Supplement of Atmos. Chem. Phys., 25, 8475–8492, 2025
https://doi.org/10.5194/acp-25-8475-2025-supplement

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

State-wide California 2020 carbon dioxide budget estimated with OCO-2 and OCO-3 satellite data

Matthew S. Johnson et al.

Correspondence to: Matthew S. Johnson (matthew.s.johnson@nasa.gov)

Supplemental Text

Text S1. Model Comparison

We compare versions of the GP inversion model using their expected log pointwise predictive density (ELPD), evaluated at unseen data points. Thus, ELPD is a measure of the predictive performance of the model (Vehtari et al., 2017). We estimate the ELPD values based on the leave-one-out cross-validation (LOO-CV) method using the ArviZ Python package (https://python.arviz.org/en/stable/). Specifically, ELPD is the sum of the log pointwise predictive density (LPPD) for each observation, which represents the log probability of the observed data under the posterior predictive distribution:

$$\mathrm{ELPD} = \sum_{i=1}^{n} \mathrm{LPPD}_i,$$

where LPPD can be expressed as:

$$\mathrm{LPPD}_i = \log \int p(y_i \mid \theta) p(\theta \mid y) d\theta$$

where the integral computes the predictive posterior probability for each observation y_i, quantifying the expected likelihood of the observed data under the model's posterior distribution.

The Bayesian LOO-CV procedure provides an estimate of the out-of-sample predictive fit, crucial for evaluating the generalizability of a statistical model. This is quantified by ELPD specifically adapted for LOO (Vehtari et al., 2017):

$$\mathrm{ELPD}_{\mathrm{LOO}} = \sum_{i=1}^{n} \log p(y_i \mid y_{-i}),$$

where $p(y_i \mid y_{-i})$ represents the predictive density of the ith observation, calculated using the remaining data:

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta) p(\theta \mid y_{-i}) d\theta$$

where $p(\theta \mid y_{-i})$ is the posterior distribution of the parameters θ, derived from the dataset excluding the ith observation. Thus, $\mathrm{ELPD}_{\mathrm{LOO}}$ represents that the ELPD is adjusted for each leave-one-out scenario, providing a cross-validated measure of the model's predictive accuracy. Figure S4 compares the computed $ELPD_{LOO}$ values, with the model performance ranked from best to worst. Overall, Model 1 showed the best performance across all months; therefore, we have used its results for subsequent seasonal and sectoral analyses.

Text S2. Graphical Description of Model Structure

We define the probability distributions for hyperparameters and likelihood-associated parameters using probabilistic programming with PyMC (Abril-Pla et al., 2023). As illustrated in Fig. S5, we treat these parameters as random variables with specified prior distributions (see Text S3 for discussion on the prior distributions). The probabilistic graph in the figure shows the hyperparameters associated with the kernel: "kernel_length_T," "kernel_length_S," and "kernel_var", which denotes the temporal length scale, the spatial length scale, and the kernel variance, respectively. The graph shows that these hyperparameters are sampled from priors of Exponential Gamma, Gamma, and Exponential distributions, respectively (see Text S3 for details). Once the model is configured, as shown in Fig. S5, we implement it through probabilistic programming using PyMC.

The graph in Fig. S5 also illustrates that we sample the scaling factors (denoted as "Lambda") from a truncated normal distribution, which serves as the prior. The prior probability distributions and the likelihood function that correspond to Fig. S5 are described in Text S3. Additionally, the bias term (denoted as "bias") is sampled from a normal distribution. For the GP noise term (denoted as "noise"), we use a Half Cauchy distribution as the prior. The graph displays the connections of the hyperparameters of the mean function—specifically, the scaling factor and the

bias—and the three hyperparameters associated with the GP kernel to the GP function (denoted as 'f'). This relationship illustrates that the behavior of a GP is determined by its mean function and its kernel.

Text S3. Prior Distributions for Parameters and Likelihood Function

This section describes the prior distributions for the hyperparameters and the likelihood function used in the GP model. Figure S5 also presents the prior distributions for the parameters and the likelihood function in a graphical format.

Scaling Factor

$$\boldsymbol{\lambda} \sim TrNormal(\mu = \boldsymbol{\mu_\lambda}, \sigma = \boldsymbol{\sigma_\lambda})$$

where *TrNormal* denotes the truncated normal distribution. For all sectors, $\boldsymbol{\mu_\lambda}$ (i.e., initial mean) is assigned a value of 1. The standard deviations $\boldsymbol{\sigma_\lambda}$ are set to 0.3, 0.5, 0.5, and 0.3 for the FF, NEE, fire, and ocean sectors, respectively. Higher $\boldsymbol{\sigma_\lambda}$ values are assigned to the NEE and fire sectors, reflecting the greater uncertainty associated with these emission sources.

Kernel Variance

$$\sigma^2 \sim Exp\ (2)$$

where *Exp* denotes the exponential distribution, and the kernel variance σ^2 is drawn from an exponential distribution with a rate parameter of 2.

Spatial Length Scale

$$\ell_s \sim \text{Gamma}(\alpha, \beta)$$

where α and β (with initial guesses of 0.5 for both) are parameters of a constrained Gamma distribution, from which random samples are drawn for ℓ_s within the range of 0.01 (i.e., a small positive value) to three times the maximum distance between any spatial points within the modeling domain.

Temporal Kernel Length Scale

$$\ell_t \sim Exp\left(N_{day}/7\right)$$

where N_{day} represents the total number of days in the year, specifically for the year 2020, and the number 7 (i.e., setting 7 days as a mean) corresponds to the standard duration of a synoptic weather cycle, measured in days (Ganesan et al., 2014).

Noise

$$\sigma_{noise} \sim HalfCauchy\ (1)$$

where $HalfCauchy$ represents the Half-Cauchy distribution with a scale parameter of 1.

Bias

$$D \sim \mathcal{N}(0, sd_{bias})$$

where we sample the bias parameter *D* from a normal distribution with a mean of zero and a standard deviation sd_{bias}, which is set to 0.5 ppm for Model 1.

Likelihood

$$\mathbf{y} \sim \mathcal{N}(\mu = \boldsymbol{f}, \sigma = \sigma_{noise})$$

where $\boldsymbol{f}$ represents the latent function, considered as latent variables, with its prior distribution defined by the GP (i.e., the GP mean and kernel).

Supplemental Figures

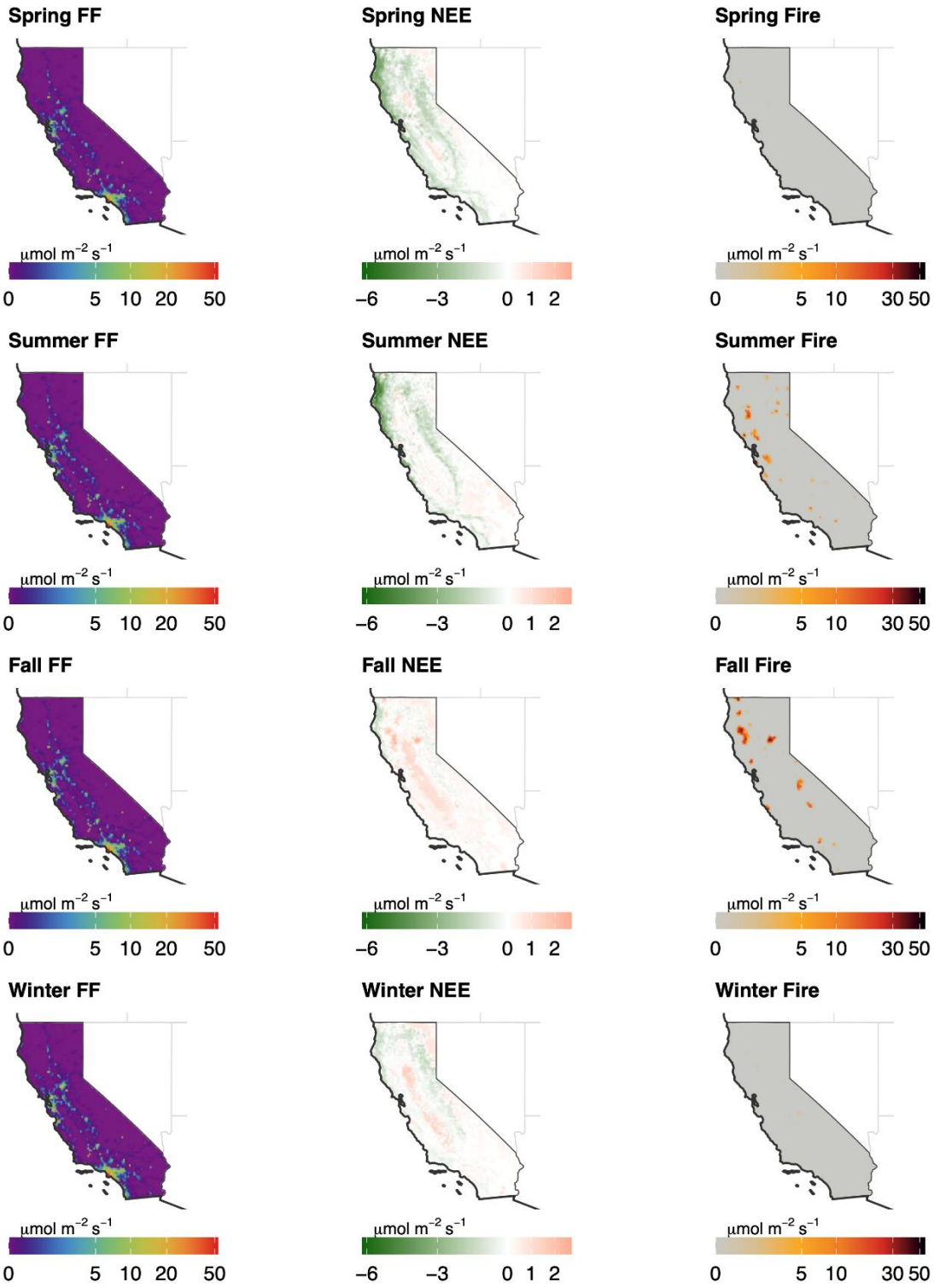

Figure S1. Seasonally-averaged 2020 a priori CO_2 emissions (μmol m^{-2} sec^{-1}) for California. Emissions from the terrestrial portion of California are shown for FF (left column), NEE (middle column), and Fire (right column) for the spring (first row), summer (second row), fall (third row), and winter (fourth row) months.

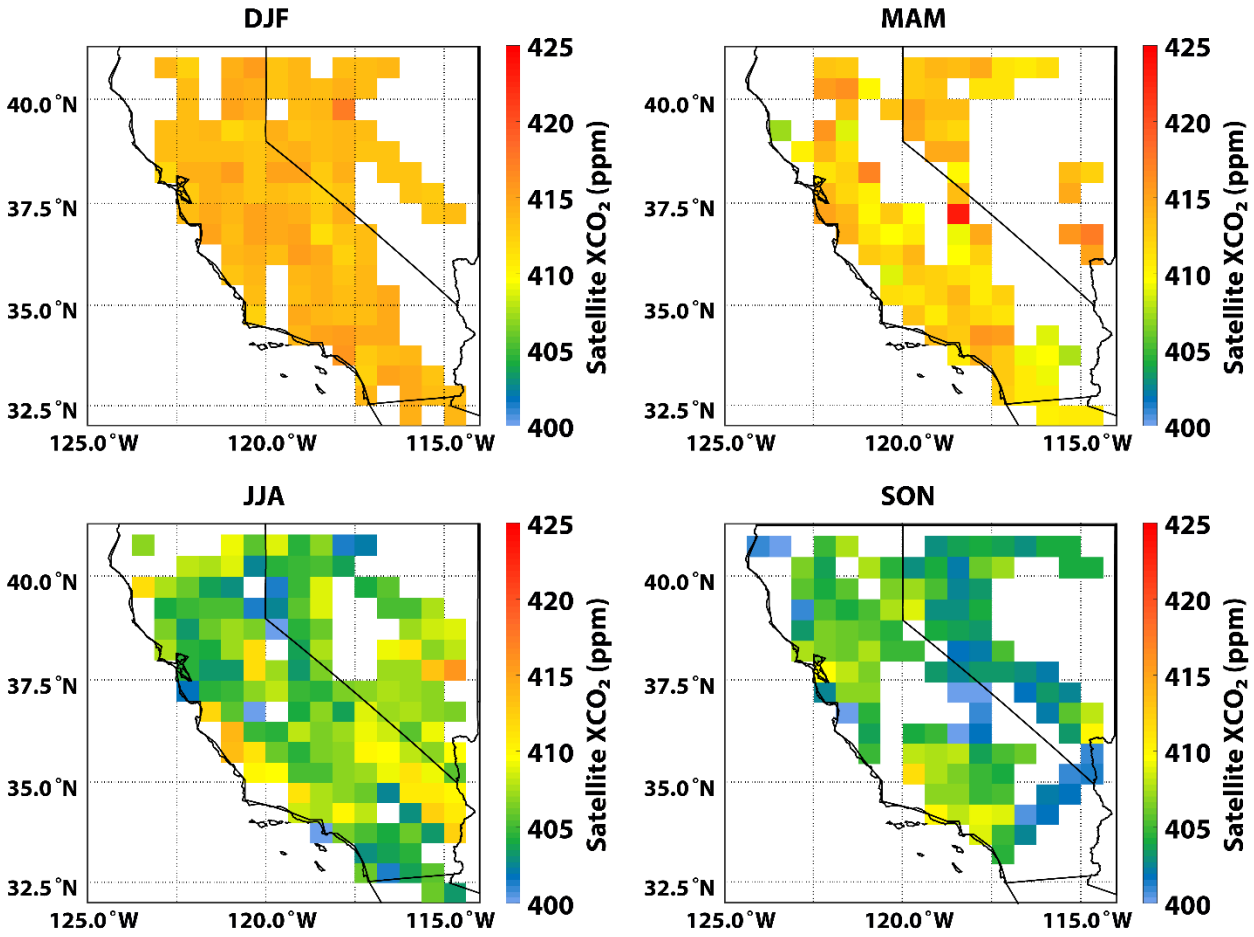

Figure S2. Seasonally-averaged XCO_2 concentrations (ppm) retrieved by OCO-2+OCO-3 during the year 2020 presented at the forward model spatial resolution of 0.5° × 0.625°.

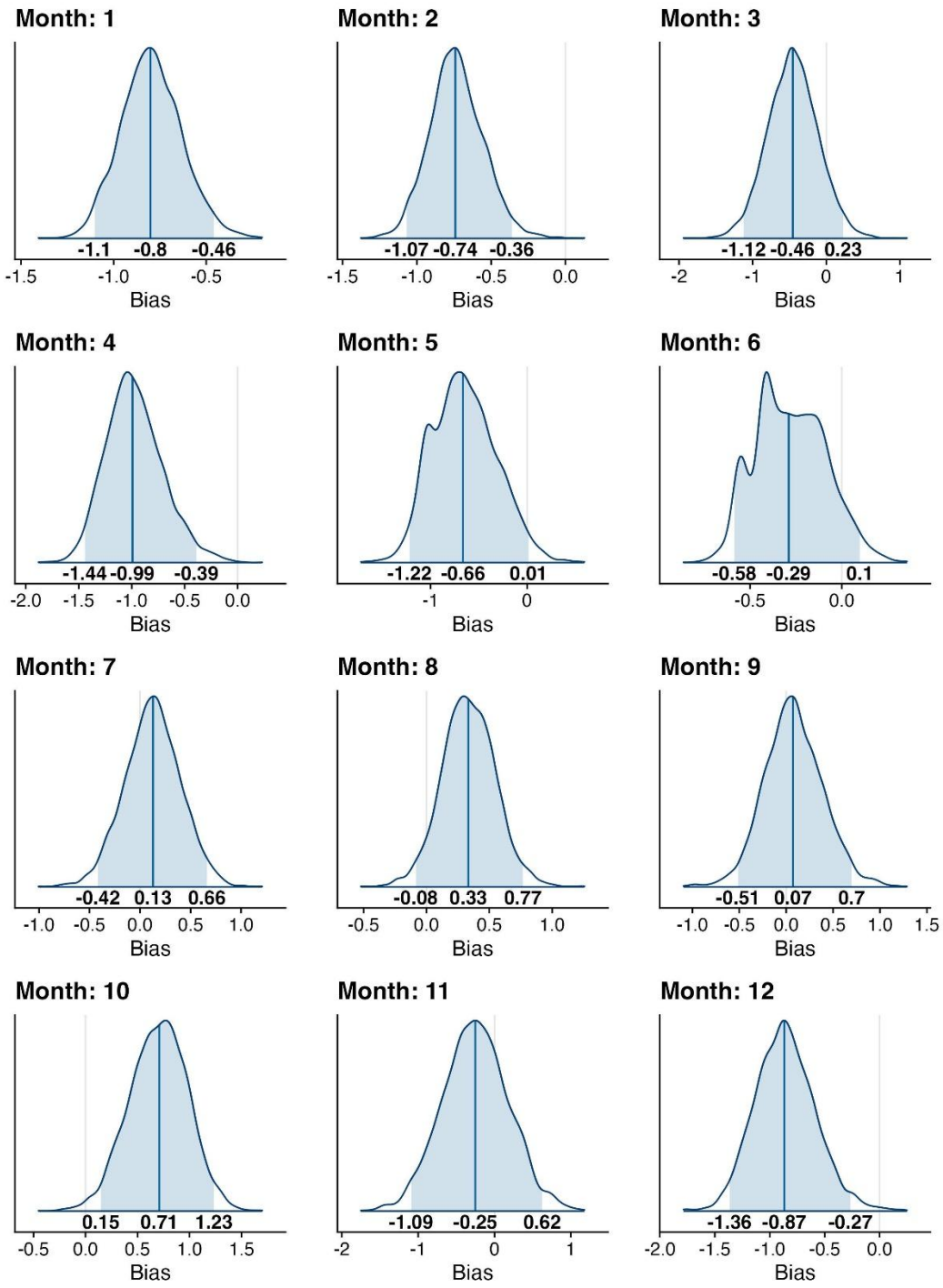

Figure S3. Probability density function for the estimated bias hyperparameter (D, in units of ppm; Model 1) in Eq. (5), shown by month. The shaded area represents the 95% confidence interval, while the vertical line indicates the median value. The bold numbers at the bottom represent the 2.5th, 50th, and 97.5th percentiles.

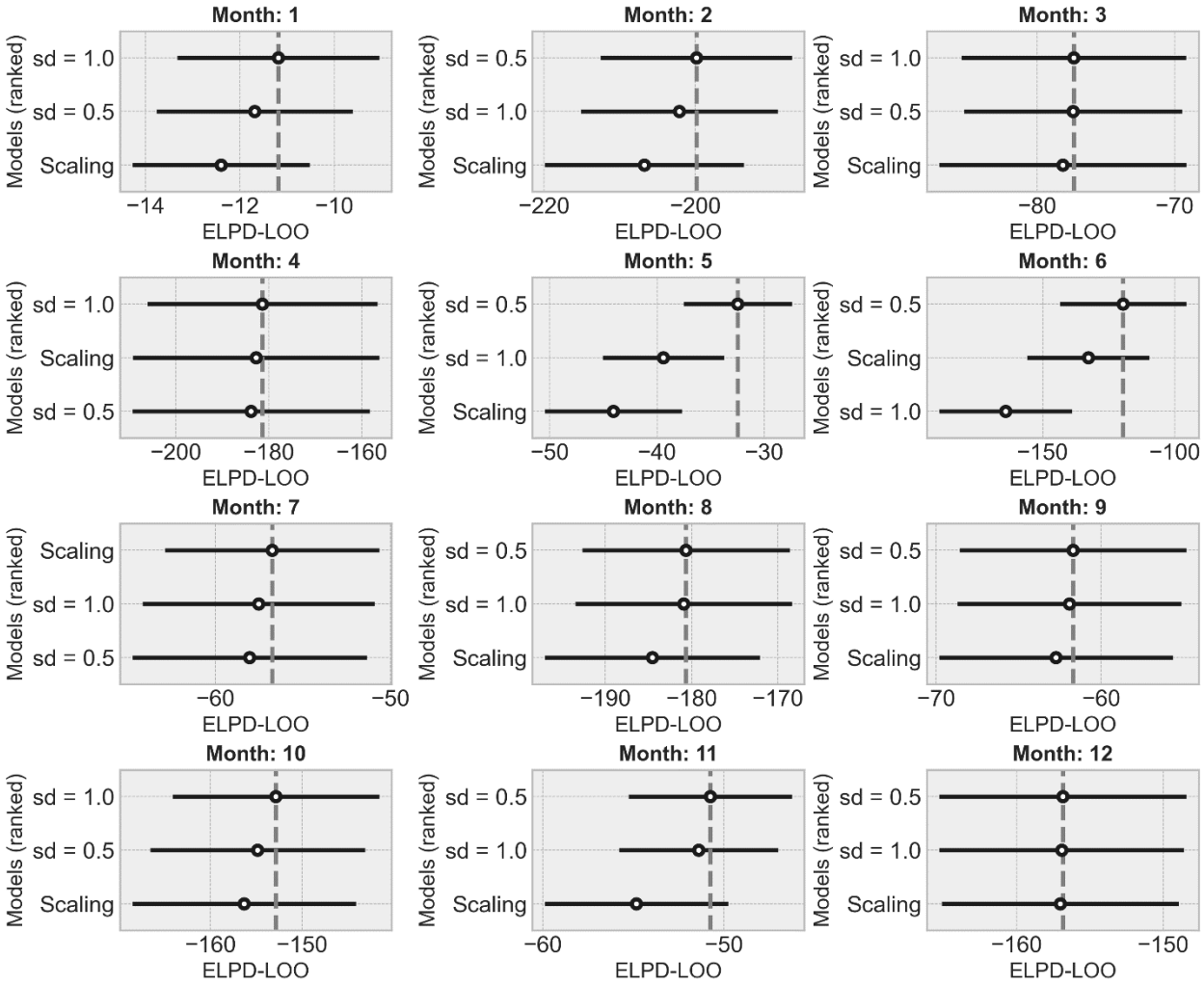

Figure S4. Model comparison using $ELPD_{LOO}$ for each month: Model 1 ("sd = 0.5"), where a standard deviation (sd) of 0.5 ppm is applied to the prior probability distribution for the systematic bias; Model 2 ("sd = 1.0"), with a standard deviation of 1.0 ppm for the prior for the systematic bias; and Model 3 ('Scaling'). The models are ranked according to their performance, with the best-performing model placed at the top. The uncertainty associated with the $ELPD_{LOO}$ value is depicted by the horizontal bars. The central estimate for the highest-ranked model for each month is indicated by the dashed vertical line.

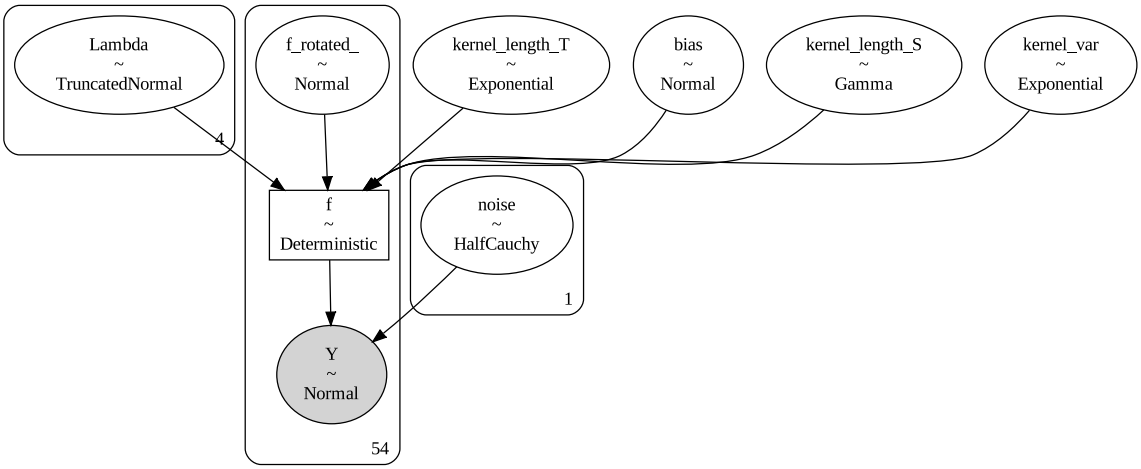

Figure S5. A graphical representation of the GP model structure for the month of September. It illustrates a probabilistic graph that outlines the dependencies among the random variables, showing the incorporation of the observed data (Y), hyperparameters, and prior distributions. It also notes the dataset size, 54, corresponding to the number of OCO observations utilized for September. This graph was generated using the ArviZ Python package (https://python.arviz.org/en/stable/).

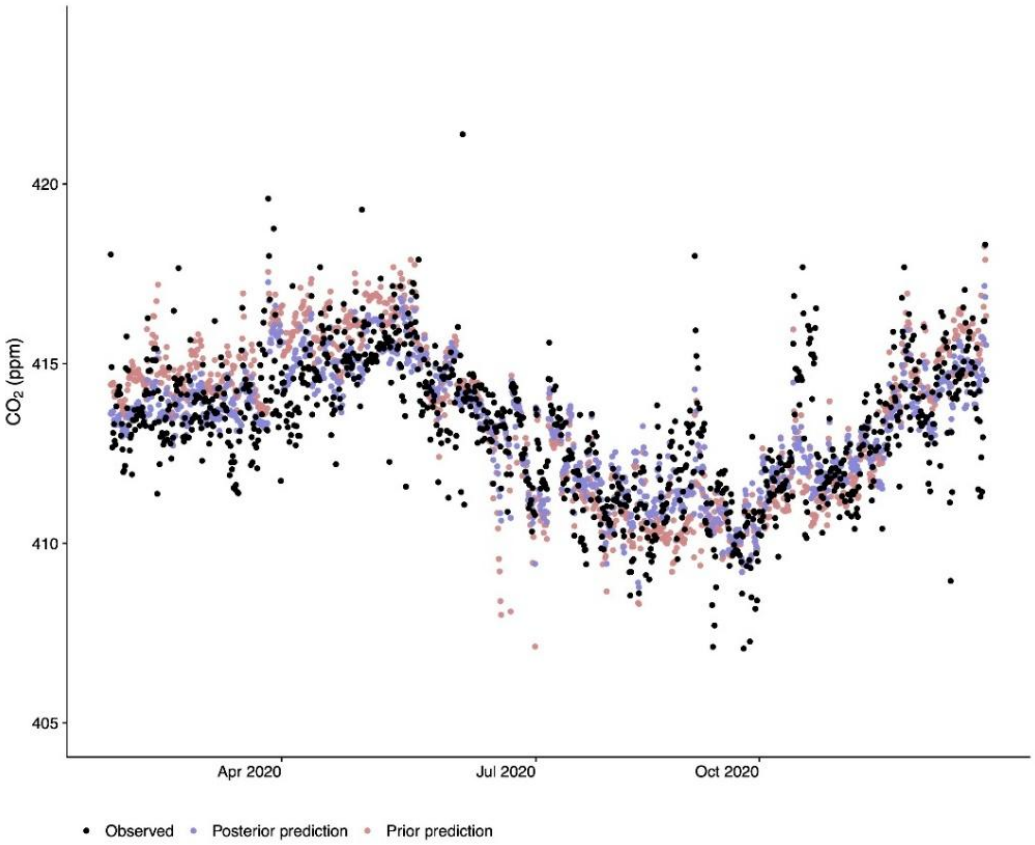

Figure S6. Timeseries of daily co-located prior (pink dots) and posterior (blue dots) model predicted XCO_2 concentrations (ppm) compared to OCO-2/3 observations (block dots) during the year 2020. Data are ordered based on unique spatiotemporal combinations, even though multiple observations were collected within a short time span on the same day.

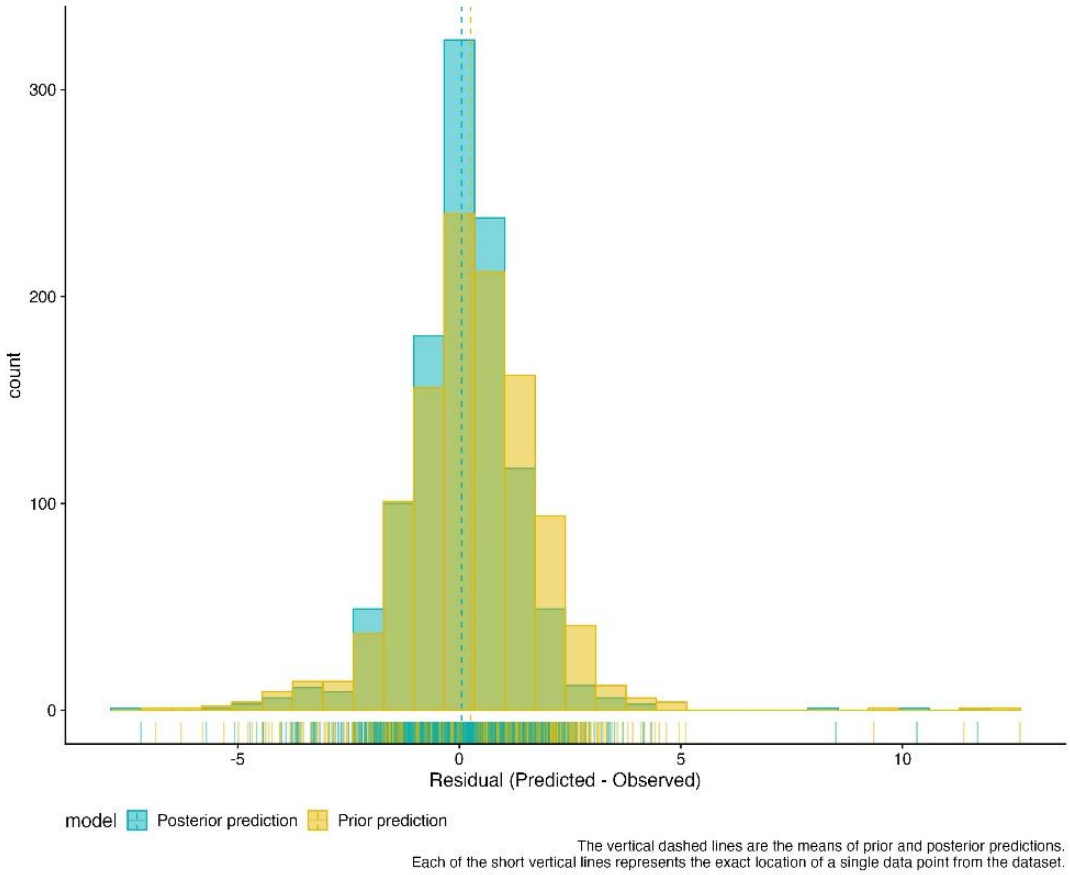

Figure S7. Annually-averaged, column-averaged XCO_2 error residuals (ppm) for prior (orange) and posterior (blue) model simulations (predicted – observed) for the year 2020.

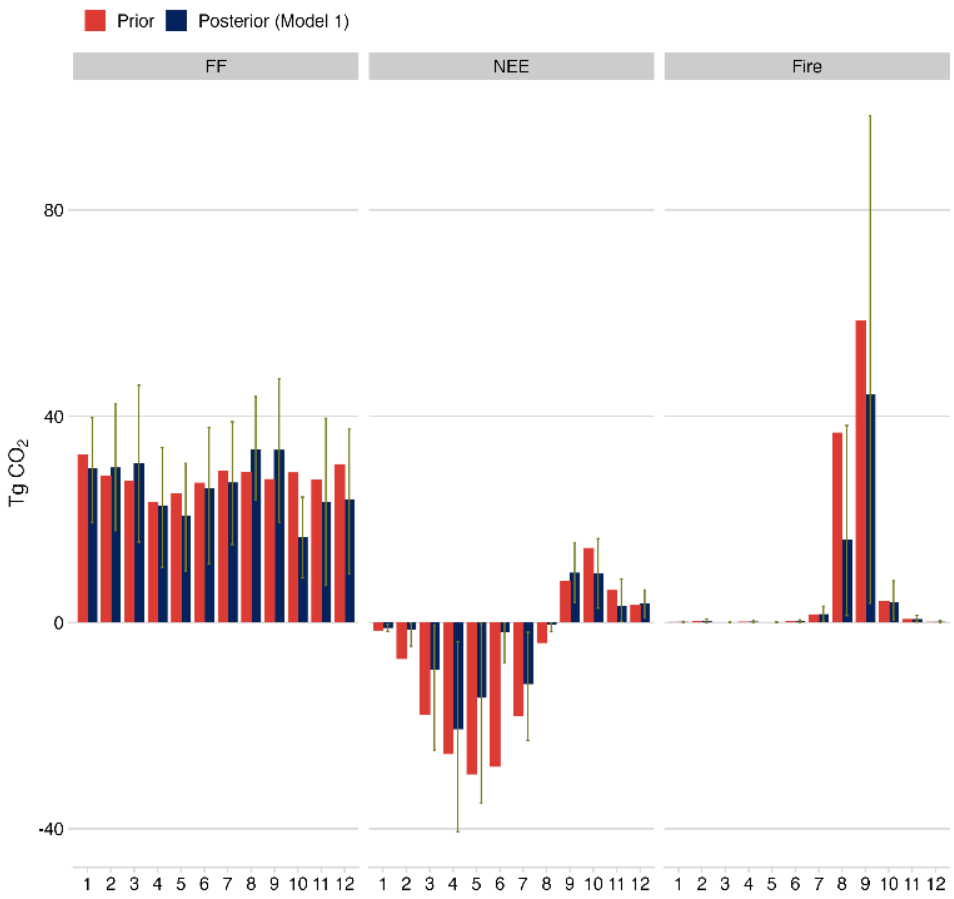

Figure S8. Monthly prior (red) and posterior (blue) CO_2 fluxes ($TgCO_2$ yr^{-1}) in California from fossil fuel, NEE, and fire during the year 2020.

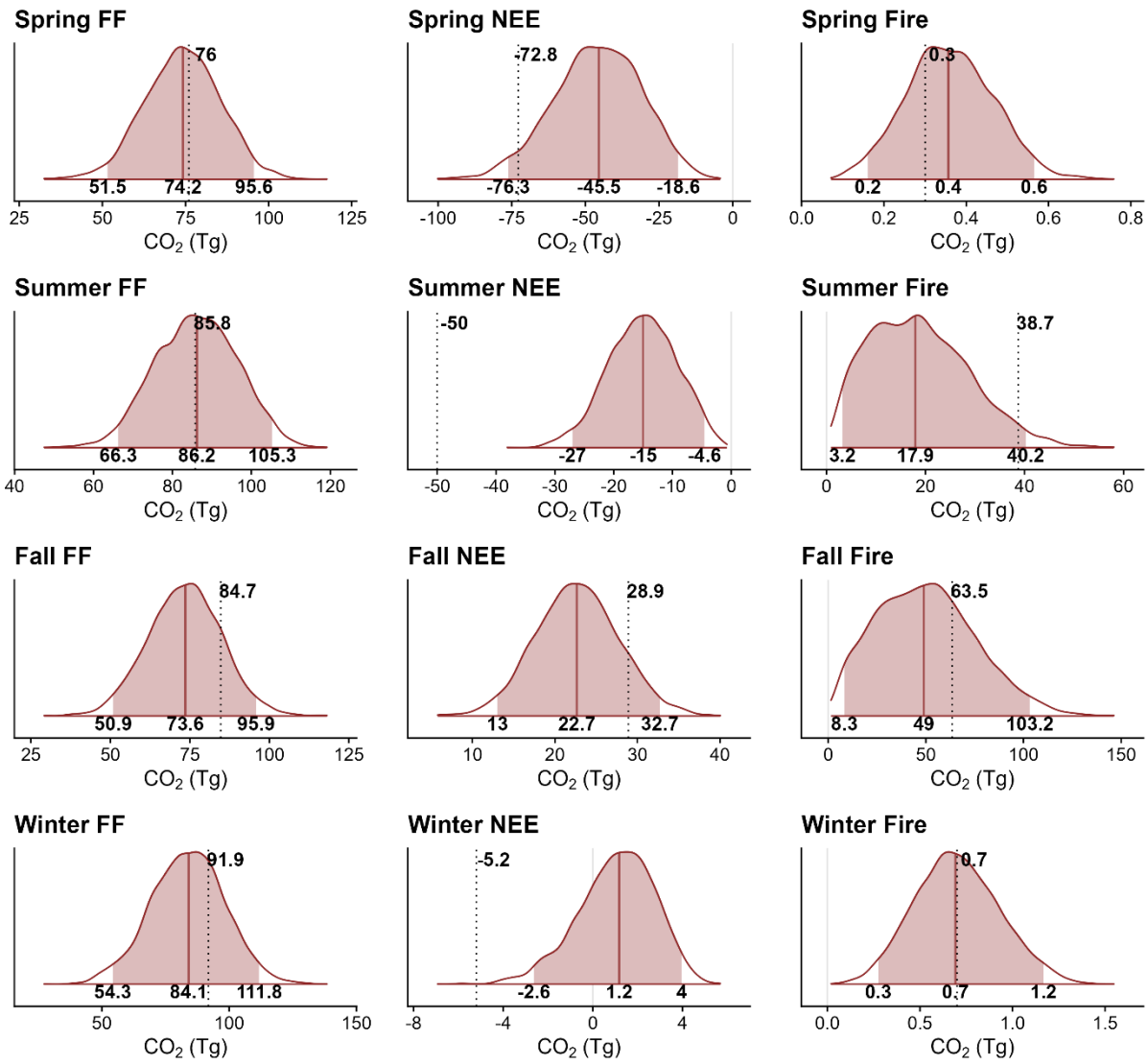

Figure S9. Seasonal total posterior GP inversion emission estimates using Model 1 by emission sector (Tg CO_2). The numerical labels at the base of each PDF denote the 2.5th, 50th (indicated by the bold vertical line), and 97.5th percentile estimates of the posterior emissions, respectively. The vertical solid line displays the median posterior flux rates, and the vertical dotted line indicates the prior emission estimate for each season, with the corresponding value displayed. Note the seasonal PDF shown here was estimated combining monthly MCMC samples for the season (4000 samples each month). This approach ensures that the final seasonal PDF accounts for variations observed across different months.