, its coarse resolution of 0.75° limited the ability to see localized features. CAMS-EGG4 XCO_2 failed to capture fine-scale gradients and exhibits spatial oversmoothing, particularly over the ocean. This smoothing was especially evident in the southern and northeastern United States (Figure 7c), as well as the densely populated eastern coastal region of China, including the Shanghai area (Figure 7f). On the other hand, our reconstructed LGBM-based XCO_2 product at 0.1° resolution accurately captured localized spatial variability and gradient structures, offering improved consistency with OCO-2 data (Figure 7b and e). This is reflected in the spatial variance values, defined as the variance of gridded XCO_2 within each region (Equation S2). The LGBM-based product

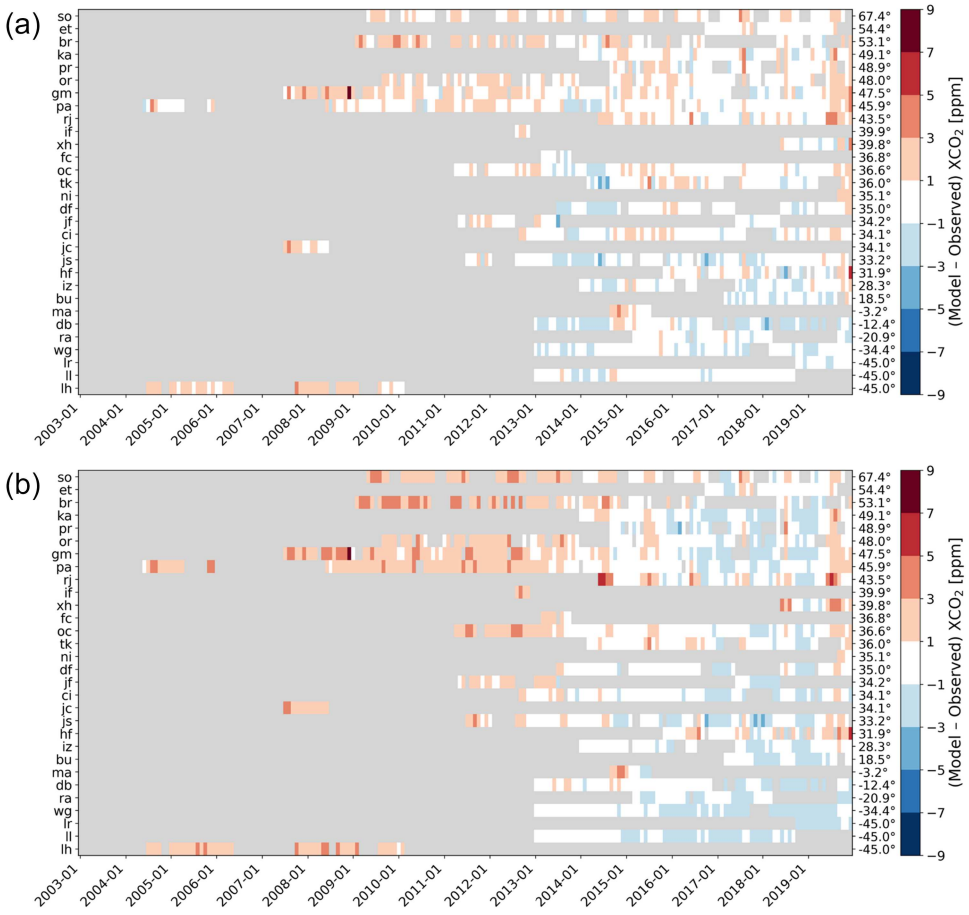

Figure 6. Mosaic plot showing the monthly XCO_2 biases (model − TCCON) across all available TCCON stations over the entire study period (2003–2019). The model denotes (a) LGBM model and (b) CAMS-EGG4 reanalysis, respectively. Each row corresponds to a TCCON station, identified by its station abbreviation as listed in Table S1, ordered in descending latitude from north to south. Colored cells indicate the bias of the model, with red (blue) colors denoting model overestimation (underestimation). Grey cells denote months with missing or unavailable TCCON observations.

shows higher spatial variance over both the United States (5.655) and East Asia (0.379) than CAMS-EGG4 (5.097 and 0.087, respectively).

To quantitatively evaluate, we compared the modeled XCO_2 values with colocated TCCON observations in each region. For the July 2019 scene over the United States (5 stations), the LGBM-based XCO_2 product exhibited a mean bias of 0.181 ppm, substantially lower than the 1.332 ppm bias observed for CAMS-EGG4 XCO_2. Similarly, for the May 2019 scene over East Asia (5 stations), the mean bias of the LGBM-based XCO_2 product was 0.002 ppm, whereas CAMS-EGG4 XCO_2 showed a negative bias of 1.075 ppm. These results confirmed that our LGBM-based XCO_2 product more accurately reproduced ground-based reference observations and substantially reduced errors compared to CAMS-EGG4 XCO_2. In summary, the LGBM model addressed both space and time issues by delivering seamless, high-resolution global maps with spatial patterns that reflect realistic atmospheric dynamics. These findings demonstrate the effectiveness of our framework in producing spatially consistent and high-fidelity reconstructions of atmospheric XCO_2.

4.4. Identification of long-term high XCO_2 anomaly regions

As illustrated in Figure 8, the global anomaly distribution map exhibits a clear hemispheric contrast. Positive anomalies were concentrated in the Northern Hemisphere, while negative anomalies dominated the Southern Hemisphere, with mean anomaly values of 1.069 ppm and -0.637 ppm, respectively. This spatial pattern was consistent with the well-established global emission asymmetry, as the Northern Hemisphere is home to the majority of anthropogenic CO_2 sources, including industrial regions and densely populated

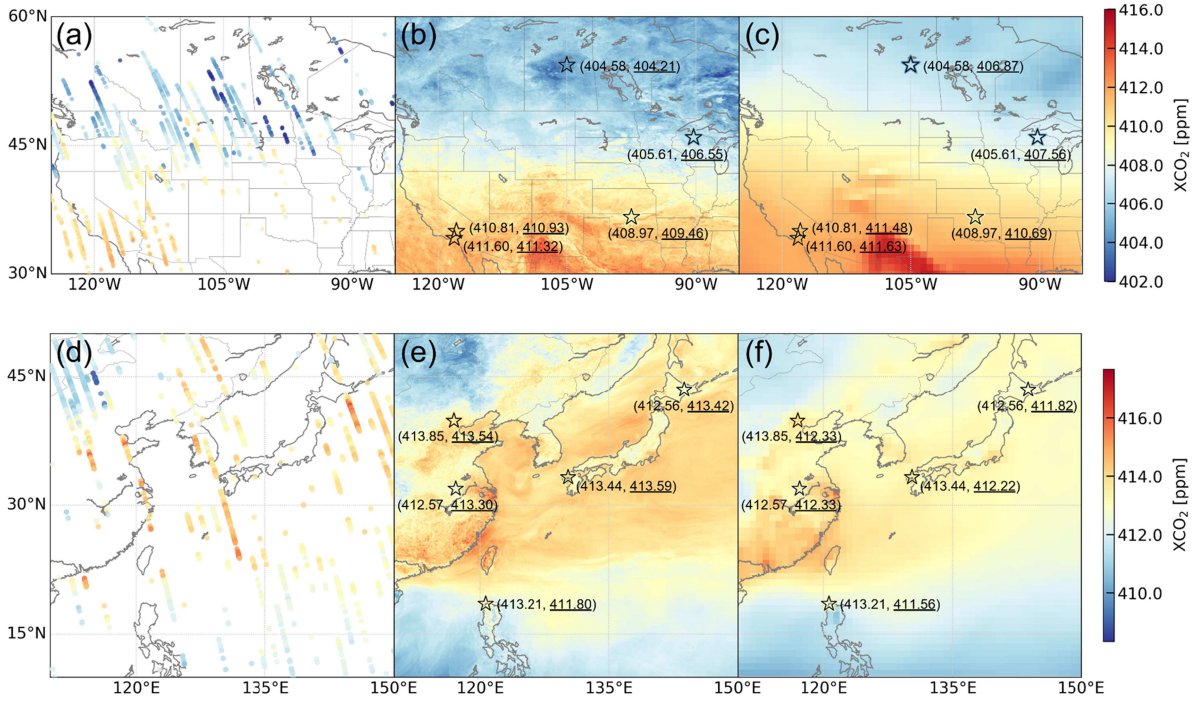

Figure 7. Spatial distributions of XCO_2 over (a–c) the United States (July 2019) and (d–f) East Asia (May 2019). For each region, panels from left to right show XCO_2 from OCO-2 observations (a, d), LGBM-based product at 0.1° resolution (b, e), and CAMS-EGG4 at 0.75° resolution (c, f). Star symbols indicate the locations of TCCON stations, with the surrounding borders representing observed XCO_2 concentrations. Color similarity with the background implies stronger model–observation consistency. Annotated values next to each station represent the TCCON measurements and the corresponding model estimate (underlined). Due to the sparse spatial sampling and small footprint of OCO-2 observations, marker sizes have been enlarged for visualization purposes only and do not represent areal coverage.

urban areas (Zickfeld et al. 2021). Interestingly, elevated anomalies were observed in topographically enclosed regions such as the Amazon basin, where surrounding orographic barriers limit atmospheric circulation. Regardless of the intensity of direct emissions, these conditions restrict horizontal mixing and exacerbate local XCO_2 accumulation (Hossain 2022). These findings suggest that long-term anomaly patterns are influenced by terrain-driven atmospheric dynamics, in addition to the strength of emissions.

To provide a detailed view of specific regions, Figure 8 (A to D) compares spatial distributions across four high-anomaly regions using the LGBM-based XCO_2 product and CAMS-EGG4 XCO_2. Although CAMS-EGG4 captured the general regional distribution of XCO_2, the LGBM-based product delineated enhanced spatial details, enabling more precise identification of localized emission hotspots. For example, in metropolitan regions such as Los Angeles (Figure 8A) and Tokyo (Figure 8D), both products capture elevated XCO_2 concentrations. However, the LGBM-based product further revealed localized hotspots that were not discernible in CAMS-EGG4. Additional comparison with sector-specific emission anomalies from EDGAR shows spatial consistency with the identified XCO_2 hotspots, providing correlational evidence that these localized enhancements may be associated with anthropogenic emission sources. In Los Angeles, enhanced anomalies were spatially co-located with the Los Angeles International Airport, suggesting a possible linkage to aviation-related emissions. Similarly, in Tokyo, elevated XCO_2 concentrations were observed near the JERA thermal power station in the Tokyo Bay area—one of the region's major coastal energy facilities—indicating a spatial association with large scale power generation infrastructure.

At the same time, these localized enhancements cannot be attributed solely to emission sources. Atmospheric transport, variability in planetary boundary-layer height, and terrain-induced flow confinement may also play important roles in shaping the observed spatial patterns. This is particularly evident in regions such as the Punjab (Figure 8B) and the Sichuan Basin (Figure 8C), where complex topography can limit horizontal ventilation and promote the accumulation of XCO_2 under atmospheric stagnant conditions (Luo et al. 2020). While CAMS-EGG4 captured broad regional enhancements, the LGBM-based product enabled the

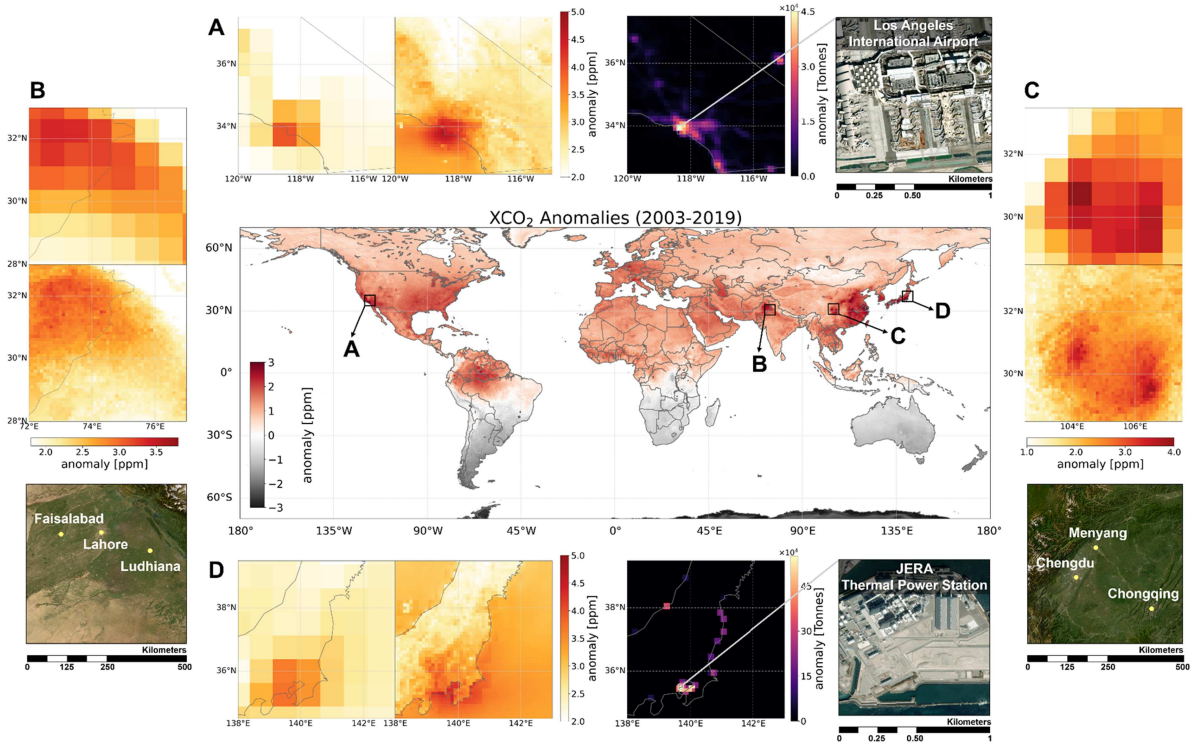

Figure 8. Global long-term XCO_2 anomaly map (2003–2019), with black boxes marking selected high-anomaly regions in the middle of figure: A. Los Angeles, B. Punjab, C. Sichuan Basin, and D. Tokyo. Oceanic regions are omitted here to enhance the visualization of global land-based variations. For each region, zoom-in comparisons are shown using CAMS-EGG4, LGBM-based product, and high-resolution Esri World Imagery basemap (0.5 m resolution for the United States and 1 m resolution for the other regions). In addition, Emissions Database for Global Atmospheric Research (EDGAR) sector-specific CO_2 emission anomaly maps are shown for regions A (Transport sector) and D (Power Industry sector) only. All regional zoom-in maps cover the same spatial extent of 500 km × 500 km. For regions A and D, the Esri World Imagery is further zoomed in to 1 km × 1 km to highlight specific emission facilities, with the corresponding locations marked by light-grey boxes in the EDGAR anomaly maps. Map scales are annotated for interpretation. Note that the XCO_2 anomaly values are identical between global and zoom-in map; the different colormaps are used for visual clarity.

identification of discrete urban hotspots, revealing localized sources with finer spatial detail. In Punjab, emissions are concentrated in major cities such as Faisalabad, Lahore, and Ludhiana (Tariq and Ali 2017; Lei et al. 2021). In the Sichuan Basin, the LGBM-based product captured distinct high-emission regions such as Chengdu, Chongqing, and Menyang (Cao and Yuan 2019; Zeng et al. 2022). In summary, these results highlight the LGBM-based product's capability to identify suburban-scale to facility-level emission features that may be influenced by nearby sources, with greater spatial precision than conventional reanalysis data.

4.5. Interannual variability and long-term trends of global XCO_2

Leveraging the global, gap-free, and high-resolution of our reconstructed LGBM-based XCO_2 product, we investigated interannual variability and long-term regional trends from 2003 to 2019. The reconstructed XCO_2 exhibits a long-term growth rate of 2.10 ppm/yr, closely comparable to the NOAA reference value of 2.16 ppm/yr, despite the NOAA product being based on a limited set of background stations rather than global, full coverage observations. As shown in Figure 9a, XCO_2 increased consistently across all regions, with global growth rates ranging from 2.016 to 2.278 ppm/yr. Notably, high growth rates (≥2.20 ppm/yr) were identified over East Asia and Central Africa. In East Asia, this pattern is consistent with concentrated anthropogenic emissions and limited atmospheric transport (Williams et al. 2007; Park et al. 2024). In contrast, the high values in Central Africa are more plausibly linked to biospheric influences such as wildfire activities, although multiple natural and anthropogenic drivers may contribute (Figure S17). Figure 8b illustrates the regional XCO_2

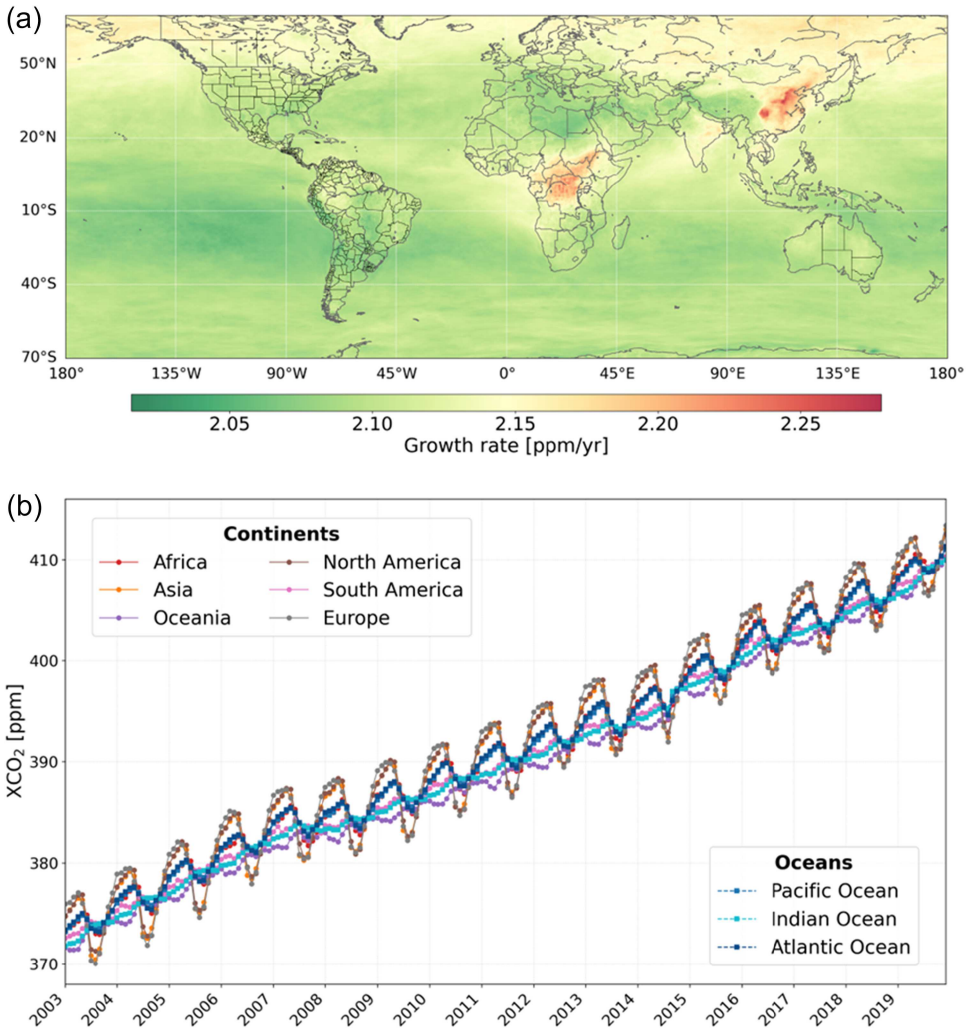

Figure 9. (a) Spatial distribution of XCO_2 growth rates (ppm/yr) from 2003 to 2019. Pixel-wise interannual trends were derived by fitting a linear polynomial to the monthly LGBM-based XCO_2 product at each 0.1° grid cell. (b) Regional mean XCO_2 time series over six continents and three oceans, computed from the reconstructed LGBM-based XCO_2 product. The continental and oceanic masks used for aggregation are defined in Figure S18.

concentrations aggregated over six continents and three ocean basins (Figure S18). All continents showed a steady increase over the 17 years, with Asia, Europe, and North America showing relatively higher concentrations due to their early industrialization and persistent emission sources (Friedlingstein et al. 2022). In contrast, oceanic regions show more stable growth and smoother seasonal cycles, which are attributed to the buffering capacity of oceanic CO_2 uptake (Gruber et al. 2019). Interestingly, the Pacific Ocean shows a stronger seasonal amplitude than some Southern Hemisphere land regions such as Oceania and South America. This enhanced amplitude may reflect a combination of long-range transportation from the Northern Hemisphere and strong seasonal ocean-atmosphere fluxes in the mid-latitudes (Landschützer et al. 2018).

To further examine the large-scale climatic influence of oceanic XCO_2 variability, we analyzed the relationship between XCO_2 anomalies and ENSO. ENSO modulates CO_2 fluxes primarily through changes in sea surface temperature (SST), vertical mixing, and biological productivity. During the El Niño phase, reduced equatorial upwelling and warmer SSTs decrease oceanic outgassing, though this is often offset by increased terrestrial emissions from droughts and fires. This is reflected in Figure 10a, where positive XCO_2 anomalies dominate the equatorial central and eastern Pacific, indicating reduced oceanic uptake (Cai et al. 2018). In contrast, La Niña events strengthen upwelling and can promote CO_2 drawdown in some oceanic regions, moderating the rate of XCO_2 increase (Sun, Liao, and Zhu 2025). As shown in Figure 10b, negative anomalies appear across much of the

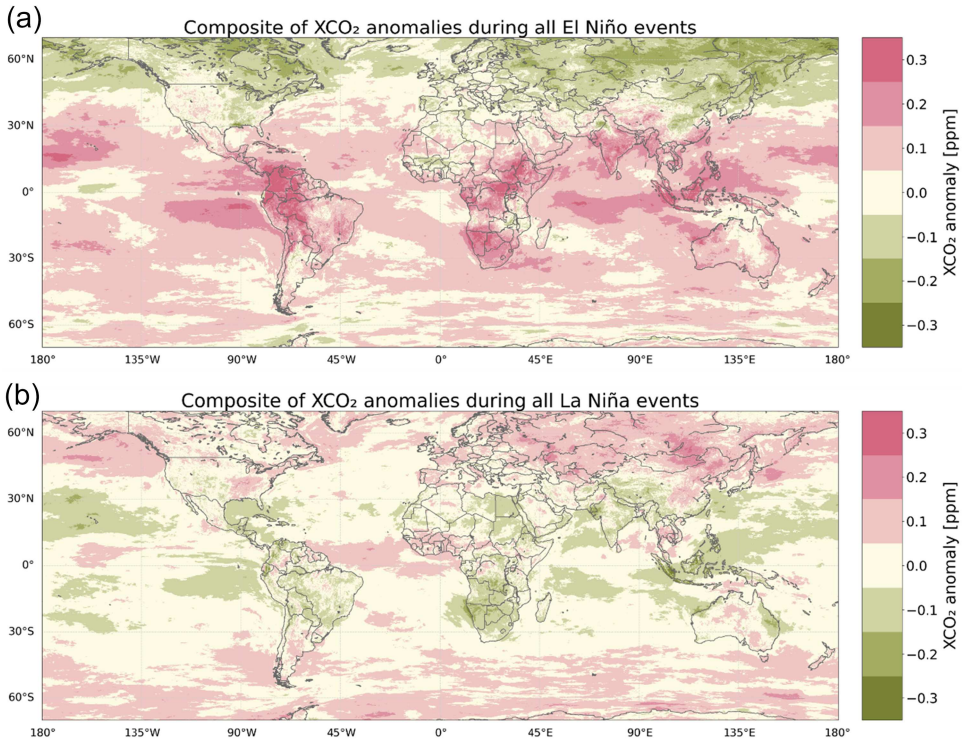

Figure 10. Spatial composite maps of detrended and deseasonalized oceanic XCO_2 anomalies during (a) El Niño and (b) La Niña periods. El Niño–Southern Oscillation (ENSO) phases were defined using the Niño3.4 index from NOAA Physical Sciences Laboratory (El Niño: > + 0.5 °C; La Niña: <−0.5 °C) over the period 2003–2019.

same region, where El Niño-related positive anomalies had reached up to 0.3 ppm, indicating strengthened oceanic uptake (Rödenbeck et al. 2015). Overall, the reconstructed XCO_2 product effectively captures both the long-term global increase in atmospheric CO_2 and interannual variability associated with large-scale climate phenomena such as ENSO. These results highlight our LGBM-based XCO_2 product's reliability in characterizing carbon cycle dynamics and its applicability for advancing research on climate-driven variability, regional emission attribution, and coupled ocean–atmosphere processes.

4.6. Comparison with previous AI-based XCO_2 studies

As shown in Table 3, many previous studies have attempted to reconstruct a gap-free XCO_2 product using various artificial intelligence (AI)-based approaches. While deep learning models have demonstrated strong performance, tree-based ML approaches also yielded comparable accuracy, despite their simpler architecture. One notable limitation is that much of the existing literature has been geographically limited to China or East Asia and covers relatively short durations, focusing on the satellite operation period (He et al. 2023; Li, Wu, and Wang 2023; Cai et al. 2024; He, Wang, and Wang 2024; Te et al. 2024; Wu et al. 2024; Li et al. 2025). Some global-scale studies, such as Zhang et al. (2023), have focused primarily on land regions, thereby excluding the oceanic domain and limiting their applicability to global carbon cycle assessments. More recently, a few studies have aimed to produce global, long-term XCO_2 products (Mustafa and Xu 2025; Guan et al. 2024a). Nonetheless, these still face limitations. While they claim to produce high-resolution outputs, the direct use of coarse-resolution CTM products (e.g. CAMS-EGG4 [0.75°] and CarbonTracker [3° × 2°]) as inputs to AI models imposes a fundamental limitation on capturing true spatial detail. For example, Guan et al. (2024a) achieved high accuracy (R^2 = 0.89–0.97, RMSE = 0.55–0.82 ppm) with a rigorous internal validation method, yet their model highly depends on CAMS-EGG4 and CarbonTracker, which together contributed over 70% of the feature importance. Similarly, while Mustafa and Xu (2025) employed a CNN-LSTM model to generate a high-quality product; their aim is to enhance the CAMS-EGG4 reanalysis itself. Consequently, the

Table 3. Summary of previous studies reconstructing full-coverage XCO_2 products using satellite, chemical transport models (CTM), and artificial intelligence (AI). The table compares model types, study domain, spatiotemporal resolutions, internal and external validation results using metrics of R^2 and RMSE, and data accessibility. The full names of model types are listed below the table. Internal validation results are annotated with superscripts indicating the validation method. External validation metrics are based on comparisons with TCCON observations; when studies reported site-level results, a range is presented.

				Internal validation		External validation			
*Model	Study area	Study period	Resolution (Temporal, Spatial)	R^2	RMSE (ppm)	R^2	RMSE (ppm)	Literature	Data accessibility
RF	China	2016–2018	Daily, 0.1°	0.88[a]	1.12[a]	–	–	He et al. (2023)	N
DNN	China	2015–2020	Daily, 0.1°	0.90–0.96[a,b,c]	0.97–1.55[a,b,c]	0.87	1.71	Li, Wu, and Wang (2023)	Y
DF	Global (Land)	2014–2020	Monthly, 0.1°	0.96[a,b]	0.98–1.07[a,b]	0.96	1.01	Zhang et al. (2023)	Y
DF	China	2015–2022	Monthly, 0.05°	0.90[a]	2.18[a]	0.72–0.81	1.77–2.01	Cai et al. (2024)	N
RF	China	2015–2020	Daily, 0.05°	0.93–0.97[a,b,c]	0.92–1.38[a,b,c]	0.82–0.94	1.4–2.1	He, Wang, and Wang (2024)	N
XGBoost	East Asia	2015–2020	Monthly, 0.1°	0.94[e]	1.27[e]	0.81–0.95	1.17–2.20	Te et al. (2024)	N
ST-ResNet	China	2015–2020	Daily, 0.01°	0.84–0.93[a,b,c]	0.66–1.00[a,b,c]	0.88	1.52	Wu et al. (2024)	N
Stacking regression	Global	2000–2020	8-day, 0.05°	0.89–0.97[a,b,d]	0.55–0.82[a,b,d]	0.95	1.06	Guan et al. (2024a)	Y
CNN-LSTM	Global	2003–2018	3-hour, 0.75°	0.99[e]	0.78[e]	0.99	0.89	Mustafa and Xu (2025)	N
DSTM	China	2015–2020	Daily, 0.1°	0.98[e]	0.72[e]	0.77–0.89	1.89–2.44	Li et al. (2025)	N
LGBM	**Global**	**2003–2019**	**Monthly, 0.1°**	**0.93–0.96[a,b,c]**	**0.80–1.11[a,b,c]**	**0.98**	**1.17**	**This study**	**Y**

*Model: RF (Random Forest), DNN (Deep Neural Network), DF (Deep Forest), ST-ResNet (Spatiotemporal ResNet), XGBoost (eXtreme Gradient Boosting), CNN-LSTM (Convolutional Neural Network–Long Short-Term Memory), DSTM (Deep learning-based Spatio-Temporal Model).

[a] Sample-based (Random) *N*-fold cross-validation results.

[b]Space-based *N*-fold cross-validation results.

[c]Time-based *N*-fold cross-validation results.

[d]Temporal hold-out validation results.

[e]Random hold-out validation results.

spatiotemporal resolution remained unchanged, limiting the product's applicability for analyzing fine-scale spatial variability in XCO_2.

In contrast, this study proposed a novel approach with distinct features compared to previous methods. Our novel ML framework reconstructs global, gap-free XCO_2 at a monthly temporal and 0.1° spatial resolution over 17 years (2003–2019), extending well prior to the operational period of the OCO-2 satellite. Unlike previous approaches, our model was developed without strong dependence on CTM-dominant predictors, enabling truly finer-scale products. Hence, the final product is publicly available, supporting diverse research directions in carbon–climate interactions, emission source attribution, and long-term carbon budget assessments.

5. Conclusions

This study developed a novel residual learning LGBM framework to reconstruct a long-term (2003–2019), high-resolution (0.1°), and gap-free global monthly XCO_2 product. By explicitly modeling the residuals between OCO-2 observations and CAMS-EGG4 reanalysis, the proposed framework effectively reduces dependence on the coarse spatial resolution of CAMS-EGG4 reanalysis and mitigates out-of-range overestimation due to mismatches between the training and reconstruction periods. Comprehensive internal validation using random, spatial, and temporal CV schemes demonstrated high predictive accuracy and generalizability, confirming its stability across diverse spatiotemporal domains. External validation with independent in situ observations from TCCON as well as WDCGG further confirmed the high quality of the reconstructed product, even in regions and times not covered by the OCO-2 satellite. Relative to the CAMS-EGG4 reanalysis, the LGBM-based XCO_2 product provides an effective enhancement in spatial resolution by a factor of approximately 7.5, enabling improved characterization of regional-scale variability and long-term global carbon cycle dynamics over both land and ocean.

Despite these advances, several limitations should be acknowledged. Polar regions above 70 ° latitude were excluded due to degraded satellite retrieval quality, and sparse sampling at high latitudes may introduce systematic uncertainties. In addition, the applied bias correction relies on a single adjustment term derived from an unevenly distributed station network, which may restrict its geographical representativeness in certain regions. The reconstructed XCO_2 product is also subject to uncertainties from multiple sources, including satellite retrieval errors, spatiotemporal preprocessing, and the ML modeling process itself. Future work could address these limitations by incorporating explainable or uncertainty-aware machine-learning approaches to better quantify regional uncertainties and attribute error sources, as well as by extending the framework to finer temporal resolutions as additional observations become available. Nevertheless, by bridging observational gaps with reduced dependence on reanalysis data, the resulting XCO_2 product offers a valuable resource for emission monitoring, carbon cycle research, and climate policy–relevant applications.

Acknowledgements

This research was supported by Korea Environment Industry & Technology Institute (KEITI) through Project for developing an observation-based GHG emissions geospatial information map, funded by Korea Ministry of Environment (MOE) (RS-2023-00232066). **Soomin Hwang** was partially supported by Basic Science Research Programme through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2024-00460619).

Author contributions

CRediT: **Soomin Hwang:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing; **Hyunyoung Choi:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing; **Yoojin Kang:** Formal analysis, Methodology, Writing – original draft, Writing – review & editing; **Jungho Im:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Research Foundation of Korea (RS-2024-00460619), and Korea Ministry of Environment (MOE) (RS-2023-00232066).

ORCID

Soomin Hwang 0009-0003-7778-0738
Hyunyoung Choi 0000-0002-5913-4607
Yoojin Kang 0000-0003-3006-6994
Jungho Im 0000-0002-4506-6877

Data availability statement

All raw data utilized in this study are publicly accessible through the links listed below. The final processed results generated in this study have been publicly available via Zenodo (Hwang et al. 2025), while the processed datasets and codes used in this study can be made available upon request to the corresponding author.

- OCO-2 satellite XCO_2 data: https://search.earthdata.nasa.gov/.
- CAMS-EGG4 reanalysis data: https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-ghg-reanalysis-egg4?tab=overview.
- ECMWF ERA5 data: https://cds.climate.copernicus.eu/.
- ODIAC: https://www.cger.nies.go.jp/en/.
- GLOBIO: https://www.globio.info/.
- SEDAC: https://sedac.ciesin.columbia.edu/.
- TCCON in situ data: https://tccondata.org.
- WDCGG in situ data: https://gaw.kishou.go.jp.
- EDGAR: https://edgar.jrc.ec.europa.eu.
- GFED: https://www.globalfiredata.org.

The final results generated in this study are available at https://doi.org/10.5281/zenodo.15771923 (Hwang et al. 2025).

References

Agustí-Panareda, A., J. Barré, S. Massart, A. Inness, I. Aben, M. Ades, B. C. Baier, et al. 2023. "Technical Note: the CAMS Greenhouse Gas Reanalysis from 2003 to 2020." *Atmospheric Chemistry and Physics* 23: 3829–3859. https://doi.org/10.5194/acp-23-3829-2023.

Cai, K., L. Guan, S. Li, S. Zhang, Y. Liu, and Y. Liu. 2024. "Full-Coverage Estimation of Co_2 Concentrations in China Via Multisource Satellite Data and Deep Forest Model." *Scientific Data* 11: 1231. https://doi.org/10.1038/s41597-024-04063-9.

Cai, W., G. Wang, B. Dewitte, L. Wu, A. Santoso, K. Takahashi, Y. Yang, A. Carréric, and M. J. McPhaden. 2018. "Increased Variability of Eastern Pacific El Niño under Greenhouse Warming." *Nature* 564: 201–206. https://doi.org/10.1038/s41586-018-0776-9.

Cao, W., and X. Yuan. 2019. "Region-County Characteristic of Spatial-Temporal Evolution and Influencing Factor on Land Use-Related Co_2 Emissions in Chongqing of China, 1997–2015." *Journal of Cleaner Production* 231: 619–632. https://doi.org/10.1016/j.jclepro.2019.05.248.

Choi, H., S. Park, Y. Kang, J. Im, and S. Song. 2023. "Retrieval of Hourly $Pm_{2.5}$ Using Top-Of-Atmosphere Reflectance from Geostationary Ocean Color Imagers I and Ii." *Environmental Pollution* 323: 121169. https://doi.org/10.1016/j.envpol.2023.121169.

Crippa, M., D. Guizzardi, F. Pagani, M. Banja, M. Muntean, E. Schaaf, W. Becker, F. Monforti-Ferrario, R. Quadrelli, and A. Risquez Martin. 2023. *GHG emissions of all world countries*. 10: 953322. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/173513

Custódio, D., C. Borrego, and H. Relvas. 2022. "Worldwide Evaluation of CAMS-EGG4 Co_2 Data Re-Analysis at the Surface Level." *Toxics* 10: 331. https://doi.org/10.3390/toxics10060331.

Cusworth, D. H., A. K. Thorpe, C. E. Miller, A. K. Ayasse, R. Jiorle, R. M. Duren, R. Nassar, J.-P. Mastrogiacomo, and R. R. Nelson. 2023. "Two Years of Satellite-Based Carbon Dioxide Emission Quantification at the world's Largest Coal-Fired Power Plants." *Atmospheric Chemistry and Physics* 23: 14577–14591. https://doi.org/10.5194/acp-23-14577-2023.

Friedlingstein, P., M. W. Jones, M. O'Sullivan, R. M. Andrew, D. C. Bakker, J. Hauck, C. Le Quéré, G. P. Peters, W. Peters, and J. Pongratz. 2022. "Global Carbon Budget 2021." *Earth System Science Data* 14: 1917–2005. https://doi.org/10.5194/essd-14-1917-2022.

Gruber, N., D. Clement, B. R. Carter, R. A. Feely, S. Van Heuven, M. Hoppema, M. Ishii, R. M. Key, A. Kozyr, and S. K. Lauvset. 2019. "The Oceanic Sink for Anthropogenic Co_2 from 1994 to 2007." *Science (New York, N.Y.)* 363: 1193–1199. https://doi.org/10.1126/science.aau5153.

Guan, Y., G. A. McKinley, A. R. Fay, S. C. Doney, and G. Keppel-Aleks. 2024b. "Ocean-Driven Interannual Variability in Atmospheric Co_2 Quantified Using OCO-2 Observations and Atmospheric Transport Simulations." *Frontiers in Marine Science* 11: 1272415. https://doi.org/10.3389/fmars.2024.1272415.

Guan, X., Z. Sun, D. Chu, G. Xie, Y. Wang, and H. Shen. 2024a. "Long-Term (2000–2020) Global 0.05° Continuous Atmospheric Carbon Dioxide Mapping Combining OCO-2 Observations and Model Simulations." *Science of the Total Environment* 957: 177051. https://doi.org/10.1016/j.scitotenv.2024.177051.

Hakkarainen, J., I. Ialongo, S. Maksyutov, and D. Crisp. 2019. "Analysis of Four Years of Global Xco_2 Anomalies as Seen By Orbiting Carbon Observatory-2." *Remote Sensing* 11: 850. https://doi.org/10.3390/rs11070850.

He, J., W. Wang, and N. Wang. 2024. "Seamless Reconstruction and Spatiotemporal Analysis of Satellite-Based Xco_2 Incorporating Temporal Characteristics: a Case Study in China during 2015&Ndash;2020." *Advances in Space Research* 74: 3804–3825. https://doi.org/10.1016/j.asr.2024.07.007.

He, S., Y. Yuan, Z. Wang, L. Luo, Z. Zhang, H. Dong, and C. Zhang. 2023. "Machine Learning Model-Based Estimation of Xco_2 with High Spatiotemporal Resolution in China." *Atmosphere* 14: 436. https://doi.org/10.3390/atmos14030436.

He, C., M. Ji, T. Li, X. Liu, D. Tang, S. Zhang, Y. Luo, M. L. Grieneisen, Z. Zhou, and Y. Zhan. 2022. "Deriving Full-Coverage and Fine-Scale Xco_2 Across China Based on OCO-2 Satellite Retrievals and CarbonTracker Output." *Geophysical Research Letters* 49(12): e2022GL098435. https://doi.org/10.1029/2022GL098435.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., and Rozum, I. 2023. ERA5 Monthly Averaged Data on Single Levels from 1940 to Present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Accessed February 4, 2026. https://doi.org/10.24381/cds.f17050d7.

Heymann, J., M. Reuter, M. Hilker, M. Buchwitz, O. Schneising, H. Bovensmann, J. P. Burrows, A. Kuze, H. Suto, and N. M. Deutscher. 2015. "Consistent Satellite Xco_2 Retrievals from SCIAMACHY and GOSAT Using the BESD Algorithm." Atmospheric Measurement Techniques 8: 2961–2980. https://doi.org/10.5194/amt-8-2961-2015.

Hong, J., J. Kim, Y. Jung, W. Kim, H. Lim, S. Jeong, and S. Lee. 2023. "Potential Improvement of Xco_2 Retrieval of the OCO-2 By Having Aerosol Information from the A-train Satellites." *GISci. Remote Sens.* 60: 2209968. https://doi.org/10.1080/15481603.2023.2209968.

Hossain, M. F.2022. "Extreme Level of Co_2 Accumulation into the Atmosphere Due to the Unequal Global Carbon Emission and Sequestration." *Water, Air and Soil Pollution* 233: 105. https://doi.org/10.1007/s11270-022-05581-1.

Hwang, S., H. Choi, Y. Kang, and J. Im. 2025. *Global gap-free monthly 0.1° XCO2 product from 2003 to 2019*. Zenodo, [data set]. https://doi.org/10.5281/zenodo.15771923

Inness, A., M. Ades, A. Agustí-Panareda, J. Barré, A. Benedictow, A.-M. Blechschmidt, J. J. Dominguez, R. Engelen, H. Eskes, and J. Flemming. 2019. "The CAMS Reanalysis of Atmospheric Composition." *Atmospheric Chemistry and Physics* 19: 3515–3556. https://doi.org/10.5194/acp-19-3515-2019.

Jin, C., Y. Xue, X. Jiang, L. Zhao, T. Yuan, Y. Sun, S. Wu, and X. Wang. 2022. "A Long-Term Global Xco_2 Dataset: Ensemble of Satellite Products." *Atmospheric Research* 279: 106385. https://doi.org/10.1016/j.atmosres.2022.106385.

Jung, J.-A., Y. Cho, S. Lee, and M.-J. Lee. 2024. "Actions to Expand the Use of Geospatial Data and Satellite Imagery for Improved Estimation of Carbon Sinks in the LULUCF Sector." *Korean J. Remote Sensing* 40: 203–217. https://doi.org/10.7780/kjrs.2024.40.2.7.

Kang, E., D. Cho, S. Lee, J. Im, D. Lee, and C. Yoo. 2024. "An Explainable AI Framework for Spatiotemporal Risk Factor Analysis in Public Health: A Case Study of Cardiovascular Mortality in South Korea." *GISci. Remote Sens.* 61: 2436997. https://doi.org/10.1080/15481603.2024.2436997.

Keeling, C. D., R. B. Bacastow, A. E. Bainbridge, C. A. EkdahlJr, P. R. Guenther, L. S. Waterman, and J. F. Chin. 1976. "Atmospheric Carbon Dioxide Variations at Mauna Loa Observatory, Hawaii, Tellus Ser. A-Dyn. Meteorol. Oceanogr." *Tellus A: Dynamic Meteorology and Oceanography* 28: 538–551. https://doi.org/10.3402/tellusa.v28i6.11322.

Keppel-Aleks, G., A. S. Wolf, M. Mu, S. C. Doney, D. C. Morton, P. S. Kasibhatla, J. B. Miller, E. J. Dlugokencky, and J. T. Randerson. 2014. "Separating the Influence of Temperature, Drought, and Fire on Interannual Variability in Atmospheric Co_2." *Global Biogeochemical Cycles* 28: 1295–1310. https://doi.org/10.1002/2014GB004890.

Kira, O., and Y. Sun. 2020. "Extraction of Sub-Pixel C3/C4 Emissions of Solar-Induced Chlorophyll Fluorescence (SIF) Using Artificial Neural Network." *ISPRS-J. Photogramm. Remote Sens.* 161: 135–146. https://doi.org/10.1016/j.isprsjprs.2020.01.017.

Landschützer, P., N. Gruber, D. C. Bakker, I. Stemmler, and K. D. Six. 2018. "Strengthening Seasonal Marine Co_2 Variations Due to Increasing Atmospheric Co_2." *Nature Climate Change* 8: 146–150. https://doi.org/10.1038/s41558-017-0057-x.

Laughner, J. L., G. C. Toon, J. Mendonca, C. Petri, S. Roche, D. Wunch, J.-F. Blavier, D. W. Griffith, P. Heikkinen, and R. F. Keeling. 2024. "The Total Carbon Column Observing Network's GGG2020 Data Version." *Earth Syst. Sci. Data* 16: 2197–2260. https://doi.org/10.5194/essd-16-2197-2024.

Leegupta, H., K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. Thorne, C. Trisos, J. Romero, P. Aldunce, and K Barret IPCC. Climate Change 2023: Synthesis Report, Summary for Policymakers In *Contribution of working groups i, II and III to the sixth assessment report of the intergovernmental panel on climate change [core writing team*. Lee, h. and j Romero, eds. geneva, Switzerland: IPCC.

Lei, R., S. Feng, A. Danjou, G. Broquet, D. Wu, J. C. Lin, C. W. O'Dell, and T. Lauvaux. 2021. "Fossil Fuel Co_2 Emissions over Metropolitan Areas from Space: a Multi-Model Analysis of OCO-2 Data over Lahore, Pakistan." *Remote Sensing of the Environment* 264: 112625. https://doi.org/10.1016/j.rse.2021.112625.

Li, T., J. Wu, and T. Wang. 2023. "Generating Daily High-Resolution and Full-Coverage Xco_2 Across China from 2015 to 2020 Based on OCO-2 and CAMS Data." *Science of the Total Environment* 893: 164921. https://doi.org/10.1016/j.scitotenv.2023.164921.

Li, Y., J. Yan, L. Zhong, D. Bao, L. Sun, and G. Li. 2025. "Full-Coverage Mapping of Daily High-Resolution Xco_2 Across China from 2015 to 2020 By Deep Learning-Based Spatio-Temporal Fusion." *IEEE Trans. Geosci. Remote Sensing* 63: 4102716. https://doi.org/10.1109/TGRS.2025.3540289.

Luo, J., J. Zhang, X. Huang, Q. Liu, B. Luo, W. Zhang, Z. Rao, and Y. Yu. 2020. "Characteristics, Evolution, and Regional Differences of Biomass Burning Particles in the Sichuan Basin, China." *Journal of Environmental Sciences China* 89: 35–46. https://doi.org/10.1016/j.jes.2019.09.015.

Meng, X., S. Li, K. Akhmadi, P. He, and G. Dong. 2024. "Trends, Turning Points, and Driving Forces of Desertification in Global Arid Land Based on the Segmental Trend Method and SHAP Model." *GISci. Remote Sens.* 61: 2367806. https://doi.org/10.1080/15481603.2024.2367806.

Muñoz-Sabater, J., E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, and H. Hersbach. 2021. "ERA5-Land: a State-Of-The-Art Global Reanalysis Dataset for Land Applications." *Earth System Science Data* 13: 4349–4383. https://doi.org/10.5194/essd-13-4349-2021.

Mustafa, F., and M. Xu. 2025. "Improving Model-Based Carbon Dioxide Datasets Using Deep Learning and Satellite Observations." *IEEE Trans. Geosci. Remote Sensing* 63: 1–14. https://doi.org/10.1109/TGRS.2025.3556309.

Oda, T., and S. Maksyutov. 2011. "A Very High-Resolution (1 km×1 Km) Global Fossil Fuel Co_2 Emission Inventory Derived Using a Point Source Database and Satellite Observations of Nighttime Lights." Atmospheric Chemistry and Physics 11: 543–556. https://doi.org/10.5194/acp-11-543-2011.

Park, S., H. Lee, and J. Park. 2025. "Estimation of Surface Nitrogen Dioxide Volume Mixing Ratios over South Korea from GEMS Observations." *Korean J. Remote Sensing* 41: 291–300. https://doi.org/10.7780/kjrs.2025.41.2.1.4.

Park, H., B.-I Lee, E.-H. Sohn, and J. Y. Kim. 2024. "Seasonal Variability of Solar-Induced Fluorescence and Correlation with Greenhouse Gas Across Different Land Cover in East Asia." *Korean J. Remote Sensing* 40: 579–588. https://doi.org/10.7780/kjrs.2024.40.5.1.13.

Pitt, J. R., I. Lopez-Coto, K. D. Hajny, J. Tomlin, R. Kaeser, T. Jayarathne, B. H. Stirm, et al. 2022. "New York City Greenhouse Gas Emissions Estimated with Inverse Modeling of Aircraft Measurements." *Elementa: Science of the Anthropocene* 10: 00082. https://doi.org/10.1525/elementa.2021.00082.

Pogson, M., and P. Smith. 2015. "Effect of Spatial Data Resolution on Uncertainty." *Environ. Model. Software* 63: 87–96. https://doi.org/10.1016/j.envsoft.2014.09.021.

Ramonet, M., Langerock, B., Warneke, T., and Eskes, H. J. 2021. Validation report of the CAMS greenhouse gas global reanalysis, years 2003-2020, Copernicus Atmosphere Monitoring Service (CAMS) report CAMS84_2018SC3_D5.1.2-2020_v0.1.pdf. https://doi.org/10.24380/438c-4597

Rödenbeck, C., D. C. Bakker, N. Gruber, Y. Iida, A. R. Jacobson, S. Jones, P. Landschützer, N. Metzl, S. Nakaoka, and A. Olsen. 2015. "Data-Based Estimates of the Ocean Carbon Sink variability–first Results of the Surface Ocean Pco_2 Mapping Intercomparison (Socom)." *Biogeosciences* 12: 7251–7278. https://doi.org/10.5194/bg-12-7251-2015.

Sha, M. K., M. De Mazière, J. Notholt, T. Blumenstock, H. Chen, A. Dehn, D. W. Griffith, F. Hase, P. Heikkinen, and C. Hermans. 2019. "Intercomparison of Low and High Resolution Infrared Spectrometers for Ground-Based Solar Remote Sensing Measurements of Total Column Concentrations of Co_2, Ch_4 and Co." *Atmos. Meas. Tech. Discuss.* 2019: 1–67. https://doi.org/10.5194/amt-13-4791-2020.

Sheng, M., L. Lei, Z.-C. Zeng, W. Rao, and S. Zhang. 2021. "Detecting the Responses of Co_2 Column Abundances to Anthropogenic Emissions from Satellite Observations of GOSAT and Oco-2." *Remote Sensing* 13: 3524. https://doi.org/10.3390/rs13173524.

Sheng, M., L. Lei, Z.-C. Zeng, W. Rao, H. Song, and C. Wu. 2023. "Global Land 1° Mapping Dataset of Xco_2 from Satellite Observations of GOSAT and OCO-2 from 2009 to 2020." *Big Earth Data* 7: 170–190. https://doi.org/10.1080/20964471.2022.2033149.

Snoek, J., H. Larochelle, and R. P. Adams. 2012. "Practical Bayesian Optimization of Machine Learning Algorithms." *Advances in neural information processing systems* 25. https://doi.org/10.48550/arXiv.1206.2944.

Sun, C., E. Liao, and X. Zhu. 2025. "Asymmetrical Ocean Carbon Responses in the Tropical Pacific Ocean to La Niña and El NiÑO." *Geophysical Research Letters* 52: e2024GL112039. https://doi.org/10.1029/2024GL112039.

Tariq, S., and M. Ali. 2017. "Spatiotemporal Assessment of Co_2 Emissions and Its Satellite Remote Sensing over Pakistan and Neighboring Regions." *Journal of Atmospheric and Solar-Terrestrial Physics* 152: 11–19. https://doi.org/10.1016/j.jastp.2016.11.001.

Te, T., C. Bao, H. Bagan, Y. Xie, M. Che, T. Yoshida, and B. Uudus. 2024. "Mapping Seamless Monthly Xco_2 in East Asia: Utilizing OCO-2 Data and Machine Learning." *International Journal of Applied Earth Observation and Geoinformation* 133: 104117. https://doi.org/10.1016/j.jag.2024.104117.

Thoning, K. W., P. P. Tans, and W. D. Komhyr. 1989. "Atmospheric Carbon Dioxide at Mauna Loa Observatory: 2. Analysis of the NOAA GMCC Data, 1974&Ndash;1985." *Journal of Geophysical Research, [Atmospheres]* 94: 8549–8565. https://doi.org/10.1029/JD094iD06p08549.

Trishchenko, A. P., L. Garand, and L. D. Trichtchenko. 2019. "Observing Polar Regions from Space: Comparison between Highly Elliptical Orbit and Medium Earth Orbit Constellations." *J. Atmos. Ocean. Technol.* 36: 1605–1621. https://doi.org/10.1175/JTECH-D-19-0030.1.

Wang, Y., Q. Yuan, T. Li, Y. Yang, S. Zhou, and L. Zhang. 2023. "Seamless Mapping of Long-Term (2010–2020) Daily Global Xco_2 and Xch_4 from the Greenhouse Gases Observing Satellite (GOSAT), Orbiting Carbon Observatory 2 (OCO-2), and CAMS Global Greenhouse Gas Reanalysis (CAMS-EGG4) with a Spatiotemporally Self-Supervised Fusion Method." *Earth Syst. Sci. Data* 15: 3597–3622. https://doi.org/10.5194/essd-15-3597-2023.

Williams, C. A., N. P. Hanan, J. C. Neff, R. J. Scholes, J. A. Berry, A. S. Denning, and D. F. Baker. 2007. "Africa and the Global Carbon Cycle." *Carbon Balanc. Manag.* 2: 1–13. https://doi.org/10.1186/1750-0680-2-3.

Wu, C., S. Yang, D. Jiao, Y. Chen, J. Yang, and B. Huang. 2024. "Estimation of Daily Xco_2 at 1 Km Resolution in China Using a Spatiotemporal ResNet Model." *Science of the Total Environment* 954: 176171. https://doi.org/10.1016/j.scitotenv.2024.176171.

Wunch, D., G. C. Toon, J.-F. L. Blavier, R. A. Washenfelder, J. Notholt, B. J. Connor, D. W. Griffith, V. Sherlock, and P. O. Wennberg. 2011. "The Total Carbon Column Observing Network." *Philos. T. Roy. Soc. A* 369: 2087–2112. https://doi.org/10.1098/rsta.2010.0240.

Yang, H., T. Li, J. Wu, and L. Zhang. 2025. "Inter-Comparison and Evaluation of Global Satellite Xco_2 Products." *Geo-Spat. Inf. Sci.* 28: 131–144. https://doi.org/10.1080/10095020.2023.2252017.

Zeng, H., B. Shao, G. Bian, H. Dai, and F. Zhou. 2022. "Analysis of Influencing Factors and Trend Forecast of Co_2 Emission in Chengdu-Chongqing Urban Agglomeration." *Sustainability* 14: 1167. https://doi.org/10.3390/su14031167.

Zhang, M., and G. Liu. 2023. "Mapping Contiguous Xco_2 By Machine Learning and Analyzing the Spatio-Temporal Variation in China from 2003 to 2019." *Science of the Total Environment* 858: 159588. https://doi.org/10.1016/j.scitotenv.2022.159588.

Zhang, L., T. Li, J. Wu, and H. Yang. 2023. "Global Estimates of Gap-Free and Fine-Scale Co_2 Concentrations during 2014–2020 from Satellite and Reanalysis Data." *Environment International* 178: 108057. https://doi.org/10.1016/j.envint.2023.108057.

Zhang, L., S. Rui, X. Zhiqiang, and X. Yao. 2025. "Assessing the Impacts of Urban Expansion and Climate Change on Net Primary Productivity over the Past Three Decades in Beijing, China." *GISci. Remote Sens.* 62: 2511503. https://doi.org/10.1080/15481603.2025.2511503.

Zheng, J., H. Zhang, and S. Zhang. 2023. "Comparison of Atmospheric Carbon Dioxide Concentrations Based on GOSAT, OCO-2 Observations and Ground-Based TCCON Data." *Remote Sensing* 15: 5172. https://doi.org/10.3390/rs15215172.

Zhu, W., Z. Xie, C. Zhao, Z. Zheng, K. Qiao, D. Peng, and Y. H. Fu. 2024. "Remote Sensing of Terrestrial Gross Primary Productivity: a Review of Advances in Theoretical Foundation, Key Parameters and Methods." *GISci. Remote Sens.* 61: 2318846. https://doi.org/10.1080/15481603.2024.2318846.

Zhu, H., T. Cheng, X. Li, X. Ye, D. Fan, T. Tang, H. Tong, and L. Zhang. 2025. "Improving Xco_2 Retrieval under High Aerosol Loads with Fused Satellite Aerosol Data: Advancing Understanding of Anthropogenic Emissions." *ISPRS-J. Photogramm. Remote Sens.* 223: 146–158. https://doi.org/10.1016/j.isprsjprs.2025.03.009.

Zickfeld, K., D. Azevedo, S. Mathesius, and H. D. Matthews. 2021. "Asymmetry in the climate–carbon Cycle Response to Positive and Negative Co_2 Emissions." *Nature Climate Change* 11: 613–617. https://doi.org/10.1038/s41558-021-01061-2.