

Supplement of

A global, temporally continuous, hourly CO₂ concentration dataset at GAW monitoring stations for 2000-2024: Enabling fine-scale carbon dynamics analysis

Tan Mi¹, Xueting Pu², Xi Zheng¹, Guangcheng Nie¹, Xinyi Liu³, Han Zhang⁴, Michael L. Grieneisen⁵, Yu Zhan¹, Maohua Wang⁶, and Fumo Yang¹

¹College of Carbon Neutrality Future Technology, Sichuan University, Chengdu, Sichuan 610065, China

²Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA 16803, United States

³Department of Environmental Science and Engineering, Sichuan University, Chengdu, Sichuan 610065, China

⁴State Grid Sichuan Electric Power Research Institute, Chengdu, Sichuan, 610041, China

⁵Department of Land, Air, and Water Resources, University of California, Davis, CA 95616, United States

⁶Institute for Disaster Management and Reconstruction, Sichuan University-The Hong Kong Polytechnic University, Chengdu 610207, China

Correspondence to: Yu Zhan (yzhan@scu.edu.cn) and Maohua Wang (wmh@scu.edu.cn)

S1. Temporal interpolation of CT and CAMS CO₂ data from 3-hour to 1-hour resolution

To obtain hourly CO₂ fields from the original 3-hour model outputs, both CT (available at 01:00, 04:00, 07:00, 10:00, 13:00, 16:00, 19:00, and 22:00 UTC) and CAMS (available at 00:00, 03:00, 06:00, 09:00, 12:00, 15:00, 18:00, and 21:00 UTC) data were temporally interpolated to 1-hour resolution using a natural cubic spline approach. For each station and each target day, a moving temporal window spanning from the afternoon of the previous day to the morning of the following day was constructed to ensure sufficient boundary constraints for spline fitting. Specifically, CT data within a window from 13:00 UTC of the previous day to 10:00 UTC of the following day were used, while CAMS data were extracted from 12:00 UTC of the previous day to 09:00 UTC of the following day. Within each window, the original 3-hour CO₂ concentrations were expressed as a function of time (in hours relative to the target day), and a natural cubic spline function was fitted. The fitted spline was then evaluated at each hour (00:00-23:00 UTC) of the target day to generate continuous hourly CO₂ estimates. To evaluate the accuracy of this interpolation, a hold-out validation was performed within the same sliding window framework. For each window, one observation point was randomly excluded from the fitting process and used as a validation sample, representing approximately 10% of the total original data points. The results demonstrated a high degree of fidelity in the interpolation. For the CT dataset, the validation yielded an R^2 of 0.98 and an RMSE of 2.66 ppm. The CAMS validation showed slightly higher uncertainty but remained robust, with an R^2 of 0.88 and an RMSE of 5.53 ppm (Fig. S2). This validation confirms that the natural cubic spline effectively captures the sub-daily variations in model-derived CO₂ concentrations.

S2. CAMS CO₂ data filling process

To ensure temporal continuity across the entire study period, we addressed data gaps in the CAMS CO₂ product (covering 2000-2003 and 2021-February 2024) using a LightGBM model based on CT CO₂ data and the environmental variables. For the period with available CAMS data, the CAMS residual CO₂ (CO_{2r}) was defined as CAMS CO₂ after removing the CT-derived trend (CO_{2t}) and seasonal (CO_{2s}) components obtained from the GAM decomposition (S4). A LightGBM model was then trained to predict CAMS CO_{2r} for all missing CAMS time steps (Eq. (S1)).

$$CO_{2r} = f(X_1, X_2, \dots, X_n) \quad (S1)$$

where CO_{2r} represents the residual CAMS CO₂, $X_1, X_2, \dots, X_n$ correspond to climatic factors, anthropogenic activities, ecosystem processes, static and spatiotemporal proxy variables (Section 2.1.3); and f denotes the LightGBM model. The residual CAMS CO₂ in 2000-2003 and 2021-February 2024 were predicted using the f , and the final CAMS CO₂ concentrations were obtained by restoring the CT-derived trend and seasonal components (Eq. (S2)).

$$CO_2 = CCO_{2r} + CO_{2t} + CO_{2s} \quad (S2)$$

To evaluate the filling performance, we implemented two distinct data-splitting strategies grouped by station and calendar month: (1) random month-based splitting, in which months for each station were randomly divided into 80% training, 10% validation, and 10% testing subsets; and (2) chronological month-based splitting, in which the first 80% of months were used for training, the subsequent 10% for validation, and the final 10% for testing. The validation results demonstrated high fidelity in the data-filling process (Fig. S3). Under the random splitting strategy, the model achieved an R^2 of 0.91 and an RMSE of 6.00 ppm at the hourly scale, with performance further improving at the daily scale ($R^2 = 0.95$, RMSE = 3.82 ppm). Even under the more stringent chronological splitting, the model maintained robust performance ($R^2 = 0.83$ for hourly; $R^2 = 0.87$ for daily), confirming the reliability of the reconstructed CAMS dataset for long-term carbon cycle analysis.

S3. Data sources and descriptions for ancillary variables

The ancillary variables employed in this study primarily comprise three categories: anthropogenic activities, ecosystem processes, and meteorological factors (Table S1). Five anthropogenic variables were used to characterize human-related CO₂ emission activities: population density (POP), land use and land cover change (LUCC), nighttime lights (NTL), the Open-source Data Inventory for Anthropogenic CO₂ (ODIAC), and the Emissions Database for Global Atmospheric Research (EDGAR). The POP data were obtained from the Land Scan Global Population Dataset with a spatial resolution of 1 km (Sims et al., 2023). The LUCC information was derived from the Climate Change Initiative land cover datasets, which provide global land cover products at a spatial resolution of 300 m (Defourny et al., 2023). The NTL data were acquired from the VIIRS VNP46A2 Daily L3 Global 500 m product, which provides lunar gap-filled BRDF-corrected nighttime light observations, and were processed and downloaded through the Google Earth Engine (GEE) platform. In addition, monthly anthropogenic CO₂ emission data were extracted from ODIAC with 1 km spatial resolution (Oda and Maksyutov, 2015) and EDGAR with 0.1° spatial resolution (Crippa et al., 2023). Because the POP, LUCC, ODIAC, and EDGAR datasets were available only up to 2022, their 2022 values were carried forward to 2023-2024 to maintain temporally complete predictor fields.

The key factors selected to represent ecosystem dynamics were normalized difference vegetation index (NDVI), gross primary productivity (GPP), net primary production (NPP), fraction of photosynthetically active radiation (FPAR), and leaf area index (LAI). All these datasets were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) products and were downloaded and processed through the GEE platform. Specifically, NDVI was derived from the MOD13Q1.061 Terra Vegetation Indices product with a spatial resolution of 250 m and a temporal resolution of 16 days. GPP and NPP were obtained from the MOD17A3HGF.061 Terra NPP Gap-Filled Yearly Global product at 500 m resolution. FPAR and LAI were extracted from the MOD15A2H.061 Terra LAI/FPAR product, which provides 8-day composites at a spatial resolution of 500 m.

The selected climatic factors affecting CO₂ were hourly 2 m air temperature (TEM), 2 m dewpoint temperature (D2M), skin temperature (SKT), 10 m zonal wind component (U 10), 10 m meridional wind component (V 10), surface pressure (SP), total precipitation (PRE), evaporation (EVA), volumetric soil water content in layer 1 (VSWL), and surface net solar radiation (SSR). These variables were primarily obtained from the ERA5-Land reanalysis products developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) at a spatial resolution of 0.1°. For stations located in coastal or oceanic regions where ERA5-Land data are unavailable, the corresponding variables were supplemented using the ERA5 reanalysis product at a coarser spatial resolution of 0.25°, to ensure spatial completeness and consistency across all stations.

S4. CT CO₂ time series decomposition process

The CT CO₂ time series was decomposed into its constituent components using a Generalized Additive Model (GAM), which provides a flexible framework for capturing non-linear temporal variations (Wood, 2017; Simpson, 2018). The model was formulated as follows:

$$Y_t = \beta_0 + f_{trend}(Time_t) + f_{seasonal}(DOY_t) + \varepsilon_t \quad (S3)$$

where Y_t represents the CT CO₂ concentration at time t ; β_0 is the model intercept, representing the overall mean level of the series; and $f_{trend}(Time_t)$ is a smooth function representing the long-term trend (CO_{2t}), implemented using a thin plate regression spline ($bs = "tp"$). The basis dimension was set to $k = 8$ to provide sufficient flexibility for capturing non-linear, multi-year trajectories. $f_{seasonal}(DOY_t)$ is a smooth function representing the seasonal component (CO_{2s}), where DOY_t is day of the year. A cyclic cubic regression spline ($bs = "cc"$) with $k = 3$ was employed to model the annual cycle, ensuring smooth continuity between the end of one year and the beginning of the next. ε_t is the residual term (CO_{2r}). The model parameters, including the smoothing parameters for each spline, were estimated using Restricted Maximum Likelihood (REML). This approach effectively penalizes over-fitting, ensuring an optimal balance between characterizing the underlying signal and maintaining model parsimony.

S5. LightGBM implementation and hyperparameter optimization

LightGBM was used as the primary algorithm for reconstructing hourly GAW CO₂ concentrations. LightGBM is an efficient gradient boosting decision tree model that sequentially combines decision trees to capture nonlinear relationships and high-order interactions among predictors. This tree-ensemble structure is well suited to hourly CO₂ reconstruction because it can represent non-smooth, event-driven temporal variability and abrupt temporal fluctuations without imposing a globally smooth response function. Similar tree-based boosting approaches have also shown strong applicability to intermittent or irregular time-series forecasting, where sharp changes and heterogeneous temporal patterns are common (Anderer and Li, 2022; Ke et al., 2017). The use of histogram-based learning, leaf-wise tree growth, gradient-based one-side sampling, exclusive feature bundling, and L1/L2 regularization allows LightGBM to achieve high predictive performance while improving computational efficiency and controlling overfitting (Ke et al., 2017). We employed a grid search strategy to optimize the key hyperparameters of the LightGBM model, with the specific hyperparameter combinations for each imputation stage detailed in Table S2.

S6. Adapted WDCGG global analysis method

To compare the reconstructed hourly CO₂ dataset with the original GAW observations, CT, and CAMS at climatological timescales, we followed the general framework of the WDCGG global analysis method (Wu et al., 2024; Tsutsumi et al., 2009), with adaptations tailored for this study. The analysis was conducted using monthly mean CO₂ series and focused on the derivation of the long-term trend, mean seasonal cycle, and zonal/global averages.

First, monthly mean CO₂ mole fractions were prepared for each station, and only stations with sufficiently long records were retained. When multiple observational records were available at the same site, they were merged into a single station series following the WDCGG principle of prioritizing continuous observations and supplementing gaps with flask data where appropriate.

Second, stations suitable for large-scale analysis were selected using the WDCGG latitudinal screening approach. Station values were normalized relative to South Pole (SPO), and their mean latitudinal distribution was fitted with a LOESS curve. Stations departing from the fitted curve by more than ± 3 standard deviations were excluded, and this procedure was repeated iteratively until the retained stations were consistent with the large-scale latitudinal pattern.

Third, for each retained station, the monthly CO₂ series was decomposed into a long-term trend and an average seasonal variation using the longest continuous segment of data. Following the WDCGG method, the seasonal cycle was represented by a Fourier polynomial with three harmonics, and the long-term trend was extracted iteratively using deseasonalization and low-pass filtering.

Fourth, missing monthly values were filled by interpolating the long-term trend component and then adding back the station-specific average seasonal cycle. To ensure temporal consistency across stations, incomplete records were further synchronized to the common period of 2000-2024 using zonal growth rates derived from long-term stations within the same latitudinal band. In this study, these interpolation and extrapolation procedures were applied only to the original GAW observations, since the reconstructed dataset, CarbonTracker, and CAMS were already temporally continuous.

Finally, the synchronized station series were aggregated into 30° latitude bands to calculate the zonal mean mole fractions. Global and hemispheric means were then derived from the zonal averages with area weighting.

Compared with the original WDCGG implementation, our analysis was adapted to the fixed study period of 2000-2024 and was used here as a common framework for comparing the reconstructed dataset with observations and model products, rather than for reproducing the official WDCGG products.

Table S1. Ancillary variables used in this study.

Category	Variable	Variable description	Spatial resolution	Temporal resolution	Product name
Anthropogenic activities	POP	Population density	1 km	Yearly	Land Scan
	LUCC	Land use and land cover change	300 m	Yearly	CCI
	NTL	Nighttime Lights	500 m	Daily	VIIRS VNP46A2
	ODIAC	Fossil-fuel CO ₂ emissions	1 km	Monthly	ODIAC
	EDGAR	Anthropogenic CO ₂ emissions	0.1°	Monthly	EDGAR
Ecosystem processes	NDVI	Normalized difference vegetation index	250 m	16-Day	MOD13Q1.061
	GPP	Gross primary productivity	500 m	Yearly	MOD17A3HGF.061
	NPP	Net primary productivity	500 m	Yearly	MOD17A3HGF.061
	FPAR	Fraction of photosynthetically active radiation	500 m	8-Day	MOD15A2H.061
	LAI	Leaf area index	500 m	8-Day	MOD15A2H.061
Climatic factors	TEM	2-m air temperature	0.1°/0.25°	Hourly	ERA5-Land /ERA5
	D2M	2-m dewpoint temperature			
	SKT	Skin temperature			
	U 10	10 m zonal wind component			
	V 10	10 m meridional wind component			
	SP	Surface pressure			
	PRE	Total precipitation			
	EVA	Evaporation			
	VSWL	Volumetric soil water content in layer 1			
Spatiotemporal variables	SSR	Surface net solar radiation			
	Doy_cos	Cosine-transformed day of year	/	/	/
	Doy_sin	Sine-transformed day of year	/	/	/
	Hod_cos	Cosine-transformed hour of day	/	/	/

Spatiotemporal variables	Hod_sin	Sine-transformed hour of day	/	/	/
	X_corrd	Spherical x-coordinate derived from latitude and longitude	/	/	/
	Y_corrd	Spherical y-coordinate derived from latitude and longitude	/	/	/
	Y_corrd	Spherical z-coordinate derived from latitude	/	/	/

Table S2. Hyperparameter combinations of the LightGBM model in this study.

Hyperparameter combinations	CAMS CO₂ reconstruction	WDCGG CO₂ reconstruction	
		Intra-day gap filling	Full-sequence reconstruction
Number of leaves in each tree	512	256	512
Maximum tree depth	20	10	20
Minimum number of samples in a leaf node	200	100	200
Learning rate	0.05	0.01	0.05
Feature sampling fraction	0.8	0.8	0.8
Minimum gain required for node splitting	0.01	0.01	0.01
L1 regularization coefficient	0.1	0.1	0.1
L2 regularization coefficient	0.1	0.1	0.1

Table S3. GAW station information used for the Canadian wildfire analysis.

Station ID	Station name	Longitude	Latitude	Elevation (m)
BCK	Behchoko	-115.9183	62.7979	184
BRA	Bratt's Lake	-104.7113	50.2016	580
CDL	Candle Lake	-105.1179	53.9871	591
CHL	Churchill	-93.8206	58.7379	16
CHM	Chibougamau	-74.3423	49.6925	383
CPS	Chapais	-74.9746	49.8216	383
ESP	Estevan Point	-126.5441	49.3829	7
EST	Esther	-110.2063	51.6706	707
ETL	East Trout Lake	-104.9869	54.3537	500
FSD	Fraserdale	-81.5167	49.8400	210
INU	Inuvik	-133.5342	68.3178	107
LLB	Lac La Biche (Alberta)	-112.4666	54.9538	548

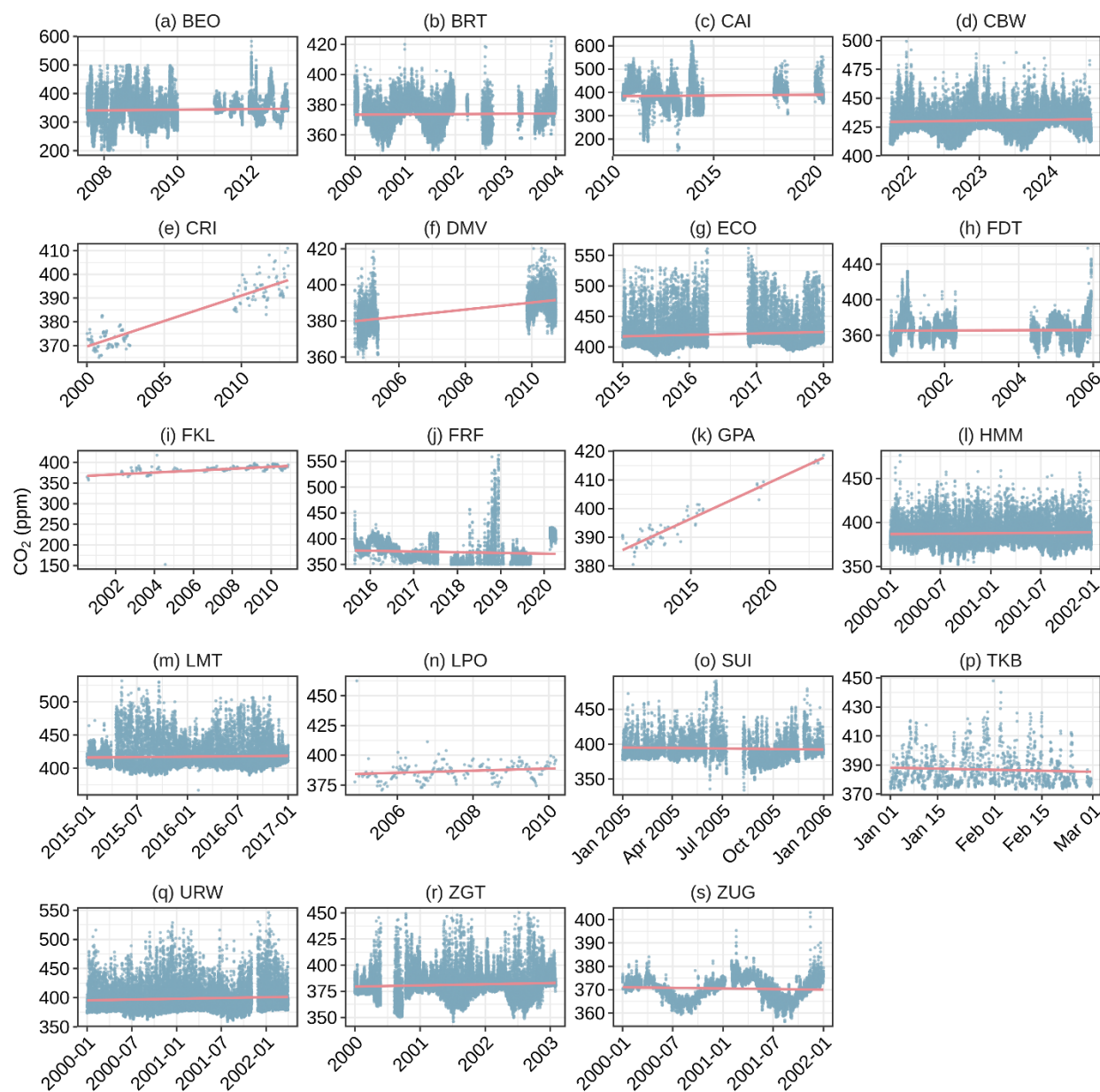


Figure S1. Time series of the original hourly CO₂ observations at the excluded GAW stations. Blue dots denote the original hourly observations, while the red lines show the station-specific linear regression trends.

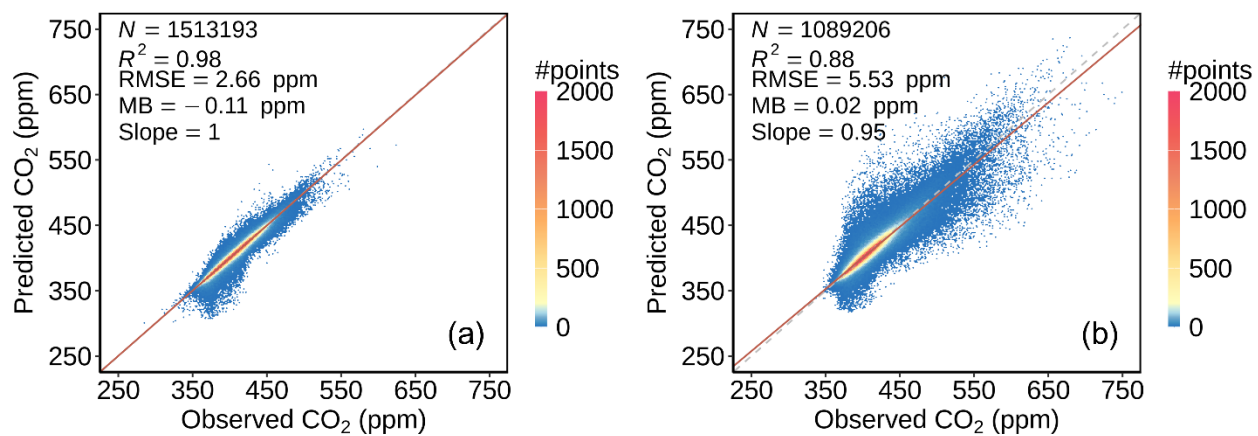


Figure S2. Density scatter plots of (a) CT and (b) CAMS CO₂ based on the hold-out validation of natural cubic spline interpolation.

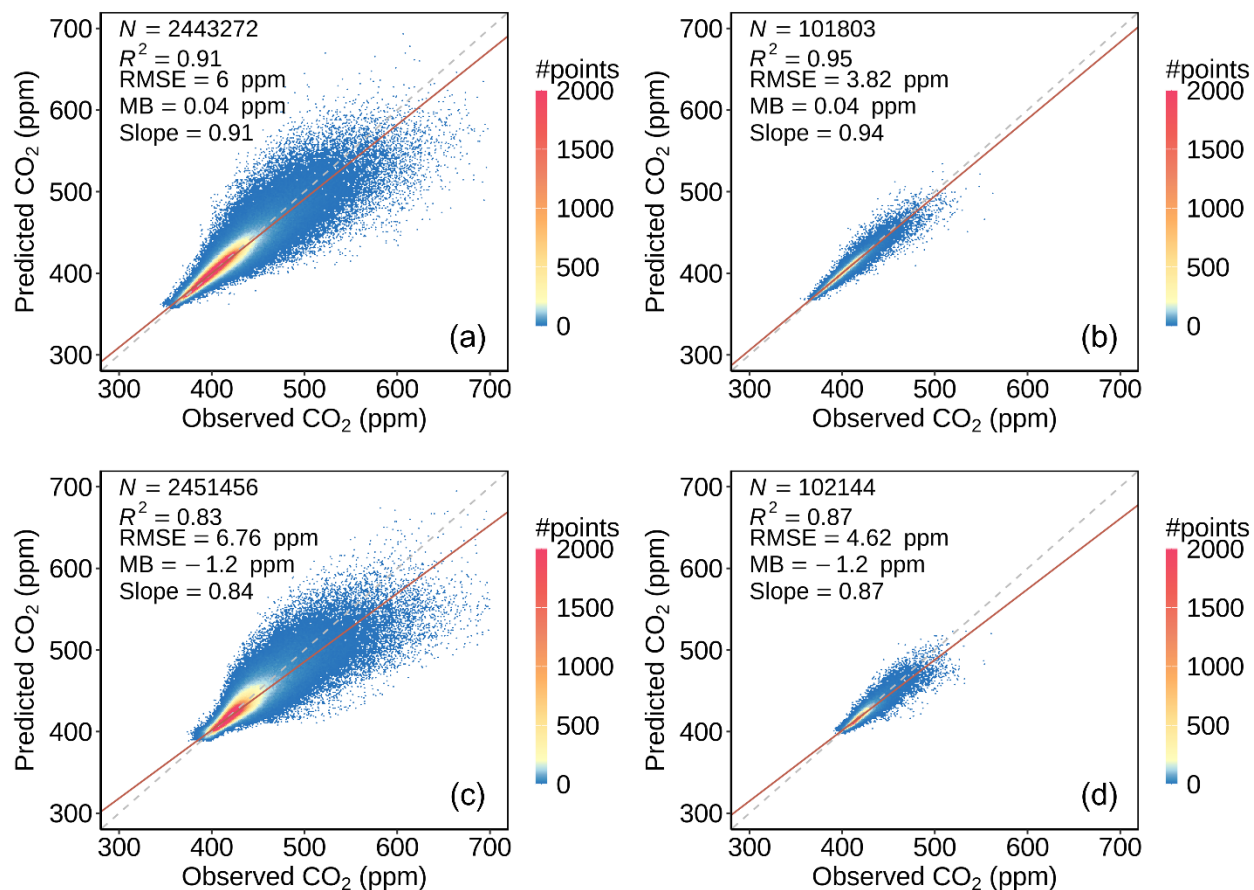


Figure S3. Density scatter plots of CAMS CO₂ filling validation. Panels (a, b) show the results based on the random month-based splitting strategy at hourly and daily scales, respectively. Panels (c, d) represent the results for the chronological month-based splitting strategy at hourly and daily scales, respectively.

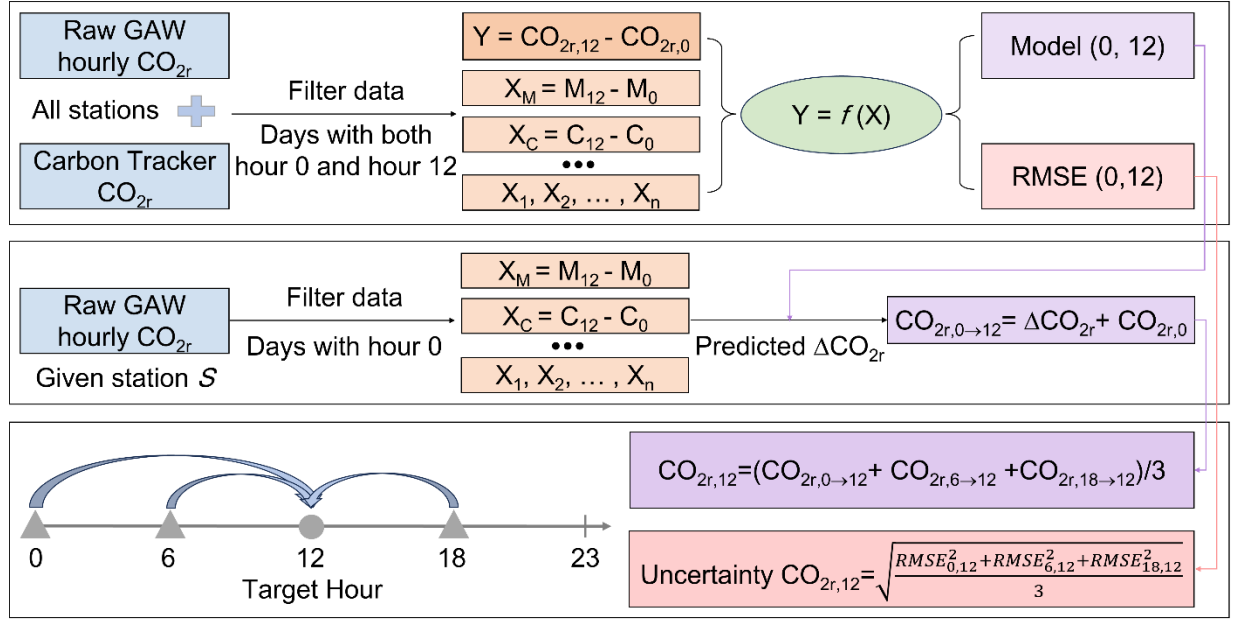


Figure S4. Schematic illustration of the hour-pair modeling framework for intra-day gap filling. For each ordered hour pair (x, y) , residual CO_2 differences between a monitoring hour x and a target hour y were used to train an independent LightGBM model. This example shown uses hour 0 as the monitoring hour and hour 12 as the target hour. During reconstruction, available observations within a given day were used as anchors to predict the residual CO_2 at missing target hours. When multiple monitoring hours were available, their predictions were averaged, and the associated uncertainty was estimated from the corresponding pair-specific RMSEs.

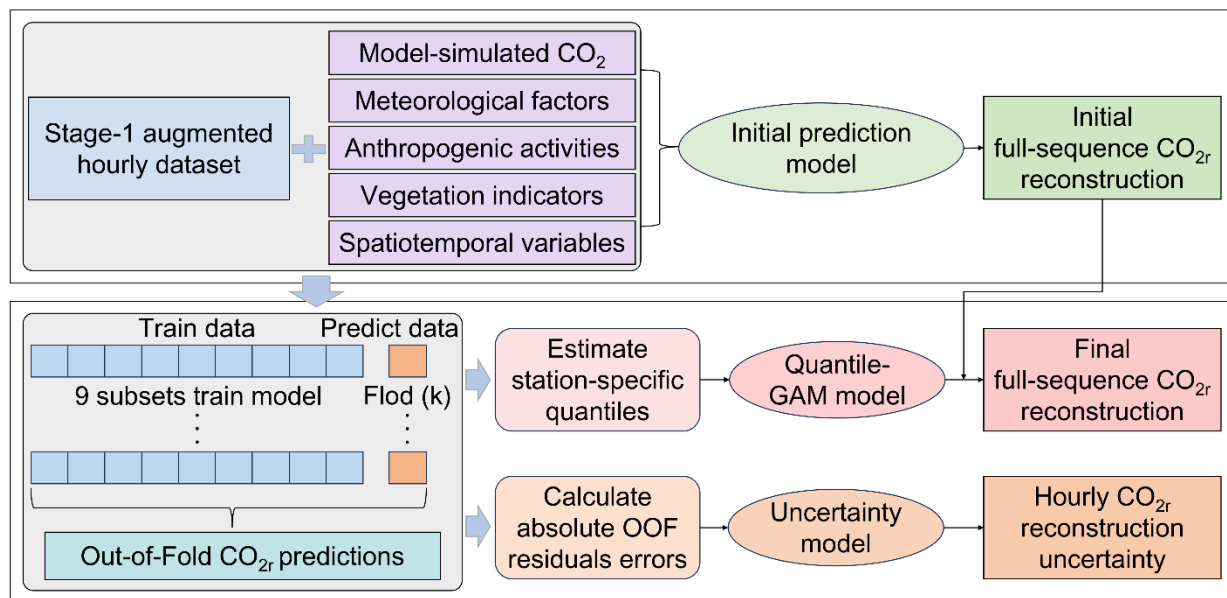


Figure S5. Schematic of Stage 2 full-sequence reconstruction. The Stage 1 augmented hourly dataset and ancillary predictors were used to train an initial LightGBM residual prediction model for all missing time steps. A 10-fold cross-fitting procedure generated out-of-fold CO_{2r} predictions, which were used to fit station-specific quantile-GAM bias correction models. Absolute out-of-fold (OOF) residual errors were further used to train an uncertainty model for estimating hourly reconstruction uncertainty.

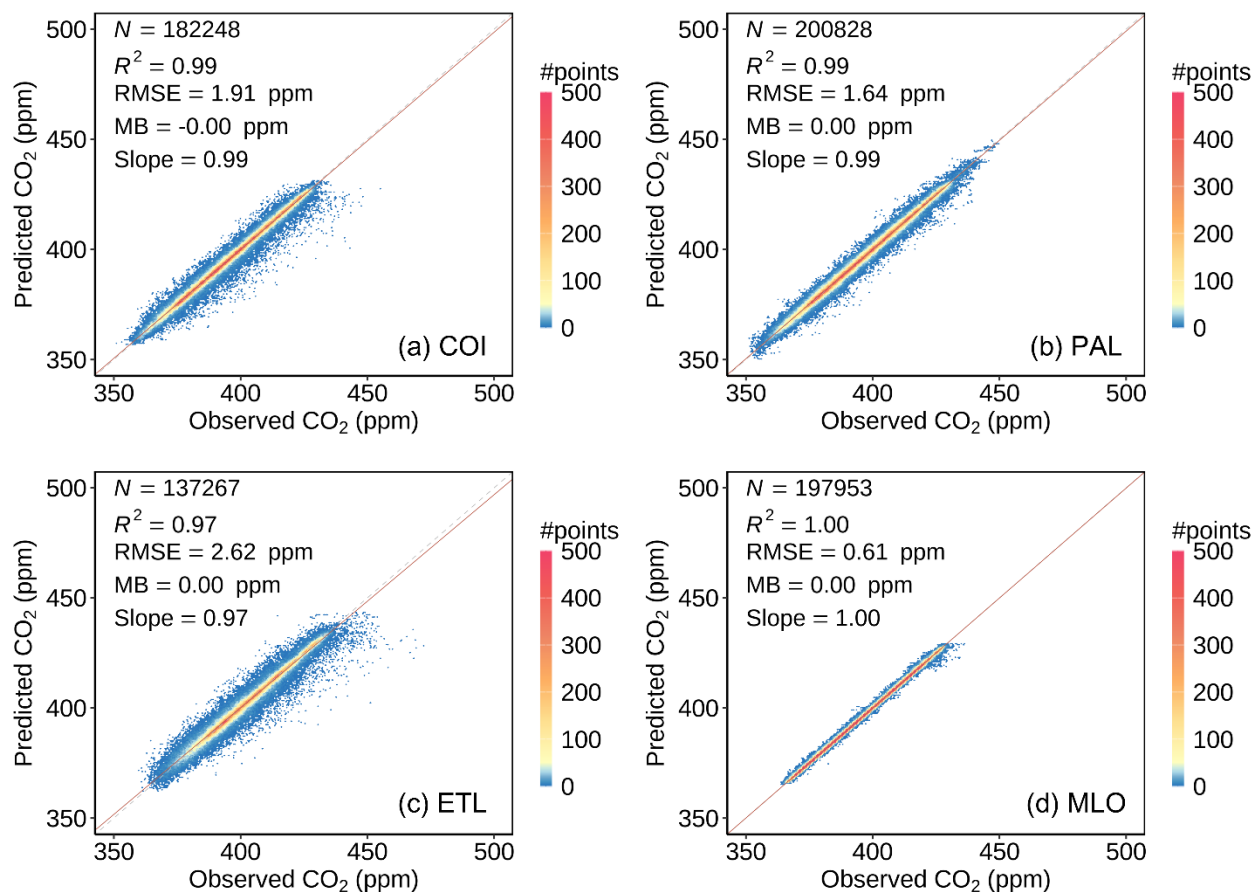


Figure S6. Validation of intra-day gap filling at representative continuous observation sites in (a) Asia (Cape Ochiishi, COI), (b) Europe (Pallas, PAL), (c) North America (East Trout Lake, ETL), and the remote marine background site (d) Mauna Loa, MLO.

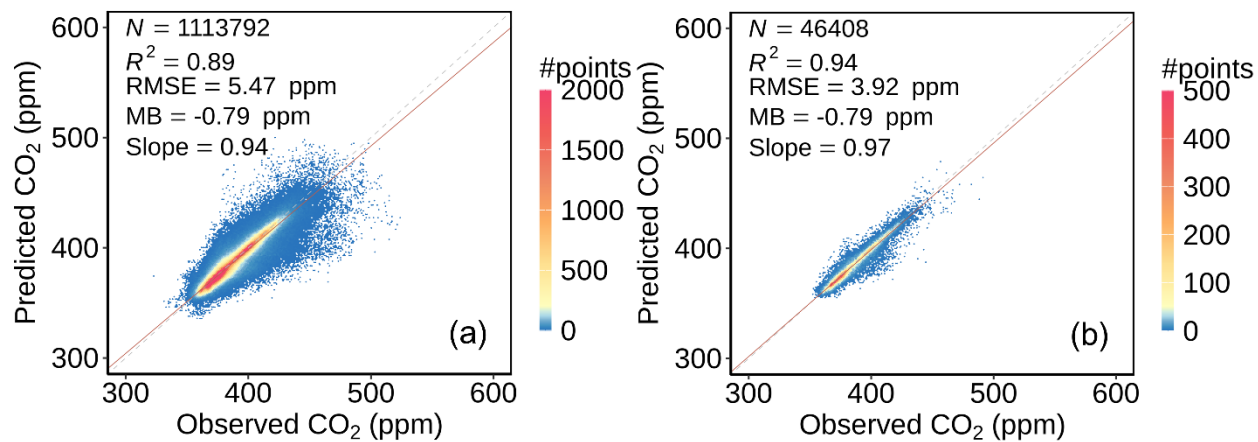


Figure S7. Comparisons between observed and reconstructed CO₂ concentrations using chronological month-based splitting at (a) hourly and (b) daily scales.

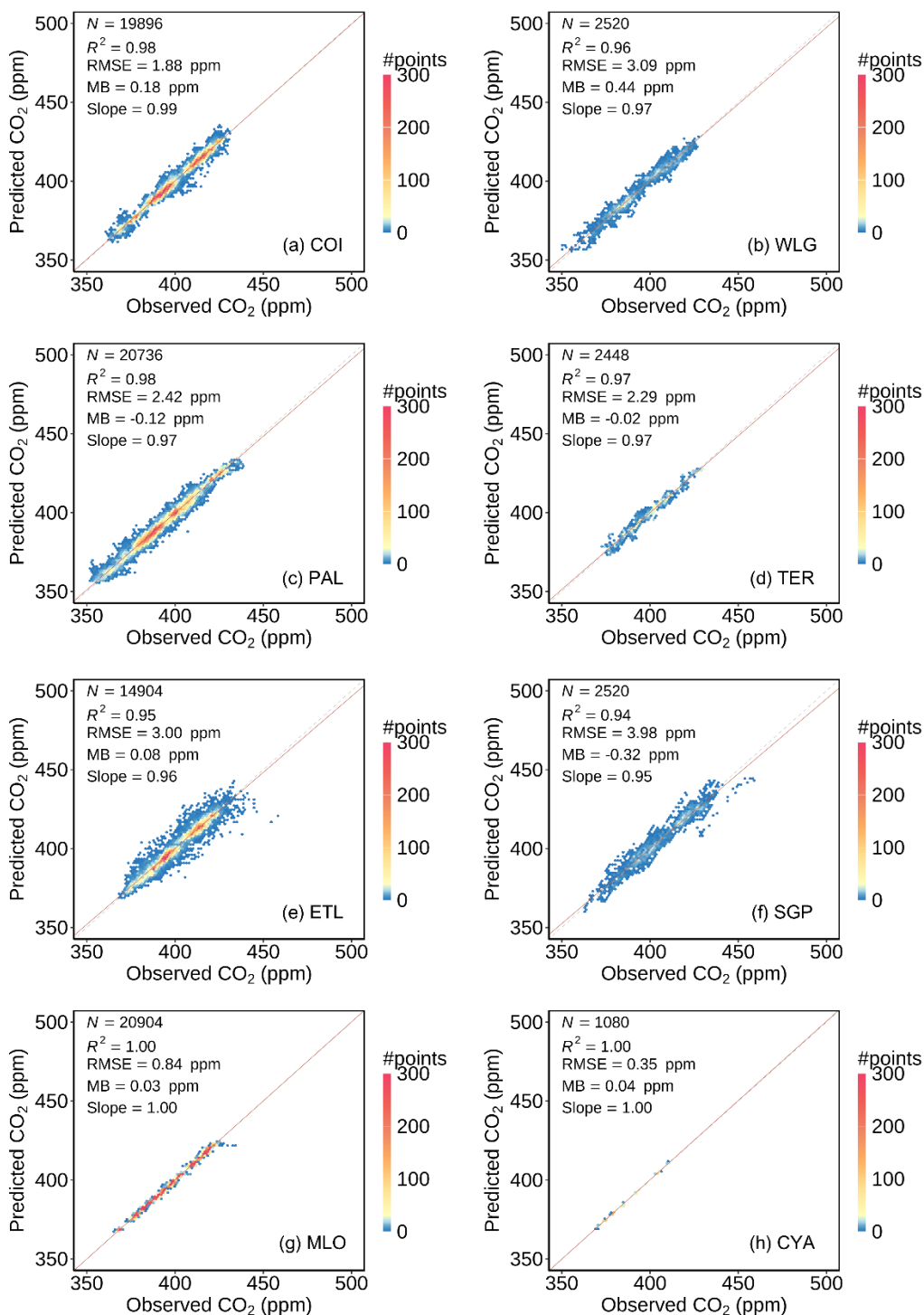


Figure S8. Validation of full-sequence CO₂ reconstruction using random month-based splitting at representative continuous observation sites in (a) Asia (Cape Ochiishi, COI), (c) Europe (Pallas, PAL), (e) North America (East Trout Lake, ETL), and (g) the remote marine background site Mauna Loa, MLO, and at representative flask sampling sites in (b) Asia (Mt. Waliguan, WLG), (d) Europe (Teriberka, TER), (f) North America (Southern Great Plains, SGP), and (h) the Antarctica background site Casey, CYA.

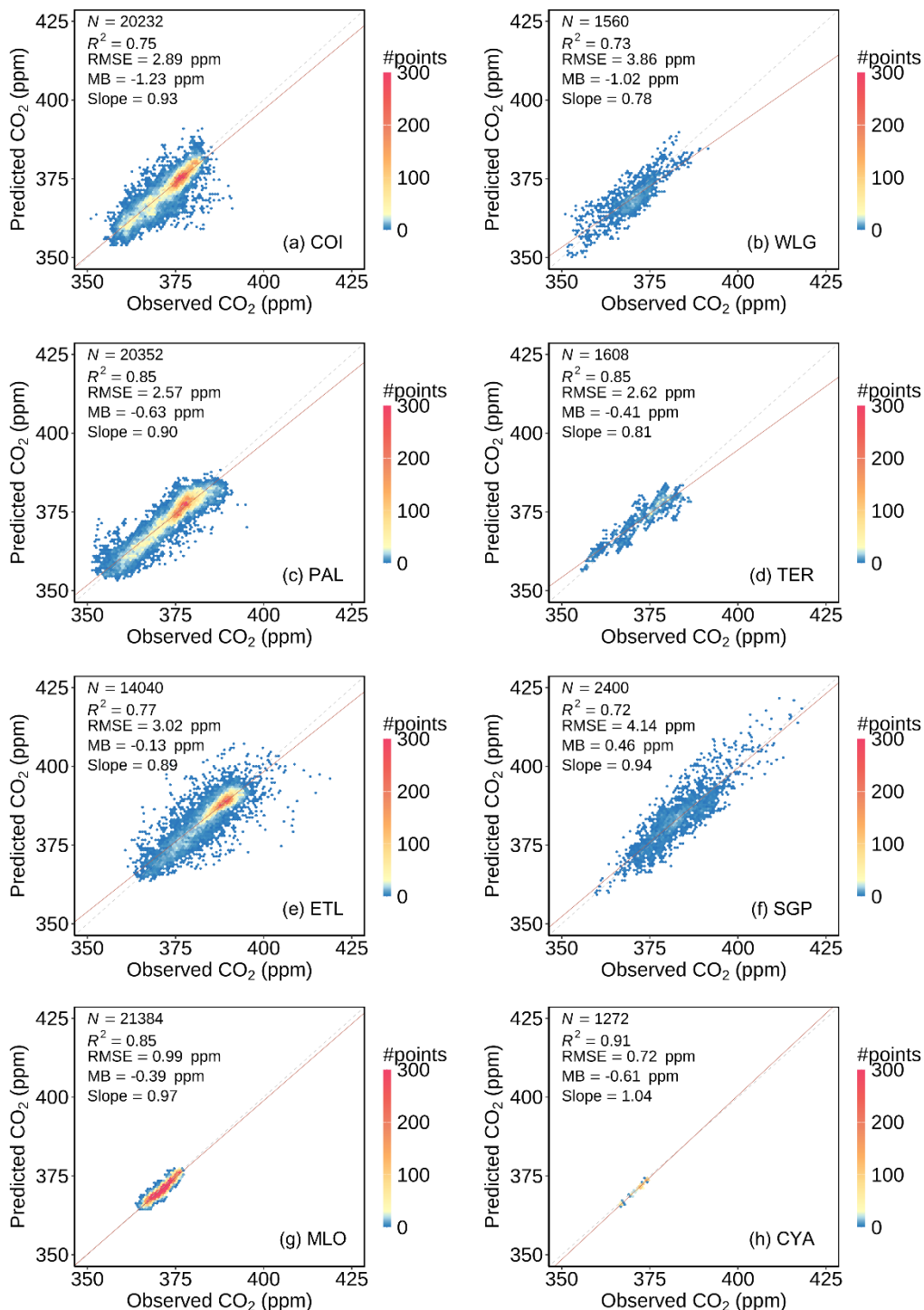


Figure S9. Validation of full-sequence CO₂ reconstruction using chronological month-based splitting at representative continuous observation sites in (a) Asia (Cape Ochiishi, COI), (c) Europe (Pallas, PAL), (e) North America (East Trout Lake, ETL), and (g) the remote marine background site Mauna Loa, MLO, and at representative flask sampling sites in (b) Asia (Mt. Waliguan, WLG), (d) Europe (Teriberka, TER), (f) North America (Southern Great Plains, SGP), and (h) the Antarctica background site Casey, CYA.

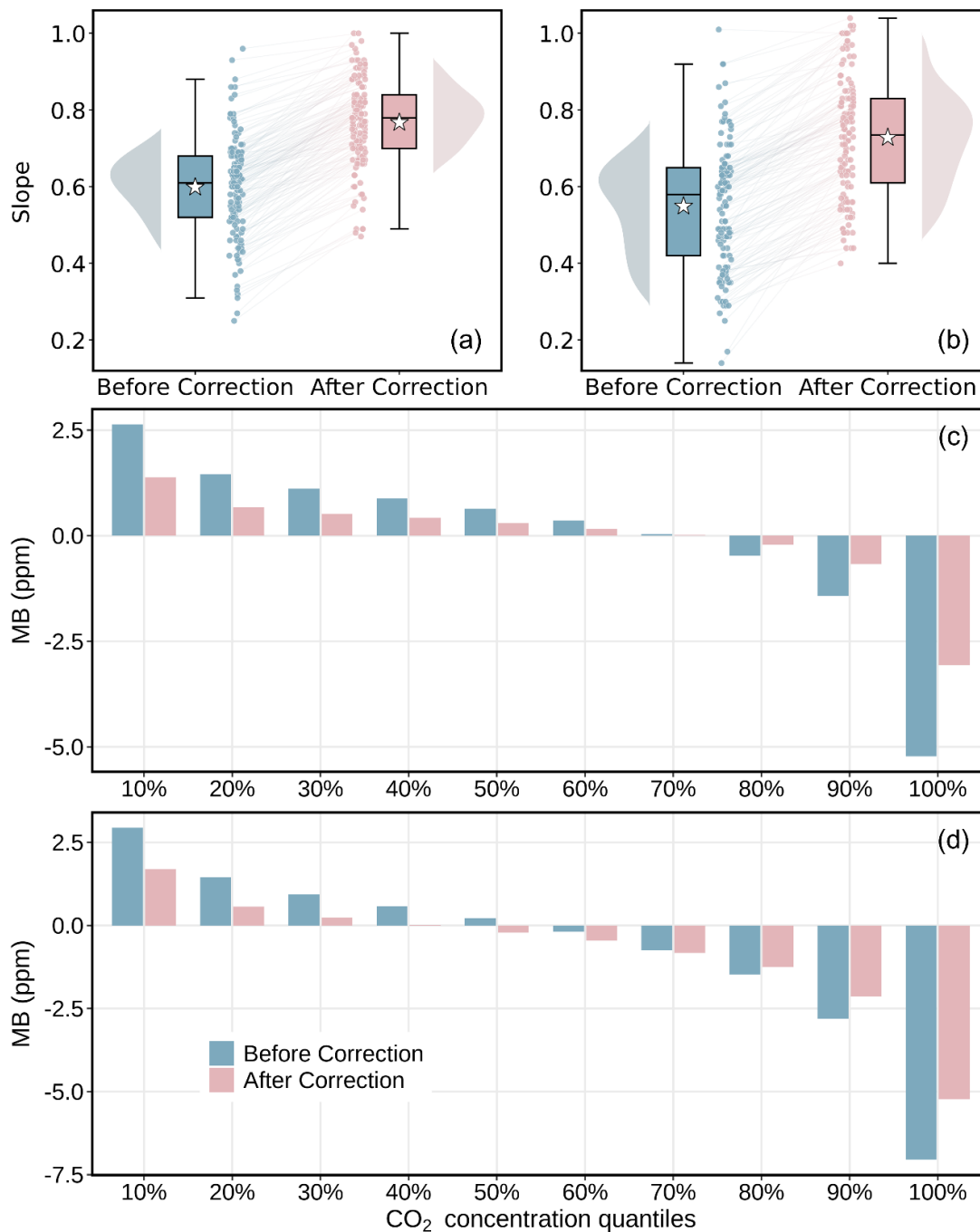


Figure S10. Effects of correction on the full-sequence CO₂ reconstruction performance under random and chronological month-based splitting. Panels (a) and (b) show station-level changes in regression slopes before and after correction, while (c) and (d) show MB across CO₂ concentration quantiles for all validation samples. Panels (a) and (c) correspond to the random month-based split, whereas (b) and (d) correspond to the chronological month-based split.

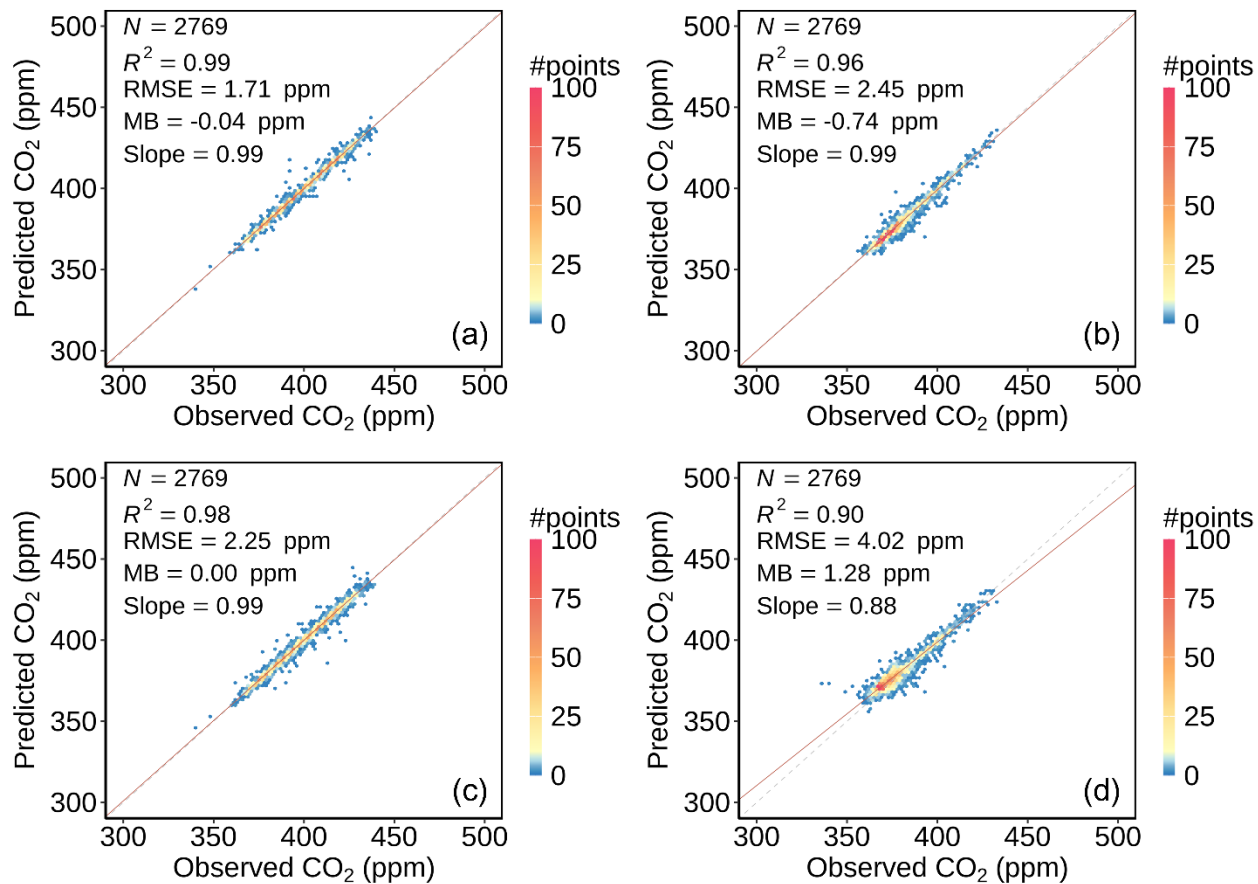


Figure S11. Comparisons of monthly-scale performance between the proposed hourly reconstruction framework (a, b) and the WDCGG monthly gap-filling procedure (c, d) under random and chronological month-based splitting. Panels (a) and (c) correspond to the random month-based split, whereas panels (b) and (d) correspond to the chronological month-based split.

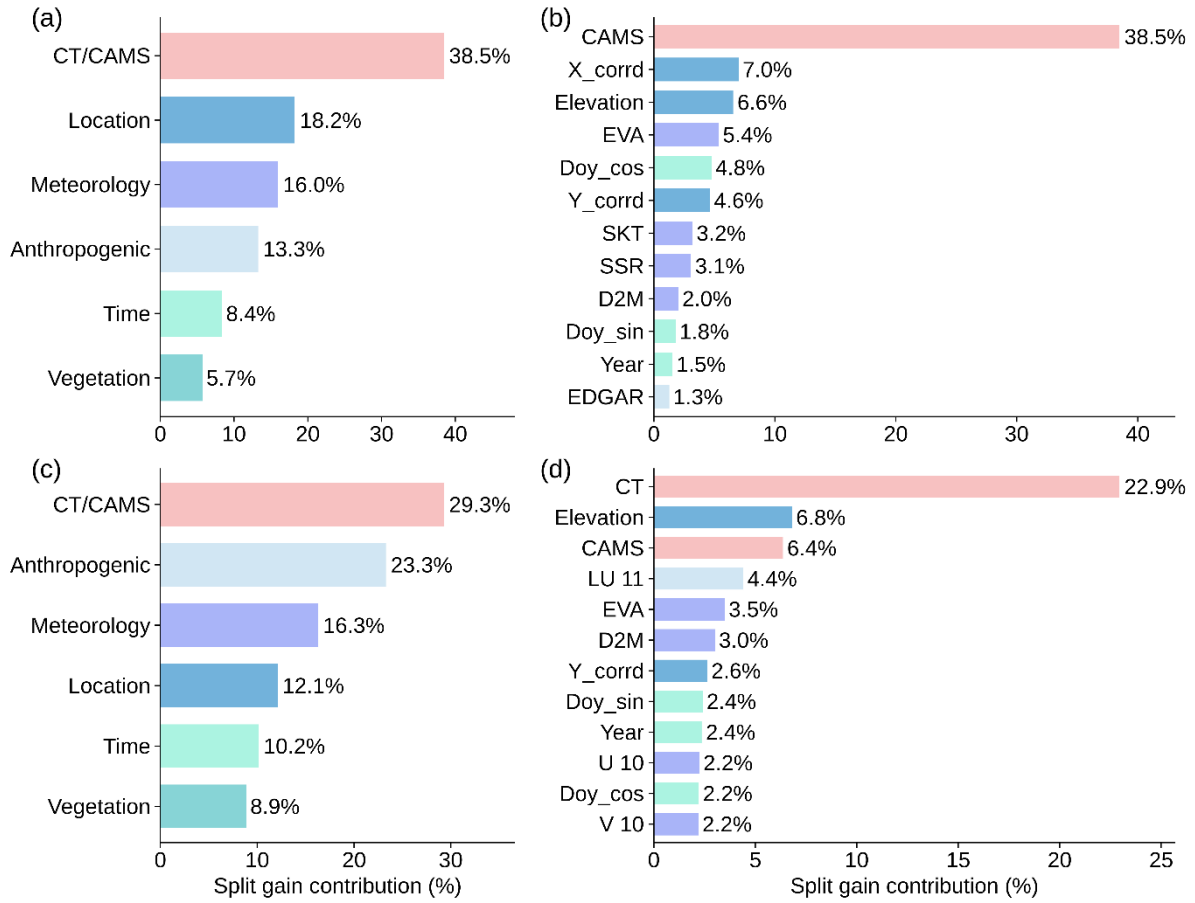


Figure S12. Feature importance of the two-stage reconstruction framework.

(a) Feature-group contribution for the intra-day gap-filling stage, calculated as the mean normalized split-gain contribution across the 552 LightGBM hour-pair models. (b) Top 12 individual predictors in the intra-day gap-filling stage. (c) Feature-group contribution for the full-sequence reconstruction stage. (d) Top 12 individual predictors in the full-sequence reconstruction stage. Detailed information on the predictor variables is provided in Table S1.

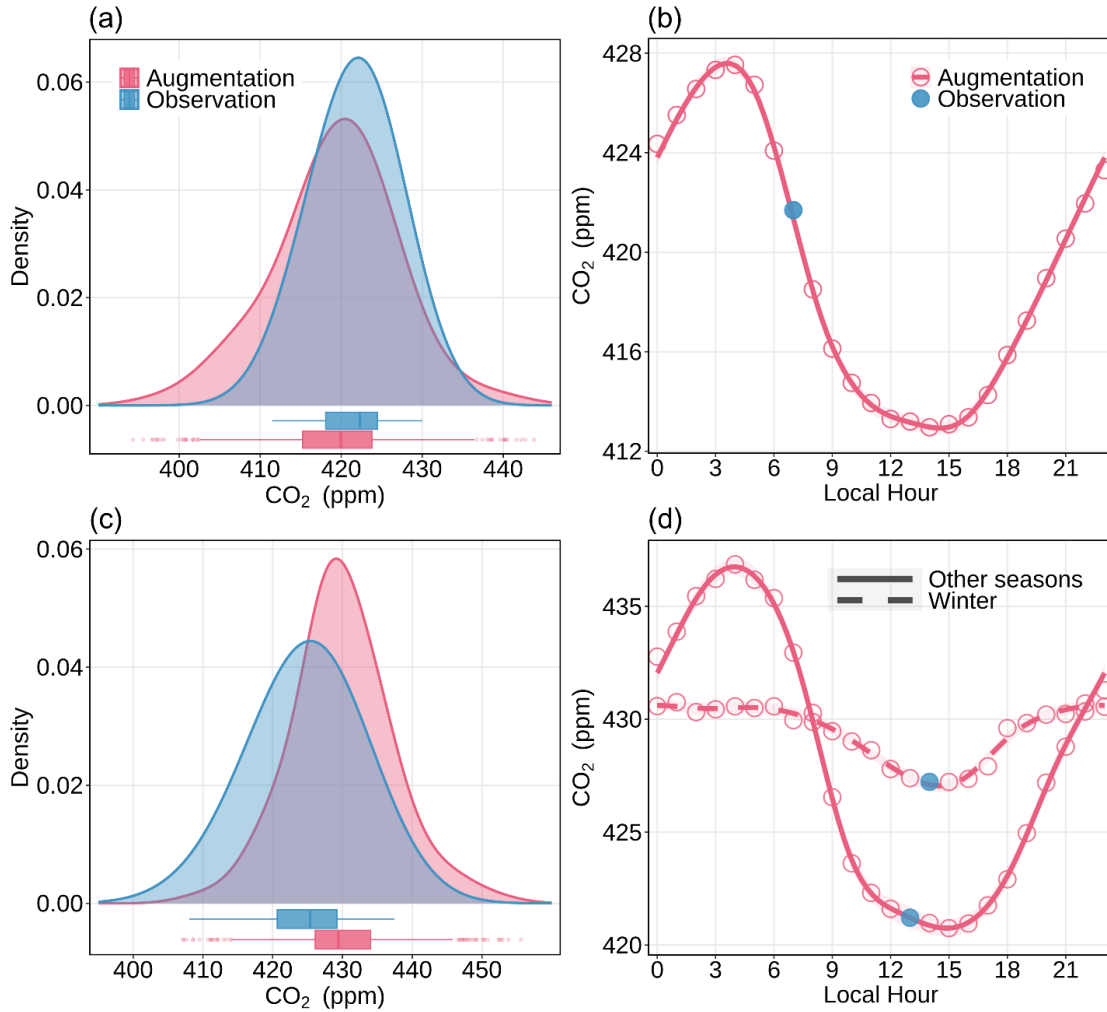


Figure S13. Comparison of CO₂ sample distributions and mean diurnal cycles between observed and intra-day filling datasets. Panels (a) and (b) show Mt. Waliguan (WLG), and panels (c) and (d) show Southern Great Plains (SGP). These two stations were selected as examples of sparse observations with fixed or seasonally varying sampling times to illustrate how intra-day gap filling improves sample representativeness and reconstructs complete hourly CO₂ variability. Panels (a) and (c) compare the concentration distributions of original observations and augmented records. Panels (b) and (d) show mean diurnal cycles after intra-day filling, with original observations marked by filled blue circles. Because SGP was sampled at different local hours in winter and other seasons, panel (d) presents winter and non-winter diurnal cycles separately. In contrast, WLG was sampled at the same local hour throughout the year, so panel (b) shows a single mean diurnal cycle.

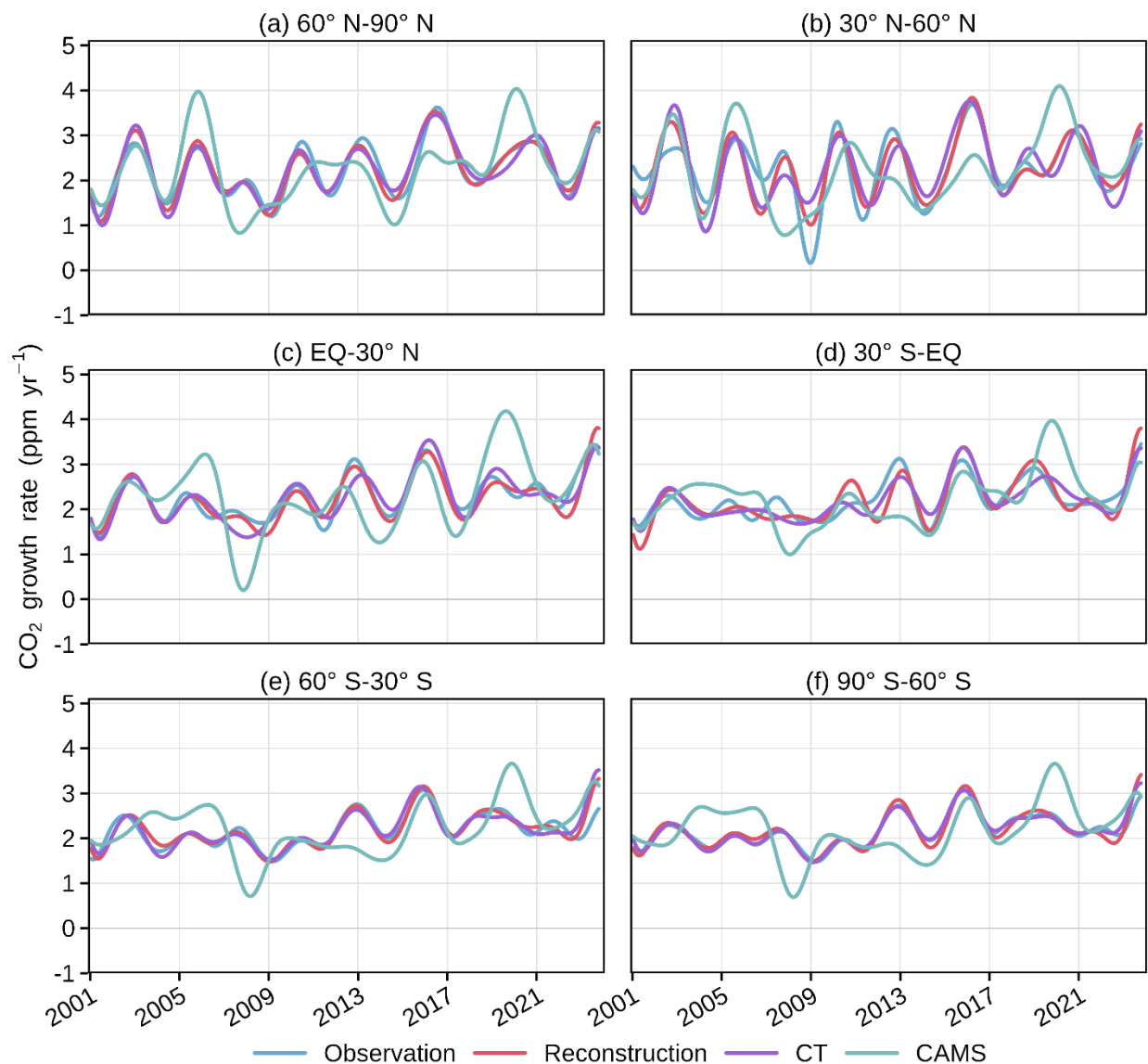


Figure S14. Latitude-band differences in monthly CO₂ growth rates using the WDCGG global analysis framework. Monthly CO₂ growth rates were derived from the WDCGG trend component for six 30° latitude bands. The four datasets are the original GAW observations, reconstructed GAW records, CT and CAMS.

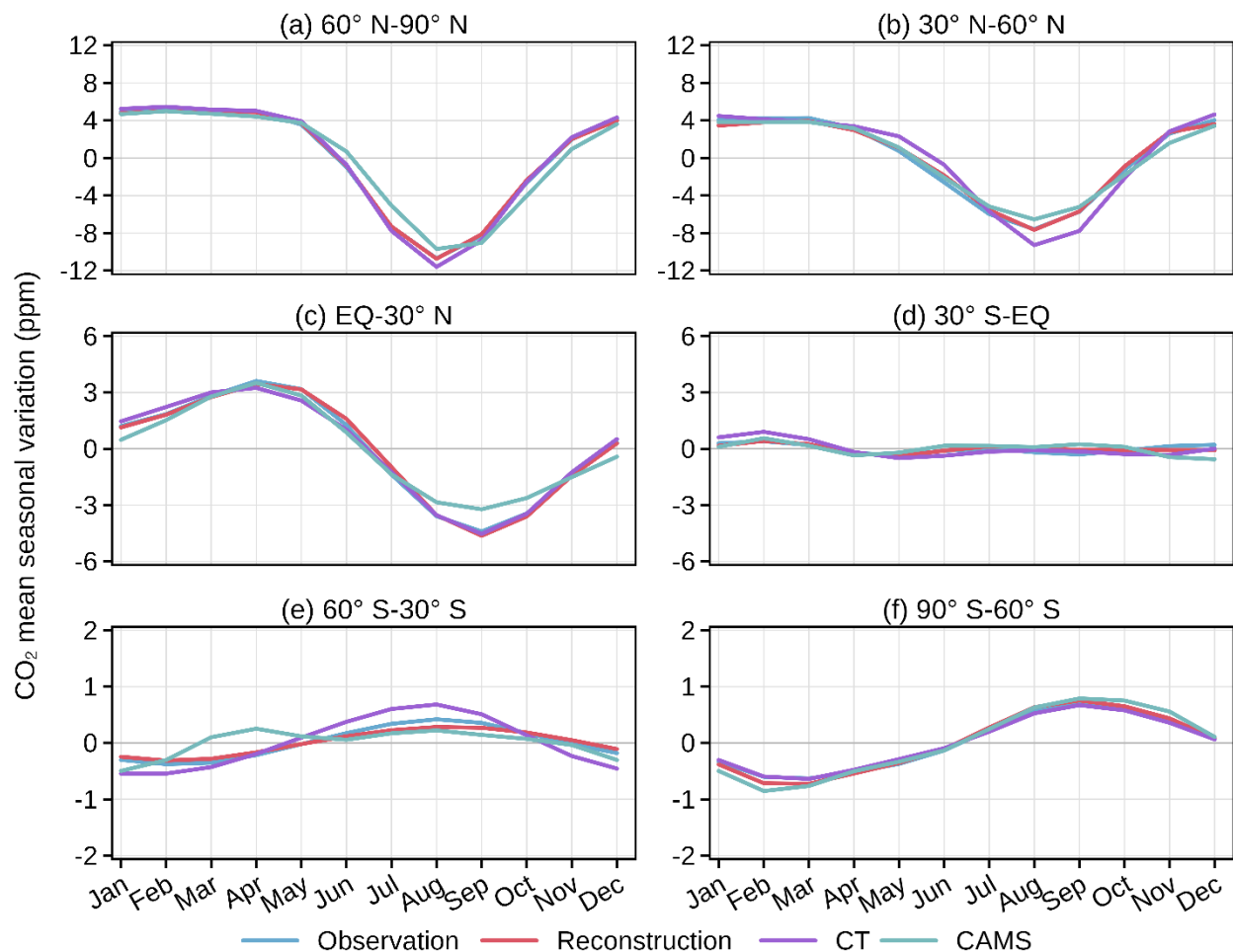


Figure S15. Zonal mean seasonal CO₂ variations derived using the WDCGG global analysis framework. The four curves show the WDCGG-derived mean seasonal components from the original GAW observations, the reconstructed GAW dataset, CT and CAMS. For each dataset, station-level seasonal components were first averaged by calendar month and then aggregated within each latitude band.

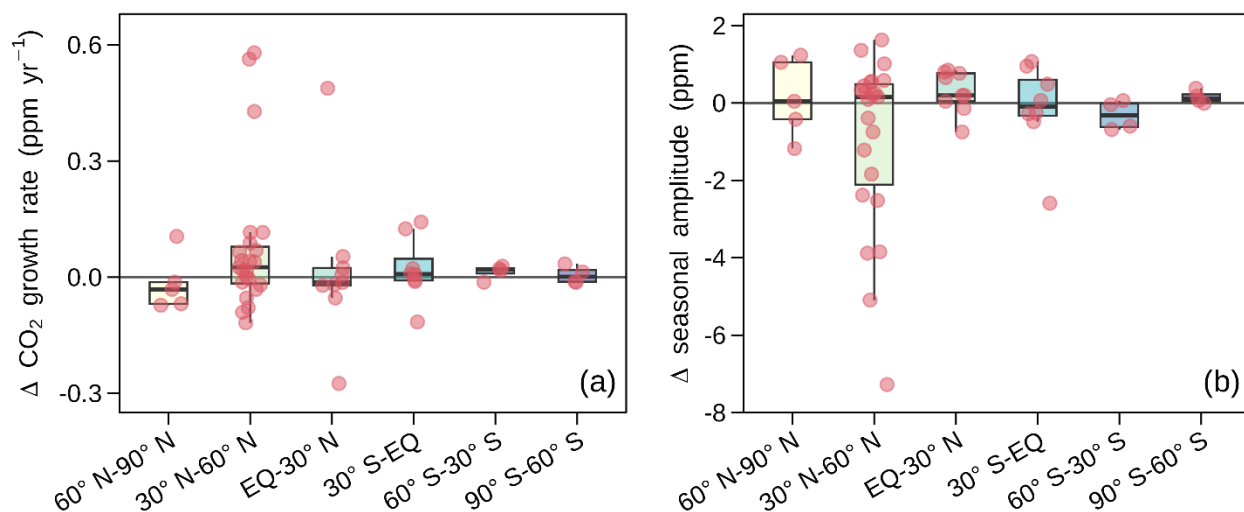


Figure S16. Reconstruction-induced changes in the WDCGG global analysis trends and seasonal amplitudes at flask-sampling stations. Box-and-jitter plots show station-level differences between the reconstructed GAW records and the original GAW observations for flask-sampling stations. (a) Difference in station-mean CO₂ growth rate, expressed as reconstruction minus observation. (b) Difference in mean seasonal-cycle amplitude, also calculated as reconstruction minus observation. Boxes summarize the distributions across stations within each 30° latitude band, and points represent individual flask-sampling stations.

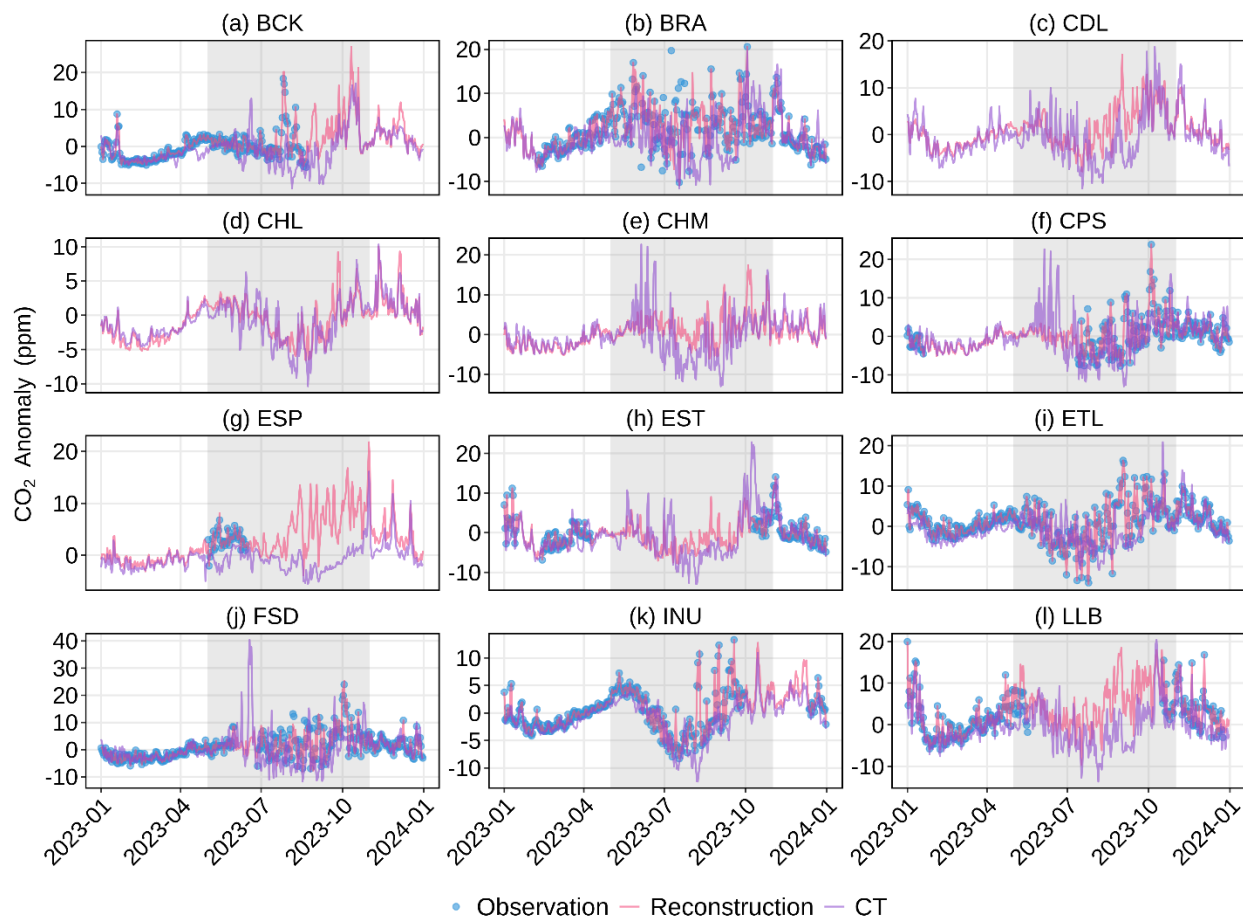


Figure S17. Time series of daily mean original observed CO₂ anomalies (blue dots), reconstructed CO₂ anomalies (red lines), and CT CO₂ anomalies at Canadian stations in 2023. The shaded periods indicate the wildfire active season from May to September. Reconstructed and CT CO₂ anomalies were calculated relative to station-specific monthly means, whereas observed anomalies used the corresponding reconstructed monthly means as the baseline because the original observations were sparse.

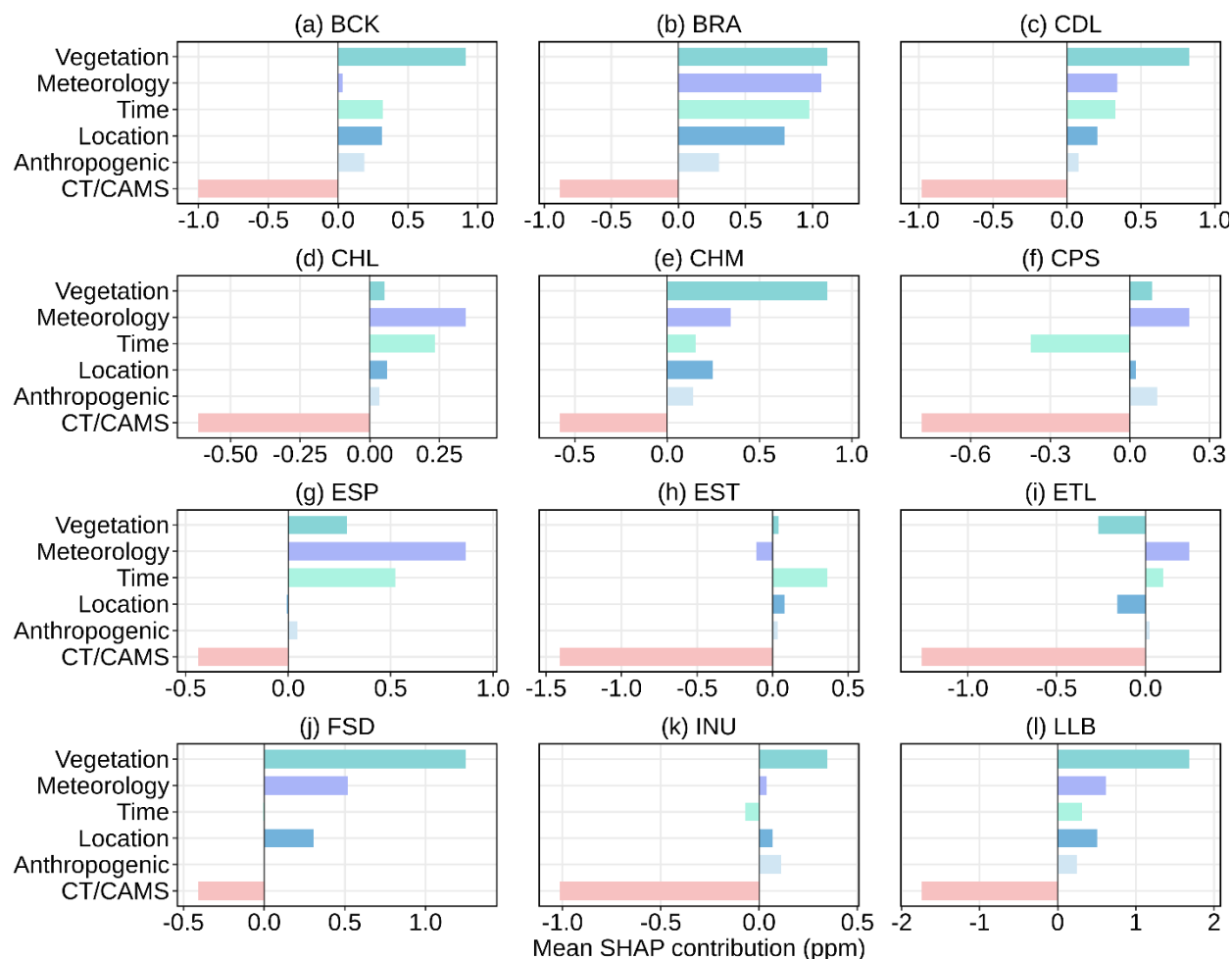


Figure S18. Group-level attribution of the fire-period changes in the reconstructed CO₂ anomaly at the 12 Canadian GAW stations. For each station, the SHAP contributions of the all predictors are aggregated into six feature groups (Meteorology, Vegetation, Anthropogenic, Location, Time, and CT/CAMS), and bars show the differences between the fire-period (May-September 2023) and non-fire mean SHAP contribution (ppm).

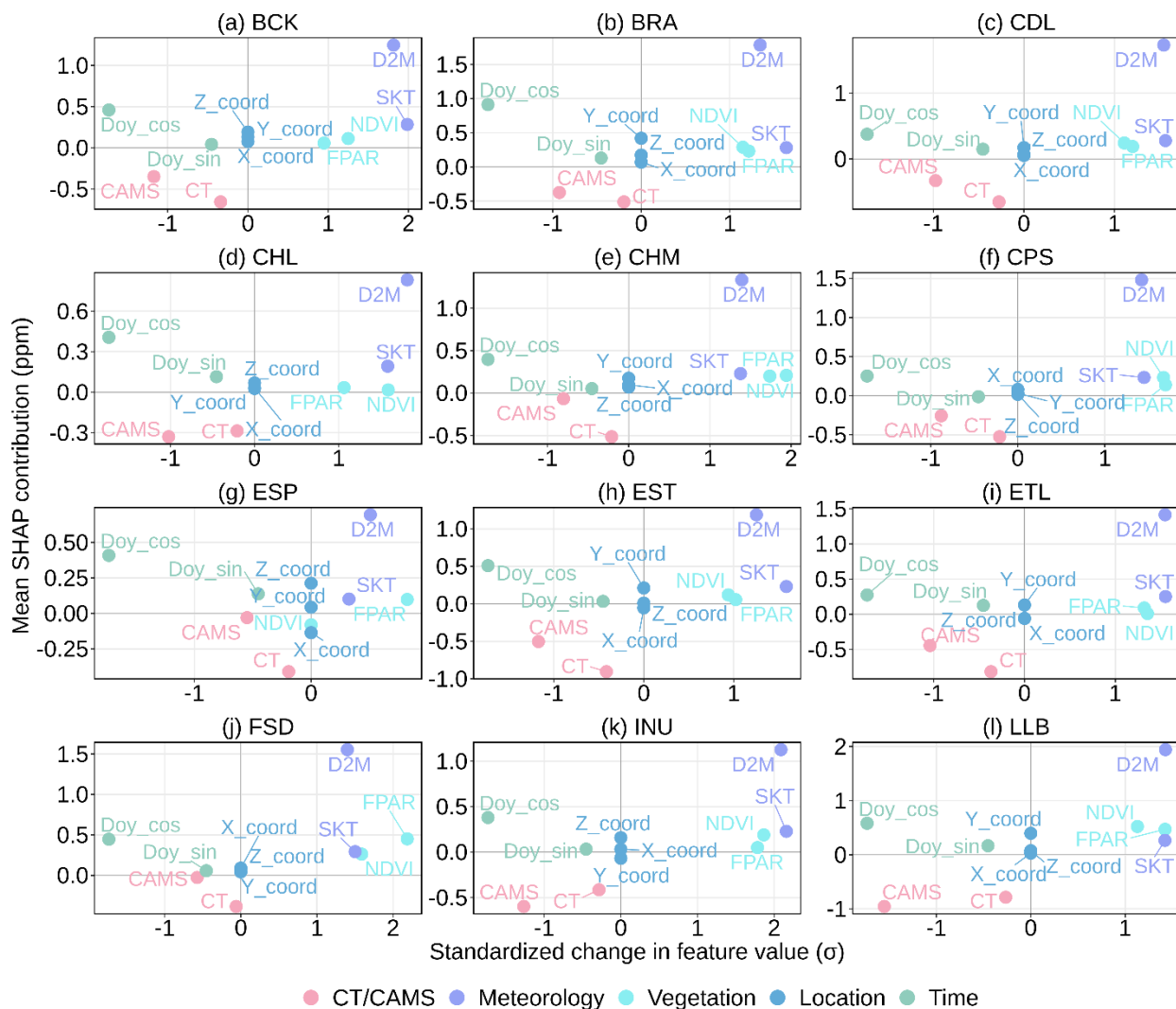


Figure S19. Mechanism diagnostic at the 12 Canadian GAW stations. For each predictor, the vertical axis shows the fire-minus-non-fire change in mean SHAP contribution (ppm) and the horizontal axis shows the corresponding change in mean feature value, standardized by each predictor's standard deviation (σ). Detailed information on the predictor variables is provided in Table S1.

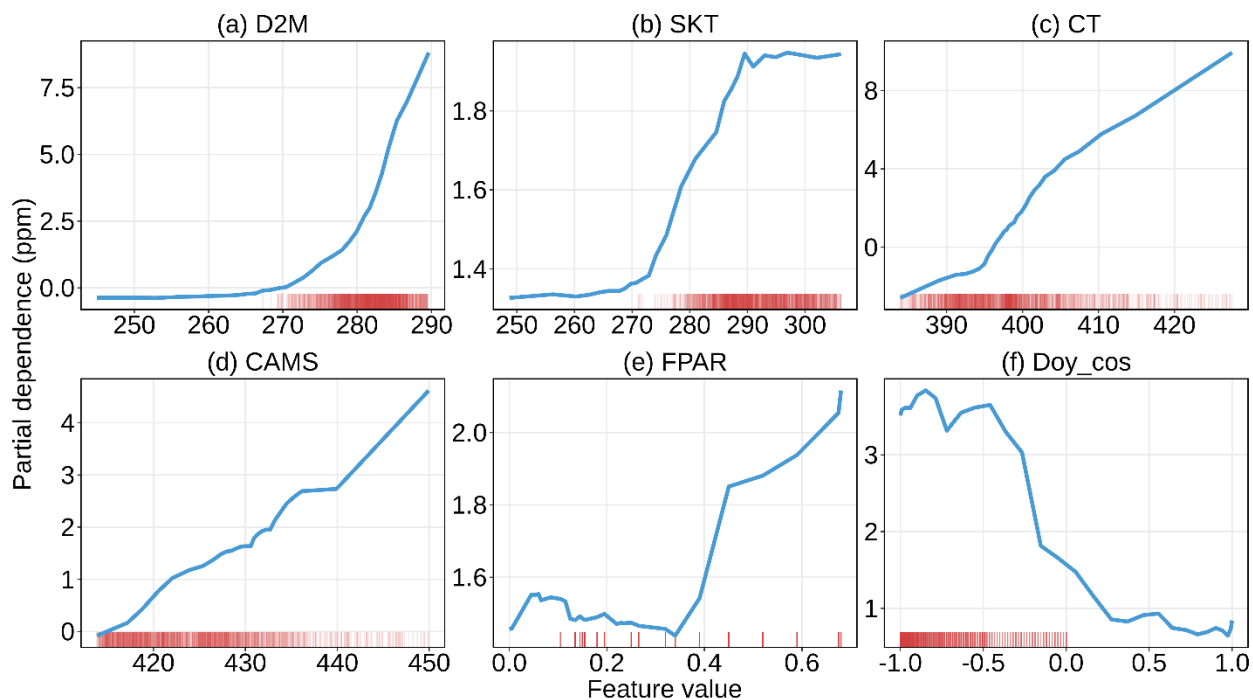


Figure S20. Partial-dependence curves (blue) of the reconstructed CO₂ anomaly on six predictors at station Bratt's Lake (BRA). Each curve sweeps one predictor across its range with the others held at the station's values. Red ticks mark the fire-period (May-September 2023) distribution of each predictor. Detailed information on the predictor variables is provided in Table S1.

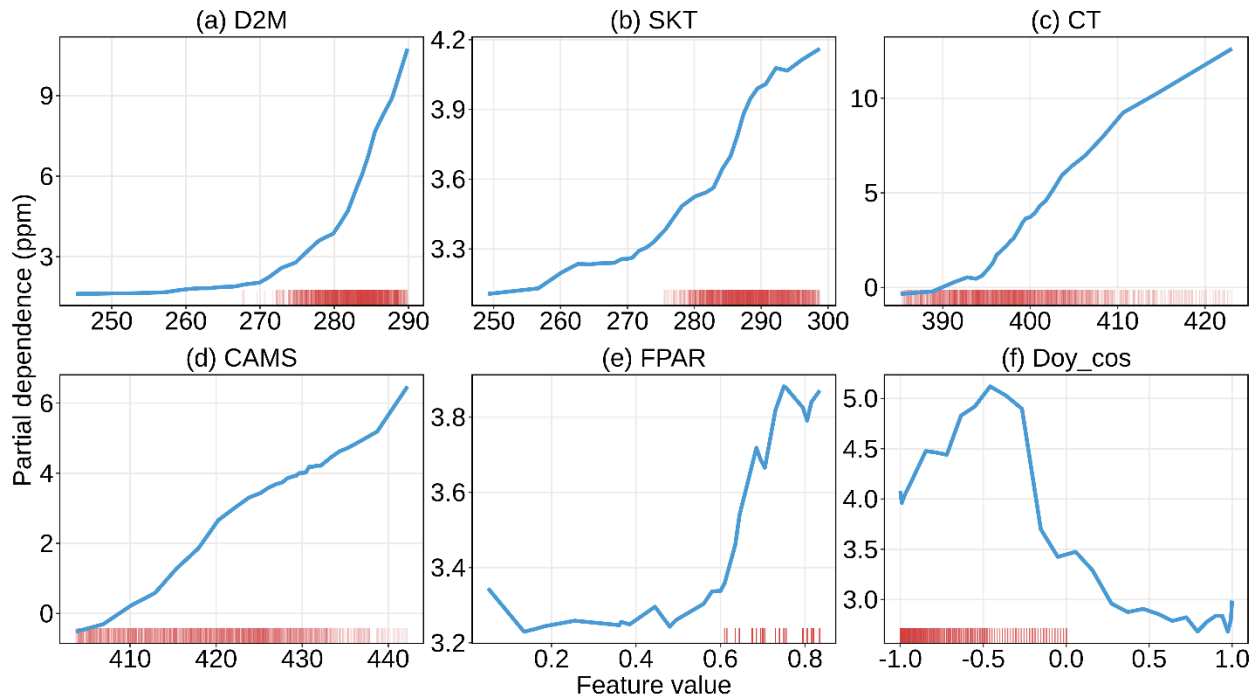


Figure S21. Partial-dependence curves (blue) of the reconstructed CO₂ anomaly on six predictors at station Lac La Biche (LLB). Each curve sweeps one predictor across its range with the others held at the station's values. Red ticks mark the fire-period (May-September 2023) distribution of each predictor. Detailed information on the predictor variables is provided in Table S1.

References

- Anderer, M. and Li, F.: Hierarchical forecasting with a top-down alignment of independent-level forecasts, *Int. J. Forecast.*, 38, 1405-1414, <https://doi.org/10.1016/j.ijforecast.2021.12.015>, 2022.
- Crippa, M., Guizzardi, D., Pagani, F., Banja, M., Muntean, M., E., S., Becker, W., Monforti-Ferrario, F., Quadrelli, R., Risquez Martin, A., Taghavi-Moharamli, P., Köykkä, J., Grassi, G., Rossi, S., Brandao De Melo, J., Oom, D., Branco, A., San-Miguel, J., and Vignati, E.: GHG emissions of all world countries, Publications Office of the European Union [dataset], <https://doi.org/10.2760/953322>, 2023.
- Defourny, P., Lamarche, C., Brockmann, C., Boettcher, M., Bontemps, S., De Maet, T., Duveiller, G. L. H., K., A., H., Kirches, G., Moreau, I., Peylin, P., Ottlé, C., J., R., Van Bogaert, E., Ramoino, F., Albergel, C., and Arino, O.: Observed annual global land-use change from 1992 to 2020 three times more dynamic than reported by inventory-based statistics, European Space Agency [dataset], <https://www.esa-landcover-cci.org>, 2023.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Adv. Neural Inf. Process. Syst.*, 30, [https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-](https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html)Abstract.html, 2017.
- Oda, T. and Maksyutov, S.: ODIAC Fossil Fuel CO₂ Emissions Dataset (Version name: ODIAC 2023), Center for Global Environmental Research, National Institute for Environmental Studies [dataset], <https://doi.org/10.17595/20170411.001>, 2015.
- Simpson, G. L.: Modelling Palaeoecological Time Series Using Generalised Additive Models, *Front. Ecol. Evol.*, 6, 149, <https://doi.org/10.3389/fevo.2018.00149>, 2018.
- Sims, K., Reith, A., Bright, E., Kaufman, J., Pyle, J., Epting, J., Gonzales, J., Adams, D., Powell, E., and Urban, M., & Rose, A.: LandScan Global 2022, Oak Ridge National Laboratory [dataset], <https://doi.org/10.48690/1529167>, 2023.
- Tsutsumi, Y., Mori, K., Hirahara, T., Ikegami, M., and Conway, T. J.: Technical Report of Global Analysis Method for Major Greenhouse Gases by the World Data Center for Greenhouse Gases (WMO/TD-No. 1473), GAW Report No. 184, Geneva, WMO, 1-23, https://library.wmo.int/index.php?lvl=notice_display&id=12631, 2009.
- Wood, S. N.: Generalized additive models: an introduction with R, 2, Chapman and Hall/CRC, Boca Raton, FL, <https://doi.org/10.1201/9781315370279>, 2017.
- Wu, Z., Vermeulen, A., Sawa, Y., Karstens, U., Peters, W., de Kok, R., Lan, X., Nagai, Y., Ogi, A., and Tarasova, O.: Investigating the differences in calculating global mean surface CO₂ abundance: the impact of analysis methodologies and site selection, *Atmos. Chem. Phys.*, 24, 1249-1264, <https://doi.org/10.5194/acp-24-1249-2024>, 2024.