UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

MACHINE LEARNING APPLICATIONS FOR ATMOSPHERIC RETRIEVAL FROM EARTH OBSERVING SPECTROSCOPY WITH ROBUST UNCERTAINTY QUANTIFICATION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the
Degree of
DOCTOR OF PHILOSOPHY

By

William R. Keely

Norman, Oklahoma
2025

MACHINE LEARNING APPLICATIONS FOR ATMOSPHERIC RETRIEVAL FROM EARTH OBSERVING SPECTROSCOPY WITH ROBUST UNCERTAINTY QUANTIFICATION

A DISSERTATION APPROVED FOR THE

GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF:

Dr. Dimitrios Diochnos, Chair

Dr. Sean Crowell, Co–Chair

Dr. Miroslav Kramar

Dr. Charles Nicholson

Dr. Dean Hougen

Contents

Acknowledgments

The contributions of this dissertation are wholly owed to my family - who always told me I could do anything I set my mind to, or as my Grandad always said, “When there is a Will there is a way.” ..I took that literally. This work is dedicated to my Grahm, Grandad, Nana and Bompa. To my Mom who taught my how to be strong, my Dad who taught me patience and Colline who taught my kindness. To my sister Olivia and my cousin Micah who have always been there for me. To The Williams for welcoming me into their family with love and open arms.

Of course, a large part of my success in recent years is owed to my wife, Paige, who supported me during some of the hardest challenges.

I also want to thank my Committee Members and my co-advisors Sean and Dimitrios. I would not be on the course I am had Sean not opened the door for me, and I would not have stayed on that course if it were not for Dimitrios’ support and guidance. Our first meeting together in Grey Owl coffee will be one of my fondest.

I also want to thank Steffen Mauceri, Otto Lamminpää, Philip Brodrick, Jake Lee, and Umaa Rebbapragada for their mentorship and friendship. I would not be the caliber of researcher I am without their guidance.

Abstract

Satellite estimates of column averaged CO_2 (XCO_2) undergo two essential post-processing steps. Bias correction suppresses systematic error introduced during retrieval, and quality filtering removes soundings whose residual error is likely unacceptable for science use. NASA's Orbiting Carbon Observatory-2 currently relies on a linear regression bias model and a rule-based filter; both requiring extensive manual tuning and degrade when biases vary non-linearly. Planned missions (e.g. ESA's CO2M) will deliver far larger data volumes, and the computational cost of physics-based optimal estimation retrieval will become a bottleneck. Automated, scalable methods with explicit uncertainty estimates are needed to keep pace.

This dissertation presents novel machine-learning (ML) applications for bias correction and quality filtering of OCO-2 XCO_2 estimates. The models output point corrections and predictive uncertainties, handling non-linear error structure better than the existing linear scheme. Quality filtering is framed as a multi-objective search guided by domain scientist expertise, balancing error reduction, spatial–temporal coverage, and extrapolation risk. Bayesian optimization explores uncertainty thresholds, spatial weightings, and model settings, while explainable ML diagnostics keep the procedure transparent. The resulting ternary quality flag provides tailored data streams for example, a high precision subset for carbon flux inversion modeling and broader coverage for plume studies.

Complementing these post-processing advances, a conditional diffusion model trained on simulated spectra from a Radiative Transfer Model, retrieves XCO_2 directly from top-of-atmosphere radiances. The approach offers competitive accuracy, supplies full predictive uncertainty, and meets the speed requirements of next-generation greenhouse gas missions. In addition to the speedup in retrieval computation time, the diffusion model applied to potentially *ill-posed* inverse problem can recover non-Gaussian predictive posteriors which can speedup the analysis of *model discrepancy* between the physics-based forward model and Nature's forward model. A novel Evidential Transformer architecture is also presented for

the retrieval of methane from NASA's EMIT instrument.

Through these applications to Remote Sensing this dissertation also makes several novel contributions to the field of Data Science & Analytics. 1.) Post-hoc uncertainty calibration of Bayesian Optimal Estimation retrievals. 2.) Spatially weighting of Pseudo-labels and loss for Machine Learning methods applied to Remote Sensing problems. 3.) A mixture of experts Bayesian multi-objective optimization framework that incorporates Domain Expert input (Human-in-the-Loop) for fast development of science enabling quality filters. 4.) Diffusion for fast posterior sampling of *ill-posed* inverse problems to enable efficient model discrepancy analysis. 5.) Injecting physics information into a conditional diffusion architecture. 6.) A novel probabilistic Transformer for the unique modality of spectroscopic data with robust uncertainty quantification.

List of Figures

List of Tables

Chapter 1

Introduction

Space-based remote sensing has had a significant impact in advancing our understanding of Earth's complex atmospheric and terrestrial processes such as the Carbon cycle. The rapid pace of improvement in instrument technology for spectroscopy must also be paralleled by the development of computational algorithms that are capable of scaling and leveraging the growing data volumes produced by modern instruments. Data-driven and machine learning (ML) methods offer potential solutions to these computational challenges. However, these algorithms must also meet the precision requirements and most importantly gain the buy-in and trust of the science community.

Successful application of machine learning approaches at the mission level requires integration of domain science knowledge and data science expertise. On the machine learning side, interpretability tools can clarify how and why a model reaches its conclusions; while rigorous uncertainty quantification (UQ) can signal when those conclusions may be unreliable. This transparency gives domain scientists the insight and confidence needed to validate results against physical principles, identify edge cases for further investigation, and iterate jointly with ML practitioners. In turn, the resulting feedback loop accelerates the transition from proof-of-concept to reliable operational capability.

In this dissertation, we will cover the development of several machine learning applications for improving the operational products of several space-based remote sensing instruments: NASA's Observing Carbon Observatory 2 & 3 (OCO-2/3), and NASA's Earth Surface Mineral Dust Source Investigation (EMIT). Through the development of the applications, we will cover integration domain science expertise with data science methodologies, robust uncertainty quantification from traditional and generative machine learning methods, and generalization of the approaches to other missions/domains. These thrusts, which have culminated in three journal papers in both the American Geophysics Union ([84]; [50]) and the European Geophysics Union [58], as well as a conference workshop paper [57] in ICLR, will form the main chapters of this dissertation. The work also spans several invited talks and presentations ([53]; [54]; [55]; [56]; [52]; [67]; [120]; [47]) and has produced several datasets ([51]; [49]; [82]). Tables (2.1; 2.2; 2.3; 2.4) further illustrate where the author's works contribute to each Chapter.

A fourth journal publication [48] presents a causal analysis of a set of predictors for estimating a religiosity metric from the World Value Survey using interpretable machine learning methods combined with domain expert analysis. Although the author of this thesis was a major contributor to this additional paper it will not appear as a chapter in this dissertation.

In the first application (Chapter 4) we will cover is the use of interpretable machine learning methods to reduce errors in atmospheric carbon dioxide (CO_2) estimates made by NASA's OCO-2 instrument, often referred to as *bias correction* (BC) by the atmospheric remote sensing community. BC is an important post-processing step that must be applied to the CO_2 estimates before they can be passed on to downstream science users. Defining such errors or *biases* is done by taking the difference between the estimated from the space-borne instrument with well-validated ground truth sources. For OCO-2, Bias Correction has operationally involved an iterative processes of *hand fitting* a *correction model*, removing outliers, and then refitting the model until acceptable error performance has been reached. This iterative process is slow as each step (both the model re/fitting and outlier removal) are hand derived by a domain science expert and is difficult to reproduce. In [58] we demonstrate how a nonlinear model can reduce errors that are poorly corrected by the operational hand-fit linear model. The non-linear model can be fit once and reduce the dependence on repeated outlier removal, thus increasing the availability of good data for the downstream science users. In [84] expands on this work by introducing an uncertainty quantification of the BC models estimate, a long desired quality from the mission community. This second paper also reduces sources of *circularity* in the BC pipeline through more rigorous selection of ground-truth sources. These thrusts, in addition to the overall correction pipeline and integration with the science team through feature importance discussions, were the primary contributions of the thesis author to the paper.

The second application (Chapter 5) covers the application of machine learning and optimization methods for *quality filtering* (QF) of the CO_2 estimates from OCO-2. The operational quality filter is co-derived with the Bias Correction during the outlier removal step of the iterative process development described above. The domain science expert sets hand-derived bounds on variables that are estimated along with CO_2 and geometry variables derived from solar and instrument position - where error remains high after the linear correction model has been applied. Where to draw such bounds is a somewhat deceptively subjective question that changes depending largely on the downstream use case. For example, climate modelers can trade off fewer observations if it means lower biased data - such biases can propagate through during the data assimilation step in carbon flux inversion models. Users who are studying regional scale, for example, the tropics, or local scale, e.g., cities or power plants, typically require more data availability, at the expense of potentially higher biases. In the third paper[50], we present a framework for combining domain scientist criteria with machine learning methods that can flexibly address the data quality needs of multiple user communities. The proposed Bias Correction in [84] and QF in [50] are slated to be operationalized in an OCO-2 algorithm update.

In the last application (Chapters 6,7), we show how generative and deep learning methodologies can drastically reduce the computational cost estimating or *retrieving* quantities of interest from satellite observations of reflected solar spectra. Current retrieval algorithms take a full year to reprocess the operational OCO-2 record even with high-performance computing resources. New missions will bring in even larger data columns, and thus the remote sensing community is actively exploring faster methods, such as machine learning, to reduce the computational bottlenecks presented by improved instrument design. Although machine learning and data-driven techniques offer the potential for orders of magnitude speed up, they often come at the cost of precision. Therefore, an active and important area of research is how to successfully train models to generalize well when deployed. In [57] we present an approach for training a generative diffusion model on simulated spectra produced by a physics model and evaluate on real observations from NASA's OCO-2 instrument. We further illustrate how the generative model can allow for exploration of non-Gaussian predictive posteriors which can be used to gain insight into challenging aerosol scattering effects ill-captured by the physics model. We also showcase work that was presented at AGU on the ongoing development of a machine learning retrieval for NASA's EMIT instrument to estimate atmospheric methane (CH_4). For EMIT, we present an alternative approach to generating training data when a high-fidelity physics model is not available by *injecting* a methane signature into observations where CH_4 is believed to be at low background levels using a simplified radiative transfer equation.

Chapter 2

Contributions to Data Science & Earth Science

This dissertation highlights multidisciplinary contributions to both Earth Science and Remote Sensing of trace greenhouse gases as well as contributions to the field of Data Science. It is my hope that this work can be read by the Remote Sensing scientist as well as by the Data Science practitioner and can apply the novel contributions presented in this thesis to their own problems and applications.

Unique data science contributions.

1. **Robust uncertainty for Bayesian Optimal Estimation via quantile-based correction and calibration. (Chapter 4,5)** A non-linear quantile random forest-based regressor estimates per retrieval a σ prediction interval alongside its point estimate of the correction values. These intervals are used to calibrated the often overconfident Laplacian uncertainty provided from Bayesian Optimal Estimation MAP estimator and evaluated against independent validation. These robust uncertainties are then propagated through the observation pipeline at data assimilation into Ensemble Kalman models.

2. **Best practices for mitigating data leakage in remote-sensing machine learning. (Chapter 4,5,6)** The work adopts robust training/validation/test splits to assess and avoid data leakage from spatial autocorrelation and mixing observations from the same orbits between training and test. The approach also uses geo spatial visualiza-

tions combined with feature importance methods to assess the physical consistency of the trained model by mitigating feature choices that enable the machine learning model to "cheat", e.g., solar zenith and azimuth can be a strong proxy for seasonality/lat/lon. The work also illustrates how methods such as Leave-One-Out cross validation can be used to assess how well a method generalizes in geo-spatial setting e.g., if we didn't have ground truth here how well would we do? Ground truth proxy aware evaluation and discussion of residual "circularity" guide the selection of independent validation sources. Together these practices realistic responses in reported skill and improve reproducibility.

3. **A human-in-the-loop multi-objective optimization framework for quality filtering. (Chapter 5)** Quality filtering is posed as a Pareto problem that (i) minimizes error against ground truth, (ii) maximizes global throughput, and (iii) constrains extrapolation risk using the correction model's uncertainty. A Bayesian multi-objective search (MOTPE) explores thresholds and weightings while domain scientist input (Human-in-the-Loop) steers final candidate filter selection. The result is a transparent ternary flag that supports multiple user communities.

4. **Spatially aware loss engineering and pseudo-label construction aligned to science objectives. (Chapter 5)** Two complementary weighting strategies explicitly balance the core trade-off between retaining data and reducing bias. First, *station weighting* gives approximately equal total influence to each ground station when training truth-proxy sub-filters, preventing heavily sampled sites from dominating the model and yielding thresholds that generalize across the network. Second, a *small-area spatial weighting* is applied during pseudo-label construction for the small-area sub-filters. Raster bins with few observations are up weighted so they are not systematically filtered out, with a single parameter controlling the coverage MAE trade off as noisier regions are admitted. These strategies are novel in that they act both

at label construction and in the loss, are tuned within a multi-objective framework using interpretable parameters, and transfer to other remote-sensing machine-learning problems with uneven station density or regional coverage (e.g., aerosol/cloud quality flags, point-source detection, urban or regional analyses).

5. **Generative modeling for ill-posed inverse problems with discrepancy analysis. (Chapter 6)** A conditional diffusion retrieval provides amortized posterior inference through fast repeated posterior sampling per sounding without re-running the physical forward model, while capturing potentially non-Gaussian structure. Such non-Gaussian posteriors accelerate model discrepancy analysis because distinct modes correspond to alternative physical explanations (e.g., elevated aerosol optical depth versus surface BRDF mismatch), thus allowing for reduced need for multiple MCMC evaluations. Low-cost sampling also fast efficient retrieval of the atmospheric state quantities of interest, also removing the repeated evaluations of the expensive forward model evaluations used by the current state of the art retrieval algorithm.

6. **Incorporating physics information in machine-learning retrievals. (Chapter 6)** Physics enters through empirical-orthogonal-function (EOF) perturbations that emulate simulator–observation mismatch, and through conditional-mean priors (e.g., Conv1D or MLP, and operational GEOS-FP prior) that guide the diffusion process. These mechanisms improve generalization over vanilla neural network regressor of diffusion approaches.

7. **Fine-tuning of machine-learning atmospheric retrievals using in-situ observations. (Chapter 6)** A large co-location set with well validated in-situ estimates (within 100 km and 1 h) is used to fine-tune the ML retrieval, yielding reduced RMSE and improved uncertainty calibration on a fully held-out year. This demonstrates a practical path to bridge simulator-to-reality gaps and to maintain performance under operational drift. I demonstrate these effects on NASA's OCO-2 instrument retrievals

of XCO2.

8. **A Novel Evidential Spectroscopic Transformer for EMIT methane retrieval.** A band aware transformer treats spectra as ordered tokens by concatenating wavelength and radiance before self-attention, and outputs Bayesian Evidential parameters to produce decomposed, the predictive posterior/uncertainty. The architecture supports high throughput XCH_4 retrieval suitable for operational screening on EMIT and related hyperspectral missions. Additionally the attention weights provide a per pixel interpretability over the spectral bands.

Earth science and remote sensing contributions.

1. **Improved post-processing of column averaged** CO_2 **via non-linear bias correction. (Chapter 4,5)** A non-linear correction reduces residual error relative to the operational hand-fit model, most notably on data previously labeled as lower quality (QF=1): a 59% reduction in error variance over land and 67% over ocean, with additional regional improvements (e.g., northern Africa and parts of tropical South America/Asia). The approach decouples filtering from the correction (no longer constrained to linear fits) and enables a relaxed filter (QF_{New}) that increases usable throughput by 14% while maintaining comparable residual error to the operational product. The workflow is designed to be extendable to future ACOS algorithm updates and to OCO-3.

2. **A flexible, multi use quality filtering approach that formalizes trade-offs among coverage, accuracy, and small-scale fidelity. (Chapter 5)** Building on the improved correction, quality filtering is framed to serve diverse user communities—global flux inversion users who prioritize bias suppression and local/regional analysts who require higher data availability—by relaxing selected operational thresholds where the non-linear correction demonstrably removes residual structure. This

recovers usable soundings that the operational filter would otherwise discard, especially in regimes where error v. feature interactions are non-linear, thereby increasing coverage without degrading error statistics.

3. **Fast, machine learning-based direct retrievals of CO_2 and CH_4 that scale to mission volumes and aid model discrepancy diagnosis. (Chapter 6)** A diffusion-based XCO_2 retrieval provides efficient posterior sampling (e.g., $\sim$0.0003 s for 10 conditional samples per sounding on laptop-class hardware) versus tens of seconds per forward-model evaluation, while capturing non-Gaussian structure that better represents retrieval uncertainty. On a held-out 2022 OCO-2–TCCON set, diffusion with finetuning improves RMSE by $\sim 10\%$ over bias-corrected ACOS and yield better calibration uncertainty. The posterior diffusion samples and posterior-predictive residuals support targeted diagnosis of forward-model discrepancy (including aerosol effects). The approach is positioned for assimilation needs and generalizes to other greenhouse gases and missions. In parallel, methane work for EMIT addresses the high-volume, hyperspectral setting and motivates fast, robust screening and retrieval under operational constraints.

Table 2.1: Author's publications/presentations/abstracts/datasets included in Chapter 3.

Citation (abridged)	**DOI / URL**	**Status**	**Ref.**
Keely *et al.* 2023 – Non-linear ML bias corr. for OCO-2 XCO_2	`https://doi.org/10.5194/amt-16-5725-2023`	Journal Pub.	[58]
Mauceri & Keely *et al.* 2025 – Uncertainty-aware ML bias corr. Part 1	`https://doi.org/10.22541/essoar.174164198.80749970/v1`	Journal Pub.[†]	[84]
Keely *et al.* 2022 – Geophys. bias corr. with XAI	`https://doi.org/10.1002/essoar.10510174.1`	Poster	[53]
Keely *et al.* 2023 – UQ for ML bias corr.	`https://doi.org/10.22541/essoar.170224547.76630014/v1`	Poster	[55]
Keely 2023 – AMT ML bias-corr. weights (dataset)	`https://doi.org/10.17605/OSF.IO/CX53S`	Dataset	[51]

[†] Accepted journal publication, awaiting production as of 10 July 2025.

Table 2.2: Author's publications/presentations/abstracts/datasets included in Chapter 4.

Citation (abridged)	**DOI / URL**	**Status**	**Ref.**
Keely *et al.* 2025 – Uncertainty-aware ML bias corr. Part 2	`https://doi.org/10.22541/essoar.174164203.37422284/v1`	Journal Pub.†	[50]
Mauceri & Keely 2025 – Bias-corr. 2014–2024 (V1)	`https://doi.org/10.5281/zenodo.15085179`	Dataset	[82]

† Accepted journal publication, awaiting production as of 10 July 2025.

Table 2.3: Chapters 5–6 – diffusion-based retrievals and related work.

Citation (abridged)	**DOI / URL**	**Status**	**Ref.**
Keely *et al.* 2025 – Conditional diffusion retrieval of CO_2	`https://climatechange.ai/papers/iclr2025/12`	Workshop Paper – ICLR 2025	[57]
Keely & Mauceri 2022 – Toward ML retrieval of OCO-2 XCO_2	`https://doi.org/10.22541/essoar.167205912.27906076/v1`	Invited talk	[54]
Keely *et al.* 2025 – Methane diffusion retrieval (EMIT)	`https://doi.org/10.22541/essoar.173627342.24139272/v1`	Poster	[52]
Keely *et al.* 2024 – DisCO diffusion method	`https://doi.org/10.22541/essoar.173542226.67396908/v1`	Poster	[56]
Lee & Keely *et al.* 2024 – Spectral Transformers for EMIT methane retrieval	`https://ui.adsabs.harvard.edu/abs/2024AGUFMA53I.2209L`	Poster	[67]
Keely 2025 – Conditional diffusion XCO_2 (dataset)	`https://doi.org/10.5281/zenodo.15115962`	Dataset	[49]

Table 2.4: Author's additional works *not* included in the dissertation.

Citation (abridged)	**DOI / URL**	**Status**	**Ref.**
Jafarigol & Keely *et al.* 2023 – Religious affiliation	`https://doi.org/10.1007/s12115-023-00838-6`	Journal Pub.	[48]
Jafarigol *et al.* 2024 – Fine-tuning foundation model for methane detection	`https://ui.adsabs.harvard.edu/abs/2024AGUFMIN52A..04J`	Invited talk	[47]
Rothenberger *et al.* 2024 – Foundation models for EMIT methane sources	`https://ui.adsabs.harvard.edu/abs/2024AGUFM.H01...08R`	Poster	[120]
Lamminpää *et al.* (in prep.) – Simulation-based Gaussian inversion	—	Journal Pub. (in prep.)	

Chapter 3

Background

This chapter provides background and establishes notation for the methods used in later chapters. It first reviews atmospheric retrieval as a Bayesian inverse problem under Optimal Estimation (OE), in which a physics-based forward model is fitted to observations under prior constraints, with Jacobians supporting iterative MAP updates. OE diagnostics - the posterior covariance, gain, and averaging kernels (and their degrees of freedom for signal) are summarized as the primary tools for uncertainty and information analysis. The chapter then discusses the persistence of systematic residuals due to forward model imperfections and outlines the standard post-hoc bias correction procedure based on collocation with trusted references and regression on explanatory variables. Next, it broadens uncertainty treatment beyond OE's Gaussian assumptions by distinguishing aleatoric (noise/ambiguity) and epistemic (model/data limitations) components as learned by modern machine-learning retrievals that estimate predictive distributions. Finally, it characterizes remote sensing as a distinct machine learning modality multi-sensor and hyperspectral, irregular in space–time, label sparse, and non-i.i.d. thereby motivating the physics-aware, multi-sensor fusion and generalization strategies employed elsewhere in the dissertation.

3.1 Atmospheric Retrieval with Optimal Estimation

Optimal Estimation (OE) frames atmospheric retrieval as a Bayesian inverse problem: given measurements $\mathbf{y}$ and a forward model $\mathbf{F}(\mathbf{x})$ (radiative transfer plus instrument effects), estimate the state vector $\mathbf{x}$ by minimizing the posterior cost $J(\mathbf{x}) = (\mathbf{y} - \mathbf{F}(\mathbf{x}))^\top \mathbf{S}_e^{-1}(\mathbf{y} - \mathbf{F}(\mathbf{x})) + (\mathbf{x} - \mathbf{x}_a)^\top \mathbf{S}_a^{-1}(\mathbf{x} - \mathbf{x}_a)$, which balances measurement residuals via the measurement-error covariance $\mathbf{S}_e$ with deviations from a prior $\mathbf{x}_a$ weighted by the prior covariance $\mathbf{S}_a$. The solution is found iteratively (e.g., Gauss–Newton/Levenberg–Marquardt) using Jacobians $\mathbf{K} = \partial\mathbf{F}/\partial\mathbf{x}$. OE naturally provides uncertainty and diagnostic outputs: the posterior covariance $\hat{\mathbf{S}}$, the gain $\mathbf{G} = \partial\hat{\mathbf{x}}/\partial\mathbf{y}$, and averaging kernels $\mathbf{A} = \mathbf{GK}$, whose trace yields the degrees of freedom for signal. By tuning $\mathbf{S}_a$ (regularization) and augmenting $\mathbf{x}$ as needed (e.g., gas columns, surface pressure, albedo, aerosols), OE delivers stable, interpretable retrievals with consistent error budgets suitable for products such as column-averaged CO_2.

3.1.1 The Inverse Problem

Remote sensors record a vector of top-of-atmosphere radiances or reflectances $\mathbf{y} \in \mathbb{R}^m$. These observations depend non-linearly on an unknown atmospheric–surface state vector $\mathbf{x} \in \mathbb{R}^n$ through a forward model $\mathbf{F}$ and are corrupted by measurement noise $\boldsymbol{\varepsilon}$:

$$\mathbf{y} \;=\; \mathbf{F}(\mathbf{x}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\Big(\mathbf{0}, \mathbf{S}_\varepsilon\Big).$$

Optimal Estimation (OE) provides regularization by introducing a Gaussian prior parameterized by by a-priori state $\mathbf{x}_a$ with covariance $\mathbf{S}_a$.

$$\mathbf{x} \sim \mathcal{N}\Big(\mathbf{x}_a, \mathbf{S}_a\Big),$$

and solving for the maximum-a-posteriori (MAP) state that minimises

$$J(\mathbf{x}) = \Big(\mathbf{y} - \mathbf{F}(\mathbf{x})\Big)^{\mathsf{T}} \mathbf{S}_\varepsilon^{-1} \Big(\mathbf{y} - \mathbf{F}(\mathbf{x})\Big) + \Big(\mathbf{x} - \mathbf{x}_a\Big)^{\mathsf{T}} \mathbf{S}_a^{-1} \Big(\mathbf{x} - \mathbf{x}_a\Big). \tag{3.1}$$

A Gauss–Newton step updates the state as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Big(\mathbf{K}_k^{\mathsf{T}} \mathbf{S}_\varepsilon^{-1} \mathbf{K}_k + \mathbf{S}_a^{-1}\Big)^{-1} \Big[\mathbf{K}_k^{\mathsf{T}} \mathbf{S}_\varepsilon^{-1} \Big(\mathbf{y} - \mathbf{F}(\mathbf{x}_k)\Big) - \mathbf{S}_a^{-1} \Big(\mathbf{x}_k - \mathbf{x}_a\Big)\Big],$$

with $\mathbf{K}_k = \partial\mathbf{F}/\partial\mathbf{x}\Big|_{\mathbf{x}_k}$.

3.1.2 The Full-Physics Forward Model

The forward model $F(\mathbf{x}; \mathbf{b})$ predicts a high-resolution radiance spectrum given the state $\mathbf{x}$ and fixed calibration parameters $\mathbf{b}$.

Atmospheric absorption. For each wavelength λ_j,

$$y_j^{\text{abs}} \;=\; f_0(\lambda_j)\,\rho_s(\lambda_j)\,g\Big(\lambda_j, \theta_0, \phi_0, \theta, \phi\Big),$$

where f_0 is the extra-terrestrial solar flux, ρ_s is surface reflectance, and g integrates absorber cross-sections (CO_2, H_2O, O_2, ...) along the incident and emergent slant paths set by solar (θ_0, ϕ_0) and viewing (θ, ϕ) angles.

Multiple scattering. Single- and multiple-scattering by molecules, aerosols, and clouds are added via a numerical solution of the vector radiative-transfer equation, yielding monochromatic radiances $I(\lambda_j)$. Surface reflection is treated as Lambertian over land and Cox–Munk plus a dark-water Lambertian term over ocean.

Instrument model. High-resolution radiances are mapped to the detector grid through

$$y_j \;=\; C_1(\lambda_j)\,I(\lambda_j) + C_2(\lambda_j), \qquad \hat{\mathbf{y}} = \mathbf{H}\,\mathbf{y},$$

where C_1 and C_2 handle Doppler shift, spectral dispersion, and pixel-wise gain/offset, and $\mathbf{H}$ convolves the spectrum with the instrument line shape (ILS) before resampling.

The full-physics model thus provides both the synthetic spectrum $\hat{\mathbf{y}}$ and its Jacobian $\mathbf{K} = \partial\hat{\mathbf{y}}/\partial\mathbf{x}$, which are the quantities required by the OE iteration in Eq. (3.1).

3.1.3 Averaging Kernels

The solution covariance $\mathbf{S}_{\hat{x}} = \left(\mathbf{K}^{\mathsf{T}}\mathbf{S}_{\varepsilon}^{-1}\mathbf{K} + \mathbf{S}_{a}^{-1}\right)^{-1}$ quantifies the retrieval uncertainty. The averaging kernel matrix $\mathbf{A} = \mathbf{S}_{\hat{x}}\mathbf{K}^{\mathsf{T}}\mathbf{S}_{\varepsilon}^{-1}\mathbf{K}$ diagnoses vertical (or component) sensitivity: an $\mathbf{A}$ close to the identity indicates information primarily comes from the measurement rather than the prior. The averaging kernels are important for downstream data assimilation into flux inversion models or for comparing to other instruments.

3.1.4 Need for Post–Hoc Bias Correction

Even after an OE retrieval has converged, residual *systematic* errors often remain because the forward model cannot perfectly reproduce reality [117, 103]. Typical contributors include

- incomplete or inaccurate spectroscopic databases;
- simplifying assumptions in aerosol, cloud, or surface models;
- computational shortcuts such as limiting the number of iterations allowed for convergence (e.g., the optimizer gets stuck in a local minima).

These errors manifest as scene-dependent biases (e.g. latitudinal trends, surface–type offsets, air-mass dependencies). Consequently, a **post-hoc bias correction step** is routinely applied:

1. Collocate the retrieved product with highly accurate "truth" measurements (often referred to as *ground truth proxies*.
2. Fit an empirical model $\Delta x = f(\mathbf{z})$ that maps a set of explanatory variables (typically retrieved) $\mathbf{z}$ (e.g. solar zenith angle, surface pressure, aerosol optical depth) to the observed residual/bias/error. $\Delta x = \hat{x} - x_{\mathrm{truth}}$.
3. Apply the correction $x_{\mathrm{corr}} = \hat{x} - f(\mathbf{z})$ to all retrievals.

3.2 Uncertainty Quantification

Forward model errors, measurement/instrumen noise, and limited training data all inject uncertainty into retrieved quantities—but they do so in different ways. Classical Optimal Estimation propagates measurement noise through the Jacobian to give a predictive posterior covariance, $\mathbf{S}_{\hat{x}} \approx (\mathbf{K}^{\top}\mathbf{S}_{\varepsilon}^{-1}\mathbf{K} + \mathbf{S}_{a}^{-1})^{-1}$. This Gaussian formula obscures two important facts: (i) errors in the forward model known as *model discrepancy*, and (ii) potential of strong non-linearities invalidate Gaussian assumption.

Modern machine-learning (ML) retrievals confront these issues by learning the *full predictive distribution* $p(\mathbf{x} \mid \mathbf{y})$. Robust UQ can often be decomposed into two sources of uncertainty:

$$\underbrace{\mathrm{Var}[\mathbf{x} \mid \mathbf{y}]}_{\text{total}} = \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\Big[\mathrm{Var}(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})\Big]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}_{\boldsymbol{\theta}}\Big[\mathbb{E}(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})\Big]}_{\text{epistemic}},$$

where $\boldsymbol{\theta}$ are the learned network weights or ensemble members. The first term captures irreducible noise (e.g. instrument precision) the second quantifies extrapolation arising from limited training data, model discrepancy, or novel atmospheric conditions. ML-based uncertainty quantification turns retrieval products from black box point estimates into insight of when the model might be wrong, thus building trust for the broader science decision pipeline.

3.3 Remote Sensing Represents a New Modality for Machine Learning

Remote sensing (satellite and airborne Earth observation) constitutes a distinct data modality for machine learning. Beyond RGB photographs, it spans multi spectral and hyperspectral optics, synthetic aperture radar, thermal infrared, elevation, and ancillary geophysical layers; data arrive as irregular spatiotemporal tensors with heterogeneous resolutions, revisit rates, viewing geometries, and radiometric properties. Labels are sparse and geographically biased; samples are strongly autocorrelated in space and time; and distributions drift with seasonality, land use change, and sensor differences. These characteristics break common i.i.d. assumptions and strain image-only architectures, motivating methods that decouple spatial and temporal dependenciees, handle multi-sensor fusion, and incorporate physical constraints [119]. As Dr. Kerner and collaborators emphasize, treating satellite imagery "like standard photos" is inadequate for analyzing a dynamic planet and new ML tools tailored to remote sensing are needed, including novel architectures and generalization strategies. [3, 59, 134].

Chapter 4

Bias Correction of Column Averaged CO_2 from OCO-2

In this chapter we explore the application of data driven approaches to correcting errors (often denoted as biases) in the retrieved column averaged trace gas. Operational methods require extensive hand derivation of a linear model and quality filter to reduce the regime of interaction between bias and co-retrieved atmospheric variables to one that is largely linear.

4.1 Chapter Foreword

Summary for the data science practitioner

This chapter shows how a non-linear machine-learning correction can replace a hand-tuned linear bias model for OCO-2. It removes more of the leftover error, adds calibrated uncertainty so each corrected value comes with a trustworthy "error bar," and uses careful train/test splits to avoid leakage. The workflow keeps more usable soundings (especially where the linear approach was discarding data) without correcting out true enhancements like power plant plumes. In practice, treating uncertainty as an output and separating "correction" from "filtering" are key design choices. The chapter also makes use of robust methods to limit data leakage from spatiotemporal autocorrelation through robust evaluation metrics and exclusion of circular proxy signals in the training labels to help ensure that reported performance gains generalize beyond the training period and locations.

4.2 A Nonlinear Data-Driven Approach to Bias Correction of XCO2 for NASA's OCO-2 ACOS Version 10

Carbon dioxide (CO2) is a key contributor to radiative forcing, and hence rising levels in the atmosphere are of concern due to their influence on future climate. Following a long history of critical in situ measurements of CO2 at key sites around the world that allowed us to better understand the carbon cycle on continental scales, the era of space-based remote sensing began with the Scanning Imaging Absorption Spectrometer for Atmospheric Cartography (SCIAMACHY) in March 2002 [8] and the Atmospheric Infrared Sounder (AIRS) launched in May 2002 [4]. These missions were followed by dedicated CO2 observers such as the Greenhouse gases Observing SATellite (GOSAT) mission in 2009 [63] and the Orbiting Carbon Observatory-2 (OCO-2) in 2014 [22]. These data have yielded substantial scientific insights, such as a much more dynamic tropical carbon cycle compared with previous understanding (e.g., [70]; [108]; [25]; [111]), as well as studies into power plant emissions and plumes [93].

OCO-2 measures reflected solar radiances, from which column-averaged CO2 dry air mole fractions (XCO2) are retrieved with the NASA Atmospheric CO2 Observations from Space (ACOS) algorithm ([23]; [104]; [20]). Radiances are measured in the near-infrared oxygen A band near 0.76µm, the shortwave infrared weak CO2 band near 1.6µm, and the shortwave infrared strong CO2 band near 2.05µm. ACOS is based on Bayesian optimal estimation [118] that adjusts input parameters (e.g., XCO2, aerosols, surface characteristics, surface pressure) to maximize agreement between a modeled spectrum (derived by a radiative transfer model) and OCO-2 measurements. The parameters that best explain the measured radiances are labeled as the "retrieved" parameters. ACOS has undergone continuous improvement since the initial version.

Since the radiances contain uncorrected calibration artifacts and the modeled representation of the atmospheric radiative transfer is not perfect, retrieved parameters contain systematic biases. The inverse problem is under-constrained and leads to posterior errors in retrieved parameters that are correlated. To correct for errors in XCO2 arising from these types of dependencies, a multiple linear regression (MLR) bias correction with co-retrieved state variables or features used as predictors is fit to the difference (ΔXCO2) between the ACOS-retrieved XCO2 and a truth proxy estimate of XCO2. This method was first introduced for ACOS retrievals applied to the GOSAT instrument [146] and later extended to OCO-2 and OCO-3 ([102]; [126]; [129]). The MLR bias correction is fit in tandem with a quality filter of empirically defined thresholds on a set of features. The bias correction and quality filter are derived iteratively, with filter thresholds chosen restricting features to a range in which the relationship between ΔXCO2 and the parameters are mostly linear, improving the goodness of fit for the multilinear regression, which is then used in turn to retune the quality filter thresholds. The combined bias correction and quality filtering process is derived manually so that the final product must be hand-tuned for each algorithm update. After the feature-based correction, a footprint correction and global Total Carbon Column Observing Network (TCCON) offset are applied. The combined bias correction and quality filter process is robust across a set of ground truth proxy metrics and greatly reduces both mean bias and error scatter of XCO2 retrieved from OCO-2. Full details of the operational bias correction and filtering can be found in [102].

A drawback of applying the quality filter is the exclusion of data due to the linear assumption of the bias correction to which the quality filter limits the regime of interaction between state vector variables and ΔXCO2, thus removing data where the bias is nonlinear. Due to loss of data, the bias correction and quality filter are often disregarded for local studies ([93]; [87]) and are too limiting for certain regions [45]. Applying nonlinear machine learning techniques has shown great promise for the task of bias correction for GOSAT and GOSAT-2 [100] and TROPOMI [121]. Specific correction of 3D cloud biases for OCO-2-retrieved XCO2 [78] using a nonlinear method fit on a small set of features correlated with 3D cloud effects in addition to the linear operational correction is demonstrated in [83].

This research demonstrates a general nonlinear bias correction approach for OCO-2 build

10 (B10; [129]) via a machine learning method and provides a post hoc explanation of the overall contribution of the selected state vector features. Our nonlinear bias correction is shown to reduce systematic errors and increase the percentage of good-quality soundings by allowing for the relaxation of the hand-tuned thresholds employed with the standard quality threshold method. The framework presented in this paper for identifying informative features for bias correction can be adapted for future OCO-2 and 3 ACOS algorithm updates.

4.2.1 Data

To develop a bias correction, we define three truth proxy data sets for the true atmospheric column mole fraction. ΔXCO2 is then set as the difference between the raw ACOS retrieval of XCO2 and the truth proxy estimate of XCO2 as shown in Eq. (2.1). For the TCCON and model mean truth proxies, the OCO-2 averaging kernel is also applied as described in Taylor et al., 2023.

$$\Delta XCO_2 = XCO_{2,ACOS} - XCO_{2,Proxy} \tag{4.1}$$

We use the same proxy data sets used in the development of the operational bias correction [106]: co-located OCO-2 soundings with Total Carbon Column Observing Network (TCCON), a collection of small-area clusters of soundings for which XCO2 is not expected to vary above the instrument noise, and a set of modeled mole fractions whose underlying surface flux is constrained by the NOAA global in situ network [76]. Data sets include soundings from November 2014 through to February 2019. Each truth proxy captures a different scale of retrieval error and as such give complementary information as described in [102]. All data sets were sampled in conjunction with corresponding locations and times in the OCO-2 B10 L2 lite files which can be found here: (https://disc.gsfc.nasa.gov/datasets/). Spatial coverage and sounding count is shown in Figure 4.1. The newest version available Level 2 product is build 11 (B11), however at the time of this writing was undergoing re-processing.

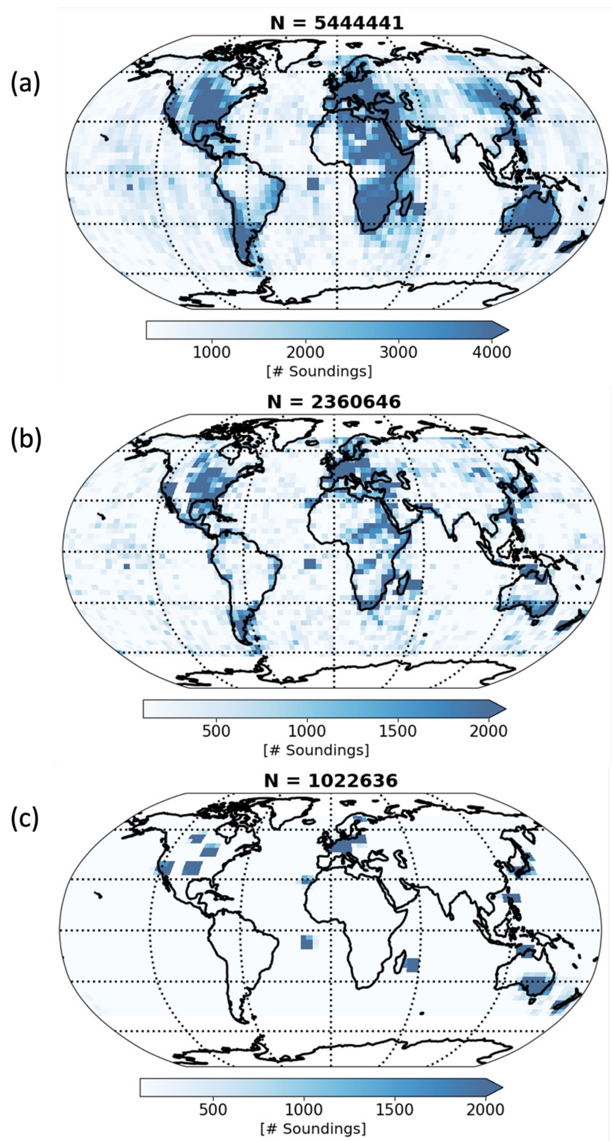

Figure 4.1: Spatial coverage for each truth proxy. The mean of a set of flux models is shown in panel (a), small area approximation is shown in panel (b), and TCCON is shown in panel (c).

4.2.1.1 Total Carbon Column Observing Network

TCCON is a system of ground-based sun-looking Fourier Transform Spectrometers with growing global coverage that retrieve dry air mole column averaged measurements of the trace greenhouse gases from radiances in similar spectral bands to OCO-2. Since each site has been extensively validated against WMO-traceable in situ observations aboard aircraft, TCCON offers the most accurate comparison for XCO2 [144]. While TCCON is well calibrated, site coverage is limited outside of North America, Europe, and Oceania. The TCCON data set therefore is spatially the sparsest of the three truth proxies and offers non-uniform point comparisons. We use the same dataset as the operational correction consisting of OCO-2 soundings co-located with TCCON GGG2014 measurements ([147]; [145]) in space (2.5º lat, 5º lon) and time (2h).

Table 4.1: TCCON sites used in bias correction and filtering for B10 ACOS. Station names marked with * denote island sites.

Station	Continent	Latitude	Altitude (m)	Date range	Citation
Saga*	Asia	33.2°N	7	201106–present	Shiomi et al. (2014) [123]
Orléans	Europe	48.0°N	130	200908–present	Warneke et al. (2019) [136]
Garmisch	Europe	47.5°N	740	200707–present	Sussmann and Rettinger (2018) [125]
Tsukuba*	Asia	36.1°N	30	200812–present	Morino et al. (2018a) [89]
Sodankylä	Europe	67.4°N	188	200901–present	Kivi et al. (2014) [61]
Rikubetsu	Asia	43.5°N	380	201311–present	Morino et al. (2018) [91]
Izaña*	Africa	28.3°N	2367	200705–present	Blumenstock et al. (2017) [6]
JPL	N. America	34.2°N	390	201103–201307; 201706–201805	Wennberg et al. (2017a) [139]
Bialystok	Europe	53.2°N	180	200903–201810	Deutscher et al. (2019) [30]
Bremen	Europe	53.1°N	27	200407–present	Notholt et al. (2019) [101]
Wollongong	Australia	34.4°S	30	200805–present	Griffith et al. (2014b) [38]
Park Falls	N. America	45.9°N	440	200405–present	Wennberg et al. (2017b) [138]
Réunion*	Africa	20.9°S	87	201109–present	De Mazière et al. (2017) [29]
Anmyeondo	Asia	36.5°N	30	201408–present	Goo et al. (2014) [36]
Darwin	Australia	12.4°S	30	200508–present	Griffith et al. (2014a) [37]
Lauder*	Australia	45.0°S	370	200406–present	Pollard et al. (2019) [113]
Lamont	N. America	36.6°N	320	200807–present	Wennberg et al. (2016) [140]
Karlsruhe	Europe	49.1°N	116	200909–present	Hase et al. (2015) [41]
Manaus	S. America	3.2°S	49.2	201408–201506	Dubey et al. (2014) [31]
Paris	Europe	48.8°N	60	201409–present	Té et al. (2014) [130]
Burgos*	Asia	18.5°N	35	201703–present	Morino et al. (2018b) [90]

4.2.1.2 Small Area Approximation

The small area approximation described in [102] offers insight into small scale drivers of bias and retrieval variability. The small area approximation truth proxy assumes that XCO2 within a 100km neighbourhood is largely uniform for a given overpass by OCO-2. This assumption is evaluated in [142], where it was found by using a high-resolution atmospheric model (GEOS-5), variance of XCO2 is around 0.1 ppm per 100km. The proxy offers improved spatial coverage compared to TCCON but struggles to capture biases with low variability over the small area.

4.2.1.3 Flux Models

A set of flux inversion models form the largest of the truth proxy data sets, both in number of soundings and in spatial coverage. The models included in this truth proxy set are found in Table 4.2. The posterior XCO2 fields produced by the models are sampled along OCO-2 tracks, the proxy is then computed as the average of the models at every sounding where there is good agreement (within 1.5 ppm) among models ([102]; [106]).

Table 4.2: Flux models used for the model-mean truth proxy. Abbreviations: TM5 – Transport Model 5; TM3 – Transport Model 3; LMDZ – Laboratoire de Météorologie Dynamique; EnKF – Ensemble Kalman Filter; 4D-Var – Four-Dimensional Variational.

Model name	**Institute**	**Transport model**	**Resolution** [lat×lon×time]	**Inverse method**	**Citation**
CarbonTracker	NOAA Global Monitoring Laboratory	TM5	2°×3°×3h	EnKF	Peters et al. (2007) [112]; CarbonTracker (2021) [99]
CarboScope	Max Planck Institute for Biogeochemistry	TM3	4°×5°×6h	4D-Var	Rödenbeck (2005) [115]; Rödenbeck et al. (2018[116]; CarboScope (2021) [85]
CAMS	Copernicus Atmospheric Monitoring Service	LMDZ	1.9°×3.75°×3h	4D-Var	Chevallier et al. (2010) [17]; CAMS (2021)[14]

4.2.2 Methods

4.2.2.1 Gradient Boosting

To model systematic error arising from co-retrieved state-vector elements, we employ *extreme gradient boosting* (XGBoost; [15]), a decision-tree ensemble that captures both linear and non-linear relationships. XGBoost sequentially fits regression trees [10], each new tree learning to predict the residuals left by the ensemble so far; the final prediction is the weighted sum of all trees.

At boosting round t the algorithm adds a new tree T_t minimizing the regularized objective.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell\Big(y_i, \hat{y}_i^{(t)}\Big) + \sum_{j=1}^{T_t} \Big(\tfrac{1}{2}\, \lambda w_j^2 + \alpha |w_j| + \gamma\Big), \tag{4.2}$$

where ℓ is the loss on sample i, w_j is the leaf weight of node j in the new tree, λ controls the ℓ_2 (ridge) penalty, α the optional ℓ_1 (lasso) penalty ($\alpha = 0$ if omitted), and γ penalizes creating additional leaves.

A candidate split of a node is selected by maximizing the *information gain*

$$\text{Gain} = \frac{1}{2}\Big(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\Big) - \gamma, \tag{4.3}$$

where G and H are the first and second derivatives (gradient and Hessian) of the loss with respect to the model output, and subscripts L and R denote the left and right child nodes.

For our experiments we use the mean-squared-error (MSE) loss between the truth-proxy bias y_i and its estimate $\hat{y}_i$:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \Big(y_i - \hat{y}_i\Big)^2. \tag{4.4}$$

Features that yield large values of (4.3) reduce residual error most effectively; summing their gains over the ensemble provides a global ranking of feature importance, offering post-hoc interpretability for correcting ΔXCO_2.

We determine optimal hyper-parameters via 10-fold cross-validation, resulting in $\lambda_{\text{LAND}} =$

2.5, $\gamma_{\text{LAND}} = 3.75$, $\lambda_{\text{OCEAN}} = 2.0$, and $\gamma_{\text{OCEAN}} = 10.0$; we found $\alpha = 0$ (i.e., no ℓ_1 term) performed best in both cases.

4.2.2.2 Quality Filtering

Soundings of the lowest quality are typically caught by the O2 A-band pre-processor (Taylor et al. 2012) and IMAP-DOAS [34] algorithms due to clouds and low SNR in the continuum and are then screened out before being run through the L2 retrieval algorithm [128]. After retrieval, an additional number of soundings are flagged and removed for which the ACOS algorithm failed to converge or for which the chi-squared difference between modelled and measured spectra is too large. Additionally, large unphysical outliers present in the tails of the conditional distributions of several atmospheric state variables are also removed by hand using domain expert selected thresholds. Finally, users can select for high-quality retrievals using the binary XCO2 quality flag (QF) with "good" data having a QF = 0, and "poor" data having QF = 1. The operational XCO2 quality flag is derived using a set of filters applied to the state vector variables found in conjunction with linear parametric bias correction. An initial linear correction is fit on soundings that have passed the pre-processing filtering steps. Each filter is then hand selected, QF = 1 data is removed, and the bias correction is re-fit until a final set of filters and linear model weights are derived that sufficiently reduce mean bias and scatter [102].

To assess the ability of the non-linear method to correct QF = 1 data and the potential for increased throughput of well corrected data, we derive a new quality flag (QFNew). Our flag is developed in a similar fashion to the B10 quality flag for use with the non-linear correction. The first step is to start with the same set of state vector variables and associated thresholds. Next, thresholds are relaxed for a selection of state vector variables that allow for higher sounding throughput, while maintaining or reducing corrected ΔXCO2 across truth proxies. Thresholds are never set to be more constraining than the B10 values in order to not remove soundings that are already considered to be of passing quality.

4.2.2.3 Training and Test Split

For training and evaluating the non-linear correction, we subset each of our truth proxy datasets into training and testing datasets. First, datasets are split by the two surface types: ocean and land. In the B10, both operation modes nadir (when the instrument is pointed *straight down*) and glint (off nadir viewing angle) are combined for the land bias correction due to low variance in feature importance between nadir and glint [102]. To compare to the operational correction, we also combine both modes for the land correction model. The land and ocean data sets are subset once more by truth proxy to identify informative features for the final land and ocean models. To ensure that model performance is indicative of how well the models generalize to unseen data, we hold out a year of data for evaluation of the final land model and ocean model. Models are trained on data from 2014, 2015, 2016, and 2017, then evaluated on data from 2018. Since data from 2019 is limited we exclude it from both training and evaluation.

4.2.2.4 Experiment Design

evaluate two methods for bias-correcting retrieved XCO2: a nonlinear machine learning model called XGBoost and, as a baseline, an MLR model trained similar to the hand-tuned model used in the operational correction. For correcting land nadir and land glint data, a single XGBoost model and MLR are trained using all three truth proxies. The predictor variables or features are the same for both model types. This allows for comparison between the nonlinear model and baseline linear method to properly assess that improved fit is coming from the captured nonlinearity and not just the inclusion of the additional predictors. A single XGBoost and MLR are derived for correcting ocean glint data, again using all three proxies and the same set of ocean features. We also compare our approach to the operational land correction and ocean correction for B10.

To identify a set of informative features to be used as inputs for the XGBoost land and ocean models, we first train a set of models independently on each truth proxy. These six models (three for land and three for ocean) are initially fit on a large set of potentially informative features, using QF=0+1 data. The resulting feature importance derived from these initial models is used to filter down the feature set to identify a subset of features that is highly informative across truth proxies. The resulting feature sets are combined to train the final proposed model pair (one for land and one for ocean), which are trained using all truth proxies.

Next, we compare the final models trained on QF=0+1 data against models trained only on good-quality data assigned QF=0 and then evaluate each model pair on QF=0 soundings that have been temporally held out. This is to ensure the ability of the nonlinear method to reproduce the linear model, which is the currently accepted community standard. Secondly, we evaluate the model trained on QF=0+1 data on the excluded regime of data labeled QF=1, where nonlinear relationships between ΔXCO2 and predictors become more pronounced. Finally, we derive a new quality flag (QFNew) used in conjunction with the nonlinear correction that increases the throughput of well-corrected data while maintaining similar error metrics as the operational filter and correction.

4.2.3 Results

4.2.3.1 Feature Selection

We select informative features for our bias correction following an iterative procedure. In the first step, we train XGBoost models for each proxy by surface type and operation mode (6 models in total). These initial models are trained using a large subset of co-retrieved state vector variables (shown in Table C1) which are potentially informative for correcting ΔXCO2 from the B10 L2Lite files. The resulting models are used to rank features according to their information gain which is defined in Eq. 2. Features that are less informative are removed from the set and new models are trained with the reduced feature set. Afterwards, feature importance is once again evaluated. To ensure robustness to correlation among features (which information gain does not account for) we calculate Pearson's correlation values between features. Features with an absolute Pearson value greater than 0.5 are included one a time and the feature with the highest importance is kept. This process is iteratively repeated until reaching a relatively small subset of maximally informative features. These features are combined to train the final bias correction models, which are trained on all proxies. Seven features are selected for land correction and five features selected for ocean as shown in Figure 4.2. The resulting features used in the final models and a brief description is shown in Table 4.3.

Features used for the operational correction are also highly informative for the proposed non-linear corrections and include the difference between the retrieved CO2 profile and prior profile used for land and ocean (`co2_grad_del`), and two surface pressure difference terms `dpfrac` for land and `dp_sco2` for ocean [60]. The `co2_grad_del` is the change in profile shape and the prior and is calculated as the difference dry air mole fraction at the surface, denoted as CO2(1), to the fraction at 0.6316 times the retrieved surface pressure, and is in units of parts per million (ppm). The calculation for `co2_grad_del` is shown in Eq. 4. For land, the `dpfrac` term is a difference ratio that considers the smaller dry air column over higher elevations and is defined in Eq. 5 where XCO2,raw is the uncorrected retrieval of the column average and $P_{\text{ap,SCO2}}$ and P_{ret} are the prior surface pressure at the strong band pointing offset and retrieved pressure respectively. For ocean, `dp_sco2` is used and is the retrieved surface pressure minus the strong band prior. The extensive use of `co2_grad_del` and surface pressure deltas for bias correction is discussed in [62].

$$\text{co2\_grad\_del} = [\text{CO2,ret}(1) - \text{CO2,ret}(0.6316)] - [\text{CO2,prior}(1) - \text{CO2,prior}(0.6316)] \quad (4.5)$$

$$\text{dpfrac} = X_{\text{CO2,raw}}\left(1 - \frac{P_{\text{ap,SCO2}}}{P_{\text{ret}}}\right) \quad (4.6)$$

For land, the `h2o_ratio` is used and is the ratio of XH2O estimated by single band

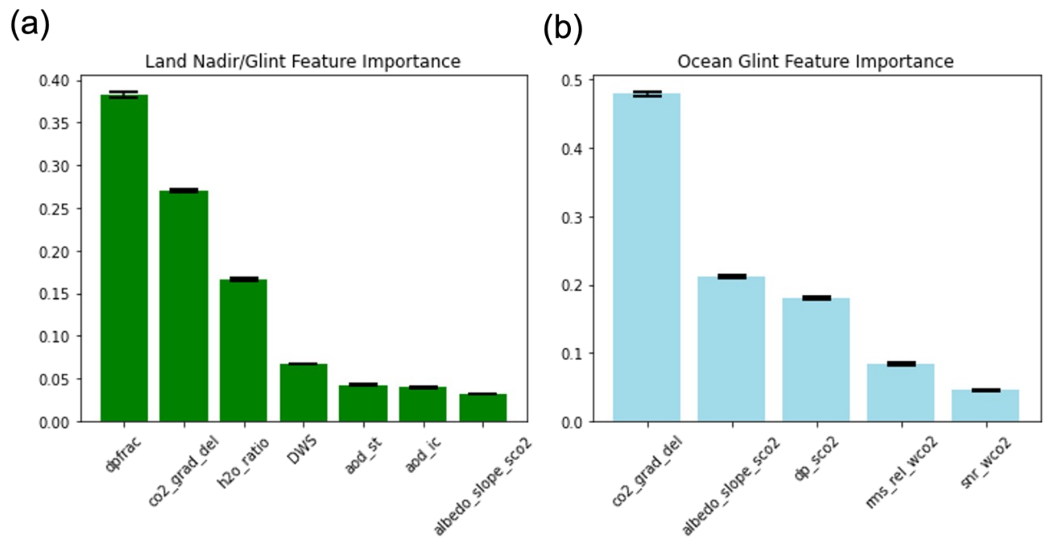

Figure 4.2: Feature importance for the final XGBoost models: (a) land; (b) ocean. Error bars show the variance across 10 random seeds.

retrievals from the strong and weak CO2 bands separately using the IMAP-DOAS algorithm, which can differ from unity in the presence of atmospheric scattering [128]. We use three aerosol features for our bias correction over land scenes. The first being the sum of dust, water, and sea salt optical thickness termed DWS. We include retrieved ice particle optical depth (`aod_ice`) and the finer stratospheric aerosol optical depth (`aod_strataer`). The last feature used for land, as well as for ocean, is the albedo slope for the strong CO2 band termed `albedo_slope_sco2`. This variable represents the slope of the reflectance across the strong CO2 spectral band for land soundings and the slope of the Lambertian component of the combined Cox–Munk and Lambertian Bidirectional Reflectance Distribution Function (BRDF) for ocean soundings [21]. In addition to the `co2_grad_del`, `albedo_slope_sco2` and `dp_sco2`, two additional variables are used for the correction of Ocean G scenes. These are `snr_wco2`, which is the estimated signal to noise ratio derived during optimal estimation; and finally, `rms_rel_wco2` which is the percent residual error from the forward modeled radiance for the weak CO2 to the measured radiance.

Table 4.3: Selected features used in the bias-correction models. N=nadir, G=glint.

State variable	Description	Surface/Mode
dpfrac	Surface-pressure difference ratio accounting for reduced dry-air columns at high elevation [60].	Land (NG)
h2o_ratio	Ratio of retrieved XH_2O in weak and strong CO_2 bands (IMAP-DOAS).	Land (NG)
DWS	Sum of dust, water, and sea-salt optical depth.	Land (NG)
aod_strataer	Stratospheric aerosol optical depth at 0.755 µm.	Land (NG)
aod_ice	Ice-cloud optical depth at 0.755 µm.	Land (NG)
co2_grad_del	Vertical gradient between retrieved and prior CO_2 profiles (Eq. 4.5).	Land (NG) / Ocean (G)
albedo_slope_sco2	Strong-band reflectance slope (land) or BRDF Lambertian slope (ocean).	Land (NG) / Ocean (G)
dp_sco2	Strong-band surface-pressure delta (retrieved minus prior).	Ocean (G)
rms_rel_wco2	Weak-band residual RMSE relative to signal.	Ocean (G)
snr_wco2	Weak-band continuum signal–to–noise ratio.	Ocean (G)

4.2.3.2 Model Evaluation for QF = 0

To ensure that the non-linear method generalizes the linear relationships largely observed for QF = 0, we evaluate two XGBoost models: one which is fit on QF = 0 + 1, and one fit on QF = 0, to a MLR fit on the same feature set as the non-linear models. As the operational quality flag is hand-tuned by re-fitting a MLR, the regime between the variables selected for correction and systematic error are reduced to mostly linear relationships. The non-linear method has only a marginal improvement over the MLR and B10 correction on soundings that are passed by the operational quality filter over land (0.02-0.04 ppm), and a slightly more substantial improvement over ocean (0.09-0.10 ppm), on the evaluation data. We found that retraining the XGBoost models on QF = 0 data does not offer a substantial reduction in error despite initial XGBoost models being trained on un-filtered data. We forgo the iterative refitting approach that is required for the MLR and operational correction by training once on QF = 0 + 1 data. Table 4.4 shows the QF = 0 RMSE results for XGBoost models trained on both QF = 0 + 1 data and QF = 0 data, alongside the MLR model fit to the filtered regime for 2018 and B10 operational correction.

Table 4.4: RMSE (ppm) on QF = 0 data for 2018.

Land (NG)

Truth proxy	XGB_{0+1}	XGB_0	MLR_0	B10
Small area	0.83	0.82	0.84	0.85
TCCON	1.15	1.14	1.19	1.20
Model mean	1.05	1.05	1.09	1.11
All	1.03	1.02	1.05	1.07

Ocean (G)

Truth proxy	XGB_{0+1}	XGB_0	MLR_0	B10
Small area	0.45	0.44	0.56	0.52
TCCON	0.83	0.81	0.89	0.95
Model mean	0.67	0.66	0.78	0.76
All	0.65	0.65	0.75	0.74

4.2.3.3 Correcting Outside the Filtered Regime (QF = 1)

Correction of systematic error outside of the quality filtered regime (QF = 1) is difficult to fit with a linear model. Strong non-linearities are observed for many of the co-retrieved state vector variables and ΔXCO2. For many variables this behavior is observed over un-physical values in a few spurious soundings and are easily filtered out. Variables such as `h2o_ratio` which are responsible for the bulk of the quality filtering (`h2o_ratio` thresholds remove 10% of soundings) exhibit such non-linear characteristics over their marginal distributions. The dependent linear correction and quality filter is prohibitive for correcting and passing data in these regions of the domain. Figures 4.3 and 4.4 illustrate the interaction between state variables chosen for correction and ΔXCO2. The non-linear model (green) improves both mean and variance of ΔXCO2 over both the raw ΔXCO2 (red) before correction, and B10 correction (blue). Table 4.5 displays the RMSE scores of the XGBoost corrected XCO2 and operational corrected XCO2 for QF = 1 data. The non-linear correction provides a large improvement in reducing the residual error for QF = 1 data over the operational correction with a 1.33-2.26 ppm improvement for land data and 1.11-1.38 ppm for ocean. These errors are still significantly larger than the corresponding QF = 0 errors.

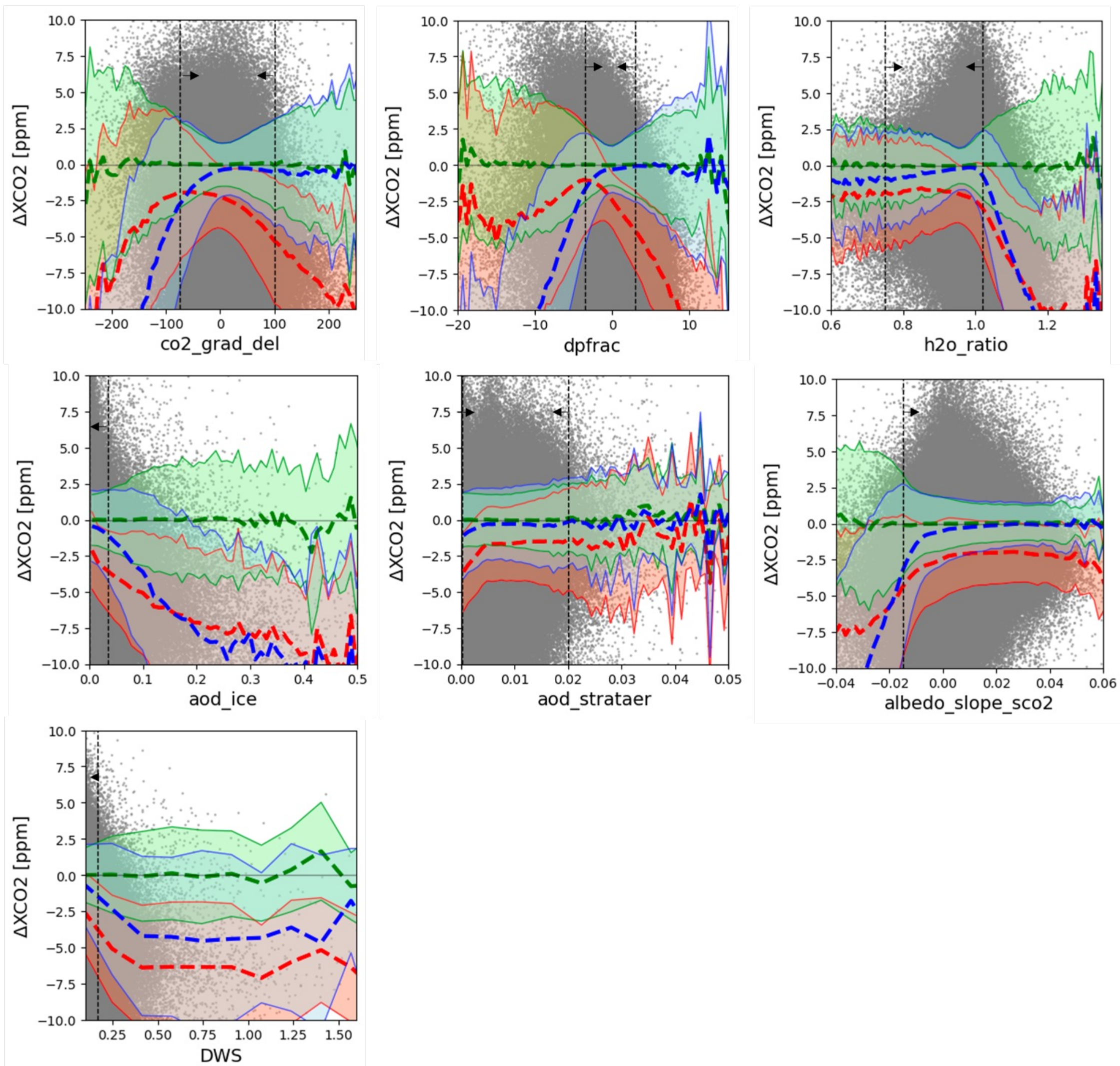

Figure 4.3: ΔXCO_2 vs. land features (2018). Red: raw; blue: B10; green: XGBoost. Vertical dashed lines denote B10 QF thresholds; arrows indicate QF = 0 region.

Table 4.5: RMSE (ppm) on QF = 1 data for 2018.

Land (NG)

Truth proxy	XGB_{0+1}	B10
Small area	1.92	3.25
TCCON	2.81	5.07
Model mean	2.46	3.95

Ocean (G)

Truth proxy	XGB_{0+1}	B10
Small area	1.25	2.36
TCCON	1.68	2.90
Model mean	1.53	2.91

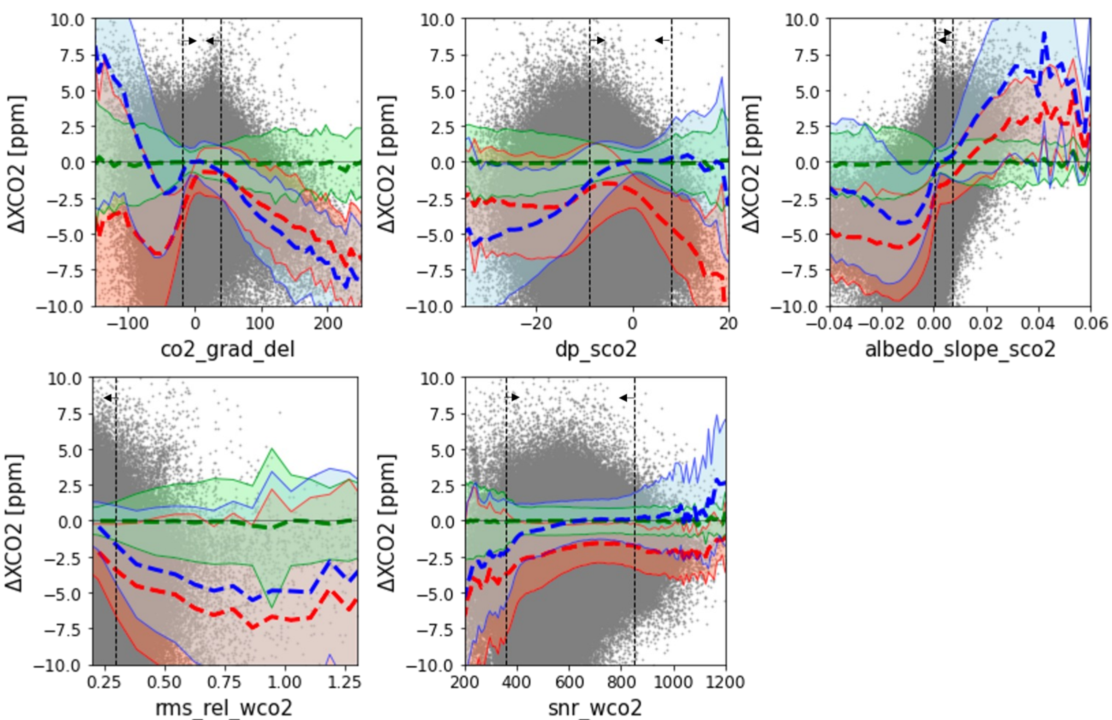

Figure 4.4: ΔXCO_2 vs. ocean features (2018). Red: raw; blue: B10; green: XGBoost. Vertical dashed lines denote B10 QF thresholds; arrows indicate QF = 0 region.

4.2.3.4 Comparison with B10

For the operational correction, regression weights for the linear model are hand-selected that have good agreement in their correction across truth proxies. The full operational correction also includes a fixed correction for each of OCO-2's eight footprints as described in [106]. To provide a fair comparison between the full correction models, we also apply the footprint correction after applying the non-linear feature correction. Table 4.6 shows the mean and 1 σ standard deviation for each bias correction and QF regime. The largest improvement in the non-linear method over B10 comes when correcting QF = 1 data. Achieving a 59% improvement in reduction of error variance for land, and a 67% improvement for ocean data respectively. The improvement in correction over B10 is less significant for QF = 0 with improvement of 8% for land and 19% for ocean.

Regionally, the non-linear correction shows up to a 0.5 ppm improvement over northern Africa, where the B10 correction appears to underestimate ΔXCO2 in comparison. A reduction in biases is also observed in large parts of South America's tropical and sub-tropical regions as well as parts of tropical Asia shown in Figure 5a. These regions also contain the largest difference in Land NG correction between the methods with an average difference (B10–XGBoost) of –0.5 ppm. There is a slight positive difference between methods over the Amazon Basin and Congo Rainforest (Figure 5e). Figures 5c and 5d illustrate the improvement of the non-linear method to correct QF = 1 data over the operational approach. For QF = 1, where the interaction between features and error is non-linear, large biases in XCO2 remain after operational correction. The XGBoost model reduces these remaining biases in many regions, indicating that there may still be usable data that is filtered out by the operational QF when paired with the non-linear correction.

4.2.3.5 Increased Sounding Throughput

One of the benefits of the non-linear bias correction is the potential for increased throughput of well-corrected QF = 1 data. Improved throughput of well corrected data would be of

Table 4.6: Comparison of combined proxy mean and standard deviation for the XGBoost-corrected XCO2, the operational B10 correction, and the un-corrected (Raw) XCO2 in 2018, split by quality-flag (QF) regime and surface type.

QF	Surface/Mode	XGBoost	B10	Raw
QF=0	Land NG	-0.04 ± 1.02 ppm	-0.13 ± 1.06 ppm	-1.90 ± 1.68 ppm
	Ocean G	0.02 ± 0.64 ppm	0.18 ± 0.71 ppm	-1.61 ± 1.10 ppm
QF=1	Land NG	0.01 ± 2.45 ppm	-1.24 ± 3.83 ppm	-2.83 ± 3.69 ppm
	Ocean G	-0.06 ± 1.50 ppm	-1.17 ± 2.59 ppm	-2.79 ± 2.75 ppm
QF=0&1	Land NG	-0.03 ± 1.75 ppm	-0.59 ± 2.64 ppm	-2.78 ± 2.73 ppm
	Ocean G	-0.01 ± 1.07 ppm	-0.36 ± 1.85 ppm	-2.09 ± 2.03 ppm

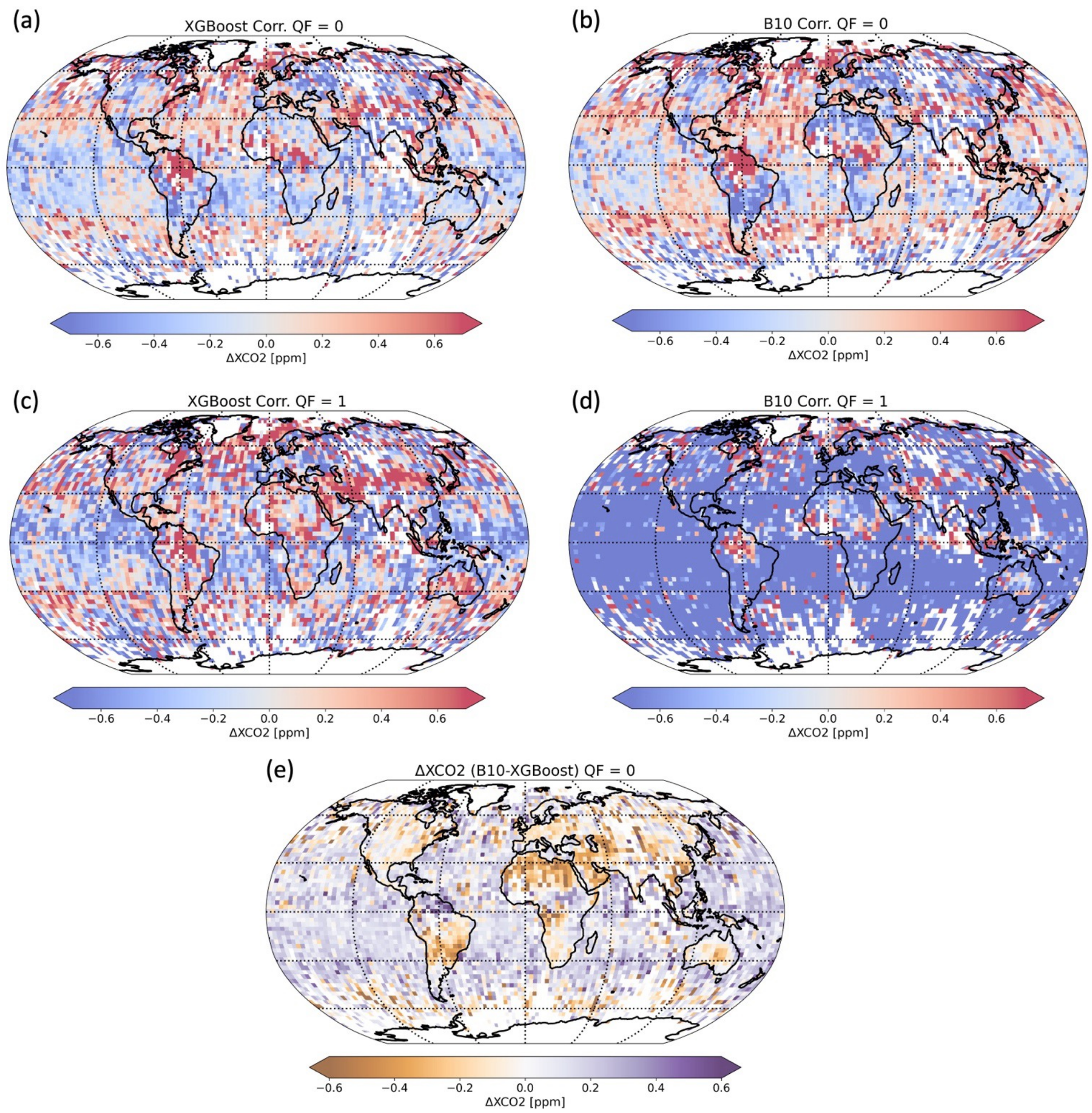

Figure 4.5: Residual ΔXCO_2 (model-mean proxy, 2018, 3°×3° bins): (a) XGBoost QF=0; (b) B10 QF=0; (c) XGBoost QF=1; (d) B10 QF=1; (e) difference (B10–XGB).

benefit to point analysis studies where data is limited by the operational QF, and potentially of benefit to flux models as well. To provide an empirical example of this, we create a modified version of the operational XCO2 quality flag utilizing our proposed ocean correction model and land correction model. We take a conservative approach where initial filter values are set equal to those of the operational quality filtering. Then, we select a few variables for which the filters are relaxed to increase sounding throughput while maintaining the RMSE of the combined operational correction and quality filter. With our new quality flag (QFNew), we are able to increase sounding throughput by approximately 14% over the B10 QF while matching the RMSE of the B10 correction as shown in Table 4.7.

For many features, the quality filters were not changed from the operational filters, as relaxing filters on variables that are already passing most of their conditional distributions would allow for only marginal improvements in throughput at the cost of large systematic errors. Therefore, we select only features for which large portions of the marginal distributions are removed by the operational flag and where the non-linear correction improves both mean and variance of Δ XCO2. The relaxed filters for these variables are shown in Figure B1 and Figure B2 by the vertical red dashed lines, and the range of data assigned QFNew = 0 shown in the red parentheses. The operational filter also minimizes the unit-less metric of the binned standard deviation of Δ XCO2 divided by the posterior XCO2 uncertainty below a value of 3 [106]. When tuning QFNew, we also aim to minimize this metric. Higher throughput of well-corrected data is observed in northern and central Africa, the Amazon basin, and in latitudes above 60^o north as seen in Figure 6. While selection of these variables and the relaxation of their filter values is subjective, this empirical result illustrates the benefit of a quality flag derived in conjunction with the non-linear bias correction. Future work will focus on the automation of defining the quality flag thresholds using a data driven approach.

Table 4.7: RMSE and data-throughput comparison for 2018.

Surface / Mode	XGBoost RMSE	B10 % pass	QF_{New} % pass
Land (NG)	1.07	59	69
Ocean (G)	0.72	60	74

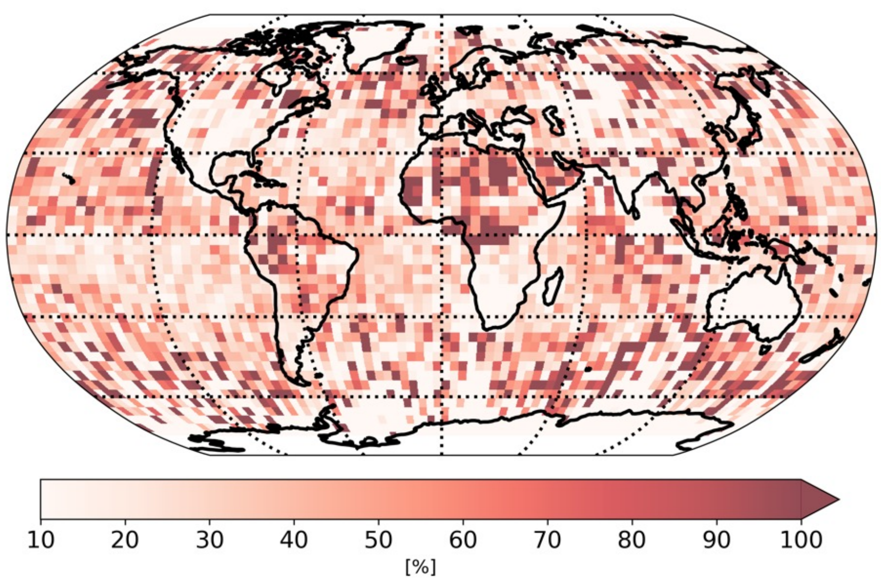

Figure 4.6: Relative increase (%) of QF_{New} over B10 QF (2018, 4°×4° bins).

4.2.4 Discussion and Future Work

4.2.4.1 Generalization Across Proxies

We acknowledge that even with a temporal training and testing split, there is still some circularity due to the lack of a truly independent truth proxy. This issue has been discussed at length for the operational bias correction in [127] and comparison and selection of independent validation data sets is still an open area of study. The risk of overfitting due to circularity becomes greater when fitting a more complex machine learning model. To evaluate generalizability to a fully independent validation proxy, we fit a set of XGBoost models on two truth proxies and evaluate on the third proxy which is held out during training. The same temporal split is used where 2018 data for the held-out proxy is used for evaluation. Results are shown in Figure 4.7, for land, and Figure 4.8, for ocean. Each column shows the residual fit for the hold out proxy, for QF = 0 (top row) and QF = 1 (bottom row). For QF = 0, increase in RMSE was minimal for both surface types and across proxies. There was some impact to performance on QF=1 data, when compared to training with all three proxies, particularly for TCCON with an increase in RMSE of 0.1 ppm for land and ocean data, indicating that the information contained in TCCON is not adequately represented by the model mean and small area approximation proxies which capture variability at larger scales. A potential approach to reducing circularity in the evaluation of the truth proxies would be to train the bias correction on TCCON and either the model mean or small area approximation, using the third proxy not chosen for validation.

4.2.4.2 Evaluating Feature Importance by Filter Regime

To understand the contribution of the features to correcting bias in QF=0 and QF=1 data, we compare the information gain between the two regimes. To perform the ablation study, we again employ the models trained on individual truth proxies and re-train and evaluate them on QF=0 and again for QF=1 data. Figure 4.9 shows the information gain for each filter regime for land and for ocean. For land, dpfrac and co2_grad_del are highly informative for

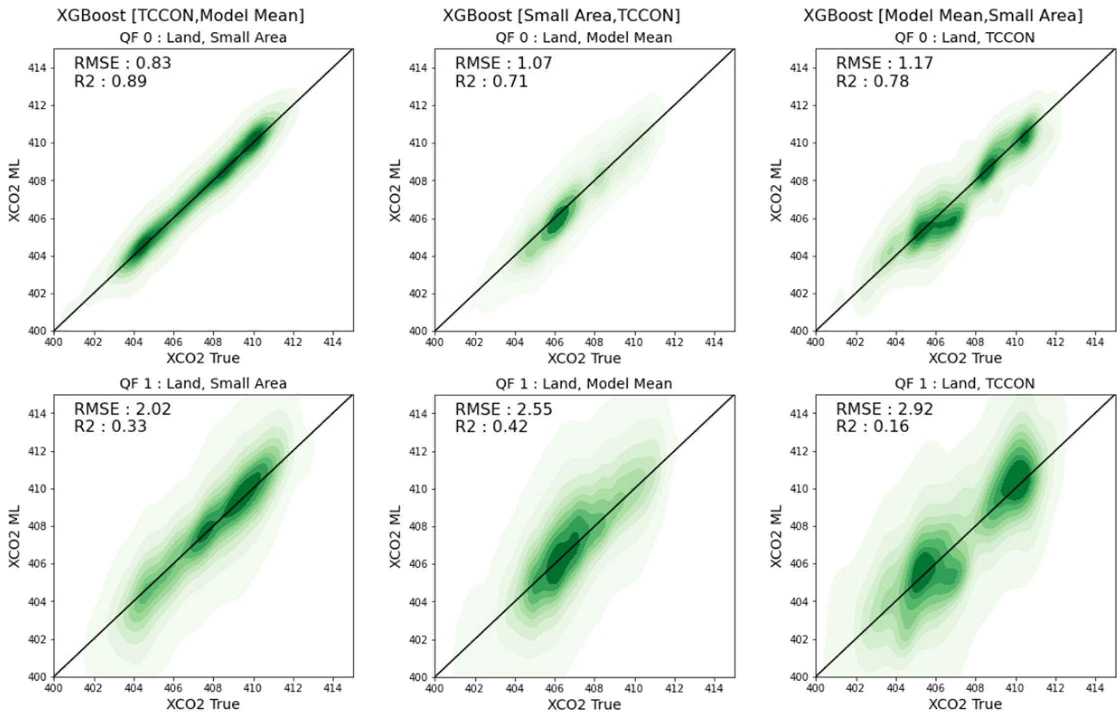

Figure 4.7: Comparison of XCO2 derived from XCO2 corrected by XGBoost (XCO2 ML) vs. truth proxy (XCO2 True) for land by hold-out proxy set and hold-out year (2018). Left-most column displays results of a XGBoost model trained on [TCCON, Model Mean] and evaluated on Small Area. Middle column displays results of a XGBoost model trained on [Small Area, TCCON] and evaluated on Model Mean. Right-most column displays results of a XGBoost model trained on [Model Mean, Small Area] and evaluated on TCCON. Generalization for the hold-out proxy and QF = 0 is shown in the top row and QF = 1 in the bottom.

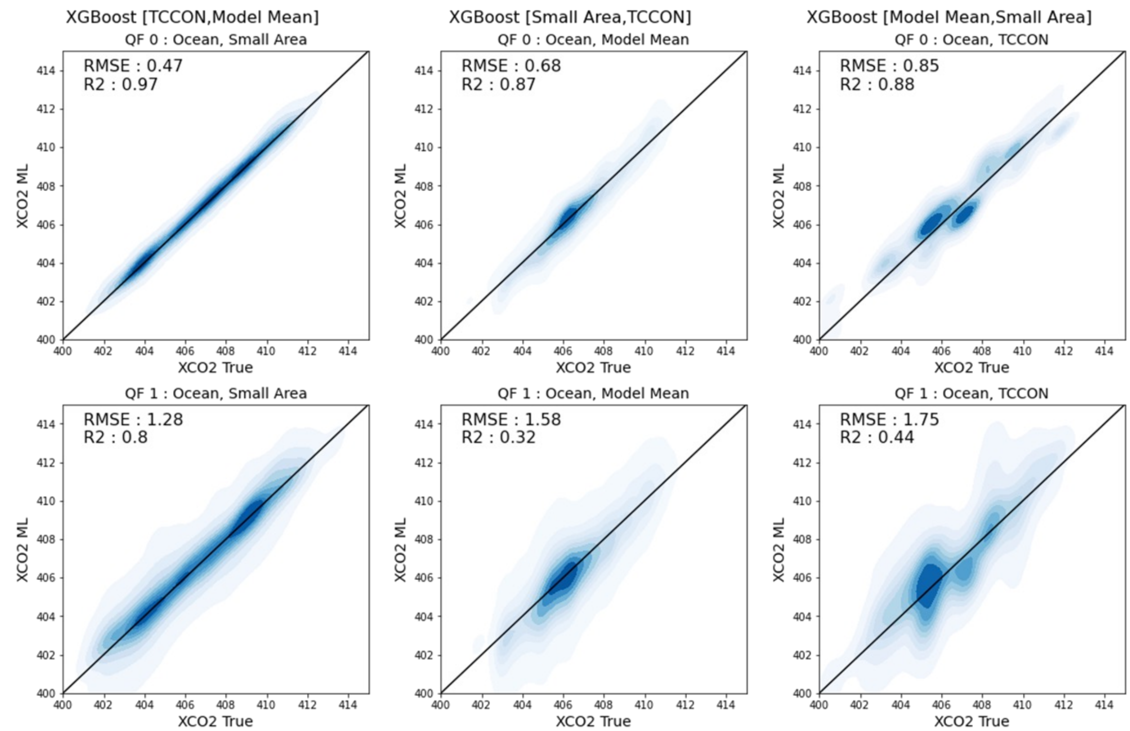

Figure 4.8: Comparison of XCO2 derived from XCO2 corrected by XGBoost (XCO2 ML) vs. truth proxy (XCO2 True) for ocean by hold-out proxy set and hold-out year (2018). Left-most column displays results of a XGBoost model trained on [TCCON, Model Mean] and evaluated on Small Area. Middle column displays results of a XGBoost model trained on [Small Area, TCCON] and evaluated on Model Mean. Right-most column displays results of a XGBoost model trained on [Model Mean, Small Area] and evaluated on TCCON. Generalization for the hold-out proxy and QF = 0 is shown in the top row and QF = 1 in the bottom.

correction of QF=0 data by the machine learning model. Similarly for ocean QF=0 data, the surface pressure delta term dp_sco2 and co2_grad_del are also highly informative. In operation, these terms are also used for bias correction in all ACOS versions (dpfrac replaced dP in B9) to date. These variables are responsible for the largest reduction in unexplained variance in the filtered regime ([109]; [106]; [102]).

For land QF=1 data, there is a drop in importance for co2_grad_del and dpfrac and large increase for h2o_ratio and relative increases for the albedo and aerosol terms. To explain the high importance for the h2o_ratio, we look to the non-linear interaction outside of the bound imposed by the operational filter which removes soundings with a h2o_ratio greater than 1.023, reducing the regime of interaction to one that is not highly correlated with ΔXCO2. In the QF=1 regime, h2o_ratio corresponds to a significant negative bias. Larger values of h2o_ratio are explained in [128], where it was shown that retrieved surface albedo from the strong CO2 band is generally lower than the weak CO2 band. In cases of larger aerosol presence, this sensitivity leads to weaking of the absorption features and a positive departure from unity. The additional albedo term for the strong CO2 band as well as the additional aerosol terms also increase in importance for QF=1.

For ocean QF=1 data, there is a significant change in information gain for several features. The surface pressure delta term dp_sco2 becomes significantly less informative for correcting QF=1 where negative values of dp_sco2 are relatively uncorrelated with ΔXCO2. Similarly to land, the albedo term for the strong CO2 band more informative for correcting outside the filtered regime along with the residual error between forward modeled radiance and measurements in the weak CO2 band.

4.2.4.3 Preservation of CO2 Enhancements

We assess the risk of the proposed bias correction to correct out and remove plume features in the data. Several features heavily utilized by the XGBoost models and in operational correction such as the CO2 gradient delta, and surface pressure terms (e.g., dpfrac, dp_o2a),

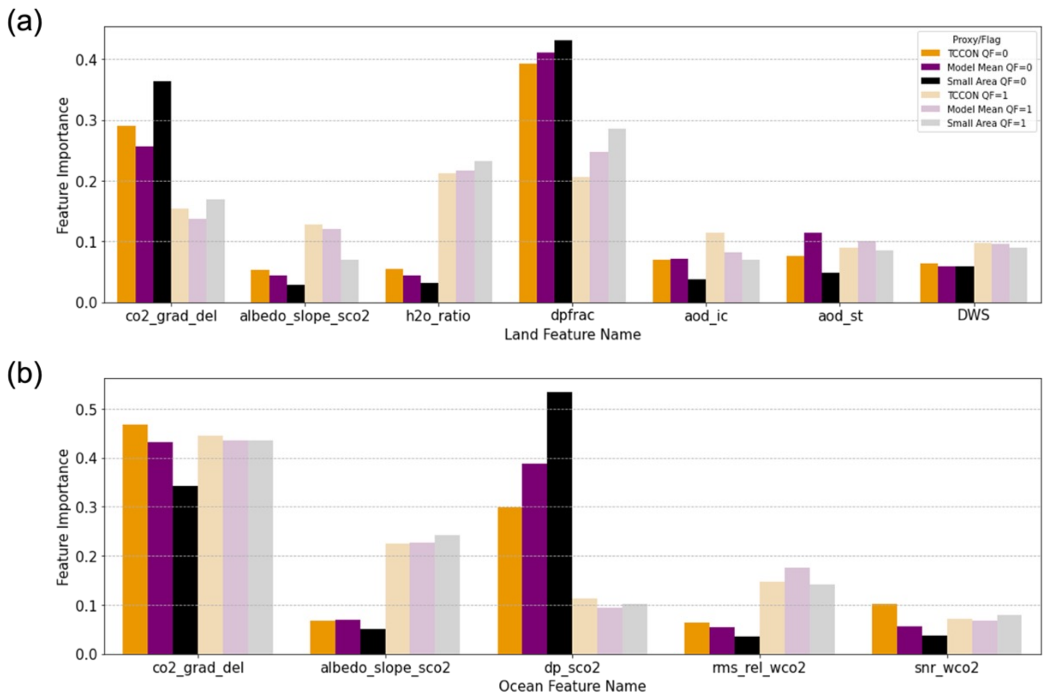

Figure 4.9: Feature importance for land is shown in (a); feature importance for ocean is shown in (b). The Y-axis displays the normalised information gain from XGBoost models with QF=0 shown in darker colours and QF=1 shown in lighter colours.

are differences between the ACOS retrieved state, and the prior. Therefore, there is potentially a risk for the bias correction to use the delta terms to over correct the retrieved XCO2 to the truth. We compare XGBoost corrected XCO2 for two known plumes first identified in [94]. The two example plumes are shown in Figure 4.10 (a) and (b): an ocean glint and land nadir plume in Taean, South Korea, and a land nadir plume observed over two co-located power plants in Ohio, US. We compare the uncorrected XCO2 retrieval (B10 Raw), the operationally corrected XCO2 (B10 Corrected) and the machine learning corrected XCO2 (XGBoost Corrected) and note that the machine learning corrected product captures enhancements not present in the training data. These results are also consistent with the findings in Mauceri et al. 2023 which include similar delta terms. This is further illustrated with the Taean plume which consists of 35% QF = 0 soundings and 65 QF = 1 soundings. QFNew = 0 improves the passing rate to 60% as shown in Figure 4.10 (c). The red stars show data that is passed by QF = 0 (and by construction QFNew = 0) and the blue stars

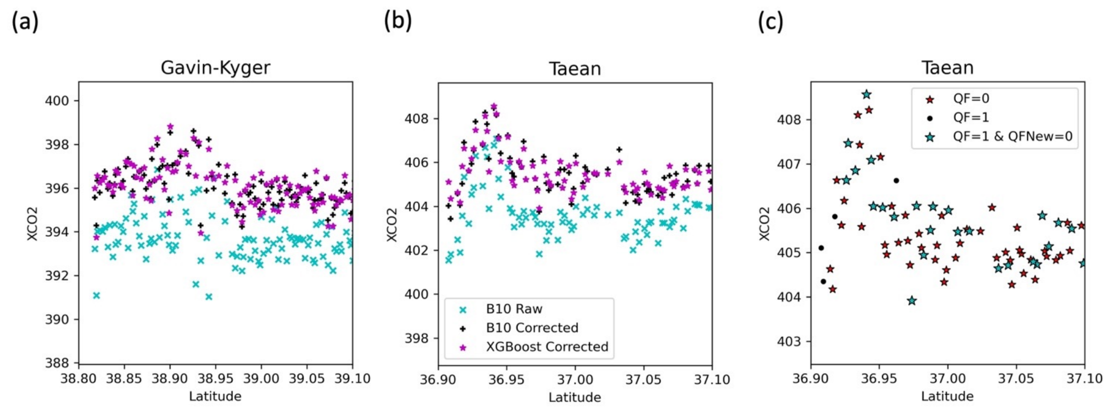

Figure 4.10: Two CO2 plumes captured downwind from power plants (Nassar et al. 2021). (a) Ocean glint and land nadir plume at Taean, South Korea (17 April 2015). (b) Land nadir plume near the J. M. Gavin and Kyger Creek power plants in Ohio, USA (30 July 2015). Plot (c) shows the increase in XGBoost-corrected data for QFNew = 0 that would be filtered by the B10 QF.

show data that would be removed by QF = 1 but is passed by QFNew = 0, indicating where the increase of available data for the plume feature. Of particular interest is the increase of data within the feature around 36.95° which includes the maximum observed enhancement value.

4.2.4.4 Potential for Further Improving Data Throughput

Figure 4.11 further illustrates how the shape of the filtering or decision surface can affect data throughput. Soundings are binned by two state vector features: h2o_ratio and dpfrac. Figure 4.11b, and Figure 4.11d show the improvement in reduction of mean ΔXCO_2 and in the error divided by the posterior uncertainty, from the non-linear correction. The QF filters for each feature are indicated by the black dashed lines and the interior of the intersection of these filters indicates the region of state space that is labeled as QF = 0 (Note: the additional filters of the QF further reduce the data that is passed in this region). Significant portions of the distribution, where the non-linear method can accurately correct, lay outside of this filtered region and are labeled QF = 1. A data driven filter can be constructed using similar interpretable machine learning techniques and produce a unified correction/filtering product.

Furthermore, moving away from the binary quality flag to a ternary ("very good", "good", "bad") will likely provide an improved data product for end users. Data driven methods for quality filtering have already proven to be useful in the northern high latitudes [87] and a genetic algorithm was previously used to derive the Warn Levels which complement the operational quality flag found in early OCO-2 data versions [75]. An important task for such future work will be to ensure that the machine learning method learns a physically consistent filter that can increase data throughput while still limiting variance of error and ΔXCO_2. We also acknowledge that while the Taean plume shown in Figure 4.24 illustrates an empirical example of the ability of a non-linear correction to improve throughput of good quality data, further evaluation of the intersection ($QF = 1$ & $QF_{New} = 0$) will be required before bringing such a method to operation.

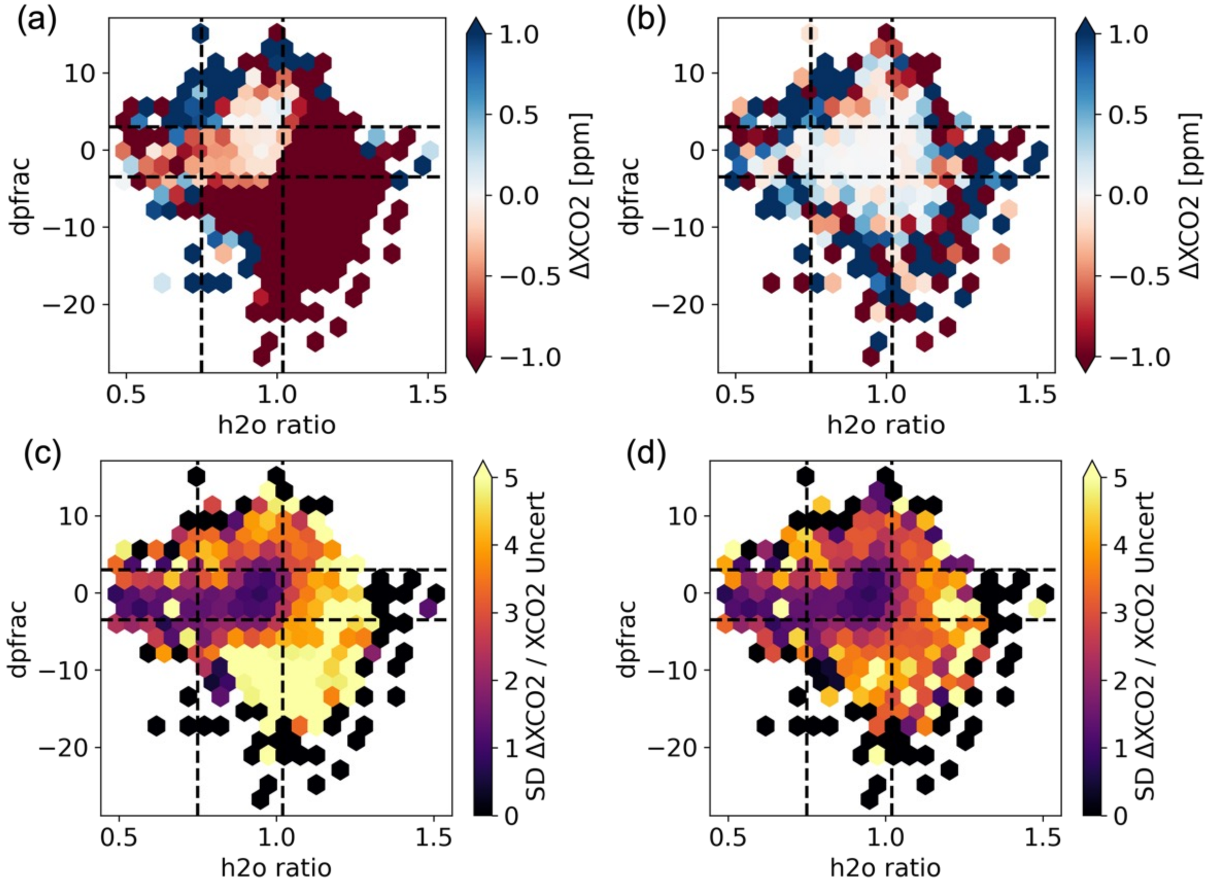

Figure 4.11: Figure 11. Hex bin plots show conditional distributions of 2018 ΔXCO_2 vs. dpfrac and h2o_ratio. Remaining ΔXCO_2 after the operational correction for B10 is shown in (a). Remaining ΔXCO_2 after the non-linear correction is shown in (b). Binned sttdev of ΔXCO_2 divided by the posteriori uncertainty from the retrieved XCO_2 is shown in (c) for the operational correction for B10 and (d) for the non-linear correction. B10 QF filter thresholds for both features are shown with black dashed lines for reference.

4.2.5 Conclusion

We demonstrate an approach for selecting co-retrieved state vector variables and other features to be used as input into a land model and an ocean model to correct biases in ACOS retrieved XCO2. The use of the non-linear method allows for decoupling of the dependent bias correction and filter used in operation, as the filter no longer needs to limit the correction function to a linear fit. By doing so, this method achieves a 59% and 67% improvement in reduction of the error variance over the operational correction on QF=1 data, for land and ocean respectively. To utilize this improvement in correction, we derive a new quality flag (QFNew), by relaxing select filter thresholds from the operational quality flag. Using the proposed QFNew flag, we increase data throughput by 14% while maintaining a comparable residual error to the operational B10 correction. The workflow outlined in this research is extendable for future ACOS algorithm updates, and for OCO-2's companion instrument, OCO-3, aboard the International Space Station.

4.2.6 Appendix

4.2.6.1 Feature Selection and Importance

To assess the robustness of our choice of features, we compare the ranking produced by the information gain feature importance generated by the gradient booster, with the ranking produced by a method called permutation feature importance (Fisher et al. 2018). Permutation feature importance captures the contribution to residual error when a feature has its values randomly shifted across observations. Permutation feature importance is a model agnostic post-hoc method that does not require the bias correction model to be retrained. In Figure A1 we compare the normalized rankings for the individual proxy/surface/mode models that were used to select variables for the final bias correction models trained on all truth proxies. Good agreement is observed in both the overall ranking and magnitude of normalized feature importance between both methods.

Figure 4.12 compares information-gain and permutation importance, while Fig. 4.13

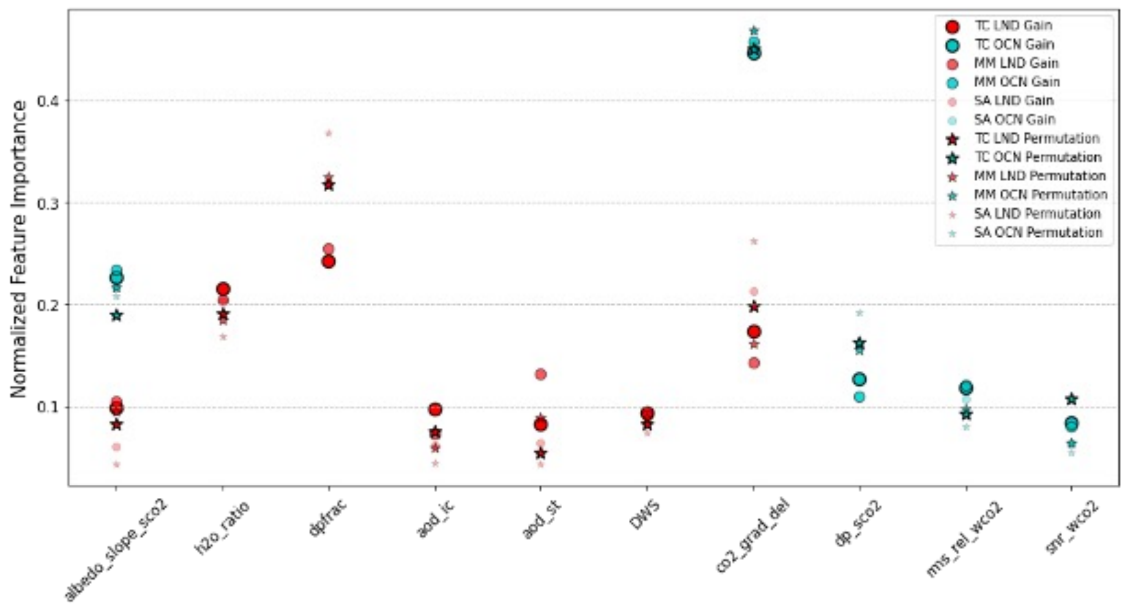

Figure 4.12: Comparison of permutation and information-gain feature importance.

shows per-proxy models.

Feature importance for models trained on individual proxies and QF = 0 + 1 data. These models were used to identify state variables to be used as input into the proposed bias correction models. While there is generally good agreement between the proxies the overall magnitude and ranking differs slightly as shown in Figure A2. For TCCON the aerosols and albedo terms contribute more to the correction while the same terms are less informative for the small area approximation, which is likely due to the small area proxy capturing biases that vary slowly over larger scales. For ocean, the `albedo_slope_sco2` is informative for the small area proxy, and all proxies exhibit better agreement in their feature importance.

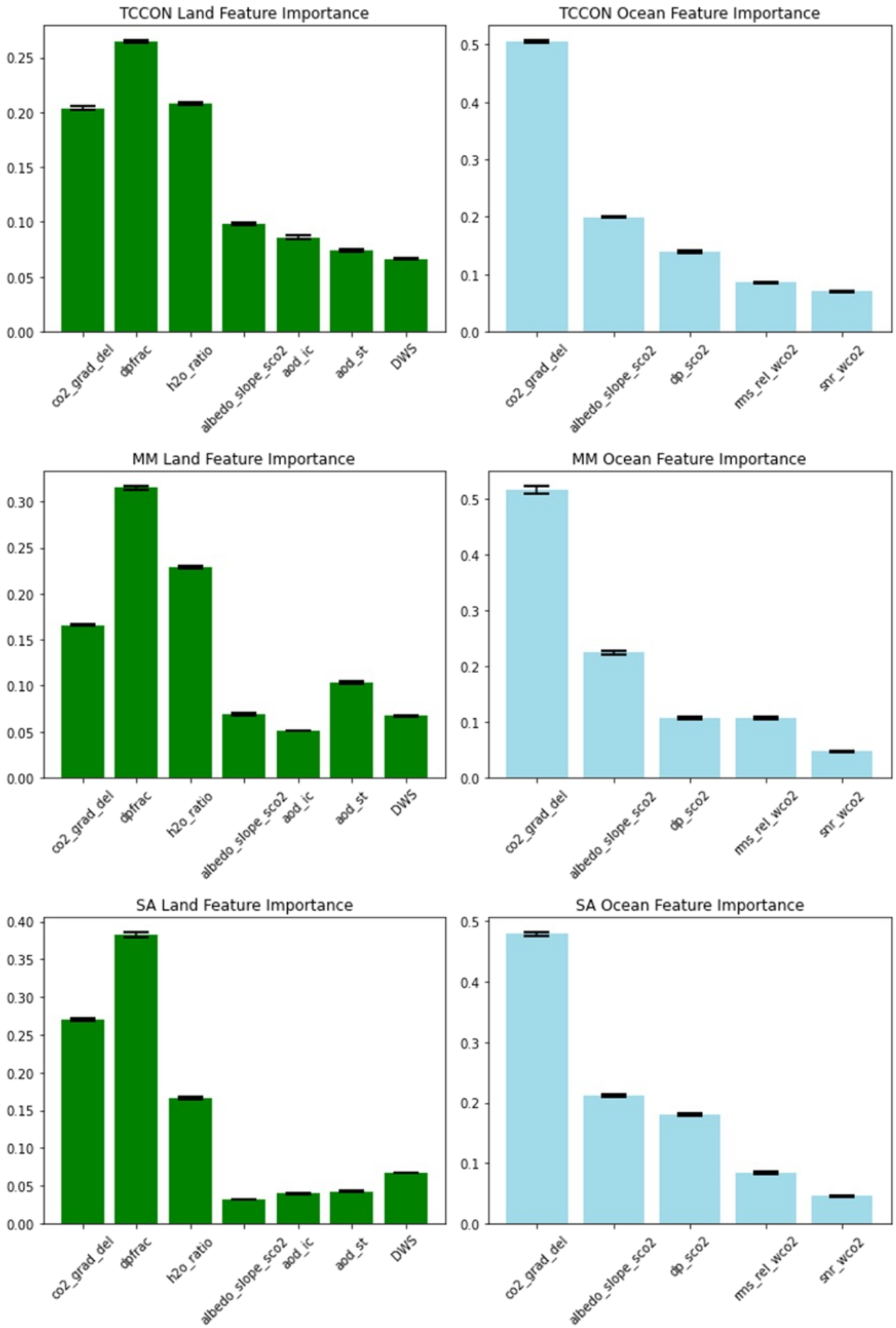

Figure 4.13: Feature importance for individual truth-proxy models. Error bars denote variance over 10 random seeds.

4.2.6.2 Threshold Values for $\mathbf{QF_{New}}$

New filter thresholds based after applying the non-linear correction models: land 4.14 and ocean 4.15.

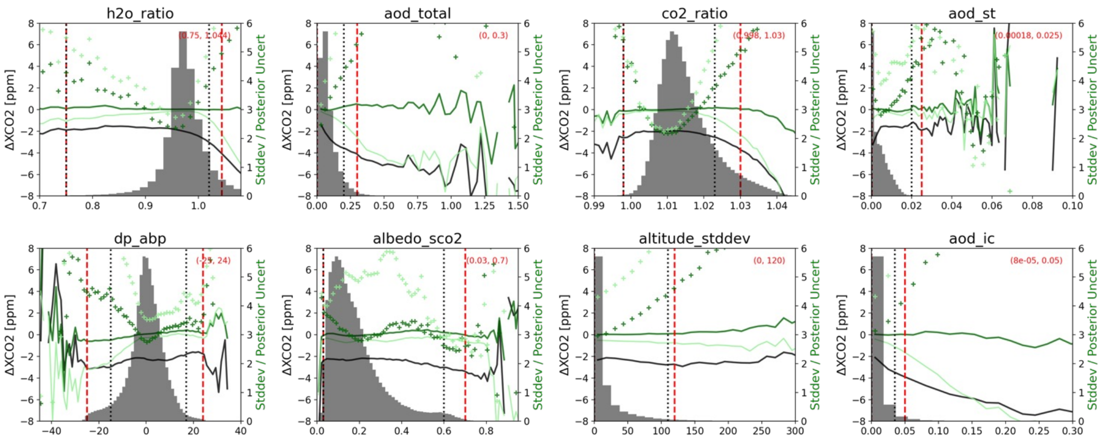

Figure 4.14: Variables selected for land QF_{New}. B10 filters (black), proposed filters (red).

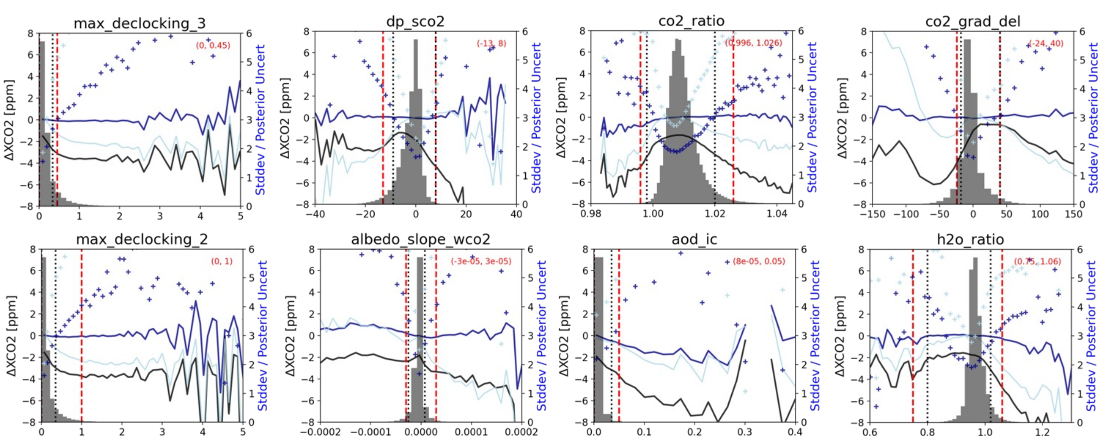

Figure 4.15: Variables selected for ocean QF_{New}. Same colour scheme as Fig. 4.14.

4.2.6.3 Lite-File Variables

All variables used or considered for correction or filtering in both the operational and proposed pipelines are shown in 4.8.

Table 4.8: Lite-file state variables used or considered for bias correction and quality filtering. Asterisks denote thresholds changed for QF_{New}.

State variable	**Description**	**Usage***
dpfrac	Surface-pressure difference ratio accounting for reduced dry-air columns at high elevation.	B10 BC; ML BC; QF; QF_{New}
h2o_ratio*	Ratio of retrieved XH_2O in weak/strong CO_2 bands (IMAP-DOAS).	ML BC; QF; QF_{New}
DWS	Sum of dust, water, and sea-salt AOD.	B10 BC; ML BC; QF; QF_{New}
aod_strataer	Stratospheric AOD at 0.755 µm.	ML BC; QF; QF_{New}
aod_ice*	Ice-cloud optical depth at 0.755 µm.	ML BC; QF; QF_{New}
co2_grad_del	CO_2 profile gradient (Eq. 4.5).	B10 BC; ML BC; QF; QF_{New}
dp_sco2*	Strong-band surface-pressure delta.	B10 BC; ML BC; QF; QF_{New}
snr_wco2	Weak-band SNR.	ML BC; QF; QF_{New}
co2_ratio*	Ratio of retrieved CO_2 columns (weak/strong bands).	QF; QF_{New}
altitude_stddev*	Std. dev. of surface elevation within FOV (m).	QF; QF_{New}
max_declocking_wco2*	Clocking error (weak band, %).	QF; QF_{New}
max_declocking_sco2	Clocking error (strong band, %).	QF; QF_{New}
dp_o2a	Surface-pressure delta to O_2A prior.	QF; QF_{New}
dp_abp*	Surface-pressure delta to fast-O_2A pre-processor.	QF; QF_{New}
albedo_slope_sco2	Strong-band reflectance slope (land) / Lambertian slope (ocean).	ML BC; QF; QF_{New}
albedo_slope_wco2*	Weak-band albedo slope.	QF; QF_{New}
albedo_sco2*	Strong-band surface reflectance.	QF; QF_{New}
albedo_quad_sco2	Quadratic term for strong-band albedo (land).	QF; QF_{New}
albedo_quad_wco2	Quadratic term for weak-band albedo (land).	QF; QF_{New}
aod_total*	Total aerosol optical depth.	QF; QF_{New}
rms_rel_sco2	RMSE of strong-band residuals / signal.	QF; QF_{New}
rms_rel_wco2	RMSE of weak-band residuals / signal.	ML BC; QF; QF_{New}
deltaT	Retrieved offset to prior temperature profile (K).	QF; QF_{New}
aod_sulfate	Sulfate AOD.	B10 BC; QF; QF_{New}
aod_oc	Organic-carbon AOD.	B10 BC; QF; QF_{New}
aod_water	Water-aerosol optical depth.	QF; QF_{New}
dust_height	Dust-layer pressure (relative).	QF; QF_{New}
aod_seasalt	Sea-salt AOD.	QF; QF_{New}
Fs_rel	Fluorescence relative to O_2A continuum.	QF; QF_{New}
chi2_wco2	Reduced χ^2 of weak-band fit.	QF; QF_{New}
windspeed	Surface-wind speed (water surfaces).	QF; QF_{New}
water_height	Cloud-water layer pressure (relative).	QF; QF_{New}

4.3 Uncertainty-Aware Machine Learning Bias Correction and Filtering for OCO-2: Part 1

4.3.1 Paper Introduction

Accurate measurement of atmospheric carbon dioxide (CO2) concentrations is critical for understanding the global carbon cycle and predicting changes in Earth's climate. CO2 is a greenhouse gas that absorbs and radiates energy, contributing to the warming of the planet [81]. The Orbiting Carbon Observatory-2 (OCO-2) [22]; [32] started operations in 2014 and makes high resolution space-based radiance measurements of reflected sunlight in the near-infrared using a three-channel grating spectrometer. Using the physics-based retrieval algorithm Atmospheric CO2 Observations from Space (ACOS) [23]; [104], based on Optimal Estimation (OE) [117], these radiance measurements, in the Oxygen A-band (0.76µm) and weak and strong CO2 bands (1.6µm and 2.05µm), are inverted to estimate column-averaged atmospheric CO2 dry-air mole fractions (XCO2). Like the results of many retrieval algorithms, the retrieved XCO2 from OCO-2 contains biases [145]. These biases stem from a variety of sources [18];[43]; [65], such as differences between the forward model used in the OE and nature (e.g., the treatment of aerosols, uncertainties in gaseous absorption coefficients) and calibration uncertainties in the spectrometer. Additionally, the iterative nature of the OE retrieval process can lead to sensitivities to initial conditions or algorithm constraints, potentially resulting in suboptimal convergence or biases [18].

In the latest version of the Level 2 Full Physics ACOS retrievals (see section 2.1.for more detail on data versions), the errors in retrieved, but uncorrected and unfiltered, XCO2 are on the order of 2 ppm and not randomly distributed. Since flux inversion models relate spatial and temporal gradients in atmospheric concentrations of CO2 to sources and sinks of CO2, systematic errors in retrieved gases can induce incorrect inference on sources and sinks. For example, multiple studies have shown that errors of a few tenths to 0.4 ppm can result in continental scale flux errors of 0.5-1 PgC/yr [16]; [88]; [105]. Since understanding and monitoring these fluxes is critical to inform decision makers and policy, both filtering and bias correction of these observations are necessary and a key processing step to provide actionable carbon cycle information to society.

The current operational bias correction exploits the correlation of errors in retrieved XCO2 with various elements of the measured radiances and the OE state vector derived from the ACOS retrieval with carefully fitted multiple linear regression models informed by expert judgment [60]. However, this linear approach may not fully capture the complex, potentially non-linear relationships between state vector variables and retrieval errors, possibly leaving systematic biases uncorrected. To guide these adjustments, multiple success metrics are evaluated. The three most important metrics are: difference between the OCO-2 retrievals and collocated Total Carbon Column Observing Network (TCCON) XCO2 [143], a reduction in variability of temporally and spatially near collocated OCO-2 retrievals from the same orbit (called Small Area Analysis), and differences between OCO-2 and an ensemble of simulations from a suite of carbon flux inversion models [103]. This last metric is particularly problematic when the models are themselves used to infer fluxes from OCO-2 data, since the assimilated data and inversion framework are no longer independent. Additionally, these models are known to have systematic artifacts from atmospheric transport [122] that could alias into the bias correction. Other metrics used are differences at land-ocean crossings, and more recently, dependencies of XCO2 to nearest cloud distance, stemming from 3D cloud effects [79]; [78]; [80]. Whereas the operational approach lacks a clearly defined independent dataset for validation, these metrics provide some guidance, but validating the bias-corrected XCO2 remains challenging due to some amount of circularity, i.e. the data that would be used for validation are employed in some manner in the bias correction.

In light of these limitations, advanced computational methods, such as machine learning (ML), have the potential to enhance the bias correction process. ML has recently gained significant traction as a transformative tool in various scientific domains, including Earth sciences. The surge in popularity has led to a plethora of literature proposing ML solutions for Earth science challenges. However, despite this proliferation of ML-based studies, there remains a notable hesitation within the Earth science community to apply these methods in an operational context. A primary concern is the lack of trust in potentially black-box methods that lack interpretability and can be overly complex, thus overfitting or producing erroneous results. An additional problem is the secular increase of CO2 in the atmosphere of roughly 2.2 ppm/year, potentially rendering ML methods inaccurate for future time periods if trained on existing observations [114]).

Building on the operational bias correction, we propose a series of improvements to the operational approach: 1) make the bias correction more traceable, transparent, and reproducible by relying on computational optimization with clearly stated objectives over manual tuning of parameters; 2) increase flexibility when modeling the retrieval errors by allowing for non-linear mapping between state vector variables and XCO2 errors. This is motivated by the recognition that the relationships between retrieval errors and state vector variables are complex and may not be adequately captured by linear models; 3) exclude the output of flux inversion models when developing the bias correction to keep the OCO-2 measurements independent and instead allow inversion models to use the measurements for flux inference; 4) quantified uncertainties of the applied correction itself at the per retrieval level; and 5) validate the bias-corrected XCO2 using k-fold cross validation for a rigorous estimate of how accurately we can measure CO2 from space. Throughout this paper we refer to this approach as the ML bias correction. We hypothesize that a more traceable bias correction is important to solidify trust in the carbon flux products they inform, that ultimately influence policy and decision making. By addressing the concerns associated with ML methods, we propose a simple, transparent, and interpretable bias correction model with quantified uncertainties that is poised for acceptance within the scientific community. Furthermore, we hypothesize that allowing for non-linear relationships between state vector variables and XCO2 errors will enable us to more accurately capture the complex dependencies inherent in the atmospheric retrieval processes. This increased flexibility can reduce residual errors, particularly in data currently flagged as unusable, thereby improving data availability in problematic regions like cloudy areas in the Amazon. Our approach builds upon recent work by [58], which introduced a machine learning-based bias correction and data quality filtering approach using XGBoost. Key differences include a stronger focus on interpretability and uncertainty quantification through Quantile Regression Forests, complete independence from flux inversion models, and extensive engagement with the science team to build trust in the methodology. Additionally, our approach incorporates rigorous uncertainty quantification at the per-retrieval level and adopts a k-fold cross-validation strategy, to better assess remaining errors. Finally, we introduce a balanced sampling strategy to enhance geographic diversity, particularly in challenging regions. The result is a more operationally-focused method than previous work.

4.3.2 Data

4.3.2.1 OCO-2

This study uses the v11.1 'Lite Files' from OCO-2 (OCO-2/OCO-3 Science Team, last access: 01/2025). This dataset covers September 2014 to April 2024 and contains approximately 55 million soundings per year. In April 2024, the OCO-2 project transitioned to a new input data stream for meteorological fields, resulting in a switch to v11.2. Further details of the impact of this switch can be found in [27] and in the OCO-2 Level 2 Data User Guide [109]. The approach presented in this study can also be applied to v11.2, but at the time of the work, v11.1 was the latest version available. For our study we consider soundings for the full years from 2015 to 2022. Besides the retrieved XCO2, these files contain numerous elements of the co-retrieved atmospheric state vector (e.g. total column water vapor, aerosols, surface pressure), as well as other proxies used for the operational bias correction (land / water mask, surface albedo, difference of retrieved and prior surface pressure) and data quality filtering (e.g. snow flag). Recently, an additional set of calculated quantities related to subscene variability have been added that correlate well with biases in retrieved XCO2 caused by 3D cloud effects, as discussed in [83]. In addition to the retrieved XCO2, the Lite files contain operationally bias corrected XCO2 and a quality flag (QF) for each retrieval. A QF of 0 is used to signify ‘good’ data and a QF of 1 is used for ‘bad’ data. In the operational product, retrievals are assigned QF=1 if they exhibit characteristics indicative of significantly biased XCO2 estimates, such as high aerosol loading, challenging surface properties, or evidence of cloud contamination. These retrievals are typically excluded from scientific analyses. However, this study includes both QF=0 and QF=1 retrievals, as we hypothesize that our non-linear ML bias correction approach can successfully correct many of the biases in QF=1 data. This is particularly important for regions like the Amazon and Southeast Asia where a large fraction of soundings are currently flagged as QF=1, creating significant data gaps. Globally, about two thirds of the data is labeled with QF=0 and one third with QF=1. The effectiveness of including QF=1 data is evaluated throughout this paper, while Part 2 explores an optimized filtering approach for the ML bias-corrected retrievals. More details of how the operational QF is determined are given in the OCO-2 data user guide [109] and Part 2.

OCO-2 has three observing modes: nadir, glint, and target. Observations in nadir mode are only processed over land because there is not enough reflected light over the dark ocean. During glint mode OCO-2 is pointed off-nadir close the surface sun glint spot, which is used over water to increase the signal, and also provides data over land. Finally, target mode points OCO-2 at a specified target, e.g. a TCCON station, and dithers across it several times over the span of about 4.5 minutes. For the purposes of the ML bias correction discussed in this paper we do not differentiate between observations by those modes but rather their underlying surfaces (land, water) as outlined below. This same assumption is used in the operational bias correction.

4.3.2.2 TCCON

TCCON is a global network of ground-based stations that measure atmospheric CO2 concentrations using Fourier transform spectroscopy [146]. TCCON stations are located at strategic sites around the world and measure XCO2 with high accuracy (< 0.5 ppm, [66]. These measurements are routinely used to provide an independent reference for space-based measurements of XCO2 from instruments including OCO-2 and Greenhouse Gases Observing Satellite (GOSAT) [69]. We make use of the latest available processing version (GGG2020) from January 2015 to December 2022 available at (Total Carbon Column Observing Network (TCCON) Team, last access: 01/2025).

4.3.2.3 Flux Models

Flux inversion models assimilate a variety of CO2 measurements, for example, from in-situ observations. Previous research has shown significant differences between CO2 concentrations derived from such flux models and the OCO-2 data record [111]; [151]. To evaluate how the ML bias correction impacts this disagreement, we compare the bias corrected XCO2 from OCO-2 to four flux inversion models: NOAA's CarbonTracker CT-NRT.v2022-1 [46], MACC v21r1 [17];[17], LoFI m2ccv1bsim [137], and the University of Edinburgh's UnivEd v5.2 ([33]. All of these models assimilate in situ CO2 observations but do not use space-based remote sensing products.

4.3.2.4 MODIS Cloud Distance

Recent work has uncovered uncorrected biases in the OCO-2 data record that stem from horizontal photon transport due to clouds, also referred to as 3D cloud effects. To investigate if the ML bias correction (and data quality filtering in Part 2) can mitigate those biases, we make use of a Moderate Resolution Imaging Spectroradiometer (MODIS) derived product that provides the distance to the nearest cloud [79]. The data is available from September 2014 to July 2019 at ([77], last accessed: 01/2025)

4.3.3 Methods

4.3.3.1 ML-based Bias Correction Overview

In this work, the OCO-2 data is corrected in two steps. First, we fit, or train, a model that aims to minimize differences between OCO-2 and TCCON retrieved XCO2, after a correction has been applied to account for the OCO-2 averaging kernel [97]; [146]. The machine learning model, trained on the difference in retrieved XCO2 from coincident observations from OCO-2 and TCCON, is then applied to the entire OCO-2 data record. This step accounts for large-scale and small-scale biases in OCO-2 as well as any systematic offsets between the two data records. In a second step, we take the TCCON corrected XCO2 data and fit a second model to reduce variability within small areas. These small areas are specially chosen collections of soundings in close proximity over which CO2 can be approximated as constant, further explained in the next section. This second step is necessary to account for small-scale biases that may exist where TCCON observations are not available, for example, most of the open ocean and tropical land. The workflow is summarized in Figure 4.16. Like the operational bias correction, we fit individual models for OCO-2 observations over land and water, respectively. We don't differentiate between the three OCO-2 observing modes: nadir, glint, and target. In contrast to the operational bias correction, we make no use of flux inversion models to develop or validate the ML bias correction. This is motivated by the desire to keep the measurements independent of flux inversion models and any biases they might have. As the final step we bring retrievals over land and ocean into agreement by adding a single offset to all ocean retrievals that removes the mean difference between land and ocean retrievals within small areas that contain both. More details are provided in Appendix A. The effect of this offset is small (reducing ocean XCO2 by 0.16 ppm) compared to the other applied corrections. The ML bias correction attempts to correct biases in all soundings independently of any filtering - a key methodological difference from the operational bias correction which develops its parameters iteratively with the filtering

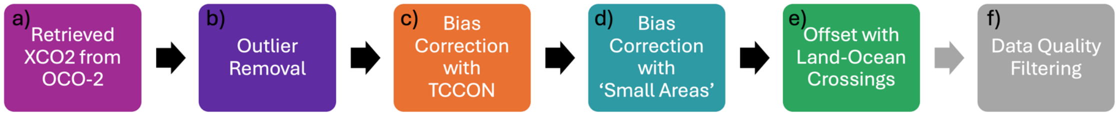

Figure 4.16: Flow chart of OCO-2 ML bias correction. a) OE retrieved XCO2 from OCO-2 observations that contain biases, b) Outlier removal, c) correction based on OCO-2–TCCON differences using a Random Forest, d) correction based on variability within small areas using a Random Forest, e) bringing land and ocean retrievals into agreement, f) data quality filtering to remove irreducible errors in OCO-2 (discussed in second paper (Part 2)).

process.

4.3.3.2 Small Areas Generation

Like the operational bias correction, the ML bias correction is in part developed on the assumption that surface fluxes impacting boundary-layer CO2 concentration can only impact column-integrated CO2 relatively slowly and therefore XCO2 should not vary by more than 0.1 ppm over spatial scales of less than 100 km [88]; [142]. Of course, there are exceptions to this assumption in the presence of strong CO2 emitters such as megacities or large power plants. Given that urban areas cover only 3% of global land surfaces, we ignore these cases and assume they will have a negligible impact on our training. To exploit this property of CO2, we group OCO-2 soundings from the same orbit that fall within 100 km of each other, the same distance used by the operational bias correction. For the remainder of the manuscript, we call these clusters small areas. After the clustering step we calculate the median of each small area and treat any differences between individual soundings and this median as biases that we wish to correct for. The median was chosen as it provides the most robust estimate of the true CO2 in a given small area.

4.3.3.3 Outlier Removal

To remove outliers from the OCO-2 data record before model training, we apply thresholds on various co-retrieved state variables. This step removes retrievals that, for example, have very high aerosol loadings (with optical thickness of greater than one) or have a surface that contains snow (early experiments showed that XCO2 biases over snow are very large compared to other surfaces and cannot be corrected as well as over other surface types). This outlier removal step removes about 5% of the data and is not to be confused with the much more thorough and optimized data quality filtering step described in the second paper (Part 2). The main motivation for this processing step was to address the desire by part of the OCO-2 science team to have a formal guarantee that certain retrievals would not be used for model training or could be classified as 'good' data. From a pure data science perspective this processing step is not required since we chose a machine learning algorithm (random forest) that is robust to outliers. We confirmed that the outlier removal step had no significant impact on the ML bias correction model (< 0.1 ppm RMSE difference).

4.3.3.4 Co-Locating TCCON and OCO-2 Estimates

To consider TCCON and OCO-2 observations as collocated we require them to be within 1 hour of each other and within a specified maximum distance. The default for the maximum distance is set to 100 km. However, some TCCON stations only provide valid reference XCO2 over a small geographic area, for example, if they are in the proximity of strong CO2 emitters such as a megacity [147]. To determine the most appropriate maximum matching distance for each TCCON station, we evaluate how the variability (root mean square error, RMSE) of OCO-2 retrieved XCO changes as the match-up distance increases. Using the RMSE for a maximum match-up distance of 20 km as the baseline for each TCCON station, we set the cutoff at the point where the RMSE increase exceeds 25%. More details about this approach and the derived maximum distance for each TCCON station are given in Appendix C. The latest TCCON GGG2020 version provides XCO2 on the X2007 and X2019 WMO scale [39]. We are using the X2019 scale so that the data can be directly assimilated together with in-situ data. Finally, following [145] we correct TCCON observations for the differences in the averaging kernels between OCO-2 and TCCON so that TCCON XCO2 mirror what OCO-2 would sense.

4.3.3.5 Subsampling and Balancing

We subsample the OCO-2 data record, selecting a more manageable subset of 5 million soundings per year from 2015 to 2022, resulting in a total of 40 million soundings. The subsampling is not performed in random fashion, rather we try to balance the sampled soundings spatially and keep all soundings that have a TCCON match up. Thus, for a given year we first sample OCO-2 soundings that have a TCCON match up, which is typically on the order of half a million soundings or 10% of the subsampled data set. Next, we sample small areas iteratively until the total number of soundings reaches 5 million. To make sure the sampled small areas, and the soundings they contain, are representative of all geographic regions and their potential sources of biases, the sampling is inversely weighted by the number of small areas contained in each 2° by 2° latitude-longitude grid cell. As a result, this sampling strategy increases data in regions such as the tropics where available observations are limited due to cloud coverage [128].

4.3.3.6 Train Test Split

To estimate the accuracy of the ML bias correction on new observations we divide the data into a training and testing set. Evaluating the accuracy on a separate testing set ensures that the model is not overfitting to the training data and can generalize to new OCO-2 observations. Since the data contain spatially and temporally correlated measurements and biases, we split the training and test sets by different years and locations, so that no test observations are spatiotemporally adjacent to training observations. Our training set contains 30 million OCO-2 soundings from 2015 to 2021. These soundings are used to determine the model architecture, feature selection, and to train the final model. The testing set comprises 5 million soundings from 2022. These soundings are used to evaluate the model performance on new observations and are representative of future OCO-2 observations. However, a pure train-test split in the temporal dimension might underestimate errors for certain geographic regions where sources of biases (e.g., unique aerosol properties) might be different than those from observations close to any of the TCCON stations. Therefore, in addition to the temporal train-test split, we perform k-fold cross validation over the TCCON stations. Holding out one station at a time for testing, we use coincident OCO-2 measurements with the other TCCON stations to train a model. Then, using the held-out station, we assess remaining errors. This process is repeated until each TCCON station was held out once. The method allows us to quantify the average expected error for geographic regions where no coincident TCCON observations are available.

4.3.3.7 Feature Selection

To develop a machine learning-based bias correction approach, a set of model inputs, or features, is required. Like the operational bias correction, we use elements of the retrieved state vector, solar and viewing geometry, instrument characteristics, and environmental ancillary data as features. The feature selection was an iterative process combining a data-driven algorithm with the input from OCO-2 scientists. The hybrid approach ensured that there exists a physical basis for the selected features, important for the robustness of the ML bias correction and to build trust with the science community. However, we recognize that incorporating human feedback in this approach may introduce some challenges in maintaining full traceability of the feature selection process. The data-driven algorithm operates as follows. First it trains a simple and efficient machine leaning model, which in our case is a random forest. Next, the values in each feature are randomly shuffled individually and the reduction in model performance is evaluated. The feature that, if shuffled, has the least impact on the model performance is removed. The algorithm is run iteratively until no feature is left, resulting in an ordering of feature importance with the most important feature being removed last. Finally, we pick the last set of features to be used in the ML bias correction model. The size of the set is chosen as a tradeoff between model performance (typically favoring more features but with diminishing return), versus the desire to reduce complexity, increase interpretability, and robustness to new data (favoring fewer features). The resulting features

Table 4.9: Summary of OCO-2 parameters used for the ML bias correction.

Variables	Description	ML BC Land	ML BC Ocean
`albedo_o2a`	Retrieved surface albedo in the O_2 A-band		×
`albedo_sco2`	Retrieved surface albedo in the strong CO_2-band	×	
`albedo_slope_sco2`	Slope of albedo in the strong CO_2 band with respect to wavenumber		×
`aod_strataer`	Aerosol optical depth of stratospheric aerosol particles	×	
`aod_ice`	Aerosol optical depth of ice particles (e.g. cirrus cloud particles)	×	
`aod_sulfate`	Aerosol optical depth of sulfate particles	×	
`co2_grad_del`	Difference in retrieved CO_2 between surface and $0.7\times$ surface-pressure level, minus same quantity for the prior	×	×
`color_slice_noise_ratio_o2a`	Ratio of continuum-radiance spatial std. dev. to noise level in the O_2A-band (Massie et al., 2021)		×
`dp`	Retrieved surface pressure (full-physics) minus prior surface pressure		×
`dpfrac`	Retrieved surface pressure (A-band pre-processor) minus forecast value adjusted for altitude effects (Kiel et al., 2019)	×	
`dust_height`	Retrieved pressure of dust aerosol layer relative to surface pressure	×	
`dws`	Combined AOD of dust, water cloud, and sea-salt particles	×	
Footprint number	OCO-2 frame footprint number (1–8)	×	×
`h2o_ratio_bc`	Ratio of total-column H_2O from WCO_2 to SCO_2 band (SCO_2 is more sensitive near surface)		×
`h_continuum_o2a`	Differences in continuum radiances to adjacent observations in the O_2A-band (Massie et al., 2021)		×
`ice_height`	Retrieved pressure of ice-cloud layer relative to surface pressure	×	×
sensor zenith angle	Line-of-sight angle relative to local vertical; solar azimuth at surface		×
`water_height`	Retrieved pressure of water-cloud layer relative to surface pressure	×	×

were then interrogated for physical consistency using various ablation studies to demonstrate the impact of removing, adding, or replacing certain features. As discussed earlier, there are separate bias correction models for land and for ocean. For both surface types the model consists of a two-step process, with the second model working on mitigating the residuals of the first model. Thus, the described feature selection algorithm was run four times in total, twice for land and twice for ocean observations. A short description of each feature selected by the algorithm is given in Table 4.9.

4.3.3.8 Model Architecture and Baseline

Different model architectures may be suitable for different applications and have unique trade-offs in terms of complexity, interpretability, accuracy, speed, how well they generalize to new data, and whether they can provide robust uncertainty estimates.

The final product of our study will be a dataset of bias-corrected OCO-2 XCO2 that will be used in the OCO Model Intercomparison Project [151] and other flux inversion modeling efforts. These modeling efforts systematically evaluate how different inverse modeling systems estimate carbon fluxes when ingesting the same XCO2 measurements and influence decisions and policies regarding greenhouse gases. Like other science applications of machine learning, this requires a reliable and interpretable approach to build trust with the user community. This is challenged by the often "black box" nature of machine learning models, which do not provide insight into how their predictions are made. To address those shortcomings, we evaluate two models that are interpretable, low complexity, and allow us to quantify uncertainties. One of the models is linear and the other one non-linear. The linear model is similar to the operational bias correction applied to OCO-2. It serves as a baseline to compare the non-linear model to and thus allows for a fair comparison when assessing how much of a possible model improvement can be attributed to its non-linearity. For the linear model we evaluate Bayesian Ridge Regression based on an algorithm proposed by [133]. Bayesian Ridge Regression is a variant of linear regression but uses Bayesian inference to estimate the regression parameters and an uncertainty of its point estimate. It is highly interpretable but not able to describe non-linear relationships between features and biases in XCO2. For the non-linear model we utilize a Quantile Regression Forest (QRF) described in [86]. The QRF combines the flexibility of random forests with the ability to estimate uncertainties via quantile regression. By quantifying uncertainty, we can provide a statistical basis of confidence in the bias corrected XCO2. Moreover, uncertainty quantification can highlight regions of heterogeneity in the training data, or where the model is forced to extrapolate, signaling where additional data collection or refinement of the model might be necessary.

This not only builds trust but also guides future model development, ensuring more reliable and transparent adoption into the operational science data processing pipeline. Similar to uncertainty quantification, explainability methods allow us to validate, at least in part, that the model does not get the right answer for the wrong reason. Post-hoc model agnostic techniques like SHapley Additive exPlanations (SHAP) [72] allow for interpretability of the QRF models' predictions. In addition to model validation, interpretability methods provide insight into the physical drivers of biases, and their correction, in the data record. This information can then be used to improve the physics-based retrieval. For example, biases that are driven by changes in sensor zenith angle might point to errors in atmospheric absorption calculations, biases driven by changes in aerosol loadings indicate that the retrieval aerosol parameterization might be flawed. Previous work has shown that model architectures like QRF (and Extreme Gradient Boosting) are well suited to perform bias correction and are competitive with more complex approaches while maintaining interpretability ([58]; [83]). For both algorithms we rely on their scikit-learn implementations [110].

4.3.3.9 ML Bias Correction Model Tuning and Training

The final model consists of four individual QRFs that differ in their hyperparameters (e.g. maximum tree depth, features). Each QRF minimizes the mean squared error (MSE) between predicted and actual XCO2 biases. The hyperparameters for each model were determined using data from 2015-2020 for training and 2021 for validation. This ensures that the model development is independent of the 2022 test data. Through multiple experiments, we evaluated the sensitivity of model performance to various hyperparameters. The final hyperparameter selection balanced minimizing the RMSE on the validation set while avoiding overfitting on the training set, as indicated by the gap between training and validation RMSE. The final set of model parameters is given in Appendix D. Finally, we had to consider the unequal number of matched soundings available for each TCCON station. The number of matched observations by TCCON stations vary by multiple orders of magnitude which would heavily and undesirably weight the correction model towards stations (and their associated atmospheric states) with more soundings. That might negatively impact the ability of our model to correct biases in the high northern latitudes, tropical forests, or large swaths of the ocean where TCCON observations are limited. Thus, we weight each collocation during the model training so that each TCCON station contributes equally to the ML bias correction by setting sample weights inversely proportional to the number of observations at each station. This contrasts with the operational bias correction that considers each OCO-2 – TCCON matchup equally.

4.3.3.10 Science Collaboration and Trust Building During Model Development

To foster acceptance of the ML-based bias correction (and data quality filtering), we engaged in a partnership and ongoing dialogue with the OCO-2 algorithm and science teams from the start. We provided regular progress updates where we explained the underlying fundamentals of our approach, solicited feedback on the methodology and processed data products, and engaged with a wider audience including the data user community at workshops and conferences. Another key aspect was an earlier demonstration [58] that a non-linear ML-based bias correction reproduces the operational linear bias correction for QF=0 data (where errors are mostly linearly correlated with features).

As a result of this collaboration, our approach differs in several key ways from many common approaches to applied ML research: 1) Instead of a single optimized cost function (e.g. RMSE), our primary goal was to develop a robust, globally applicable model that is interpretable, allows us to quantify uncertainties, allows for quick iterations, and is not overly sensitive to the choice of hyperparameters. Compared to a typical ML approach this moved us towards a simpler model with conservative hyperparameters to avoid overfitting. 2) We conducted extensive model ablation studies to demonstrate the impact of including or excluding specific features, guided by the OCO-2 scientists to ensure all included features made physical sense and the model responded in a physically consistent manner. Thus, the final set of features is derived from a combination of domain expertise and a data-driven feature selection algorithm biasing the list of features towards a smaller set than if we had purely optimized for RMSE on the validation set. 3) We removed outliers in the data before fitting the ML bias correction model, based on thresholds recommended by domain experts. While outliers can be removed with various data-driven approaches, it was important to have a formal guarantee that certain data would not be erroneously labeled as good quality in any downstream process. 4) Finally, we adopted a spiral development approach with multiple iterations of feedback from the scientists. This highly iterative process was instrumental in aligning the model's outputs with the operational requirements of the OCO-2 mission, and to gain trust and buy-in. However, it also significantly increased the development time of the approach. We note that this iterative approach was independent of the filtering derived in Part 2, in contrast to the operational approach where there is a certain interdependence between the filtering and bias correction [102].

4.3.4 Results and Discussion

After training the model we correct the OCO-2 data record (2015-2022) and compare the resulting XCO2 to the operational approach, ground-based observations from TCCON, and carbon flux models. Additionally, we investigate the inner working of the model with uncertainty quantification and explainability methods.

4.3.4.1 Comparison to the Operational Correction

Let ΔXCO2 be defined as:

$$\Delta XCO_2 = XCO_{2,MLBC} - XCO_{2,OpsBC} \tag{4.7}$$

To investigate how the proposed bias corrected XCO2 differs from the operational product, Figure 4.17 shows ΔXCO2 on a 2° latitude by 2° longitude grid separated by QF for data from 2015 to 2022. For QF=0 (good quality data), over most of the ocean, ΔXCO2 is

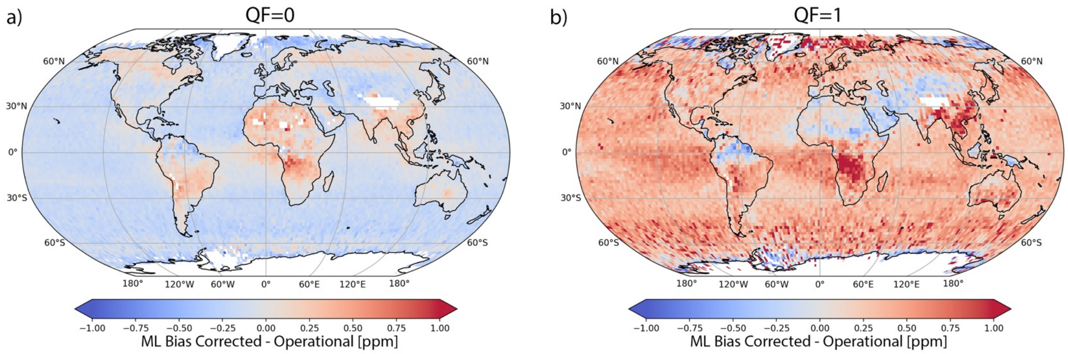

Figure 4.17: Average difference in XCO2 of ML bias correction – operational bias correction, binned over a 2° latitude by 2° longitude grid for OCO-2 data from 2015 to 2022 with a) QF=0 and b) QF=1. Bins with fewer than 10 retrievals are shown in white. Areas where the ML bias corrected XCO2 is higher (lower) than the operational product are shown in red (blue).

approximately -0.2 ppm. Above 70°N ΔXCO2 is about -0.4 ppm. Over land, ΔXCO2 is approximately -0.2 ppm over large parts of the United States, Europe, and northern Asia. The largest decrease in ΔXCO2 is over the Amazon of approximately 0.5 ppm. Positive ΔXCO2 values are found over parts of Brazil, the Andes, and most of Africa. For QF=1, differences are significantly larger. This is expected, as the operational bias correction does not consider QF=1 data when tuning its parameters. Over most of the ocean ΔXCO2 is positive. Over land ΔXCO2 is positive over southern Africa, Southeast Asia, and central South America and negative over the Amazon, Middle East, and Northwest Asia. Comparing the QF=0 and QF=1 differences, the reversed sign for differences over the ocean highlights how a more or less conservative quality filter can lead to higher or lower reported XCO2 over the ocean, respectively.

4.3.4.2 Validation with TCCON Observations

We validate the ML bias correction with independent observations from TCCON using k-fold cross validation as outlined in Section 3.2.5. This train-test split provides a robust estimate of the model's performance at locations and times where no TCCON observations are available, which is the case for most of the OCO-2 record. Table 4.10 shows the results of this evaluation separated by land and ocean for QF=0 and QF=1. The table also shows how well the bias corrected OCO-2 compares to TCCON over the training period (not an independent comparison) (ML – TCCON for 2105 to 20221) and how the operational bias correction compares over both periods. The remaining errors on the training set and the held-out data are similar. That implies that the ML bias correction model generalizes well to new data and did not overfit to the training set. Compared to the operational bias correction for the held-out test year (2022), we find similar errors over land and ocean. Keep in mind that these errors cannot be directly compared as we used k-fold cross validation for the proposed approach while the operational approach uses all TCCON stations for training and testing.

Table 4.10: Comparison of bias-corrected OCO-2 XCO_2 to TCCON. Data are separated by surface type and quality flag. Values are the mean difference ± one-sigma standard deviation. The first two columns show the performance of the proposed machine learning approach and the third and fourth column of the operational OCO-2 approach.

	ML – TCCON [ppm]		**Operational – TCCON [ppm]**	
	K-fold 2022	2015–2021	2022	2015–2021
Land, QF = 0	−0.03 ± 1.10	−0.19 ± 1.11	−0.19 ± 1.07	0.09 ± 1.16
Land, QF = 1	0.01 ± 2.27	−0.21 ± 1.73	−0.38 ± 2.32	−0.22 ± 1.94
Ocean, QF = 0	−0.36 ± 0.67	−0.16 ± 0.67	−0.06 ± 0.59	0.04 ± 0.84
Ocean, QF = 1	−0.69 ± 0.96	−0.20 ± 1.06	−1.08 ± 1.09	−0.22 ± 1.41

To better understand how remaining errors are distributed over individual TCCON stations, we show the median and quartiles of their differences in Figure 4.18 and Figure 4.19 for retrievals over land and ocean, respectively. Over land the largest systematic errors (median difference) are present for Reunion (+0.7 ppm), Harwell (-0.6 ppm), Pasadena (+0.5 ppm), Xianghe (+0.5 ppm), and JPL (+0.4 ppm). Reunion is located on a small island east of Madagascar and far away from any large CO2 sources. However, the island has been formed by volcanos and thus has steep terrain. The biases are likely due to the ACOS OCO-2 retrieval algorithm not accounting for sloped surfaces where within a given instrument footprint the altitude changes. Similarly, the ML bias correction does not contain variables that could correct for those effects, such as slope and aspect. Including those variables in the bias correction could be investigated in the future. The Harwell TCCON site is located approximately 100 km from London which could be responsible for some of the biases. Otherwise, there are no clear contenders to explain the negative bias. Similarly, Pasadena, JPL, and Xianghe are located close to the megacities of Los Angeles and Beijing, respectively. This can cause strong XCO2 gradients in time and space and, thus, make differences between OCO-2 and TCCON larger. Comparing the operational data product to TCCON we find that differences are of similar magnitude and direction than the ML bias corrected product.

Over ocean the biggest systematic difference exists for Darwin (-0.6 ppm) in northern Australia. In the absence of strong CO2 sources or complex terrain these differences might be indicative of larger scale negative biases over the ocean in this area. Other TCCON stations that provide observations over the ocean have comparatively small median biases of less than 0.2 ppm. Differences between TCCON and the operational product follow a similar pattern with the largest differences at Darwin.

When comparing retrievals over ocean and land, ocean observations consistently demonstrate both smaller biases and reduced scatter compared to land observations. This is likely due to the inherent challenges of land retrievals, which must account for complex surface effects (albedo and topography) and typically higher aerosol loading. This higher precision and accuracy of ocean retrievals contrasts with their comparatively greater scrutiny by the carbon flux inversion modeling community [129].

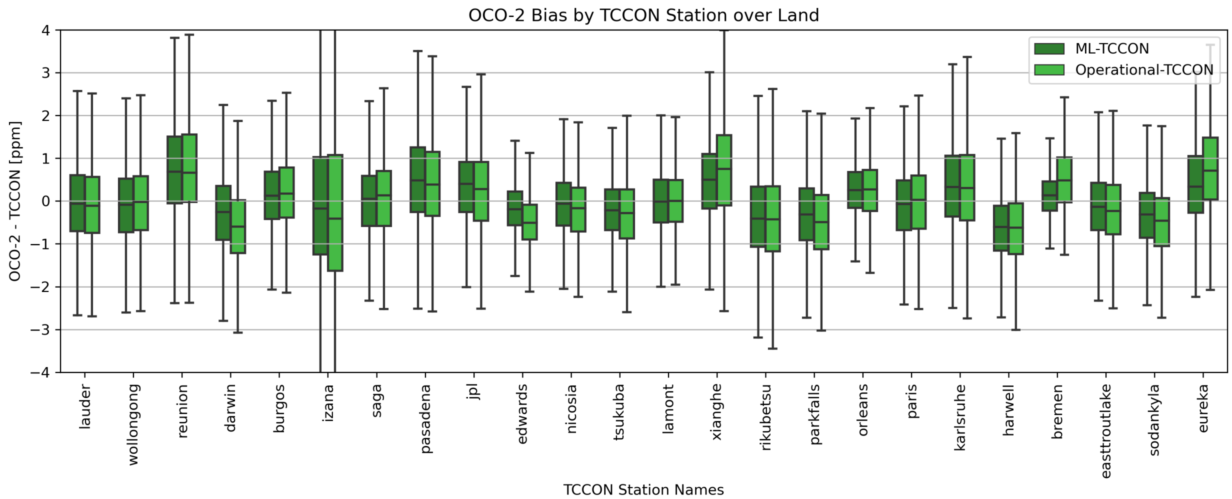

Figure 4.18: Box and whisker plot of ML bias corrected and operational OCO-2–TCCON for each station for QF=0 data from 2015 to 2022 over land. The boxes show the interquartile range (IQR) between the 25th and 75th percentiles, with the median indicated by the horizontal line. The whiskers extend to the most extreme data points that are not considered outliers (within 1.5 times the IQR). Stations are ordered by latitude from south to north. Stations with fewer than 50 colocated soundings are excluded.

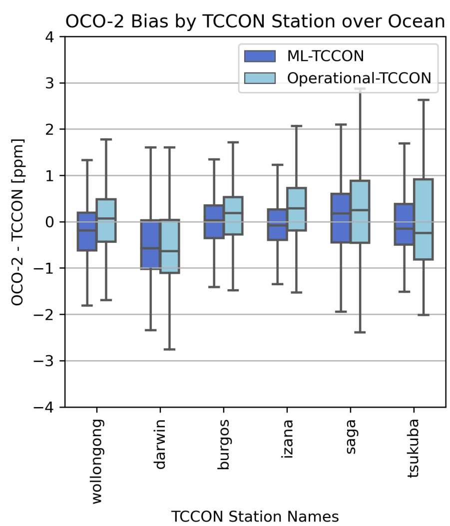

Figure 4.19: Box and whisker plot of ML bias corrected and operational OCO-2–TCCON for each station for QF=0 data from 2015 to 2022 over ocean. The boxes show the interquartile range (IQR) between the 25th and 75th percentiles, with the median indicated by the horizontal line. The whiskers extend to the most extreme data points that are not considered outliers (within 1.5 times the IQR). Stations with fewer than 50 colocated soundings are excluded.

4.3.4.3 Quantifying Uncertainties

Our modeling approach based on QRF allows us to estimate uncertainties (Meinshausen & Ridgeway, 2006) for the ML bias correction on a retrieval-by-retrieval basis. The QRF provides uncertainty estimates by calculating prediction intervals across the ensemble of decision trees, where the spread between the upper and lower quantiles reflects the confidence in each prediction. This uncertainty estimate includes both systematic components (from the bias correction) and random components (like instrument noise), since both contribute to the observed differences between OCO-2 and TCCON measurements. To evaluate how well the estimated uncertainties capture the empirically observed error, as compared to TCCON, we count how often the remaining XCO2 errors fall within various quantiles of the estimated uncertainty. For example, a well-calibrated one sigma uncertainty estimate would capture the true error 68% of the time for Gaussian distributed errors. If it would capture 90% of the true errors it would mean that the uncertainty is under confident, if it would capture 30% it would be overconfident. Figure 4.20a shows how the estimated uncertainties compare to the observed errors versus TCCON for the ML bias correction. The observed error and estimated uncertainty are in general agreement (miscalibration area = 0.18) over the whole interval with the estimated uncertainties being slightly overconfident. To put these results into perspective we also show the expected calibration error for the operational ACOS retrieval algorithm in Figure 5b. The analysis shows that the ACOS provided uncertainty is overconfident as well. For example, the 60% prediction interval from ACOS only covers approximately 20% of the observed error.

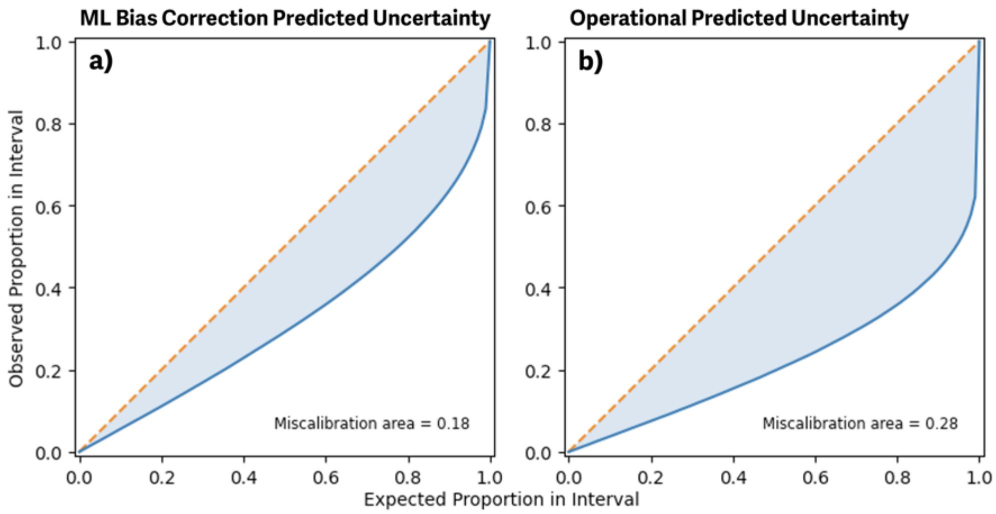

Figure 4.20: Expected Calibration Error curve and miscalibration area of the predicted uncertainty from (a) the ML bias correction, (b) operational predicted XCO2 uncertainty (for comparison). Uncertainty estimates are for the independent evaluation year 2022. The x-axis shows the ideal or expected fractional coverage of an uncertainty estimate with respect to the ground truth (TCCON). The y-axis shows the observed coverage of the predicted uncertainty estimates. A perfectly calibrated uncertainty estimate would fall on the one-to-one line (orange dashed). The miscalibration area is shown with blue shading.

After validating the provided uncertainty estimates with independent TCCON observations, we use those estimates to better understand how remaining errors in XCO2 are distributed globally. Figure 4.21 shows the distribution of spatially aggregated uncertainty from 2015 to 2022. Overall, observations over the ocean contain lower uncertainties than over land. The notable exception being near the poles, where it exceeds 2 ppm. A possible explanation for the increased errors in those regions could be soundings containing sea ice that were not successfully flagged as such by the operational processing. Over land, uncertainties are increased for regions with substantial amounts of aerosols (Northern Africa, Middle East, East Asia) and over the tropics where TCCON observations are limited, and OCO-2 observations are sparse due to clouds.

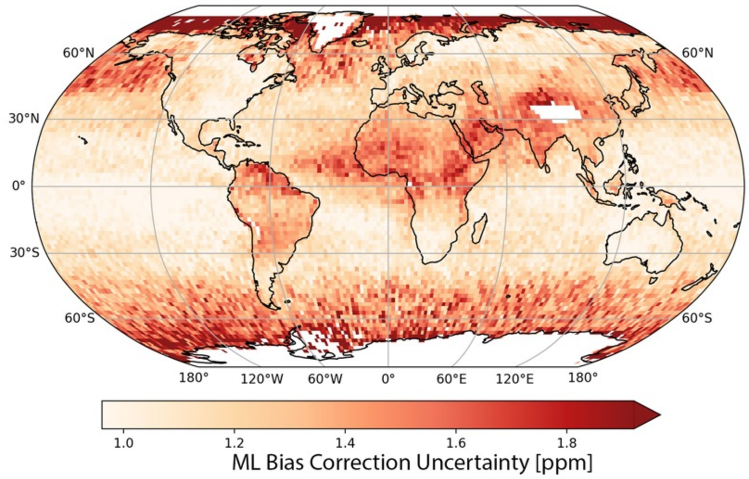

Figure 4.21: Uncertainty of ML bias correction binned over a 2° latitude by 2° longitude grid for OCO-2 data from 2015 to 2022. Shown is the 1-sigma uncertainty derived from the QRF. Uncertainty is for the combination of systematic and random errors.

4.3.4.4 Comparison to In-Situ Flux Inversion Models

One of the main applications of OCO-2 measurements is their use in flux inversion models to derive spatially and temporally resolved carbon sources and sinks. To inform those models with independent measurements it is important that the XCO2 retrieval algorithm and any post processing (such as bias correction and quality filtering) are independent to those models. Therefore, the model training of the ML bias correction is completely independent from those models. Nevertheless, to highlight possible impacts of the ML bias correction on flux inversions, we compare the bias corrected XCO2 to the mean of the XCO2 computed from the CO2 concentration fields of three flux inversion models using data from 2015 to 2022. Figure 4.22 shows the difference of ML bias corrected XCO2 to the mean of four carbon flux inversion models (CarbonTracker CT-NRT.v2022-1, MACC v21r1, LoFI m2ccv1bsim, UnivEd v5.2) by season. Significant spatial and seasonal differences are apparent, which could result from remaining OCO-2 biases or deficiencies in the flux inversions. During December, January, February (Figure 4.22a), we find that bias corrected XCO2 from OCO-2 is higher than from the in-situ flux inversion models over the Sahel region of Africa, where dry season conditions and high aerosol loading exist. This difference has been noted previously (Gaubert et al., 2023) and is an ongoing discussion topic within the flux inversion science community.

For most of the ocean OCO-2 is lower than the models, most notably over the Tropical Pacific, South Atlantic, and Southern Ocean. Because the inversions are matching marine boundary-layer observations, such a difference could result from vertical mixing biases in the transport models used in the inversions. In March, April, May, higher OCO-2 XCO2 are present over parts of the North Atlantic and North Pacific, whereas lower values appear over most of the ocean for the southern hemisphere. For June, July, August, OCO-2 XCO2 is notably higher over northern high-latitudes over the ocean and over the Amazon rainforest, but lower over most other regions. During September, October, November, the higher values for OCO-2 persist over the Amazon and northern ocean. Compared to March, April, May, the band where OCO-2 reports lower values shifts from the Southern Hemisphere tropics to the Northern Hemisphere tropics. These differences are not necessarily residual errors in the OCO-2 data, since there are few measurements constraining these models over the ocean as well as in the Tropics. The magnitudes of these differences make it clear that inversions using only in-situ or OCO-2 observations will infer very different regional and zonal-mean fluxes. To evaluate the significance of differences between the ML bias correction and model mean, we compared them to the spread among the flux inversion models themselves (not shown). Over most ocean regions, the ML bias correction differs from the model mean by less than 2 standard deviations of the inter-model spread. However, near the equatorial Pacific, these differences exceed 4 standard deviations. Over land, we observe a similar pattern, with differences typically around 2 standard deviations, but reaching up to 4 standard deviations in regions including northern Africa, the Middle East, the Himalayas, the Andes, and Australia. For QF=0 data from 2015 to 2022, the average ratio of ML-to-model mean differences relative to the inter-model standard deviation is 2.3 over ocean and 2.6 over land.

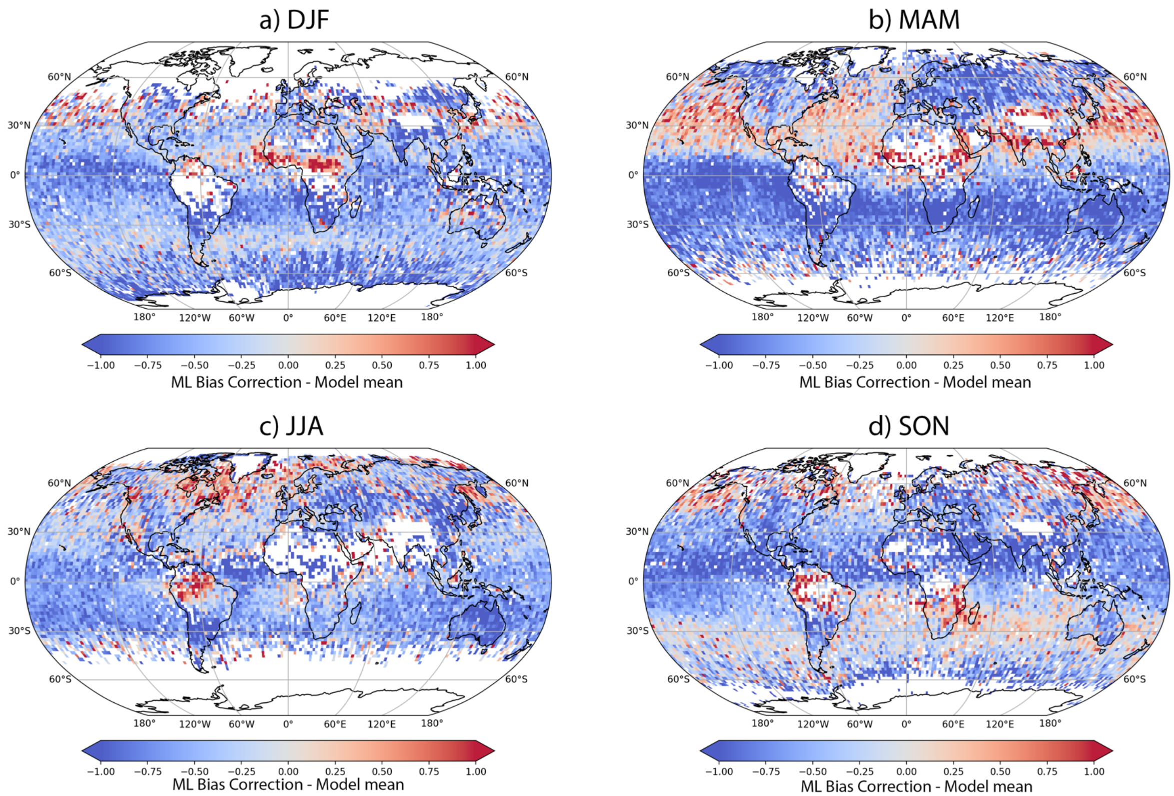

Figure 4.22: XCO2 after the ML bias correction minus the mean of four in situ flux inversion models (CarbonTracker CT2022, MACC v21r1, LoFI m2ccv1bsim, UnivEd v5.2) across four seasons: a) DJF (December-February), b) MAM (March-May), c) JJA (June-August), and d) SON (September-November), binned over a 2° latitude by 2° longitude grid by season for QF=0 data from 2015 to 2022. Bins with fewer than 10 retrievals are shown in white. Areas where the ML bias corrected XCO2 is higher (lower) than the flux inversion models are shown in red (blue).

4.3.4.5 Correction of 3D Cloud Biases

An earlier study had found uncorrected negative biases in OCO-2 retrieved XCO2 in the proximity of clouds [78]. These biases have been attributed to 3D cloud effects (horizontal photon scattering) that is not accounted for in the operational retrieval. Since directly modeling the 3D cloud effect during the XCO2 retrieval would be too computationally expensive, an empirical bias correction was developed [83]. A key conclusion of that study was that 3D cloud biases are non-linear and mitigating them can be achieved with hand-engineered features (CSNoiseRatio and H_continuum) that rely solely on OCO-2 observations and not on external cloud information, e.g., from MODIS. The operational bias correction started using those features with a linear bias correction in the latest version (B11). We include those same features in the feature selection process outlined in Section 3.3 and both features were selected for the bias correction over ocean, but not land.

To evaluate how successful the ML bias correction is in mitigating 3D cloud biases, we compare nearest cloud distance to XCO2 biases within small areas. However, the small area mean is only calculated across soundings that are further than 5 km away from the nearest cloud. Figure 4.23a shows the removal of cloud-dependent biases after applying the proposed correction. For comparison, the operational OCO-2 product is shown in Figure 4.23b and still contains remaining biases in the proximity of clouds even after correction. The spread of the distribution (5th to 95th percentile) is reduced as well with the ML bias correction from 3.9 ppm to 3.2 ppm at a nearest cloud distance of 1 km and from 2.2 ppm to 2.0 ppm at nearest cloud distance of 5 km. It is noteworthy that most of this bias in the operational product is for QF=1 data; QF=0 data have a much smaller cloud bias, which is nevertheless slightly improved with the ML bias correction. In general, 3D cloud biases play a significant role of OCO-2 (and similar missions) retrievals as approximately 40% of QF=0 and 70% of QF=1 retrievals are within 4 km of clouds).

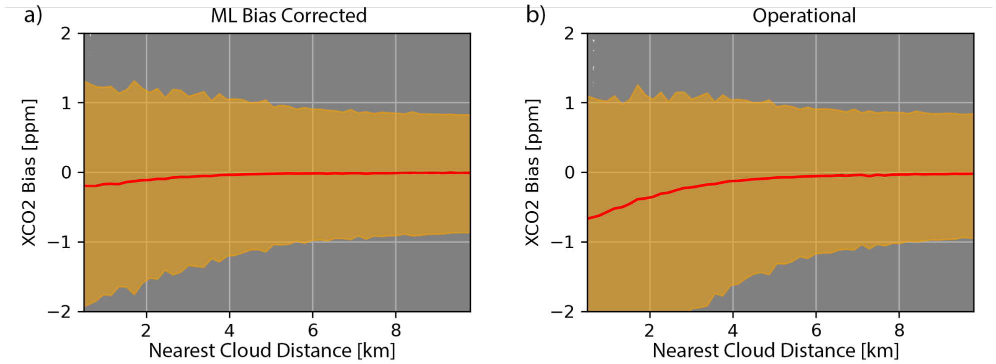

Figure 4.23: 3D cloud biases in XCO2 versus nearest cloud distance in km for OCO-2 retrievals. Our ML bias correction is shown in a) and the operational product in b). Biases are derived from QF=0 and QF=1 data from 2016 to 2019. The orange shading shows the 5th to 95th percentile of the biases and the red line shows the mean for each cloud distance bin.

4.3.4.6 Plumes

A potential concern of fitting a more complex bias correction model that in part relies on the small area approximation is the removal of real CO2 enhancements from the record, e.g., CO2 plumes in the proximity of power plants. To verify that the ML bias correction does not remove real CO2 plumes, we show the corrected and operational retrieved XCO2 for the overpass of two CO2 plumes from active power plants in Figure 9. The first plume originates from the Matimba power plant in South Africa and was recorded November 18th, 2018. The second plume stems from the Ghent power plant in the USA and was recorded on August 13th, 2015. Both plumes are clearly visible in the OCO-2 data, with enhancements on the order of 2 and 4 ppm, respectively. As shown in Figure 4.24, the ML bias correction maintains the overall shape and magnitude of the plumes.

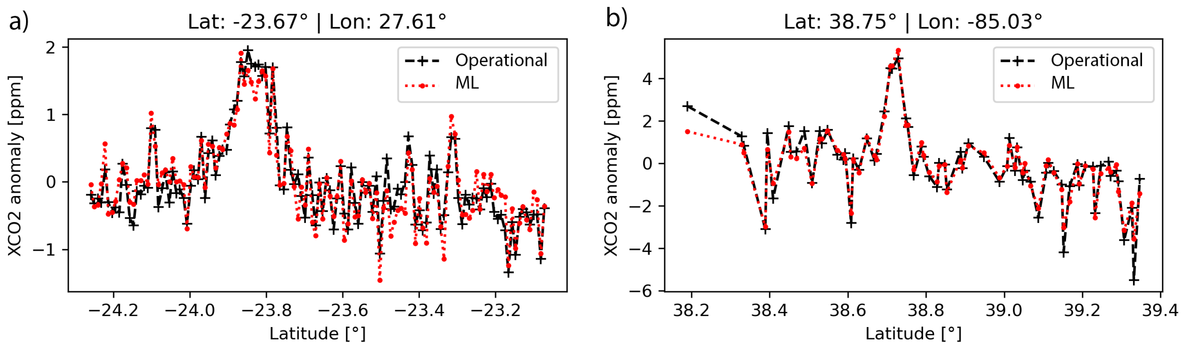

Figure 4.24: Two XCO2 plumes measured by OCO-2 processed operationally (black) and by the ML bias correction (red). a) Matimba power plant in South Africa on 11/18/2018; b) Ghent power plant in the USA on 08/13/2015.

4.3.4.7 Correction Model Interpretation

To further investigate how much of the reduction in OCO-2 biases stems from the correction model being a QRF, we trained a second, simpler model for comparison. This model is based on linear ridge regression and uses the same data preprocessing, train-test split, and features as our proposed approach. However, it can only represent linear relationships and does not use the ensemble approach of the QRF. The results are compared to the ML bias correction and the operational product in Table 4.11. Over land for QF=0 and QF=1 data the three models have similar agreement with TCCON with a slight improvement in scatter by the ML bias correction. Over ocean we find again similar performance between the models. Most notable differences are within the scatter. The ML bias correction reduces the scatter in the operational product from 0.82 ppm to 0.67 ppm for QF=0 and 1.42 ppm to 1.10 ppm for QF=1. Given that the linear model has comparable scatter to the operational product we can attribute this improvement to the non-linear nature of the ML bias correction.

Next, we examine the inner workings of the model by assessing the feature importance and ensure that we don't get the right answer for the wrong reason. We investigate global and local feature importance and contribution. Globally, we assess the relative importance of each feature by shuffling one feature at a time and evaluating how much this affects the model accuracy. Features that, when shuffled, lead to a significant drop in accuracy are

Table 4.11: Carbon-flux models from the inversion ensemble used to evaluate the combined bias correction and quality filter.

Model name	**Institute**	**Transport model**	**Resolution** [lat × lon × time]	**Inverse method**
CarbonTracker v2022-1	NOAA Global Monitoring Laboratory	TM5	$2° \times 3° \times 3$ h	EnKF
MACC v21r1	Copernicus Atmosphere Monitoring Service	LMDZ	$1.9° \times 3.75° \times 3$ h	Variational
UoE v5.2	University of Edinburgh	GEOS-Chem	$2° \times 2.5° \times 3$ h	EnKF
LoFI m2ccv1bsim	NASA Goddard GMAO	GEOS GCM	$0.5° \times 0.63° \times 3$ h	—

deemed important and vice versa. Figure 4.25 shows the feature importance for the ML bias correction model over land (left) and ocean (right) minimizing errors relative to TCCON (top) and small areas (bottom). To interpret the plots, it's important to remember that the bias correction based on TCCON is applied to the data before it is used to train the second bias correction model based on small area differences. Thus, the importance of features shown for the second bias correction model cannot be interpreted in isolation of the first bias correction model.

For the TCCON-based bias correction the two most important features over land are pressure differences (dpfrac) and the change in the retrieved vertical CO2 gradient relative to the prior (`co2_grad_del`) (Figure 14.25a). Errors in retrieved surface pressure are themself a proxy for cloud and aerosol contamination and their impact on the path lengths. Both features are also part of the operational bias correction, illustrating the consistency between both approaches. In third place is the retrieved pressure height of water clouds (`water_height`), that is not used in the operational bias correction. For the small area-based bias correction over land (see Figure 4.25c) we find the same two features, `co2_grad_del` and dpfrac, to be significant indicators of bias. Differences between the individual instrument footprints are important as well. In the operational product, these footprint differences are believed to arise from residual calibration errors that vary by footprint and are removed as a post-processing step after the parametric bias correction. For the ML bias correction, we use footprint number as a predictor instead of performing the operational post-processing adjustment. This permits the ML bias correction model to capture calibration dependencies as well as interactions with the other predictors. Finally, retrieved aerosol loading of dust, water and salt (`dws`) play an important role which has been noted previously (Bell et al., 2023).

Over the oceans there are two notable features of high importance: pressure difference (dp) and sensor_zenith_angle (Figure 4.25b). The sensor_zenith_angle term addresses a systematic bias of OCO-2 relative to TCCON as a function of latitude, further discussed below. The dp term over the ocean is like dpfrac over land in that it indicates that errors in retrieved surface pressure are the most pressing challenge for OCO-2 retrievals. The small area-based bias correction (Figure 4.25d) is dominated by footprint biases and 3D cloud biases captured by color_slice_noise_ratio_o2a, a measure of the subscene spatial variability in the observed radiances.

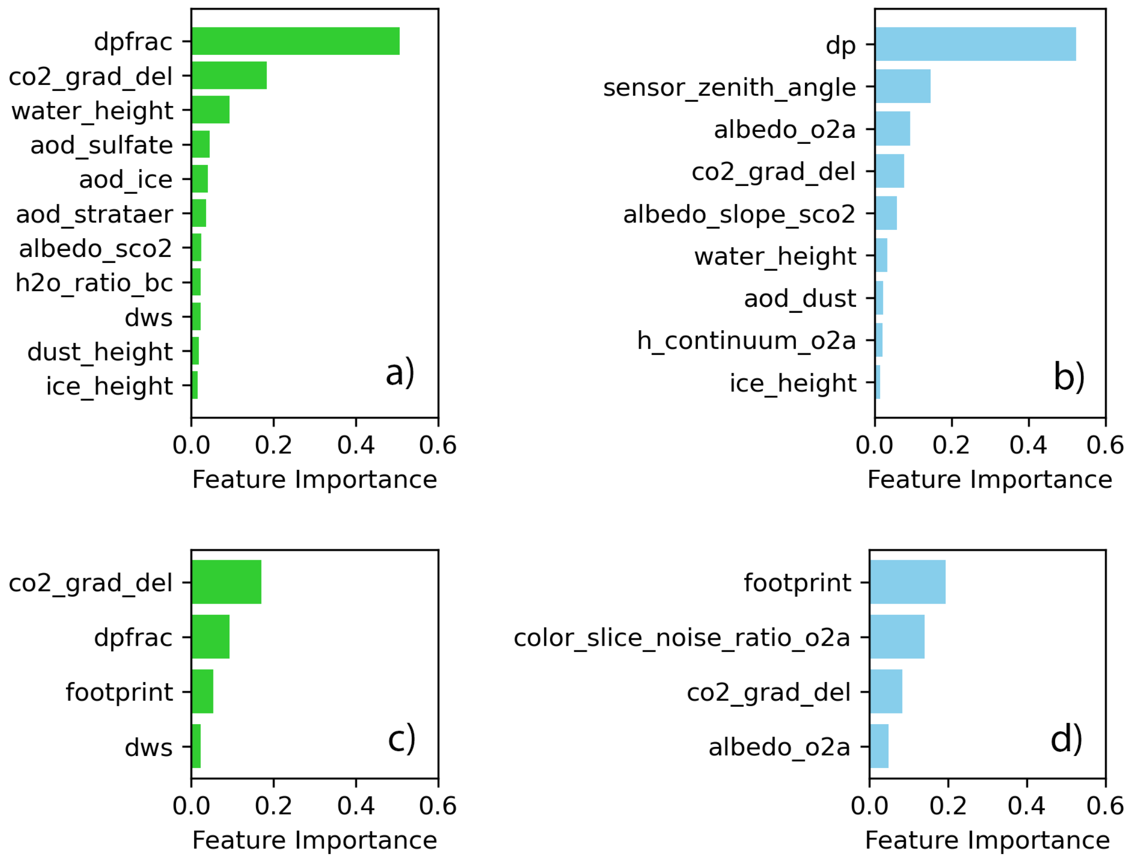

Figure 4.25: Feature importance for OCO-2 observations from 2015 to 2022 for ML bias correction based on comparisons to TCCON over a) land and b) ocean and correction based on small areas over c) land and d) ocean. Feature importance is calculated as the reduction in coefficient of determination of the ML bias correction model when shuffling values of individual features.

For the local analysis we use SHAP to investigate how each feature contributes to the estimated bias and how this contribution varies spatially. SHAP uses a game theoretic approach to decompose a prediction, in our case XCO2 bias, into the sum of contributions from each feature. This allows us to quantify how each feature contributes to the overall prediction. For each retrieval and feature, the SHAP value indicates both the direction and magnitude of that feature's contribution to the predicted XCO2 bias, which is then subtracted during correction. Using our ML bias correction models, we calculate SHAP values for each retrieval for 2015 to 2022. The results are averaged globally over a 2° latitude by 2° longitude grid. The aggregated maps of the four most important features are shown in Figure 4.26.

For land the contribution by dpfrac is mostly negative in the tropics (20°S to 20°N) and increasingly positive at latitudes above 45° N. Additionally, dpfrac is an indicator of biases over the Andes in South America, which is likely due to surface pressure errors in the prior over steep terrain. The spatial pattern of where `co2_grad_del` contributes to predicted biases seems to be correlated with regions of high aerosol loading in Northern Africa, the Middle East, and East Asia. The `water_height` is mostly a predictor of negative XCO2 biases and again is correlated with regions of high aerosol loading. The model contribution by sulfate aerosols follows a more complex spatial pattern. Areas where the ACOS retrieval algorithm indicates moderate to high sulfate aerosol loadings (not shown) show a positive SHAP value while areas where there are limited or no sulfate aerosols show a negative bias.

Over ocean, `dp` and `color_slice_noise_ratio_o2a` seem to pick up on negative biases collocated with clouds. These biases could stem from thin clouds that were not filtered out by the operational cloud filtering preprocessor ([128] or result from 3D cloud effects [78]. The `sensor_zenith_angle`, a proxy for the distance that sunlight travels through the atmosphere, shows a clear latitudinal dependence. The model uses `sensor_zenith_angle` to correct strong positive biases towards the poles and weak negative biases towards the equator. There are multiple other variables that are highly correlated with this feature and could replace it in the bias correction. Those are airmass (another proxy for how far sunlight travels through the atmosphere) and T700 (temperature at 700 hPa from GEOS5FPIT). All three variables point towards errors in the atmospheric absorption coefficients used in the forward model during the OE retrieval (ABSCO v5.2). Finally, footprint-dependent biases do not show regionally coherent spatial variability.

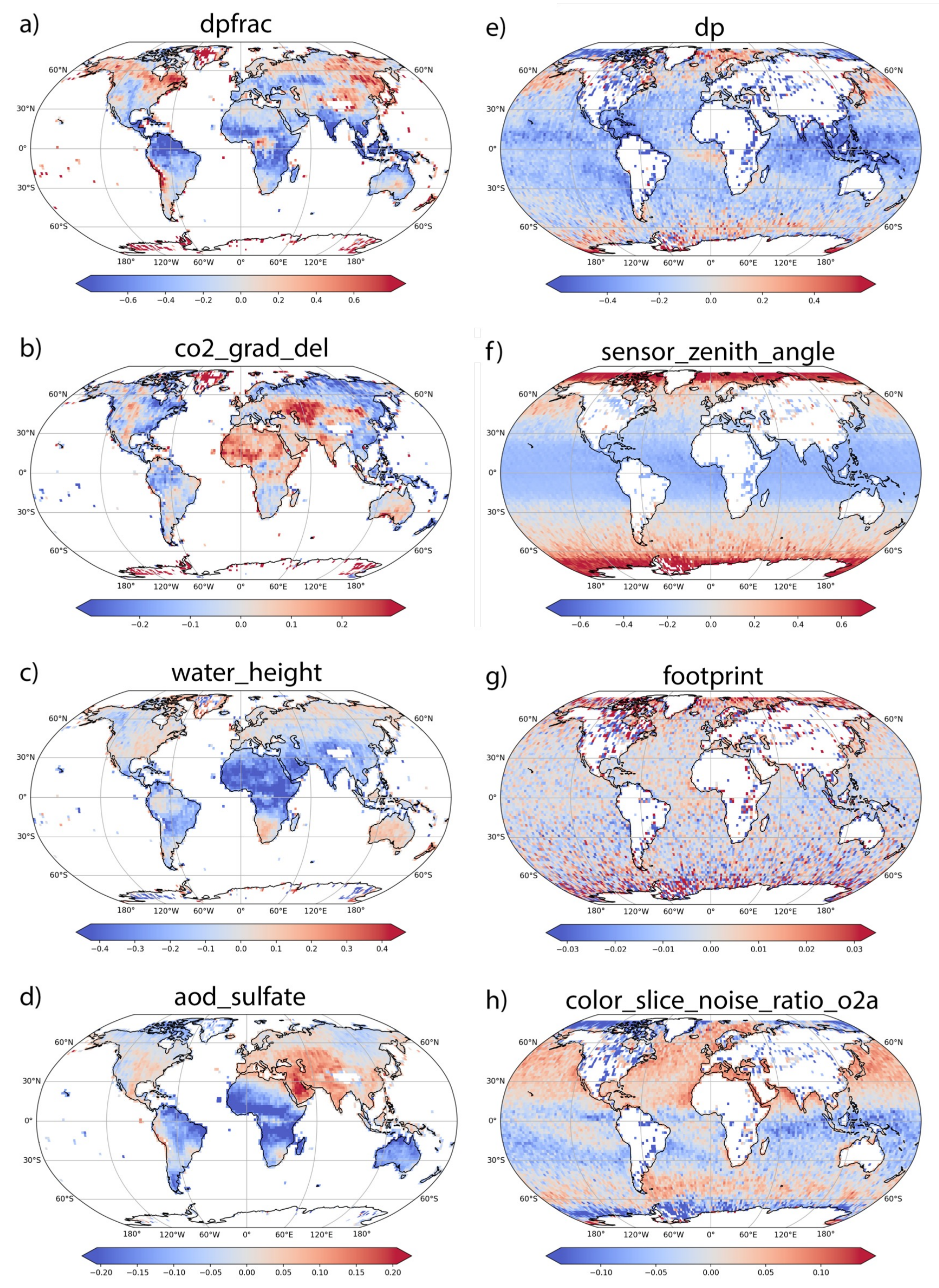

Figure 4.26: Feature attribution of the ML bias correction for OCO-2 observations over land (a-d) and ocean (e-h) for their four most important features. Feature attribution is calculated with SHAP at the per-retrieval level and binned over a 2° latitude by 2° longitude grid for OCO-2 data from 2015 to 2022. While SHAP values are unitless they can be interpreted as the contribution of each feature to the bias in approximate ppm. Bins with fewer than 10 retrievals are shown in white. Areas where a feature is associated with a positive XCO2 bias are shown in red and areas with a negative bias in blue. Note, color bars differ for each panel.

4.3.5 Conclusions

We present a new ML framework for bias correcting OCO-2 XCO2 retrievals that increases the utility of space-based CO2 measurements in several ways. By transitioning from a manual tuning process to a machine learning-based approach, we provide an improved bias correction while maintaining transparency and traceability of the approach. Furthermore, this methodological shift simplifies the transition towards non-linear bias modeling that captures complex non-linear relationships between predictor variables and biases. OCO-2 data corrected using the newly developed ML bias correction has a similar RMSE compared to TCCON for high quality (QF=0) data but significantly improves the RMSE for lower quality (QF=1) data. The most significant XCO2 bias predictors we identified are surface pressure retrieval errors caused by aerosols and clouds. Additional important predictors include aerosols, surface characteristics, and viewing geometry parameters, all of which are known to affect the accuracy of the full physics retrieval algorithm [5].

Using Quantile Regression, we derive uncertainties of the ML bias correction at the per-retrieval level. The uncertainties indicate when the ML bias correction should not be trusted, for example, when the model extrapolates or there is irreducible error in the retrieved XCO2. The derived uncertainties can be used for data quality filtering, which we explore in [50]. Overall, uncertainties were lowest near the equator, with values increasing poleward. Especially large uncertainties were found for high northern latitude oceans and in a spatial pattern that resembles global maps of aerosol loading. Our approach removes all flux model dependencies from the bias correction development process, ensuring that OCO-2 measurements can be properly assimilated into flux inversion models without circular dependencies. This independence is critical for accurately estimating carbon fluxes that inform policy and decision making.

The ML framework was developed with tight collaboration between members of the calibration, algorithm, and science teams. Together with the use of simple, explainable models, this fostered trust and buy-in from the OCO-2 team. By aligning our ML practices with the needs of the science community, we developed an approach that not only enhances the accuracy and utility of space-based XCO2 measurements but also earns the trust and support of domain experts. This case study exemplifies how collaborative approaches can bridge the gap between advanced computational techniques and traditional Earth science methods, paving the way for more integrated and trusted solutions in Earth system monitoring. These efforts align with the broader goal of making scientific data more accessible and actionable, thereby playing a crucial role in addressing global climate challenges.

The approach is generalizable to other missions that fulfill two criteria: their retrieval is a joint retrieval of parameters other than XCO2 (like Optimal Estimation) and for a subset of the retrievals there exists a proxy for the unbiased quantity of interest (e.g., ground-based measurements or airborne campaigns like ATom [96] and other global survey campaigns). For such a mission, our approach provides a framework that allows for the correction of systematic retrieval biases in a reproducible and traceable manner.

Chapter 5

Combining Domain Science Input with Data-Driven Techniques for Multi-Use Quality Filtering of Column Averaged CO_2

5.1 Chapter Foreword

Quality filtering represents a deceptively subjective challenge that uniquely differs from the purely quantitative task of the bias correction in which we simply aim to reduce the RMSE of OCO-2 observations to ground truth sources. While we still aim to further reduce error in a quality filtered product, we must contend with a diverse set of end use cases. For example: studies that analyze local scale systems such as power plants [94] or regional scale such surface flux in the Amazon Rain-forest Basin [7]; often have sparse data availability due to limited overflights or increased cloud presence. These users have often opted to forgo employing the operational quality filter as it drastically reduces available soundings.

In [50], a novel quality filter that targets these additional use cases while still provided the standard option for flux inversion modeling is presented. This new quality filter scheme utilizes several machine learning classifiers which are developed using a Bayesian Multi-Optimization and several competing objective functions (e.g., minimizing RMSE; maximizing data availability; minimizing overall retrieved uncertainty) along with additional constraints. The development of these objectives and constraints involved long discussions with the algorithm and science teams. These interactions also helped to increase trust and buy-in from the domain science experts in data science methodologies. The multi-disciplinary collaboration focus of this work was presented at the *Science Understanding from Data Science (SUDS)*

Conference, hosted by CalTech (https://sudsconf.com/) (August 21-23rd, 2024).

Summary for the data science practitioner

This chapter turns quality filtering into a transparent, tunable machine-learning system. Three complementary sub-filters—one aligned to ground truth, one that preserves small-area coverage, and one driven by model uncertainty—are combined into a simple ternary flag. Domain-science input is encoded directly through multi-objective optimization, making the coverage–bias trade-off explicit and adjustable for different use cases (for example, flux inversions versus local plume studies). Station weighting and small area spatial weighting counter uneven regional availability so that thresholds do not overfit to well observed areas and balances minimizing systematic errors with throughput in under served regions. Removing features that correlate with autocorrelation spatially and temporally avoids smoothing out true CO2 enhancments from the record. The result is a filter that recovers data the operational approach would discard while keeping systematic errors in check and providing clear knobs that science teams can tune.

5.2 Uncertainty-Aware Machine Learning Bias Correction and Filtering for OCO-2: Part 2

5.2.1 Introduction (Paper)

The Orbiting Carbon Observatory-2 (OCO-2) has produced one of the longest running data records of global column-averaged dry-air mole fraction of carbon dioxide (XCO2) estimates from 2014 to the present day [32]. The OCO-2 instrument makes measurements of reflected solar spectra in three viewing modes (nadir, glint, and target) and utilizes NASA's Atmospheric Carbon Observations from Space (ACOS) [19]; [23]; [104] algorithm to derive estimates of the atmospheric state including XCO2. These observations are used in downstream tasks, such as estimating carbon fluxes from natural and anthropogenic sources and sinks [111];[93].

Space-based estimates of XCO2 have been shown to be biased when compared to ground-based estimates such as from the Total Carbon Column Observing Network (TCCON) [146] and flux inversion models [129]. This requires a post-hoc correction to the XCO2 retrieved by ACOS [102]. Operationally, this correction is iteratively fit in tandem with a quality filter that flags low-quality retrievals. The current operational quality filter relies on manually setting thresholds on select co-retrieved atmospheric state variables as well as variables derived from pre-processors. A variety of optimization criteria have been designed to reduce the mean and variance of the remaining biases in XCO2. The thresholds are also chosen so that the regime of interaction between the state variable and mean error is largely linear, thus improving the fit of the multilinear bias correction. Soundings that fall within the accepted range for all selected state variables are designated as 'good' and denoted as (QFOp=0) while low-quality data is indicated as QFOp=1. Recent work has shown that a non-linear bias correction model can further reduce biases in previously filtered data from 3D cloud effects [83] and that fitting a non-linear correction can allow for the relaxation of the hand-selected filters used in the operational quality filtering, thus increasing the availability of good quality data [58].

Constraining global carbon flux inversion models requires high-accuracy estimates of XCO2 and is one of the primary use cases of the OCO-2 XCO2 data record. The operational quality flag was developed with this in mind. However, application of the filter reduces data availability by 40%, which is not suitable for local [95] or regional [87] studies where data availability is already limited due to clouds or snow. Additionally, the OCO-2 operational bias correction and quality filter are partially determined using co-sampled XCO2 estimates from a set of carbon flux models that in turn are used to estimate sources and sinks of CO2 by assimilating OCO-2 XCO2. The current operational bias and filtering methodology therefore contains some degree of circularity [127]. Early in the OCO mission, a genetic-based filter, Data Ordering Genetic Optimization [75], was used to create warn levels in an attempt to address the wider user community. Warn levels consisted of 20 flag values that could be selected based on requirements. Although DOGO results were made available in earlier ACOS implementations, but its use was discontinued in part due to its complexity.

In this work we propose a new quality filtering approach developed in conjunction with the OCO-2 science team that addresses the issues outlined above and further improves upon the operational filter by: 1.) utilizing a data-driven machine learning approach that combines uncertainties from the bias correction with agreement between ground truth sources, 2.) removing the carbon flux model data from the fitting of the quality filter thus removing a potential source of circularity, 3.) introducing a new ternary quality flag that provides an improved reduction of remaining biases for carbon flux modeling and that allows for greater flexibility in meeting the needs of user communities while remaining simple to use. This paper covers the development of the quality filtering approach and builds upon the non-linear bias corrections described in Part 1 [84]. Section 2 introduces the training and evaluation data sources, which we term truth proxies. Section 3 discusses methodologies used to develop the quality filter using a multi-objective optimization of domain expert defined constraints. In Section 4 we present and interpret the results of the proposed quality filter and compare to the operational approach. Lastly, we conclude with future work plans.

5.2.2 Data

5.2.2.1 ACOS Retrieved XCO2

NASA's OCO-2 instrument makes measurements of reflected solar spectra in three bands: the near infrared O2 A-band at 0.765 µm, the shortwave infrared weak CO2 band at 1.61 µm and the short-wave infrared strong CO2 at 2.06 µm. These measurements are made every third of a second at a ground resolution of 2.25 km 1.3 km at nadir for each of the eight footprints across the 10 km-wide orbital swath. Thus, amounting to approximately 105 cloud free soundings per day. OCO-2 utilizes the Atmospheric Carbon Observations from Space (ACOS; [102]) algorithm (current version 11.2) which retrieves a maximum a-posteriori estimate of the atmospheric state, several surface parameters, and other factors including a 20-layer profile of CO2, by employing an optimal estimation-based approach [118]. The retrieved state, along with additional variables such as viewing geometry and the operational quality flag, are provided by the OCO-2 "Lite" files and are available at (https://disc.gsfc.nasa.gov/datasets/OCO2$_L 2_L ite_F P_1 1.1r/summary, lastaccess : 01/2024$).

In addition to the state variables, we incorporate bias-corrected XCO2 that stems from a new non-linear bias correction described in Part 1. The novel bias correction is based on machine learning and trained in a two-step process, constraining XCO2 with ground-based estimates from TCCON and XCO2 variability of collocated soundings. The machine learning methods used provide a robust uncertainty estimate of their predictions. Unless otherwise stated, this study uses the newly corrected XCO2 to derive the discussed quality filter while the operationally corrected XCO2 is only used for comparisons.

Observations are made in one of three viewing modes (nadir, glint, or target) depending on surface type. Nadir is the primary viewing mode for land surface types however glint and target is also employed e.g., at TCCON stations or calibration sites. Over the ocean, OCO-2 utilizes only the glint viewing mode due to the dark reflective properties. Similarly to the operational the ML bias correction combines all three viewing modes when fitting the land models.

5.2.2.2 TCCON XCO2

TCCON is a network of ground-based Fourier Transform Spectrometers that make full column retrievals of CO2 using the GGG2020 retrieval algorithm [66]. The well-calibrated global network is a truth proxy for the OCO-2 and OCO-3 missions [129] that is used to train and fit the current operational bias correction and filter as well as for validation. The full TCCON data record can be found at (https://tccondata.org, last access: 01/2024). It is important to note that TCCON sites are not uniformly distributed and are spatially biased toward the northern hemisphere, particularly in North America and Europe.

5.2.2.3 Carbon Flux Inversion XCO2 Fields

We utilize XCO2 fields provided by flux inversion model operational runs to evaluate the proposed bias correction+quality filter. The model-reported XCO2 are aggregated into a median value and assigned to collocated OCO-2 soundings. A similar model average XCO2 was used to fit the OCO-2 v11 operational bias correction and quality filter; however, we only use the model averaged XCO2 for comparison and to visualize any differences. The models used in the evaluation dataset are shown in Table 5.1.

Table 5.1: Carbon-flux models from the inversion ensemble used to evaluate the combined bias-correction and quality-filter approach.

Model name	**Institute**	**Transport model**	**Resolution** [lat×lon×time]	**Inverse method**
CarbonTracker v2022-1	NOAA Global Monitoring Laboratory	TM5	2°×3°×3 h	EnKF
MACC v21r1	Copernicus Atmosphere Monitoring Service	LMDZ	1.9°×3.75°×3 h	Variational
UoE v5.2	University of Edinburgh	GEOS-Chem	2°×2.5°×3 h	EnKF
LoFI m2ccv1bsim	NASA Goddard GMAO	GEOS GCM	0.5°×0.63°×3 h	n/a

5.2.3 Methods

To derive the ternary data quality flag, we construct three sub-filters. Two of the sub-filters are based on machine learning classifier models and are trained using different ground truth proxies. The third sub-filter is based on the bias correction uncertainty that was discussed in Section 4.3 of Part 1. The agreement or union between the three sub-filters is used to define the quality flag where the best quality data (QFTernary=0) passes all three filters. This data targets the flux inversion community that preferentially values high accuracy XCO2 measurements over high data throughputs. Soundings that pass one or two of the three sub-filters are assigned a quality flag value of QFTernary=1. Users requiring more data can select QFTernary=1 data in addition to QFTernary =0, which we will denote as

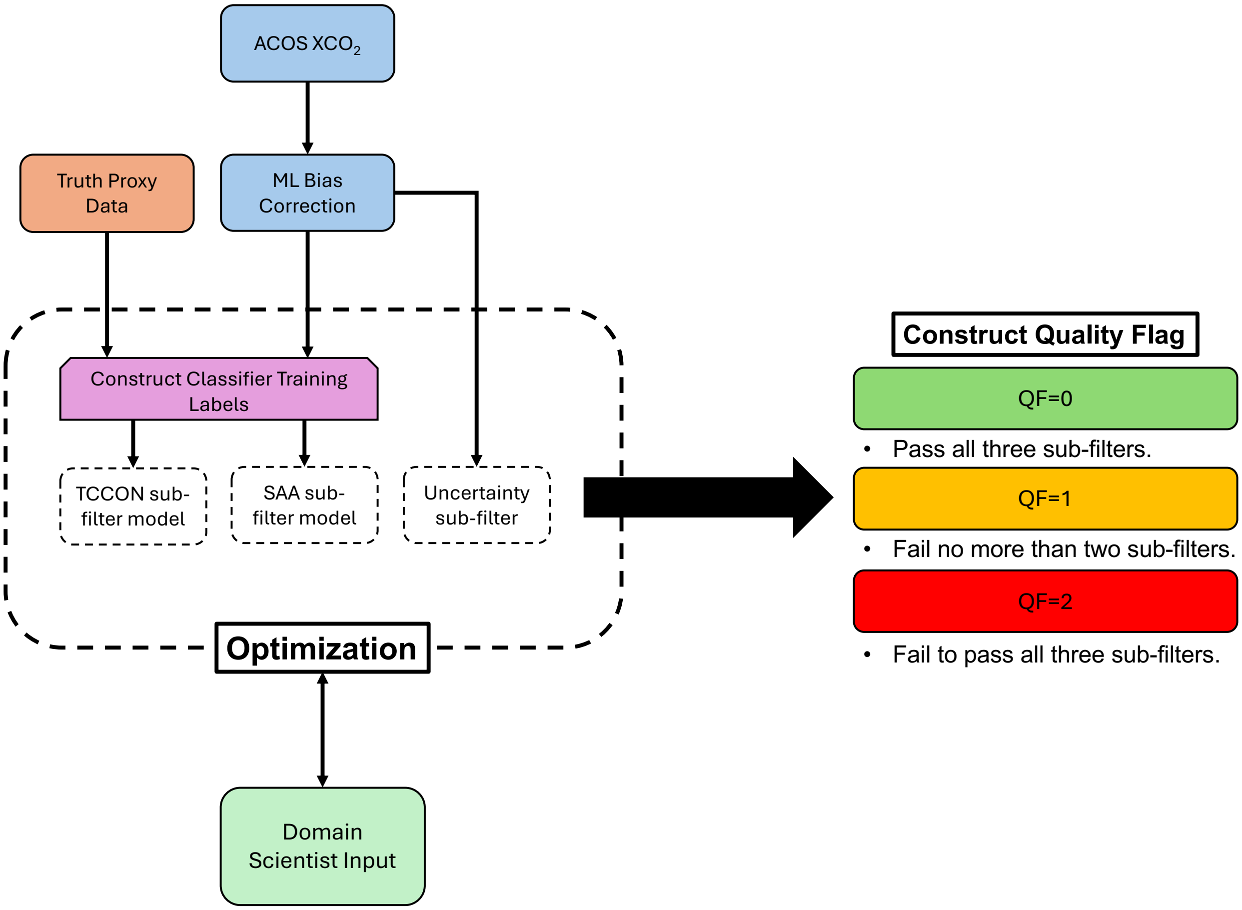

Figure 5.1: Flow of operations to derive the proposed ternary quality flag. ML truth proxy & uncertainty sub-filters, domain scientist feedback and ternary quality flag logic.

QFTernary=0+1. Soundings that do not pass any of the filters are assigned the quality flag value (QFTernary=2). The two machine learning sub-filters and uncertainty-based sub-filter are optimized using a Multi-Objective Tree-Structured Parzen Estimator (MOTPE; [107]) that balances competing requirements of minimizing biases and maximizing data availability. An overview of the filter development process is shown in the flow chart in Figure 5.1 and discussed below.

5.2.3.1 ML-based Truth Proxy Sub-Filters

The operational quality flag (QFOp) is binary, separating the available data into 'good' (QFOp=0) and 'low-quality' (QFOp=1) and is primarily tailored for global carbon flux analysis. For users interested in plume analysis or regional or local scale flux analysis, the operational flag is often not utilized, which results in the use of both good and poor-quality data [94] – an unintentional result of the binary QF design.

Similarly to the operational quality filtering approach, we utilize truth proxies to develop the ML component of our proposed quality filter. Like the operational filter we also separate filtering by surface type and combine the land and nadir viewing modes over land scenes. For each surface type we train a pair of ML-based sub-filters one for each truth proxy source. These proxies are independent XCO_2 values that are used in the construction labels for training the ML sub-filter models which use the difference or bias between the ML corrected XCO_2 value for OCO-2 and the truth proxy:

$$\Delta XCO_{2,\text{Proxy}} \;=\; XCO_{2,\text{MLBC}} \;-\; XCO_{2,\text{Proxy}}.$$

Here $XCO_{2,\text{MLBC}}$ is the machine learning bias-corrected column-averaged CO_2 from OCO-2 as described in Part 1 and $XCO_{2,\text{Proxy}}$ is the column-averaged CO_2 from the data quality proxy. Further construction of the training labels is described below.

To derive the labels used for model training we first bin $|\Delta XCO_{2,\text{Proxy}}|$ by percentile. A percentile is then selected. Values below the selected percentile are labeled as 0 ('good') and labels greater than or equal to the threshold are labeled as 1 ('low-quality). This is performed independently for each data quality proxy by surface type/observation mode (land/nadir+glint +target and ocean/glint). The labels are only calculated for the training datasets $\{(x_{\text{State}}, y_{\text{Proxy}})\}$ used for fitting the Random Forest filter models.

The proxy sub-filters are Random Forest classifiers [10], which are a commonly employed ensemble-based supervised machine learning method that approximates a function from input features to a target set of labels by building multiple weak learners (decision trees) during training and outputs the mode of the predictions of the individual members. Each decision tree is constructed from a bootstrapped sample drawn from the training set. At each node, instead of using the best split among all features, the algorithm selects the best split from a random subset of features, leading to diversity in the ensemble members and thus enhancing the model's robustness and accuracy when applied to new data. Each of the Random Forest sub-filter models are trained on a set of preprocessing features and retrieved atmospheric state variables x_{State}, that serve as input features, paired with the discrete binary training labels y_{Proxy} where $y_{\text{Proxy}} \in \{0, 1\}$.

5.2.3.2 TCCON Sub-Filter Classifier

For the TCCON proxy we match OCO-2 soundings to TCCON retrievals that are collocated temporally within 1 hour and spatially within 100 km. Any differences between the two measurements are attributed to remaining errors in OCO-2 after applying the averaging kernel correction as described in [129]. The calculated bias is then used to construct a binary training label by thresholding the difference and taking the values that fall below the threshold as 'good' and values that fall above the threshold as 'low-quality'. Choosing a higher threshold allows for increased data availability at the cost of increased bias, while a stricter threshold reduces bias as well as available data. These binary labels are then used for training a machine learning classifier model that we refer to as the TCCON sub-filter. The sub-filter model takes atmospheric state variables co-retrieved with the ACOS CO_2 column and outputs a binary estimate of 0 if a given sounding's remaining bias falls below the selected threshold and 1 if the remaining bias is above the selected threshold. To address the variability in available OCO-2 matchups among different TCCON stations we can optionally weight the discrete training labels y_{TCCON} available for a given station inversely by the total number of available matchups for that station N_{Station} as shown in Eq. 5.1.

$$y_{\text{TCCON}} = \frac{1}{N_{\text{Station}}} y_{(\text{TCCON},\text{Station})} \tag{5.1}$$

5.2.3.3 Small Area Approximation Sub-Filter Classifier

The second of the two truth proxies used for ML filter training is the small area approximation. This truth proxy is derived by calculating the difference of OCO-2 XCO2 within areas of 100 km within in the same orbit to the median XCO2 within the same area. Atmospheric CO2 at these scales can often be approximated as constant due to CO2 being a well-mixed gas. This assumption does not hold near large anthropogenic emission sources such as coal-fired power plants. The same approximation is used in the creation of the operational bias correction and quality filter [102]. We again generate binary labels by thresholding the bias similarly to the TCCON proxy. A binary classifier, which we will refer to as the small area filter, is then trained to estimate if a given sounding's bias falls above or below the threshold.

To balance bias reduction with data availability in regions with few soundings such as the tropics, we apply a tunable spatial weighting to the absolute ΔXCO_2 as shown in Eq. 2 prior to percentile binning.

$$\Delta XCO_{2,\text{Proxy Weighted}} = |\Delta XCO_{2,\text{Proxy}}| \left[1 + \gamma \left(\frac{N_{i,j} - N_{\min}}{N_{\max} - N_{\min}} \right) \right] \tag{2}$$

Soundings in the training set are binned into a latitude and longitude (i^{th}, j^{th}) raster. Given a sounding count N for the (i^{th}, j^{th}) bin, we normalize by the maximum sounding count $N_{\max}$ and the minimum sounding count $N_{\min}$ so that values for sounding counts are in the closed interval $[0, 1]$. γ is a real-valued tuning term that is set greater or equal to 0. When $\gamma = 0$, no weighting is applied and the quantity that is split upon during training is simply the unweighted remaining absolute bias. For larger values of γ, more emphasis is placed on balancing data availability in regions where soundings are sparse by penalizing bins that have higher sounding counts more steeply.

5.2.3.4 Sub-Filter Model Training and Validation

The OCO-2 data record does not have a uniform distribution over the globe and is limited in regions with large cloud presences such as the tropics. We subsample the OCO-2 catalogue to 5×10^6 soundings per year from 2015–2022. We aim to balance the sampled set spatially and give priority to OCO-2 soundings that are collocated with a TCCON station. Thus, for a given year approximately 10% of the soundings contain a collocated TCCON station. Next, we sample small areas by inversely weighting by the number of small areas contained in each $2° \times 2°$ latitude–longitude grid cell. This sampling strategy ensures good representation of geographical areas such as the tropics where OCO-2 observations are often sparse.

To develop the machine learning sub-filters and to evaluate the proposed ternary quality flag, we split the data into training and validation sets. To train the models, we use data from 2015–2021. To understand how the filters generalize to future observations as well as to the rest of the OCO-2 data record, we use a separate validation set comprising of data from 2022. For the validation set, we randomly select another 5×10^6 soundings and do not apply the balancing to accurately evaluate statistics based on the unbalanced OCO-2 distribution.

5.2.3.5 Classifier Feature Selection

To decide which features to include in the sub-filter models we start with nearly the full set of ACOS retrieved state variables available in the OCO-2 Lite files. However, we omit latitude, longitude and temporally dependent variables since to first order the spatiotemporal distribution of atmospheric CO2 can be predicted by these quantities. We train the initial pair of Random Forest filter models with this feature set and then rank the features by their SHapley Additive exPlanations (SHAP;[72]. SHAP allows for both local scale (single sounding) and global scale (full OCO-2 record) explanations of each feature's contribution to the machine learning sub-filters' binary prediction. The method has become the state of the art for machine learning explainability, especially for models that are not inherently interpretable.

We remove features with low contribution (we choose a cutoff of $< 1\%$), which simplifies the overall model architecture and makes the filters robust to data not used in the training process. This is done for both land and ocean filter models and for each truth proxy. Additionally, we rely on domain expertise to augment and approve the final list of features. We further discuss the SHAP feature importance in Section 4. Appendix B Table 5.5 shows all features considered and which filters they are utilized in.

5.2.3.6 Uncertainty-Based Sub-Filter

Extrapolation by a non-linear machine learning model on data outside of the distribution observed for training can potentially result in large errors that are challenging to quantify. This is particularly true in geospatial problems, where validation sources are often sparsely distributed around the globe, providing a limited set of observed atmospheric states that do not completely span the space of those observed by the satellite. For example, TCCON stations are mostly located in the northern hemisphere and over land. There may also be regions of the training data that are highly variable despite good representation in the training set. If the model provides a robust uncertainty estimate it can be beneficial to allow the model to abstain from prediction when that uncertainty is large. Given the calibrated uncertainty provided by the ML bias correction ($\sigma XCO_{2,MLBC}$) discussed in Part 1 (Fig 5.2a), we construct the following quantity in Eq. 3. We aim to isolate the systematic error component from random error by subtracting out the squared uncertainty from the uncorrected OE retrieval ($\sigma XCO_{2,OE}$) shown in Fig 5.2b–5.2c. This is then thresholded by selections t_{Land} and t_{Ocean} for land and ocean soundings respectively. The choice of t can either be hand-selected or optimized as described in the following sections. Smaller values of t reduce the variance of the remaining biases at the cost of throughput whereas larger values of t allow for more data availability.

$$\sqrt{\left(\sigma XCO_{2,MLBC}\right)^2 - \left(\sigma XCO_{2,OE}\right)^2} < t \tag{3}$$

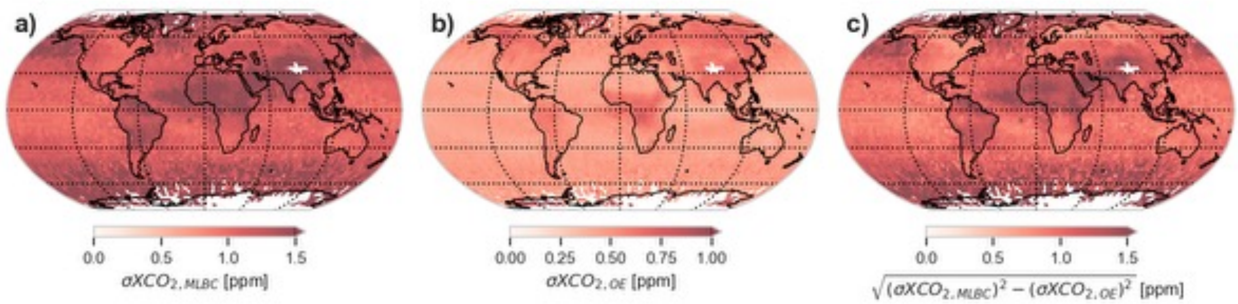

Figure 5.2: Global uncertainties from a) the ML bias correction ($\sigma XCO_{2,\mathrm{MLBC}}$), b) the OE XCO_2 uncertainty ($\sigma XCO_{2,\mathrm{OE}}$), and c) our derived quantity used to construct the uncertainty sub-filter shown in Eq. 3.

5.2.3.7 Deriving the Ternary QF from Sub-Filter Agreement

The logic for constructing the ternary flag is based on agreement between the three filters: the two ML data quality proxy filters and the uncertainty filter. For a sounding to be considered 'good' ($QF_{\mathrm{Ternary}} = 0$), all filters must be passed e.g., the bit string formed by the binary outputs of the three sub-filters should be $\{0, 0, 0\}$. The proposed $QF_{\mathrm{Ternary}} = 0$ quality flag value retains the same role as that of the current operational $QF_{\mathrm{Op}} = 0$ where the best quality data is identified for carbon flux inversion model, which is the primary science goal of OCO-2. For users requiring more data availability, $QF_{\mathrm{Ternary}} = 1$ is defined by soundings that pass one or two of the three filters e.g., the bit string should contain no more than two non-zero elements. $QF_{\mathrm{Ternary}} = 1$ data may be selected in addition to $QF_{\mathrm{Ternary}} = 0$ data to allow for additional availability of reasonably well-corrected data. $QF_{\mathrm{Ternary}} = 2$ indicates soundings that fail all three sub-filters, e.g., sub-filter bit strings with $\{1, 1, 1\}$.

5.2.3.8 Optimizing Sub-Filter Parameter Selection

In general, quality filtering represents a multi-objective optimization problem where a domain expert must balance minimizing error variance with maximizing data availability both temporally and spatially. Thus, evaluating the performance of a quality filter involves some degree of subjectivity in assessing the tradeoff between these objectives, as reducing RMSE reduces the sounding throughput of the filter and vice versa. Sub-filter parameters need to be selected that find an optimal tradeoff between these competing objectives. We employ a Bayesian Optimization (BO) method to find an optimal set of potential sub-filter parameter selections. Each possible parameter selection is a unique combination of uncertainty threshold, spatial weightings, and choice of percentile selection for training label split that lead to different tradeoffs between minimizing remaining errors and data availability. We also include the hyperparameters of the two Random Forest classifier sub-filter models, such as tree depth in the search. We show in the final selected parameters in Table 5.2.

We use a state-of-the-art BO method known as Multi-Objective Tree-structured Parzen Estimator or MOTPE [107]. Given a dataset of pairs $\{(x^*, y^*)\}$, MOTPE models $p(x^* \mid y^*)$ where x^* is a realization of the parameters and y^* is a vector of corresponding optimization function values. $p(x^* \mid y^*)$ is defined using the probability density functions $l(x^*)$ and $g(x^*)$ as shown in Eq. (4). $l(x^*)$ is the probability that the realization produces optimization function values less than a set of thresholds Y^* and $g(x^*)$ is the probability that the realization produces optimization function values greater than the threshold set. In BO, an acquisition function is maximized to direct the search for optimal candidate parameter realizations. MOTPE calculates the Expected Hyper Volume Improvement (EHVI) as defined in Eq. (5). Here,I_H is the hypervolume indicator defined by the objective functions and the difference in volume between $I_H(Y^*) \cup \{y^*\}$ and $I_H(y^*)$ are integrated over the range of thresholds. Using Bayes' Theorem, Eq. (6) can be reduced to Eq. (7), where α represents quantile threshold selections on the objective functions. Thus, by maximizing the ratio, the search is directed towards promising candidates as a larger $l(x^*)$ indicates an increased probability that x^* is a Pareto optimal solution in the Pareto set.

$$p(x^* \mid y^*) = \begin{cases} l(x^*) & y^* < Y^*, \\ g(x^*) & Y^* \leq y^*. \end{cases} \tag{4}$$

$$\underset{x^*}{\arg\max}\ \text{EHVI} = \int_{-\infty}^{Y^*} \Big(I_H(Y^*) \cup \{y^*\} - I_H(y^*)\Big)\, p(y^* \mid x^*)\, dy^* \tag{5}$$

$$= \int_{-\infty}^{Y^*} \Big(I_H(Y^*) \cup \{y^*\} - I_H(y^*) \Big) \frac{p(x^* \mid y^*)\, p(y^*)}{p(x^*)}\, dy^* \tag{6}$$

$$\propto \left(\alpha + (1-\alpha)\, \frac{g(x^*)}{l(x^*)} \right)^{-1} \tag{7}$$

We define three objective functions as filtering criteria for optimizing QFTernary=0. The first is to minimize the root mean square error (RMSE) (Eq. 8) between XCO2 from OCO-2 (XCO2_{OCO-2}) and collocated XCO2 derived from TCCON (XCO2_{TCCON}). The second is maximizing the global throughput of QFTernary=0 data as shown in Eq. 9. Lastly, we wish to reduce biases introduced by extrapolation of the machine learning correction by limiting the uncertainty of its estimates for QFTernary=0 data. We minimize the global average bias correction uncertainty as shown in Eq. 10. Selections that may optimally minimize/maximize one objective function may perform poorly for another objective function, e.g., filters that maximize throughput may have poor RMSE. MOTPE will sample the region loosely bounded by the individual optima which have high probability of being drawn from $l(x^*)$. These solutions approximate the Pareto optimal set which contain filter parameter selections that balance the trade-off between the three competing objectives.

$$\min(\text{RMSE}) \;=\; \min\left(\sqrt{\frac{1}{N_{QF=0}} \sum_{i=1}^{N_{QF=0}} \Big(XCO_{2,\text{OCO2}} - XCO_{2,\text{TCCON}} \Big)^2} \right) \tag{8}$$

$$\max(\text{Global Throughput}) \;=\; \max\!\left(100 \times \frac{N_{QF=0}}{N_{\text{Total}}} \right) \tag{9}$$

$$\min(\text{Bias Correction Uncertainty}) \;=\; \min\left(\frac{1}{N_{QF=0}} \sum_{i=1}^{N_{QF=0}} \sigma_i^{(\text{bias})} \right) \tag{10}$$

5.2.4 Results

5.2.4.1 Sub-Filter Parameter Selections

Together with the OCO-2 science team we performed several rounds of evaluation and feedback to improve and guide the QFTernary development. By integrating domain expertise, the Pareto set of potentially optimal filter solutions was further reduced by applying additional criteria to meet critical science goals. For example, over land the science team emphasized the need to not pass less data than the current operational flag on a global and seasonal level. This requirement is less strict for glint scenes over the ocean, where the community determined that further reducing biases was critical for potential ingestion of OCO-2 ocean XCO_2 estimates into flux inversion models. This led to the optimal sub-filter set shifting towards filter parameters that prioritize reduction in RMSE but that also avoid introducing large gaps into the seasonal spatial distribution which can cause spurious local biases. The hyperparameter selections for the ML filters are shown in Table 5.2, values for t_{Land} and t_{Ocean} were set to 1.23 and 1.1 respectfully. In addition to incorporating subject matter expertise for the Pareto optimization, we also incorporate scientists' feedback for feature selections into each ML truth proxy filter, rather than using a purely automated approach Appendix Table 5.5. Final feature selections and their importance's are discussed in depth below.

Table 5.2: Hyperparameter selections for the TCCON and small area approximation (SAA) ML sub-filters found through MOTPE optimization. RF hyperparameters include `max_depth` is the number of branch layers for each decision tree estimator and `N_estimators` is the number of ensemble members in the RF. Class label construction and weighting parameters include the station weighting bool value for the TCCON sub-filters, the spatial weighting Gamma for the SAA sub-filters and the percentile cut for binning training soundings as 'pass' or 'fail'.

Surface/Sub-filter		max_depth	Station weighting	Spatial weighting Gamma	Label cut (Percentile)
Land	TCCON RF	12	True	N/A	'pass' $<$ 55
	SAA RF	7	N/A	0.03	'pass' $<$ 55
Ocean	TCCON RF	8	True	N/A	'pass' $<$ 45
	SAA RF	12	N/A	0.00	'pass' $<$ 50

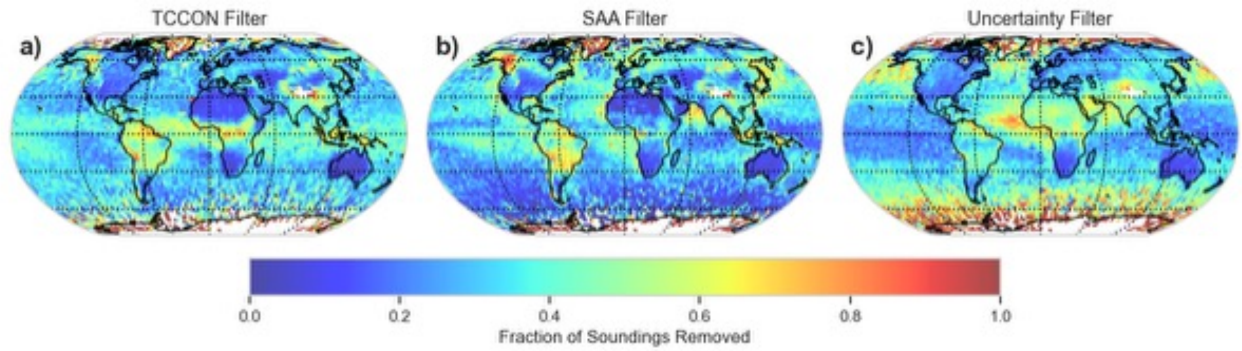

Figure 5.3: Fraction of retrievals removed relative to the number to the total number of retrievals that converge and pass pre-processors for each sub-filter given a 2ox2o latitude-longitude binning. a.) TCCON ML sub-filter, b.) small area approximation (SAA) ML sub-filter, c.) uncertainty sub-filter.

The two truth proxy ML sub-filters and the uncertainty sub-filter exhibit some degree of disagreement with each other regionally. The TCCON sub-filter (Fig 5.3a) exhibits stricter throughput over southern hemisphere ocean, and particularly over tropical Africa, compared to the other two filters. These are regions heavily affected by the presence of clouds year-round. The small area approximation sub-filter (Fig 5.3b) is more restrictive in the regions where the TCCON filters are more lenient but less restrictive in general over southern hemisphere ocean and around coastal crossings. The uncertainty sub-filter (Fig 5.3c) removes data at the extreme northern and southern latitudes and is stricter over northern and central Africa as well as over the Atlantic Ocean than the other two filters. All three sub-filters agree in regions with higher surface elevations, towards latitude extremes, and over the Atlantic coast of Africa where there is a significant presence of aerosols.

5.2.4.2 Comparison to the Operational QF

The new ternary quality flag introduces three flag values: QFTernary $= 0, 1, 2$. QFTernary $=$ 0 serves a similar function as the current operational QFOp $= 0$, which is to provide the best quality data for carbon flux inversion models. The QFTernary $= 0$ filter is now more restrictive on ocean glint data (lower throughput) in favor of reducing biases, as recent versions of OCO-2 ocean glint observations have been shown to exhibit latitudinally dependent biases that can be erroneously propagated onto land by flux inversions (Crowell et al., 2019). The QFTernary = 0 flag value for land produces global data availability similar to that of the current OCO-2 v11 operational filter (QFOp), but further reduces remaining biases. Table 5.3 shows the RMSE and data availability for retrievals over land and ocean as compared to TCCON and the Model Median for both training and holdout years. For land, QFTernary = 0 has a similar throughput relative to QFOp = 0 (54% vs 57%, respectively), with comparable RMSEs against the training and the 2022 holdout data against both the TCCON and the Model Median. For ocean, QFTernary = 0 passes $\sim$ 20% less of the total data volume than QFOp = 0 (43% vs 64%, respectively) and improves the RMSE across all truth proxy metrics for both training set and hold year. To separate the effects of the proposed quality filter in this work to the ML bias correction proposed in Part 1, we also compare to ML bias correction with the operational filter applied (QFOp = 0 + MLBC). Overall, QFOp = 0+MLBC is competitive with the operational bias correction and filter for TCCON land and ocean, and is in closer agreement to the ocean Model Median which is not used for training the ML bias correction but is used in the development of the operational bias correction.

Table 5.3: RMSE scores for TCCON and flux model median evaluation datasets and percentage of data passed for record.

Proxy	TCCON [ppm]		Model Median [ppm]		Percent QF=0 2015-2022
	2022	2015 to 2021	2022	2015 to 2021	
Land					
QFTernary =0	0.90	0.87	1.22	1.11	54%
QFOp =0	1.03	1.02	1.24	1.15	57%
QFOp =0+MLBC	0.96	0.95	1.23	1.16	57%
Ocean					
QFTernary =0	0.66	0.62	0.71	0.69	43%
QFOp =0	0.80	0.83	0.85	0.81	64%
QFOp =0+MLBC	0.84	0.70	0.78	0.78	64%

Figure 5.4b and 5.4d shows the difference between the bias corrections for the $QF_{\mathrm{Ternary}} = 0$ and $QF_{\mathrm{Op}} = 0$ (Fig 5d shows the difference remaining absolute bias between Fig 5.4a and

Fig 5.4e). The largest differences between filters and bias corrections are seen over the Amazon, tropical Africa, as well as over the Atlantic Ocean. It is possible that the ML derived bias correction and filtering methodology are doing an improved job of correcting and/or screening aerosol contaminated soundings, which are known to induce biases in the retrieved XCO_2 [5].

Figures 5.4f shows a spatial map of the differences in availability between $QF_{\text{Ternary}} = 0$ and $QF_{\text{Op}} = 0$. Over the ocean QF_{Ternary} passes less data overall, particularly poleward of about 45 degrees latitude, as well as over the Atlantic Ocean due to wind-blown dust and biomass burning aerosols originating from northern Africa [92]. Over land, the QF_{Ternary} filter passes approximately the same amount of data as $QF_{\text{Op}} = 0$, but with some significant spatial differences. The $QF_{\text{Ternary}} = 0$ passes slightly less data over parts of the Continental United States (CONUS), southeast Asia and parts of central Africa and northern South America. The most significant difference in throughput for QF_{Ternary} compared to QF_{Op} is over the Sahara. The OCO-2 v11 operational filter uses an albedo term that is responsible for the flagging of a large fraction of this data. For tropical land in Africa and parts of South America, the $QF_{\text{Ternary}} = 0$ passes soundings that are removed by $QF_{\text{Op}} = 0$, likely due to the QF_{Op} filtering scheme using an abledo filter that may be too aggressive over dark forests. The passing of additional soundings in these regions that are currently under sampled and important for studying the terrestrial carbon cycle might be of great value.

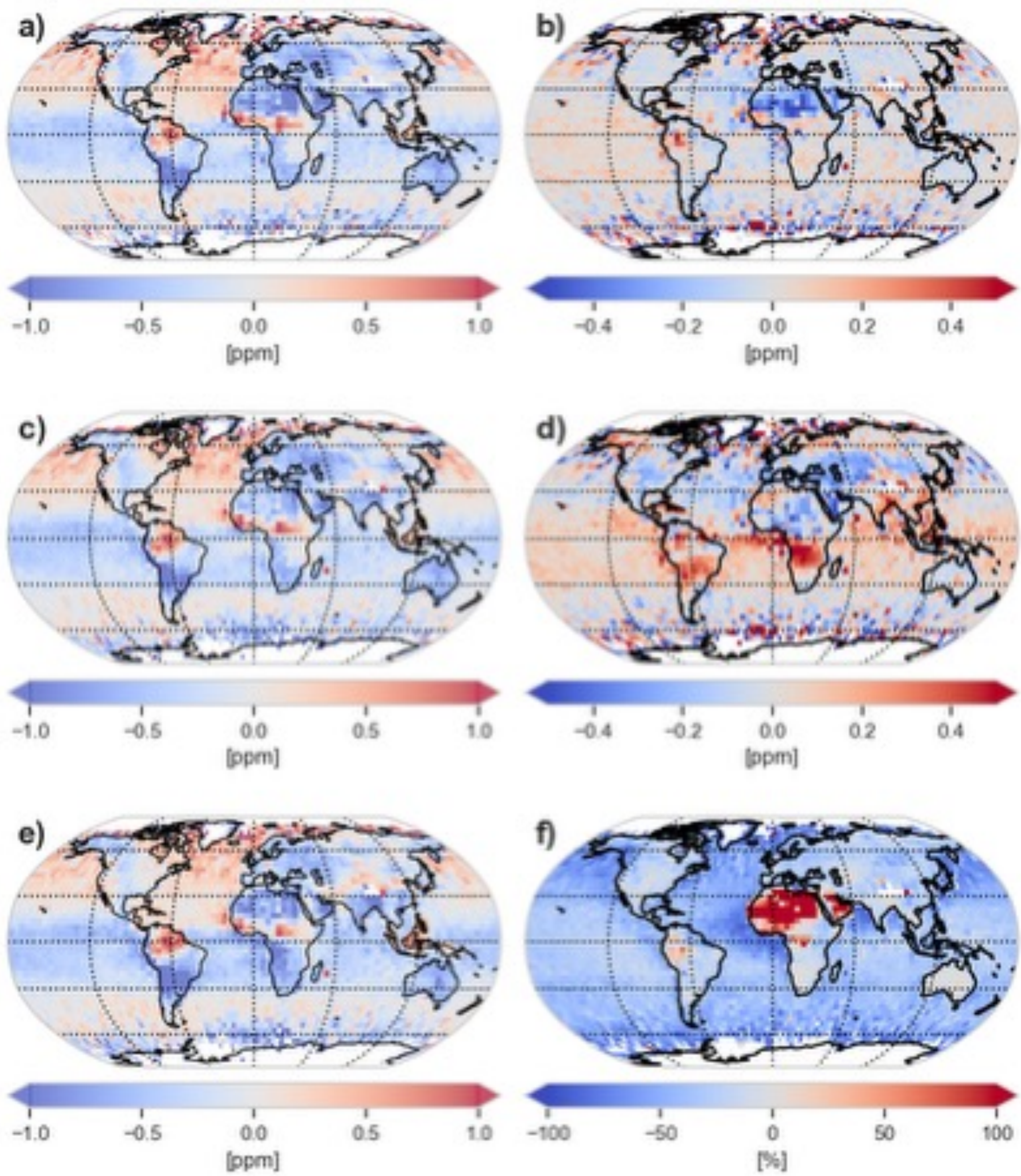

Figure 5.4: Comparison of data availability and remaining biases $\left(\Delta XCO_2 = XCO_{2,\text{OCO-2}} - XCO_{2,\text{Model Median}}\right)$ for QFTernary = 0 and QFOp = 0. a) Global ΔXCO_2 using the ML bias correction and QFTernary = 0. b) Absolute magnitude of differences using the ML bias correction between QFOp = 0 and QFTernary = 0. This is calculated as $|\Delta XCO_{2,QF_{\text{Op}}=0}| - |\Delta XCO_{2,QF_{\text{Ternary}}=0}|$. c) Global ΔXCO_2 using the ML bias correction and QFOp = 0. d) Absolute magnitude of differences using operational bias correction for QFOp = 0 and the ML bias correction for QFTernary = 0. This is calculated as $|\Delta XCO_{2,QF_{\text{Op}}=0}| - |\Delta XCO_{2,QF_{\text{Ternary}}=0}|$. e) Global ΔXCO_2 using the operational linear bias correction and QFTernary = 0. f) Percent difference in data availability between QFTernary = 0 and QFOp = 0. Red indicates regions where QFTernary = 0 is passing more data than the operational flag.

5.2.4.3 Increasing Data Availability

In previous ACOS algorithm versions, $QF_{\mathrm{Op}} = 1$ indicated low-quality data to be excluded from carbon flux inversion models. For analysis where higher data availability is required, such as quantification of power plant emissions [94], the QF_{Op} is often not used to quality filter data. Soundings that pass only one of the sub-filters are designated as $QF_{\mathrm{Ternary}} = 1$. Users who require more data in their analysis can select $QF_{\mathrm{Ternary}} = 1$ in addition to $QF_{\mathrm{Ternary}} = 0$, denoted $QF_{\mathrm{Ternary}} = 0 + 1$ (Fig 5.5b). $QF_{\mathrm{Ternary}} = 0 + 1$ is comparable to the $QF_{\mathrm{Op}} = 0$ with operational bias correction as seen in Table 5.4. For $QF_{\mathrm{Ternary}} = 0 + 1$ land there is an approximately 0.10–0.17 ppm increase in RMSE observed across evaluation sets in exchange for a $\sim$ 30% increase in sounding throughput relative to $QF_{\mathrm{Ternary}} = 0$ (Table 5.4). For observations over the ocean, $QF_{\mathrm{Ternary}} = 0 + 1$ shows a 0.10 ppm increase RMSE for TCCON ocean and a 0.20 ppm increase for the held-out year of 2022 as compared to $QF_{\mathrm{Ternary}} = 0$. $QF_{\mathrm{Ternary}} = 0 + 1$ exhibits slightly worse RMSE scores for the Model Median particularly for 2015–2021.

Table 5.4: RMSE scores and percentage in data availability over the catalogue for the proposed $QF_{\mathrm{Ternary}} = 0 + 1$.

Proxy	TCCON [ppm]		Model Median [ppm]		Percent $QF = 0 + 1$ 2015-2022
	2022	2015 to 2021	2022	2015 to 2021	
Land	1.07	1.03	1.33	1.27	85%
Ocean	0.93	0.74	0.94	0.92	88%

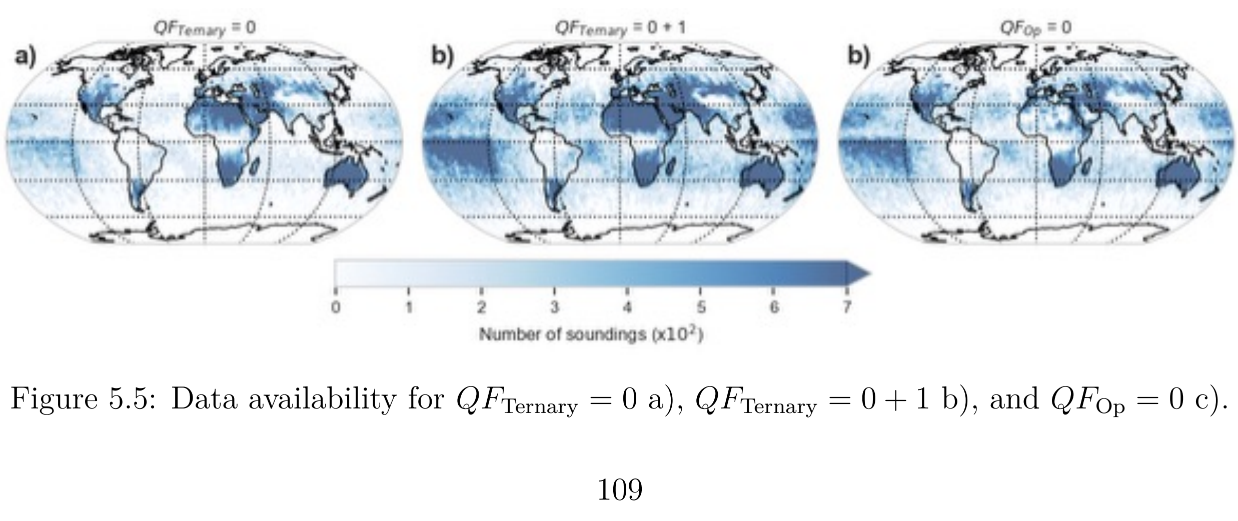

Figure 5.5: Data availability for $QF_{\mathrm{Ternary}} = 0$ a), $QF_{\mathrm{Ternary}} = 0 + 1$ b), and $QF_{\mathrm{Op}} = 0$ c).

5.2.4.4 Removing Cloud Effected Soundings

3D cloud effects have been shown to cause significant biases in retrieved XCO_2 [79]; [80]; [83]. Newly derived variables that capture such effects have been incorporated into the operational QF as well as our proposed approach. In Part 1, we show how these variables are useful for correcting 3D cloud biases. However, irreducible errors from clouds remain. To evaluate the role of the combined bias correction and proposed ternary QF in reducing cloud biases, we again use calculated distances between OCO-2 soundings to the nearest cloud using collocated Moderate Resolution Imaging Spectroradiometer (MODIS) observations. The collocated MODIS observations are taken approximately six minutes after OCO-2 in NASA's A-Train constellation [73]. This cloud distance is then used to evaluate the ability of the proposed approach to reduce potential cloud-induced biases. If for a given retrieval the associated cloud distance is relatively small then we can assume 3D cloud effects are potentially present. Ideally the quality filter will remove soundings that are effected by such irreducible biases especially in regions affected by high cloud cover such as the ocean and tropics as seen in Figure 5.6c.

We use the computed cloud distances for OCO-2 soundings from September 2014 to July 2019 to assess the ability of the combined bias correction and QFTernary to potentially reduce and remove cloud biases. A larger mean cloud distance is observed for QFTernary = 0 data over land (10.43 km $\rightarrow$ 12.62 km) and ocean (7.30 km $\rightarrow$ 9.15 km) scenes indicating that soundings with clouds in or near the scene are being filtered out by the two ML truth proxy filters and the uncertainty sub-filter particularly for tropical land and middle and northern latitude ocean. Soundings over the tropical ocean with QFTernary = 0 (Fig 5.6a) still show areas of potentially cloud-affected soundings. QFOp = 0, however, shows a reduced ability to remove cloud-affected soundings over the ocean and tropical land as compared to the non-linear ML approach (Fig 5.6b).

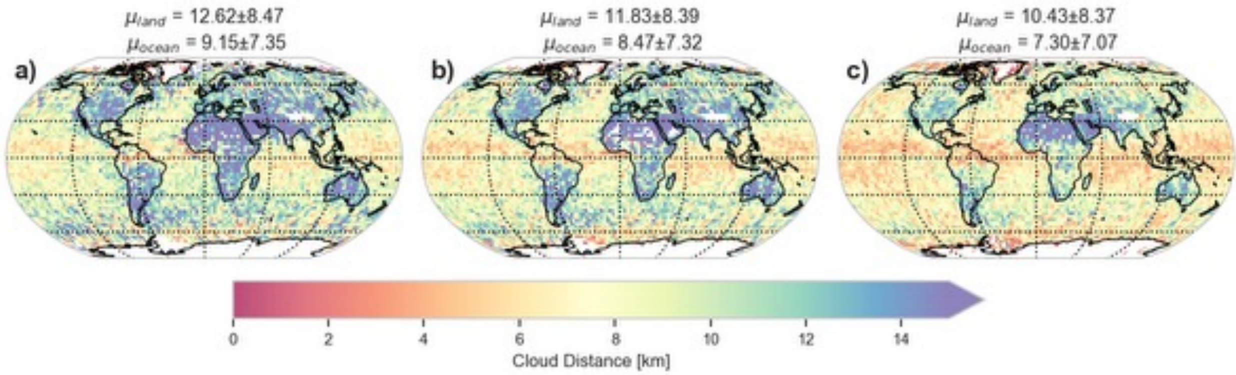

Figure 5.6: Cloud distance calculated from collocated OCO-2 and MODIS for a 2° × 2° latitude–longitude binning: a) remaining cloud distance for $QF_{\text{Ternary}} = 0$, b) remaining cloud distance for $QF_{\text{Op}} = 0$, c) no filtering applied. Mean plus/minus 1-sigma of cloud distance for land and ocean is also shown for each surface type above each subplot.

5.2.4.5 Interpretation of Sub-Filter Classifiers

To better understand the predictions made by the ML sub-filter models, we utilize SHAP feature importance. SHAP values allow for post-hoc insight into the models' decision-making process through ranking of the features or covariates that have the most impact on the estimate. We look at the ranked feature importance at a global scale (absolute ranking) for the top six most informative features for each of the four truth proxy/surface type models (Fig 5.7).

For land, we find that the most important features are similar between both models. The uncertainty value from the bias correction (bias_correction_uncert) is the most informative covariate for both TCCON and small area approximation models. The h_continuum_wco2 and co2_ratio_bc appear in the top features for both land models. The h_continuum_wco2 is the variability of the continuum intensity in the weak band across nearby scenes and can be indicative of 3D cloud effects [79]. Difference from unity in the co2_ratio_bc which is the ratio between CO2 retrieved from the weak and strong bands from the IMAP-DOAS retrieval used as part of the OCO-2 prescreening [128] can indicate the presence of aerosols. The TCCON land sub-filter also shows high importance for the variance in surface elevation across the scene in meters (altitude_stddev) and the SAA sub-filter model favors

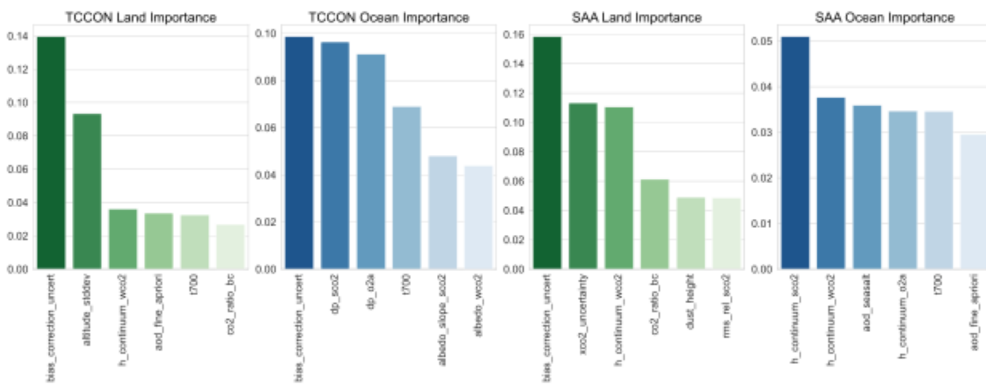

Figure 5.7: SHAP feature importance plots for each of the truth proxy-based Random Forest filters. Each plot shows the top 6 most informative features.

a second uncertainty term which captures the gaussian posterior of the retrieved XCO2 (xco2_uncertainty).

For ocean, we see less agreement between the feature sets. The most important features for the TCCON ocean filter are the bias correction uncertainty (bias_correction_uncert), and two surface pressure difference terms: dp_o2a and dp_sco2 which represent the difference between retrieved surface pressure and the prior surface pressure for the A-band and strong CO2 band, respectively. The ACOS L2FP algorithm requires a prior surface pressure (taken from GEOS5-IT for v11) as it retrieves surface pressure for each spectral band due to slight misalignment in the spectrometers. A temperature estimate at 700 hPa (t700) is the only shared feature between both ocean models. The SAA ocean sub-filter's top features include all three of the h_continuum_band variables indicating that the Small Area Approximation truth proxy may capture information about cloud effects over water scenes.

5.2.5 Conclusions

In this work, we present a novel quality filtering approach for OCO-2 XCO_2 that combines domain scientist input with trustworthy machine learning methods, such as uncertainty quantification and explainability and builds upon improvements brought by a ML bias correction presented in Part 1. We employ a novel spatial weighting method to guide the training of the machine learning models that can be tuned to adjust the tradeoff between spatial availability of data and reducing XCO_2 biases in OCO-2 retrievals. Filter development is done with a multi-objective optimization that allows for fast iteration and feedback with the domain scientists. Our combined machine learning bias correction and sub-filter agreement quality flag exhibit the potential to flag and/or correct soundings that are likely contaminated by aerosols and show a 15% improvement in RMSE for land scenes and an improvement of 25% for ocean when comparing to the operational product. Furthermore, the ability to relax agreement between the sub-filters allows for greater flexibility when selecting soundings by users, thus enabling support for multiple downstream use cases such as plume analysis or regional flux inversions that require more data. For land $QF_{Ternary} = 0 + 1$ provides a 28% increase in data availability over land and a 24% increase over ocean scenes. The quality filtering approach presented in this manuscript is generalizable to current and future missions that make maximum a posteriori estimates of atmospheric column states and have access to ground truth proxies. Future work includes evaluation of the combined bias corrected XCO_2 and both $QF_{Ternary} = 0$ and $QF_{Ternary} = 0 + 1$ data in carbon flux inversion models and development of sub-filters specifically for QF=1 data that further exploit spatial and temporal weightings.

5.2.6 Appendix

Table 5.5: Table A1. Features considered for truth-proxy ML filters. "Used In" indicates which filter the feature was employed in. *TC_Lnd* = TCCON land filter, *TC_Ocn* = TCCON ocean filter, *SA_Lnd* = small-area land filter, *SA_Ocn* = small-area ocean filter.

State Feature	**Description**	**Used In**
albedo_sco2	Retrieved reflectance of the strong CO_2 band.	TC_Lnd, TC_Ocn, SA_Lnd
albedo_wco2	Retrieved reflectance of the weak CO_2 band.	TC_Lnd, TC_Ocn, SA_Lnd
altitude_stddev	Standard deviation of the surface elevation within the scene.	TC_Lnd, SA_Lnd
aod_diff	Difference between retrieved total AOD and its a-priori value (*aod_total – aod_total_apriori*).	SA_Lnd, SA_Ocn

Continued on next page

Table A1 – continued from previous page

State Feature	**Description**	**Used In**
aod_dust	Retrieved dust aerosol optical depth at 0.755 µm.	TC_Lnd, TC_Ocn, SA_Lnd
aod_sulfate	Retrieved sulfate aerosol optical depth at 0.755 µm.	TC_Lnd, TC_Ocn
aod_total	Retrieved total cloud + aerosol optical depth at 0.755 µm.	TC_Lnd, TC_Ocn, SA_Ocn
aod_water	Retrieved water-cloud optical depth at 0.755 µm.	TC_Lnd, TC_Ocn, SA_Lnd
bias_correction_uncert	Uncertainty estimate from the Part 1 bias correction (future Lite-file field).	TC_Lnd, TC_Ocn, SA_Lnd, SA_Ocn
co2_ratio_bc	Footprint-corrected ratio between strong and weak band CO_2.	TC_Lnd, TC_Ocn, SA_Lnd, SA_Ocn
color_slice_noise_ratio_o2a	Sub-scene continuum variability (O_2 A-band).	SA_Lnd, SA_Ocn
color_slice_noise_ratio_sco2	Sub-scene continuum variability (strong CO_2).	SA_Lnd, SA_Ocn
color_slice_noise_ratio_wco2	Sub-scene continuum variability (weak CO_2).	SA_Ocn
dp_abp	O_2 A-band pre-processor surface-pressure minus a-priori value.	TC_Lnd, TC_Ocn, SA_Ocn
dp_o2a	Retrieved–prior surface-pressure difference (O_2 A-band).	TC_Lnd, TC_Ocn, SA_Lnd
dp_sco2	Retrieved–prior surface-pressure difference (strong CO_2 band).	TC_Lnd, TC_Ocn
dpfrac	Elevation-adjusted surface-pressure difference (see Kiel et al., 2019).	TC_Lnd, SA_Lnd
dust_height	Pressure of dust aerosol layer relative to surface pressure.	TC_Lnd, TC_Ocn, SA_Lnd
dws	$aod\_dust + aod\_water + aod\_seasalt$.	TC_Lnd, TC_Ocn, SA_Ocn
glint_angle	Angle between local sensor direction and specular-glint direction.	TC_Lnd, TC_Ocn
h2o_ratio_bc	Footprint-corrected ratio between strong and weak band H_2O retrievals.	TC_Lnd, TC_Ocn, SA_Lnd
h2o_scale	Retrieved water-vapor scale factor.	TC_Lnd, TC_Ocn
h_continuum_o2a	Frame-normalized continuum SD (O_2 A-band).	SA_Ocn
h_continuum_sco2	Frame-normalized continuum SD (strong CO_2).	SA_Ocn
h_continuum_wco2	Frame-normalized continuum SD (weak CO_2).	SA_Ocn

Continued on next page

Table A1 – continued from previous page

State Feature	**Description**	**Used In**
ice_height	Central pressure of ice-cloud layer relative to surface pressure.	TC_Lnd, TC_Ocn, SA_Lnd
max_declocking_o2a	Sub-scene radiance variability (O_2 A-band).	SA_Ocn
max_declocking_sco2	Sub-scene radiance variability (strong CO_2 band).	SA_Ocn
max_declocking_wco2	Sub-scene radiance variability (weak CO_2 band).	SA_Ocn
rms_rel_sco2	RMS of fit residuals (strong CO_2 band).	TC_Lnd, TC_Ocn, SA_Ocn
rms_rel_wco2	RMS of fit residuals (weak CO_2 band).	SA_Ocn
snr_sco2	Signal-to-noise ratio (strong CO_2 band).	TC_Lnd, TC_Ocn, SA_Lnd, SA_Ocn
t700	Estimated temperature at 700 hPa from GEOS-5 FP-IT (linear interpolation).	TC_Lnd, TC_Ocn, SA_Lnd, SA_Ocn
tcwv_uncertainty	Uncertainty of retrieved total-column water vapor.	TC_Lnd, TC_Ocn, SA_Lnd
water_height	Pressure of water-cloud layer relative to surface pressure.	TC_Lnd, TC_Ocn, SA_Lnd, SA_Ocn
xco2_uncertainty	Uncertainty estimate of retrieved column-averaged CO_2.	TC_Lnd, TC_Ocn, SA_Lnd, SA_Ocn
zlo_o2a	Zero-level offset proxy for ice build-up on detector (O_2 A-band).	TC_Lnd, TC_Ocn

The spatial and TCCON station weightings described in Section 3.1.6, which are used in the creation of the training labels, can lead to potential trade-offs between regional data throughput and bias reduction. We explore this trade-off in an ablation study on a hold-out dataset and several filter realizations in which parameters are fixed except for the spatial weighting tuning parameter γ and the TCCON station weighting. In Figure 5.8 we show the results of the data availability and mean absolute error (MAE) for several γ selections, no TCCON station weighting in (a–b), and the same selections with TCCON station weighting applied (c–d). For higher γ there is a notable increase in throughput of soundings in the northern latitudes and tropics that increases further when station weighting is also applied. However, higher weightings also introduce larger biases, particularly in the middle to northern latitudes. We note that a similar weighting strategy could also be applied temporally in

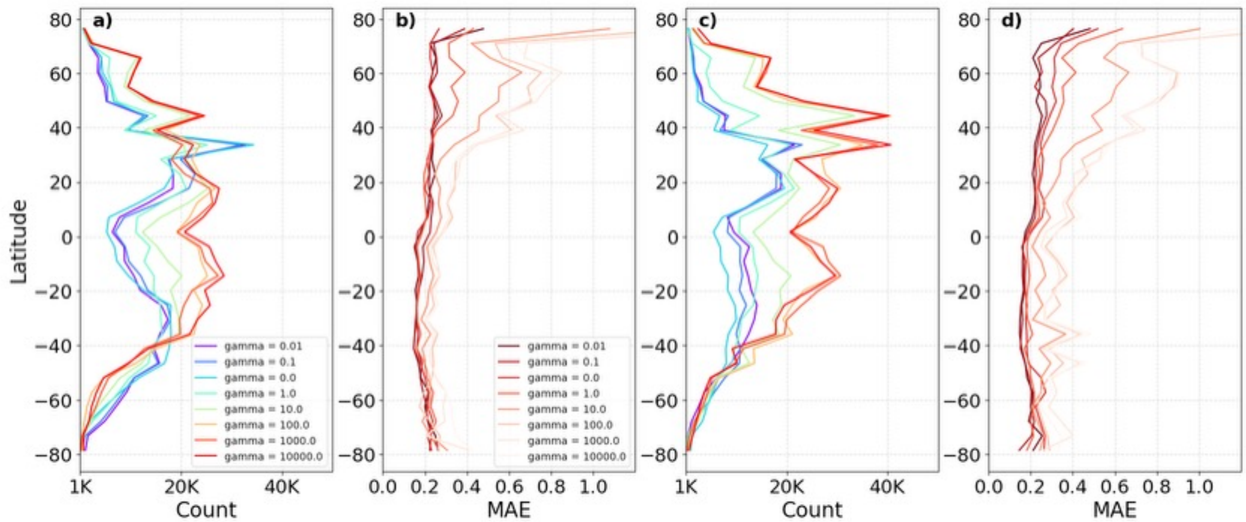

Figure 5.8: Ablation results for spatial weighting and station weighting strategies. a-c.) the binned sounding counts for each of the hyperparameter combinations (gamma + station_weighting_bool). b-d.) the mean absolute error (MAE) for each hyperparameter combination.

addition to spatial weighting to increase the availability across seasons.

Chapter 6

Direct Retrieval of Trace Gases from Earth Observing Spectroscopy

6.1 Chapter Foreword

Summary for the data science practitioner

This chapter replaces slow, per scene physics inversions with amortized inference from a diffusion generative model. The conditional diffusion model generates many posterior samples quickly and can flexibly capture Gaussian or non-Gaussian uncertainty that helps diagnose when and why the forward model disagrees with observations. Incorporating physics via priors and spectral perturbations stabilizes learning and improves generalization. Finetuning on co-locations with trusted ground truth improves both RMSE and calibration of the posterior uncertainty. We also present a novel band-aware evidential transformer provides fast XCH_4 estimates with decomposed, calibrated uncertainty. Training data are created by injecting methane signatures into background spectra. The attention weights also provide a per pixel interpretability.

6.2 Conditional Diffusion-Based Retrieval of Atmospheric CO2 from Earth Observing Spectroscopy

6.2.1 Introduction (Paper)

Satellite observations of greenhouse gases (GHGs) are essential for monitoring and quantifying the impact of anthropogenic activities (e.g., fossil fuel combustion, deforestation, urbanization) on the climate of the planet particularly in rapidly changing regions like the high latitudes and Tropics which are difficult to measure *in situ*. Satellite-bourne spectrometers such as NASA's Orbiting Carbon Observatory-2 and 3 (OCO-2/3; [24]), and JAXA's Greenhouse Gases Observing Satellite (GOSAT; [150]) make high spectral resolution measurements of reflected solar spectra in order to infer changes in GHGs. A physics-based algorithm is used to estimate or *retrieve* the atmospheric state including the column averaged dry air-mole fraction of atmospheric carbon dioxide (XCO_2) from these spectra. XCO_2 observations are used to constrain surface fluxes of GHGs, which are further used to study the Earth's complex carbon cycle. [26].

Estimating GHG concentrations from spectral observations is typically posed as an inverse problem, where we aim to recover the state of the atmosphere, denoted $\mathbf{x}$, from observed spectra $\mathbf{y}$ with instrument viewing geometries $\mathbf{b}$. The relationship between state and observation can be written as: $\mathbf{y} = F(\mathbf{x}, \mathbf{b}) + \nu$, where F is the forward operator, and noise $\nu \sim \mathcal{N}(0, \mathbf{R})$ which is scene and wavelength-dependent. OCO-2 uses an Optimal Estimation (OE) based algorithm called Atmospheric Carbon Observations from Space (ACOS) detailed in [20]; [104]. OE methods solve the inverse problem using Bayes Rule: $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, where estimates of $\mathbf{x}$ are obtained as a *maximum a posteriori* (MAP) point estimate which maximizes the posterior probability $p(\mathbf{x}|\mathbf{y})$ [118] with an associated Gaussian approximation to the posterior (co-)variance as an uncertainty estimate. The ACOS retrieval employs a Radiative Transfer Model forward model (FM) coupled with an instrument model for the OCO-2 spectrometer as F to produce simulated spectra, $\mathbf{y_{Sim}}$, and partial derivatives for a given $\mathbf{x}$. The algorithm iteratively seeks the MAP estimate that minimizes a cost function that represents the sum of squared errors between the observations $\mathbf{y_{Obs}}$ and $\mathbf{y_{Sim}}$.

[44], [65] and [9] illustrate how realizations of the true conditional posterior for XCO_2 can be non-Gaussian or bimodal due to the nonlinear nature of the forward model, arising from varying aerosol and surface conditions. The classical formulation does not apply in these cases. The Gaussian assumed uncertainty estimate produced by ACOS is often over confi-

dent, e.g., the prediction interval often does not contain the true value of XCO_2. Applications of these data to constrain CO_2 emissions can be sensitive to the assumed uncertainties and often incorporate ad hoc adjustments to the uncertainties prior to their use. This introduces additional uncertainties in the CO_2 emissions estimates themselves [111]. The uncertainty in predicted emissions is critical for inference at policy relevant (e.g., national) scales necessary for treaties and enforcement [13].

OE requires repeated evaluations of the computationally expensive forward FM per observation. OCO-2 currently makes 10^5 cloud-free observations per day and has a record that spans from 2014 to present. The decade-long catalog has undergone several updates, with reprocessing taking around a year to complete with the current ACOS version and modern HPC architectures. This computational bottleneck will be prohibitive for future missions such as ESA's upcoming Carbon Dioxide Monitoring Mission [124], which will generate tens of millions of observations daily.

It is critical to develop new retrieval techniques that are fast and accurate while also providing robust uncertainties. In this work, we apply a diffusion-based approach for accurate atmospheric CO_2 retrieval that avoids the need for repeated FM evaluations after training is complete, while also providing a robust uncertainty for NASA's OCO-2 instrument. Diffusion is a generative machine learning method that allows for efficient sampling of potentially complex posteriors and has seen application in solving inverse problems in the natural sciences. Diffusion's success comes from the ability to define a prior distribution which is typically a standard Gaussian $\mathcal{N}(0, I)$ and then training a *denoising* model that iteratively learns to shift samples drawn from the prior to the true data distribution using a Markov process. This allows for efficient sampling to recover either a Gaussian uncertainty by taking the first two statistical moments of the draws or to explore non-Gaussianity in the retrieved posterior.

6.2.2 Methods

To train the retrieval we use the L2 product [109] of version 11 ACOS FM simulated spectra denoted as $\mathbf{y_{Sim}}$. OCO-2's wavelength ranges are the near-infrared Oxygen A band near 0.76 µm, the shortwave infrared "weak" CO_2 band near 1.6 µm, and the shortwave infrared 'strong' CO_2 band near 2.05 µm. Additional retrieval covariates include the solar zenith angle (SZA) and a weather model predicted surface pressure term P_{Surf} that comes from analysis produced by the GEOS5-FP-IT data assimilation system [71]. The concatenated covariate vector for training point i is denoted as $\mathbf{\bar{y}_i} = \{\mathbf{y_i}, SZA_i, P_{Surf_i}\}$. The XCO_2 scalars from the L2 state vector $\mathbf{x}$, in parts per million (ppm), serve as the training label x_i. In total, we construct a training dataset of 3×10^6 $\{\mathbf{\bar{y}_i}, x_i\}$ pairs. Dependence on the FM for generating training data can introduce errors into the retrieval due to differences between simulated training radiance and real observations seen at inference time: $(\mathbf{y_{Sim}} - \mathbf{y_{Obs}})$. We account for this difference by careful perturbation of the simulated radiance prior to model training, as shown in Appendix 4.2.6.3 and resample both $\mathbf{y_{Sim}}$ and $\mathbf{y_{Obs}}$ to a common wavelength grid. We use the FM for training despite the abundance of OCO-2 observations with an eye to current and future missions that may lack sufficient temporal or spatial coverage of observations for an ML-based retrieval. These missions cannot adopt ML algorithms trained directly on observations to generalize well across the globe, thus requiring learning on simulated training data.

Given realizations of the covariates, we aim to retrieve the posterior distribution $p(x|\mathbf{\bar{y}})$, using the conditional diffusion method described in [40]. This paper suggests the use of a conditional mean prior, x_{prior}, which is a pre-trained neural network f_ϕ with weights ϕ to parameterize the forward diffusion distribution as $\mathcal{N}(x_{prior}, 1)$ instead of the more standard choice of $\mathcal{N}(0, 1)$. This introduces conditioning information into the forward process and potentially speeds up inference by reducing the number of diffusion steps required. We explore different methods to specify x_{prior} including a neural network retrieval of XCO_2 as well as XCO_2 values from GEOS-FP-IT in Appendix 4.2.6.3. The denoising model,

$\epsilon_\theta(\mathbf{\bar{y}_i}, x_{prior_i}, x_{i,t}, t)$ is then trained as described in [42] using the mean squared error loss, where x_t is the noised label at diffusion step t.

The OCO-2 operational XCO_2 product includes an extensive post hoc bias correction to reduce systematic errors or *biases* in the raw OE retrieved XCO_2 [58, 102]. This process involves fitting a regression model on errors in XCO_2 with predictors selected from the retrieved atmospheric state. XCO_2 errors are calculated using ground truth labels of XCO_2 from the Total Carbon Column Observing Network (TCCON; [146]) that are collocated with OCO-2 measurements in space and time. In this spirit, we evaluate the potential of reducing bias in the trained diffusion model using a similar dataset of collocated TCCON XCO_2 labels to finetune the retrieved value. We collocate OCO-2 and TCCON samples within 100km spatially and 1 hour temporally, for which we also have FM produced radiance for years 2015 to 2021. This results in a finetuning dataset of 25×10^4 pairs. To finetune the retrieval, we first unfreeze the last layers of f_ϕ and train an additional 10 epochs. We then continue training ϵ_θ using the finetuned conditional mean backbone until the loss for a small validation set stabilizes.

To evaluate the performance of the proposed diffusion retrieval relative to ACOS, we use a holdout set of data of OCO-2 observations collocated with TCCON sites and retrieve on the OCO-2 measured reflected solar radiance ($\mathbf{y_{Obs}}$) for the year of 2022. This test dataset contains 20,000 collocated pairs, and contains examples from all operational TCCON sites for the year. We assess the quality of the point estimates of XCO_2 for ACOS and the diffusion retrieval with and without finetuning/bias-correction, as well as the quality of their uncertainty estimates.

6.2.3 Results

6.2.3.1 Efficient Posterior Sampling

The reverse diffusion process allows us to generate samples from $p(x|\mathbf{\bar{y}})$ by first drawing each $x_T \sim \mathcal{N}(x_{prior}, 1)$ and then iteratively denoising with ϵ_θ from $t = T$ to $t = 0$ to recover the

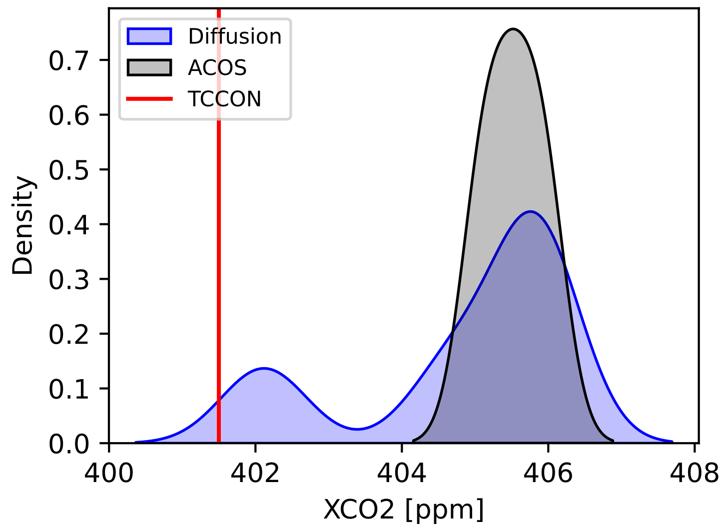

Figure 6.1: Example of bi-modal distribution recovered through 10 conditional samples from the Diffusion posterior (blue) without finetuning compared to the operational OE posterior which is derived from the point estimate ± an estimate of the variance (black) without bias correction. The Diffusion posterior covers the ground truth value from TCCON (red) within the first mode which is not captured by the Gaussian operational posterior.

denoised XCO_2 samples, thus enabling efficient density estimation of potentially complex posteriors as shown in Figure 6.1. The diffusion retrieval takes on average 0.0003 seconds to draw 10 conditional samples for a single observation on a 2017 Apple MacBook Pro. [64] showed that FM evaluations on similar hardware take approximately 33-55 seconds.

6.2.3.2 Comparison to the Operational OCO-2 Retrieval

In Figure 6.2, we evaluate the XCO_2 point estimate and uncertainty for both the operational ACOS retrieval and our proposed diffusion-based retrieval on a holdout year of OCO-2 observations collocated with TCCON sites for 2022. The diffusion-based retrieval without finetuning shows a 3% improvement to ACOS prior to bias correction. When we compare to the operational product with additional finetuning for the diffusion retrieval, the finetuned diffusion retrieval shows a 10% improvement in RMSE over bias-corrected ACOS retrieved XCO_2 for the test year. We also compare the uncertainty estimate of both methods. Since the ACOS retrieval enforces a Gaussian form, we also assume a Gaussian prediction interval

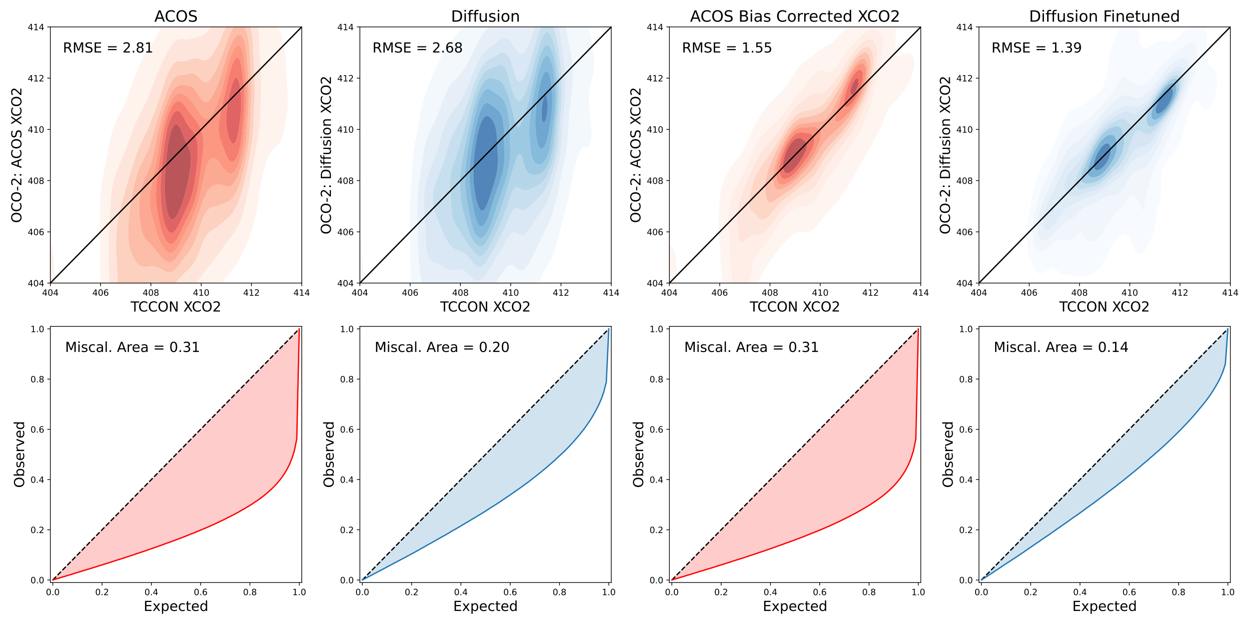

Figure 6.2: Comparison of ACOS and Diffusion with and without bias correction/finetuning for the holdout year of 2022. We compare the point estimate of each method (top row) by evaluating the RMSE. We compare the Gaussian uncertainties of each method (bottom row) by evaluating the Miscalibration Area.

for the diffusion retrieval by taking the first two moments over the diffusion samples. The performance of the uncertainties is assessed by calculating the Miscalibration Area which is the integrated region between the calibration curve and the $x = y$ line. The smaller the area, the better the coverage of the prediction interval for a given percentile. The assumed Gaussian prediction interval from the diffusion retrieval contains the collocated TCCON XCO_2 value more frequently than the operational retrieval. Finetuning provides an added benefit of further calibrating the uncertainty and reducing the Miscalibration Area on the holdout year. In Section 4.2.6.3, we train several additional types of deep learning models that can provide uncertainty quantification of their estimate and compare them to diffusion and ACOS.

6.2.4 Related Work

Machine learning approaches to greenhouse gas retrieval from Earth observing spectroscopy have made positive impact in reducing the computational cost over full physics based approaches. For OCO-2, [28] and [11] offered a first attempt at directly retrieving column CO_2 by pairing XCO_2 labels from the carbon inversion model CAMS directly with observations. Their approach improved computational speed but did not improve on precision due to the coarse scale of the CAMS labels. [54] introduced an initial attempt at providing a robust uncertainty quantification for a deep learning retrieval through ensembling but on a limited simulated training data set. [148] illustrated the limitations of training an ML retrieval for OCO-2 using observations only, and enhanced the training data set using a fast FM emulator. [64] proposes emulating the FM with a Gaussian Process to speed-up the iterative forward model evaluations in OE. The non-Gaussianity of posterior distribution of OCO retrieval has been explored using MCMC ([12], [65]) and Simulation Based UQ ([135], [9]). In this work, we propose a diffusion-based approach to efficiently sample the conditional posterior and train directly on radiance produced by the high fidelity FM used by the current operational retrieval algorithm for OCO-2. We also show how finetuning the retrieval using co-sampled ground truth labels from TCCON [146] can make ML retrievals competitive with the current state-of-the-art.

6.2.5 Experiments

6.2.5.1 Empirical Orthogonal Functions

A matrix of training residuals M_j is given for each band j. M_j is decomposed into its eigenvectors using the singular value decomposition $M_j = U_j W_j V_j^T$. The $k = 1$ eigen vector explains the largest fraction of the total variance. For the OCO-2 operational retrieval, coefficients of these fixed EOFs are retrieved to minimize the effect of errors from forward model discrepancies during retrieval. Hence, they serve as a proxy for the forward model error in the simulated radiances. To reduce similar errors in the ML retrieval we retrain using the EOFs to perturb the simulated training set ideally making the model more robust to sim-to-real differences. Before application to each i^{th} training point, each EOF $\mathbf{e_{j,k}}$ is randomly scaled by a coefficient $c_{i,j,k}$ drawn from the distributions formed by the operational scaling

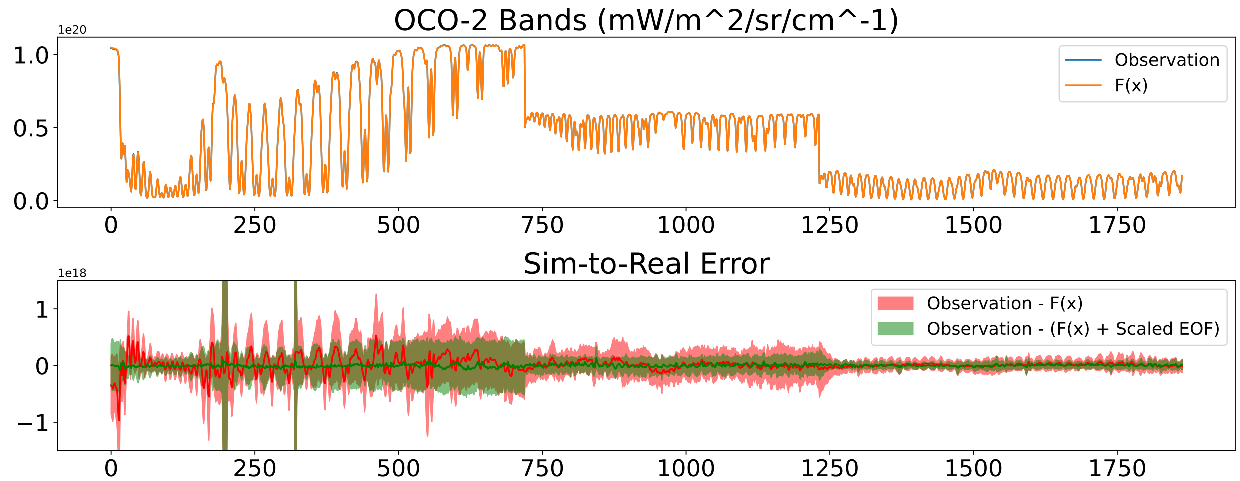

Figure 6.3: The top figure shows an example of an observed radiance (blue) and the FM modeled radiance (orange) for a single sounding. The bottom figure shows the mean and variance of remaining error before and after application of the randomly scaled EOFs.

terms. Thus, the training radiance for band $\mathbf{y_{Sim}}$ become:

$$\mathbf{y}'_{\mathbf{Sim_{i,j}}} = \mathbf{y_{Sim_{i,j}}} + \sum_{k=1}^{K} c_{i,j,k}\mathbf{e_{j,k}} \tag{6.1}$$

Figure 6.3 shows the mean difference between the simulated radiance and observation pairs withheld from fitting of the EOFs. ACOS includes the coefficients of the EOFs in the OE state vector. In future work we will look at retrieving the scaling term in a similar fashion using the diffusion-based retrieval as part of an expanded retrieved state. In Table 6.1, we perform a simple ablation to quantify the effects of including the Sim-to-Real mitigation during training of an MLP retrieval. By multiplying the EOFs by a random scaler drawn from the scaling distributions of those from the OE retrieval we see a 8% reduction in error versus the 2022 TCCON evaluation set.

Table 6.1: Ablation of retrieval with EOF perturbed training radiances and without. RMSE in parts per million [ppm] is calculated for the 2022 TCCON set.

Retrieval	**RMSE [ppm]**
MLP + EOF	3.15
MLP + Randomly Scaled EOF	**3.11**
MLP Baseline	3.37

6.2.5.2 The Conditional Mean Prior

In Table 6.2, we evaluate several choices for the conditional mean prior, x_{prior}, in the forward and reverse diffusion processes. First, we follow the method outlined in [40], where $x_{prior} = f(\mathbf{y})_\phi$ and f_ϕ is a pre-trained MLP model. For the second approach we modify the MLP to use 1-dimensional convolution layers with max-pooling. Lastly, we evaluate using the GEOS-FP-IT [71] XCO_2 labels which are also used as priors in the operational OE-based retrieval. We find that the convolutional model offers the best performance, where the other two priors provide a poorer RMSE as compared to the ACOS retrieval.

Table 6.2: Comparison of diffusion retrievals with different conditional mean priors. Mean and Standard Deviation of bias in parts per million [ppm] is calculated for the 2022 TCCON set.

Cond. Mean Prior	**Bias [ppm]**
MLP	-0.70 $\pm$ 2.88
Conv1D	**-0.23 $\pm$ 2.77**
GEOS-FP $XCO_{2,apriori}$	-1.03 $\pm$ 2.95

6.2.5.3 Comparison of Diffusion with Alternative ML Retrievals

In Table 6.4, we compare the conditional diffusion retrieval to several other deep learning-based methods that provide an uncertainty of their point estimates. The choice of methods are state of the art in several different paradigms of UQ for deep learning models: ensemble,

Table 6.3: Comparison of diffusion retrievals with different conditional mean priors. Mean and Standard Deviation of bias in parts per million [ppm] is calculated for the 2022 TCCON set.

Cond. Mean Prior	**Bias [ppm]**
MLP	-0.80 ± 2.92
MLP+$P_{Surf,apriori}$	**-0.33 ± 2.81**
GEOS-FP $XCO_{2,apriori}$	-1.03 ± 2.95

kernel learning, and hierarchical Bayesian.

Table 6.4: Mean and Standard Deviation of bias in parts per million [ppm] and Miscalibration Area (Miscal. Area) for ML retrieval calculated for the 2022 TCCON set.

Method	**Bias [ppm]**	**Miscal. Area**
ACOS	-0.62 ± 2.82	0.31
MLP	-0.8 ± 3.33	N/A
SWAG	-0.56 ± 3.14	0.23
Deep Kernel	-1.54 ± 4.50	0.38
Evidential	-0.49 ± 2.80	0.26
Diffusion	**-0.23 ± 2.72**	**0.21**

SWAG: Stochastic Weight Averaging-Gaussian (SWAG) [74] turns the final trajectory of Stochastic Gradient Descent (SGD) into a tractable Bayesian approximation:

1. *Collect SGD iterates.* After ordinary training converges, continue SGD for T extra steps, saving every k-th weight vector $\{\mathbf{w}_i\}_{i=1}^{T}$.

2. *Fit a Gaussian in weight space.* Compute the sample mean $\bar{\mathbf{w}}$ and covariance

$$\boldsymbol{\Sigma} = \frac{1}{T-1}\sum_{i=1}^{T}(\mathbf{w}_i - \bar{\mathbf{w}})(\mathbf{w}_i - \bar{\mathbf{w}})^{\top},$$

retaining only a diagonal plus low-rank update $\boldsymbol{\Sigma} \approx \mathbf{D}+\mathbf{U}\mathbf{U}^{\top}$ to stay memory-efficient.

3. *Bayesian model averaging.* Draw M samples $\mathbf{w}^{(m)} \sim \mathcal{N}(\bar{\mathbf{w}}, \boldsymbol{\Sigma})$, evaluate $p\Big(y \mid \mathbf{x}, \mathbf{w}^{(m)}\Big)$, and average:

$$p(y \mid \mathbf{x}) \;\approx\; \frac{1}{M}\sum_{m=1}^{M} p\Big(y \mid \mathbf{x}, \mathbf{w}^{(m)}\Big).$$

Evidential Regression: Deep Evidential Regression (DER) [1] embeds a conjugate Normal–Inverse-Gamma (NIG) prior inside a neural net, yielding both aleatoric and epistemic uncertainty in a single forward pass.

- *Hierarchical likelihood*

$$y \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2), \qquad (\mu, \sigma^2) \mid m, \nu, \alpha, \beta \sim \mathrm{NIG}(m, \nu, \alpha, \beta).$$

  The network outputs evidence parameters (m, ν, α, β), enforcing $\nu, \alpha, \beta > 0$ (e.g. via `softplus`).

- *Predictive variance* decomposes as

$$\underbrace{\frac{\beta(1+\nu)}{\alpha - 1}}_{\text{aleatoric}} + \underbrace{\frac{\beta(m-\mu_*)^2}{(\alpha-1)\nu}}_{\text{epistemic}}, \quad \text{with } \mu_* = m.$$

Deep Kernel Learning: Deep Kernel Learning (DKL) [141] couples neural feature learning with Gaussian Process (GP) inference.

- *Kernel construction*

$$k_{\boldsymbol{\theta}}\Big(\mathbf{x}, \mathbf{x}'\Big) = k_{\text{base}}\Big(f_{\boldsymbol{\phi}}(\mathbf{x}), f_{\boldsymbol{\phi}}(\mathbf{x}')\Big),$$

  where $f_{\boldsymbol{\phi}}$ is a deep network and k_{base} (e.g. Matérn 5/2) is a standard kernel.

- *Scalable training* uses inducing points or kernel interpolation (KISS-GP) to reduce cubic GP cost to nearly linear, while jointly optimizing $\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2$ via the marginal likelihood.

- *Prediction* for a test input $\mathbf{x}_*$:

$$p(y_* \mid \mathcal{D}) = \mathcal{N}\Big(\mu_*, \sigma_*^2\Big),$$

where σ_*^2 captures epistemic uncertainty and inflates outside the training distribution.

The diffusion model is implemented using MLP architectures with *Conv1D* layers for both f_ϕ and ϵ_θ. We use a $T = 50$ with a cosine noise schedule [98]. The remaining parameters for the start and end values for the schedule and EMA parameters are left as suggested in [40]. The retrieved XCO2 from the diffusion model is closer to a mean bias of 0.0 parts-per-million than the operational ACOS retrieval without any finetuning or bias correction with a similar error variance. We also note that both the Evidential and SWAG retrievals provide an improved comparison to the TCCON ground truth than ACOS. These methods do not require a conditional mean prior, as observed in Table 6.2. which can have a substantial impact on the diffusion-based retrieval.

6.2.6 Conclusions

We present a diffusion-based approach for retrieving atmospheric XCO_2 from satellite observations that is competitive with the current state-of-the-art approach while offering the for an increase of several orders of magnitude in speed during inference. The ability to quickly sample the conditional posterior also allows for improved uncertainty quantification compared to the operational product, thus allowing for an improved constraint of flux estimates by inversion models. In future work, we will further benchmark the computational speed up against the operational retrieval as well as derive an averaging kernel which is essential for assimilation of XCO_2 observations into carbon flux inversion models. It will also be important to assess the potential for the finetuning process to introduce biases in regions where TCCON is not available, since regionally coherent systematic errors in XCO_2 propagate directly into inferred CO_2 surface fluxes. The approach is highly generalizable to retrieval of other important GHGs such as methane (CH_4) and to future GHG missions such as Carbon-I [35] and CO2M, which will produce large data sets directly impacting future climate policy.

6.3 Methane Retrieval for EMIT

6.3.1 Operational Challenges of Methane Retrieval for Hyperspectral Imagers

Hyperspectral Imagers such as NASA's Earth Mineral Dust Source Investigation (EMIT) make high spatial resolution observations leading to high data volumes. Processing these data require fast methods such as Column-wise Matched Filters (CMF)([131]; [132]). For every spatial pixel we compute a filter score

$$\mathrm{MF} = \frac{\mathbf{s}^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\mathbf{s}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{s}},$$

where

- $\mathbf{x}$ – measured radiance spectrum,
- $\boldsymbol{\mu}$ – column–mean radiance,
- $\mathbf{C}$ – spectral covariance,
- $\mathbf{s}$ – methane "signal" vector.

Covariance. A shrinkage estimator stabilizes the empirical covariance $\mathbf{S}$:

$$\mathbf{C} = (1-a)\mathbf{S} + a\,\mathrm{diag}(\mathbf{S}), \qquad a = 10^{-9}.$$

Signal vector. Here, $\mathbf{s}$ is the element-wise product of the mean spectrum and a finite–difference radiance derivative:

$$\mathbf{t} = \frac{1}{L}\frac{\Delta L}{\Delta c}, \qquad L \text{ from MODTRAN 6}, \; \mathbf{s} = \mathbf{t} \odot \boldsymbol{\mu}.$$

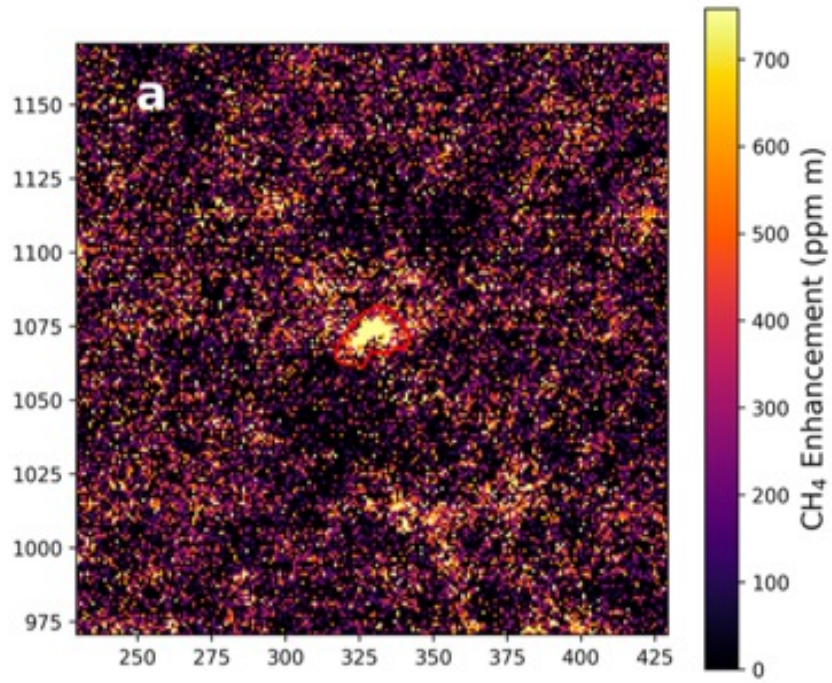

Figure 6.4: CMF with annotated plume.

Here ΔL is the radiance change between two MODTRAN runs that differ by Δc in methane column abundance, using water vapor profiles from the L2 retrieval. Whitening by $\mathbf{C}^{-1}$ maximizes sensitivity to the methane absorption signature while suppressing correlated spectral noise.

A drawback of the CMF is that it exchanges speed of retrieval for predictive performance, resulting in high noise and error over highly reflective confuser materials such as roads and clouds. This makes it difficult to reduce false positive identification of plume sources.

6.3.2 Injecting a Methane Signature into Background Radiance

For EMIT we lack a true simulator based forward model similar to that available for OCO-2. Instead we exploit the wealth of *background* pixels—spectra whose total column methane is assumed to equal the background—and inject a methane enhancement (greater than 0.0 ppmm). Pixels are randomly drawn from the annotated EMIT catalog, excluding any locations flagged as enhancements or containing known confusers (e.g. roads, clouds, infrastructure).

Step 1: Isolate Surface Radiance. The at-sensor spectral radiance for a background pixel can be given as the simplified form:

$$L_{\mathrm{bg}}(\lambda) = L_{\mathrm{path}}(\lambda) + T_{\mathrm{bg}}(\lambda)\, L_s(\lambda),$$

with

$$L_{\mathrm{path}}(\lambda) = \frac{\rho_{\mathrm{atm}}(\lambda)\, E_0(\lambda)\, \mu_0}{\pi}, \qquad T_{\mathrm{bg}}(\lambda) = \exp\Big[-\tau_{\mathrm{bg}}(\lambda)\Big]. \tag{6.2}$$

Here E_0 is the top-of-atmosphere solar irradiance, $\mu_0 = \cos\theta_0$ the solar zenith cosine, ρ_{atm} an effective atmospheric back-scatter reflectance, and τ_{bg} the total transmissivity. Solving for the unknown surface term:

$$L_s(\lambda) = \frac{L_{\mathrm{bg}}(\lambda) - L_{\mathrm{path}}(\lambda)}{T_{\mathrm{bg}}(\lambda)}. \tag{6.3}$$

Step 2: Inject Methane. Adding a column enhancement $\Delta \mathrm{XCH}_4$ increases the optical depth by $\Delta\tau_{CH4}(\lambda)$, so that the new transmission is

$$T_{\mathrm{inj}}(\lambda) = T_{\mathrm{bg}}(\lambda)\, \exp\Big[-\Delta\tau_{CH4}(\lambda)\Big].$$

The corresponding path radiance is recomputed as $L_{\mathrm{path,inj}}(\lambda)$ (typically a weak function of τ in the shortwave). The injected at-sensor radiance is therefore

$$L_{\mathrm{inj}}(\lambda) = L_{\mathrm{path,inj}}(\lambda) + T_{\mathrm{inj}}(\lambda)\, L_s(\lambda). \tag{6.4}$$

Look-up tables (LUTs). Both $T(\lambda)$, $\rho_{atm}(\lambda)$ and $L_{\mathrm{path}}(\lambda)$ are obtained from pre-computed MODTRAN 6 look-up tables indexed by water vapor, aerosols, viewing geometry, and a CH4 label ranging from 0.0–3000.0 ppmm.

6.3.3 Evidential Spectroscopic Transformer (SpecTf–DER) for $\mathrm{XCH_4}$

We adapt the spectroscopic transformer encoder of [68]—originally used for classification—to scalar regression of $\mathrm{XCH_4}$ with calibrated uncertainty via Deep Evidential Regression (DER) [2]. Treat the injected spectrum as a sequence of S bands:

$$\mathbf{r} = [r_1, \ldots, r_S]^\top, \qquad \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_S]^\top.$$

Band-aware embedding. Instead of sinusoidal position encodings, we concatenate the (normalized) wavelength to each band value and embed per band:

$$\mathbf{x}_s = \begin{bmatrix} r_s \\ (\lambda_s - \mu_\lambda)/\sigma_\lambda \end{bmatrix} \in \mathbb{R}^2, \quad \mathbf{z}_s = \phi(\mathbf{W}\,\mathbf{x}_s + \mathbf{b}) \in \mathbb{R}^d, \tag{6.5}$$

where ϕ is `tanh` (empirically stabilizing training), yielding $\mathbf{Z} \in \mathbb{R}^{S\times d}$.

Pre-LN self-attention encoder. Apply one or more pre-LayerNorm encoder layers (stability per [149]):

$$\tilde{\mathbf{Z}} = \mathrm{MHA}\Big(\mathrm{LN}(\mathbf{Z}), \mathrm{LN}(\mathbf{Z}), \mathrm{LN}(\mathbf{Z})\Big), \tag{6.6}$$

$$\mathbf{Z}' = \mathrm{FFN}\Big(\mathrm{LN}(\tilde{\mathbf{Z}})\Big), \tag{6.7}$$

then aggregate across bands with $\mathrm{Agg} \in \{\mathrm{mean}, \mathrm{max}, \mathrm{flatten}, \texttt{[CLS]}\}$ to obtain a fixed-size representation $\mathbf{h} \in \mathbb{R}^d$.

6.3.3.1 Self-Attention for Spectral Tokens

Let $\mathbf{Z} \in \mathbb{R}^{B\times S\times d}$ be the embedded spectral sequence for a batch of B pixels, with S bands and embedding size d. Self-attention forms queries, keys, and values from the same sequence:

$$\mathbf{Q} = \mathbf{Z}\mathbf{W}_Q, \qquad \mathbf{K} = \mathbf{Z}\mathbf{W}_K, \qquad \mathbf{V} = \mathbf{Z}\mathbf{W}_V, \quad \mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}. \tag{6.8}$$

Scaled dot-product (one head). For head size d_h,

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{B\times S\times S}, \qquad \mathbf{U} = \mathbf{A}\mathbf{V} \in \mathbb{R}^{B\times S\times d_h}, \tag{6.9}$$

where each band attends to all other bands. **Multi-head.** With H heads ($d = H\, d_h$), split $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ along d, compute $\mathbf{U}_h$, concatenate, and project:

$$\mathrm{MHA}(\mathbf{Z}) = \mathrm{Concat}_h(\mathbf{U}_h)\, \mathbf{W}_O, \qquad \mathbf{W}_O \in \mathbb{R}^{d\times d}. \tag{6.10}$$

Intuition for spectroscopy. Self attention yields a data-adaptive similarity over bands, emphasizing methane sensitive regions and down weighting confusers/artifact heavy bands [68]. **Cost.** Per layer: time $\mathcal{O}(B\, H\, S^2 d_h)$, memory $\mathcal{O}(B\, S^2)$ for the attention map.

Evidential head and predictive distribution. Map $\mathbf{h}$ to Normal–Inverse-Gamma (NIG) parameters $m = (\gamma, \nu, \alpha, \beta)$ with positivity via softplus and $\alpha > 1$ enforced by $1 + \mathrm{softplus}$. The latent model is

$$(\mu, \sigma^2) \sim \mathrm{N}\left(\mu \,\middle|\, \gamma, \frac{\sigma^2}{\nu}\right), \qquad \sigma^2 \sim \mathrm{InvGamma}(\alpha, \beta),$$

whose marginal predictive for $y = \mathrm{XCH}_4$ is Student-t:

$$p(y \mid m) = \mathrm{St}\left(y;\ \mu_t = \gamma,\ \sigma_t^2 = \frac{\beta(1+\nu)}{\nu\alpha},\ \nu_t = 2\alpha\right). \tag{6.11}$$

Uncertainty decomposes into aleatoric and epistemic parts:

$$\mathbb{E}[y] = \gamma, \qquad \underbrace{\mathbb{E}[\sigma^2]}_{\text{aleatoric}} = \frac{\beta}{\alpha - 1}, \qquad \underbrace{\mathrm{Var}[\mu]}_{\text{epistemic}} = \frac{\beta}{\nu(\alpha - 1)}, \quad \alpha > 1. \tag{6.12}$$

A Student-t evidential variant (SDER) instead parameterizes (γ, ν, β) directly and trains with t-NLL.

Evidential NLL loss (with evidence regularization). Maximizing the marginal likelihood in (6.11) yields the per-sample NIG evidential NLL:

$$\mathcal{L}_{\mathrm{NLL}} = \tfrac{1}{2}\log\frac{\pi}{\nu} - \alpha\log\Omega + \Big(\alpha + \tfrac{1}{2}\Big)\log\Big((y-\gamma)^2\,\nu + \Omega\Big) + \log\Gamma(\alpha) - \log\Gamma\Big(\alpha + \tfrac{1}{2}\Big), \quad \Omega = 2\beta(1+\nu). \tag{6.13}$$

To discourage unwarranted confidence under large residuals, we add an evidence penalty that scales with total evidence $(2\nu + \alpha)$:

$$\mathcal{L}_{\mathrm{R}} = \Big|y - \gamma\Big|\,(2\nu + \alpha), \qquad \mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \lambda\,\mathcal{L}_{\mathrm{R}}. \tag{6.14}$$

For SDER, one can train with the Student-t NLL

$$\mathcal{L}_{t\text{-NLL}} = \frac{1}{2}\log(\nu\,\pi) + \log\frac{\Gamma\Big(\frac{\nu+1}{2}\Big)}{\Gamma\Big(\frac{\nu}{2}\Big)} + \frac{1}{2}\log s^2 + \frac{\nu+1}{2}\log\left(1 + \frac{(y-\gamma)^2}{\nu\,s^2}\right), \quad s^2 = \sigma_t^2. \tag{6.15}$$

Training and inference summary. We train on pairs $\Big(L_{\mathrm{inj}}(\lambda),\, y\Big)$ drawn from the injected dataset (Sec. 2), minimizing (6.14) (or (6.15)). At inference, the point estimate is $\widehat{y} = \gamma$ and calibrated uncertainty is reported via (6.12) or by the full Student-t in (6.11).

6.3.4 DER Retrieval

We implement the above architecture as a SpecTf–DER model trained on the injected EMIT dataset and evaluated on EMIT scenes. The model outputs a posterior predictive Student-t

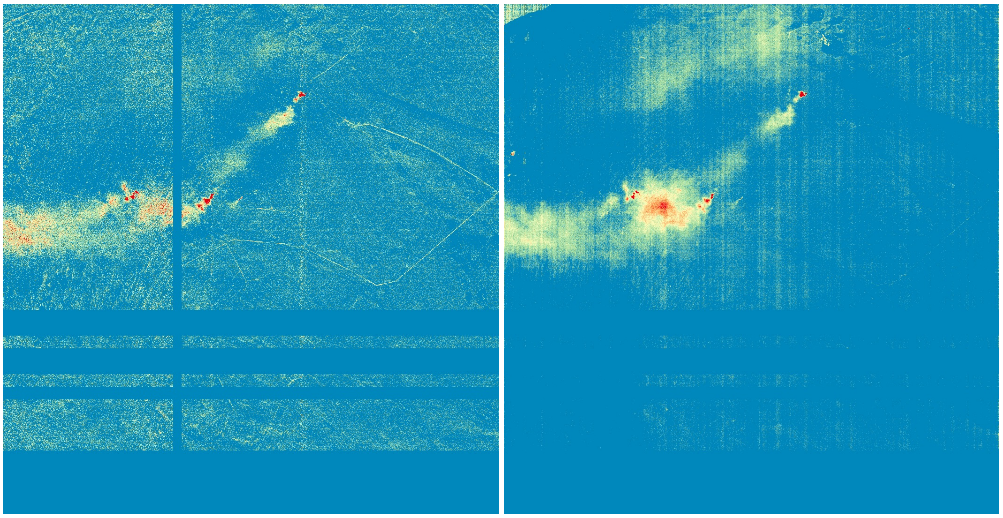

Figure 6.5: CMF and DER retrievals for `emit20230620t084426`.

distribution whose mean γ serves as the XCH_4 estimate and whose variance decomposes into aleatoric and epistemic components (Eq. 6.12). Initial results are promising, with reduced false enhancements over high-reflectance surfaces (e.g., roads) relative to CMF (Fig. 6.5). Instrument artifacts (e.g., striping) remain visible and are a target for future work. Planned directions include stronger artifact suppression, and further integration of transformer encoders with evidential heads for other retrieval tasks.

Bibliography

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

[2] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[3] ASU News. A career in orbit. *https://news.asu.edu/20250814-sun-devil-community-career-orbit*, 2025.

[4] H. H. Aumann, M. T. Chahine, C. Gautier, M. D. Goldberg, E. Kalnay, L. M. McMillin, H. Revercomb, P. W. Rosenkranz, W. L. Smith, D. H. Staelin, L. L. Strow, and J. Susskind. Airs/amsu/hsb on the aqua mission: Design, science objectives, data products, and processing systems. *IEEE Transactions on Geoscience and Remote Sensing*, 41:253–264, 2003.

[5] E. Bell, C. W. O'Dell, T. E. Taylor, A. Merrelli, R. R. Nelson, M. Kiel, A. Eldering, R. Rosenberg, and B. Fisher. Exploring bias in the oco-3 snapshot area-mapping mode via geometry, surface, and aerosol effects. *Atmospheric Measurement Techniques*, 16(1):109–133, 2023.

[6] T. Blumenstock, F. Hase, M. Schneider, O. E. Garcia, and E. Sepulveda. TCCON data from izaña (es), release ggg2014.r1, 2017. Dataset.

[7] Santiago Botía, Saqr Munassar, Thomas Koch, Danilo Custodio, Luana S Basso, Shujiro Komiya, Jost V Lavric, David Walter, Manuel Gloor, Giordane Martins, et al.

Combined co 2 measurement record indicates decreased amazon forest carbon uptake, offset by savannah carbon release. *EGUsphere*, 2024:1–55, 2025.

[8] H. Bovensmann, J. P. Burrows, M. Buchwitz, J. Frerick, S. Nöel, V. V. Rozanov, K. V. Chance, and A. Goede. Sciamachy—mission objectives and measurement modes. *Journal of the Atmospheric Sciences*, 56:127–150, 1999.

[9] Amy Braverman, Jonathan Hobbs, Joaquim Teixeira, and Michael Gunson. Post hoc uncertainty quantification for remote sensing observing systems. *SIAM/ASA Journal on Uncertainty Quantification*, 9(3):1064–1093, 2021.

[10] L. Breiman. *Classification and Regression Trees.* Routledge, New York, 1984.

[11] François-Marie Bréon, Leslie David, Pierre Chatelanaz, and Frédéric Chevallier. On the potential of a neural-network-based approach for estimating xco2 from oco-2 measurements. *Atmospheric Measurement Techniques*, 15(18):5219–5234, 2022.

[12] J. Brynjarsdóttir, J. Hobbs, A. Braverman, and L. Mandrake. Optimal estimation versus mcmc for co_2 retrievals. *Journal of Agricultural, Biological and Environmental Statistics*, 23:297–316, 2018.

[13] Brendan Byrne, David F Baker, Sourish Basu, Michael Bertolacci, Kevin W Bowman, Dustin Carroll, Abhishek Chatterjee, Frédéric Chevallier, Philippe Ciais, Noel Cressie, et al. National co 2 budgets (2015–2020) inferred from atmospheric co 2 observations in support of the global stocktake. *Earth System Science Data Discussions*, 2022:1–59, 2022.

[14] CAMS (Copernicus Atmosphere Monitoring Service). Validation report for the co_2 fluxes estimated by atmospheric inversion, v18r2. Technical report, European Centre for Medium-Range Weather Forecasts, 2021.

[15] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, 2016. ACM.

[16] F. Chevallier, F.-M. Bréon, and P. J. Rayner. Contribution of the orbiting carbon observatory to the estimation of co_2 sources and sinks: Theoretical study in a variational data-assimilation framework. *Journal of Geophysical Research: Atmospheres*, 112:D09301, 2007.

[17] F. Chevallier, P. Ciais, T. J. Conway, T. Aalto, B. E. Anderson, P. Bousquet, E. G. Brunke, L. Ciattaglia, and *et al.* co_2 surface fluxes at grid-point scale estimated from a global 21-year reanalysis of atmospheric measurements. *Journal of Geophysical Research: Atmospheres*, 115:D21307, 2010.

[18] B. Connor, H. Bösch, J. McDuffie, T. Taylor, D. Fu, C. Frankenberg, C. O'Dell, V. H. Payne, M. Gunson, R. Pollock, J. Hobbs, F. Oyafuso, and Y. Jiang. Quantification of uncertainties in oco-2 measurements of xco_2: Simulations and linear error analysis. *Atmospheric Measurement Techniques*, 9(10):5227–5238, 2016.

[19] B. J. Connor, H. Bösch, G. Toon, B. Sen, C. Miller, and D. Crisp. Orbiting carbon observatory: Inverse method and prospective error analysis. *Journal of Geophysical Research*, 113:A05305, 2008.

[20] Brian J Connor, Hartmut Boesch, Geoffrey Toon, Bhaswar Sen, Charles Miller, and David Crisp. Orbiting carbon observatory: Inverse method and prospective error analysis. *Journal of Geophysical Research: Atmospheres*, 113(D5), 2008.

[21] C. Cox and W. H. Munk. The measurement of the roughness of the sea surface from photographs of the sun's glitter. *Journal of the Optical Society of America*, 44:838–850, 1954.

[22] D. Crisp, R. M. Atlas, F. M. Bréon, L. R. Brown, J. P. Burrows, P. Ciais, B. J. Connor, S. C. Doney, and *et al.* The orbiting carbon observatory (oco) mission. *Advances in Space Research*, 34(4):700–709, 2004.

[23] D. Crisp, B. M. Fisher, C. O'Dell, C. Frankenberg, R. Basilio, H. Bösch, L. R. Brown, R. Castano, and *et al.* The acos co_2 retrieval algorithm – part ii: Global xco_2 data characterization. *Atmospheric Measurement Techniques*, 5(4):687–707, 2012.

[24] David Crisp. Measuring atmospheric carbon dioxide from space with the orbiting carbon observatory-2 (oco-2). In *Earth Observing Systems*, volume 9607, page 960702. SPIE, 2015.

[25] S. Crowell, D. Baker, A. Schuh, S. Basu, A. R. Jacobson, F. Chevallier, J. Liu, F. Deng, L. Feng, K. McKain, and *et al.* The 2015–2016 carbon cycle as seen from oco-2 and the global in situ network. *Atmospheric Chemistry and Physics*, 19:9797–9831, 2019.

[26] Sean Crowell, David Baker, Andrew Schuh, Sourish Basu, Andrew R Jacobson, Frederic Chevallier, Junjie Liu, Feng Deng, Liang Feng, Kathryn McKain, et al. The 2015–2016 carbon cycle as seen from oco-2 and the global in situ network. *Atmospheric Chemistry and Physics*, 19(15):9797–9831, 2019.

[27] S. Das, M. Kiel, J. Laughner, and G. Osterman. Three ways to validate the xco_2 product from the orbiting carbon observatory-2. manuscript in preparation, 2024.

[28] Leslie David, Francois Marie Breon, and Frederic Chevallier. Xco2 estimates from the oco-2 measurements using a neural network approach. *Atmospheric Measurement Techniques*, 14:117–132, 1 2021.

[29] M. De Mazière, M. K. Sha, F. Desmet, C. Hermans, F. Scolas, N. Kumps, J.-M. Metzger, V. Duflot, and J.-P. Cammas. TCCON data from réunion island (re), release ggg2014.r1, 2017. Dataset.

[30] N. M. Deutscher, J. Notholt, J. Messerschmidt, C. Weinzierl, T. Warneke, C. Petri, and P. Grupe. TCCON data from bialystok (pl), release ggg2014.r2, 2019. Dataset.

[31] M. K. Dubey, R. Lindenmaier, B. G. Henderson, D. Green, N. T. Allen, C. M. Roehl, J.-F. Blavier, Z. T. Butterfield, S. Love, J. D. Hamelmann, and D. Wunch. TCCON data from four corners (us), release ggg2014.r0, 2014. Dataset.

[32] A. Eldering, C. W. O'Dell, P. O. Wennberg, D. Crisp, M. R. Gunson, C. Viatte, C. Avis, A. Braverman, R. Castano, A. Chang, L. Chapsky, C. Cheng, B. Connor, L. Dang, G. Doran, B. Fisher, C. Frankenberg, D. Fu, R. Granat, J. Hobbs, R. A. M. Lee, L. Mandrake, J. McDuffie, C. E. Miller, V. Myers, V. Natraj, D. O'Brien, G. B. Osterman, F. Oyafuso, V. H. Payne, H. R. Pollock, I. Polonsky, C. M. Roehl, R. Rosenberg, F. Schwandner, M. Smyth, V. Tang, T. E. Taylor, C. To, D. Wunch, and J. Yoshimizu. The orbiting carbon observatory-2: first 18 months of science data products. *Atmospheric Measurement Techniques*, 10:549–563, 2017.

[33] L. Feng, P. I. Palmer, H. Bösch, and S. Dance. Estimating surface co_2 fluxes from space-borne co_2 dry-air mole-fraction observations using an ensemble kalman filter. *Atmospheric Chemistry and Physics*, 9(8):2619–2633, 2009.

[34] C. Frankenberg, U. Platt, and T. Wagner. Iterative maximum a posteriori (imap)–doas for retrieval of strongly absorbing trace gases: Model studies for ch_4 and co_2 retrieval from near-infrared spectra of sciamachy. *Atmospheric Chemistry and Physics*, 5:9–22, 2005.

[35] Christian Frankenberg, Anna Michalak, Daniel J Jacob, Andrew K Thorpe, Yi Yin, Lori Bruhwiler, Ermias Kebreab, Alison Hoyt, Alex J Turner, Paul O Wennberg, et al. Carbon-i, a nasa earth system explorer mission concept for greenhouse gas observations. *AGU24*, 2024.

[36] T.-Y. Goo, Y.-S. Oh, and V. A. Velazco. TCCON data from anmeyondo (kr), release ggg2014.r0, 2014. Dataset.

[37] D. W. T. Griffith, N. M. Deutscher, V. A. Velazco, P. O. Wennberg, Y. Yavin, G. Keppel-Aleks, R. A. Washenfelder, G. C. Toon, and *et al.* TCCON data from darwin (au), release ggg2014.r0, 2014. Dataset.

[38] D. W. T. Griffith, V. A. Velazco, N. M. Deutscher, C. Murphy, N. Jones, S. Wilson, R. Macatangay, G. Kettlewell, R. R. Buchholz, and M. Riggenbach. TCCON data from wollongong (au), release ggg2014.r0, 2014. Dataset.

[39] B. D. Hall, A. M. Crotwell, D. R. Kitzis, T. Mefford, B. R. Miller, M. F. Schibig, and P. P. Tans. Revision of the world meteorological organization global atmosphere watch (wmo/gaw) co_2 calibration scale. *Atmospheric Measurement Techniques*, 14(4):3015–3032, 2021.

[40] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

[41] F. Hase, T. Blumenstock, S. Dohe, J. Gross, and M. Kiel. TCCON data from karlsruhe (de), release ggg2014.r1, 2015. Dataset.

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[43] J. M. Hobbs, B. J. Drouin, F. Oyafuso, V. H. Payne, M. R. Gunson, J. McDuffie, and E. J. Mlawer. Spectroscopic uncertainty impacts on oco-2/3 retrievals of xco_2. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 257:107360, 2020.

[44] Jonathan Hobbs, Amy Braverman, Noel Cressie, Robert Granat, and Michael Gun-

son. Simulation-based uncertainty quantification for estimating atmospheric co2 from satellite data. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):956–985, 2017.

[45] N. Jacobs, W. R. Simpson, D. Wunch, C. W. O'Dell, G. B. Osterman, F. Hase, T. Blumenstock, and *et al.* Quality controls, bias, and seasonality of co_2 columns in the boreal forest with oco-2, tccon, and em27/sun measurements. *Atmospheric Measurement Techniques*, 13:5033–5063, 2020.

[46] A. R. Jacobson, K. N. Schuldt, P. Tans, A. Andrews, J. Miller, and T. Oda. Carbontracker ct2022, 2023. Dataset.

[47] Elaheh Jafarigol, Sean Crowell, Christopher O'Dell, Gregory McGarragh, Jay Rothenberger, and William Keely. Fine-tuning a Foundation Model for Methane Emissions Detection in the Middle East. In *AGU Fall Meeting Abstracts*, volume 2024, pages IN52A–04, December 2024.

[48] Elaheh Jafarigol, William Keely, Tess Hortag, Tom Welborn, Peyman Hekmatpour, and Theodore B Trafalis. Religious affiliation in the twenty-first century: A machine learning perspective on the world value survey. *Society*, 60(5):733–749, 2023.

[49] W. Keely. Conditional diffusion-based retrieval of atmospheric co_2 from earth observing spectroscopy oco-2/tccon 2021, 2025. Dataset, Tackling Climate Change with Machine Learning (ICLR 2025).

[50] W. Keely, S. Mauceri, R. Nelson, J. Laughner, C. O'Dell, S. Massie, D. Baker, M. Kiel, O. Lamminpää, J. Hobbs, A. Chatterjee, T. Taylor, P. Wennberg, S. Crowell, B. Stephens, and V. Payne. Uncertainty-aware machine-learning bias correction and filter for oco-2: Part 2. *Authorea Preprints*, 2025.

[51] W. R. Keely. Amt ml bias correction oco-2, 2023. Dataset.

[52] William Keely, Philip G. Brodrick, Andrew K. Thorpe, Jake Lee, and Jay Fahlen. Methane emission quantification using diffusion applied to imaging spectroscopy time series for EMIT. In *AGU Fall Meeting Abstracts*, volume 2024, pages A05–07, December 2024.

[53] William Keely, Sean Crowell, Berrien Moore, Dimitrios Diochnos, and Christopher O'Dell. Geophysical bias correction of trace green house gas satellite retrievals using explainable machine learning methods. In *AGU Fall Meeting Abstracts*. AGU, 2021.

[54] William Keely and Steffen Mauceri. Towards a machine learning retrieval of xco2 from oco-2 radiances with uncertainty quantification. In *AGU Fall Meeting Abstracts*, volume 2022, pages A14E–06, 2022.

[55] William Keely, Steffen Mauceri, Otto Lamminää, Amy Braverman, Jonathan Hobbs, Joaquim Teixeira, and Vivienne Payne. Uncertainty quantification for a machine learning bias correction of oco-2 column averaged co2. *Authorea Preprints*, 2023.

[56] William Keely, Steffen Mauceri, Otto Lamminpää, Gregory McGarragh, Christopher O'dell, and Sean Crowell. Disco: Diffusion for conditional retrieval of atmospheric co2 from satellite observations. In *AGU Fall Meeting Abstracts*, volume 2024, pages A05–06, 2024.

[57] William R Keely, Otto Lamminpää, Steffen Mauceri, Sean MR Crowell, Christopher W O'Dell, and Gregory R McGarragh. Conditional diffusion-based retrieval of atmospheric co2 from earth observing spectroscopy. *arXiv preprint arXiv:2504.17074*, 2025.

[58] William R Keely, Steffen Mauceri, Sean Crowell, and Christopher W O'Dell. A non-linear data-driven approach to bias correction of xco2 for nasa's oco-2 acos version 10. *Atmospheric Measurement Techniques*, 16(23):5725–5748, 2023.

[59] Hannah Kerner, Marc Rußwurm, Hamed Alemohammad, et al. 3rd iclr workshop on

machine learning for remote sensing, 2025. Workshop overview highlighting unique challenges and methods for ML in remote sensing.

[60] M. Kiel, C. W. O'Dell, B. Fisher, A. Eldering, R. Nassar, C. G. MacDonald, and P. O. Wennberg. How bias correction goes wrong: measurement of xco_2 affected by erroneous surface-pressure estimates. *Atmospheric Measurement Techniques*, 12(4):2241–2259, 2019.

[61] R Kivi, P Heikkinen, and E Kyrö. Tccon data from sodankylä (fi), release ggg2014. r0. *(No Title)*, 2014.

[62] S. S. Kulawik, C. O'Dell, R. R. Nelson, and T. E. Taylor. Validation of oco-2 error analysis using simulated retrievals. *Atmospheric Measurement Techniques*, 12:5317–5334, 2019.

[63] A. Kuze, H. Suto, M. Nakajima, and T. Hamazaki. Initial onboard performance of tanso-fts on gosat. In *Advances in Imaging (OSA Technical Digest)*. Optica Publishing Group, 2009.

[64] Otto M Lamminpää, Jouni I Susiluoto, Jonathan M Hobbs, James L McDuffie, Amy J Braverman, and Houman Owhadi. Forward model emulator for atmospheric radiative transfer using gaussian processes and cross validation. *Atmospheric Measurement Techniques Discussions*, 2024:1–32, 2024.

[65] Otto Lamminpää, Jonathan Hobbs, Jenný Brynjarsdóttir, Marko Laine, Amy Braverman, Hannakaisa Lindqvist, and Johanna Tamminen. Accelerated mcmc for satellite-based measurements of atmospheric co_2. *Remote Sensing*, 11(17), 2019.

[66] J. L. Laughner, G. C. Toon, J. Mendonca, C. Petri, S. Roche, D. Wunch, J.-F. Blavier, D. W. T. Griffith, and *et al.* The total carbon column observing network's ggg2020 data version. *Earth System Science Data*, 16:2197–2260, 2024.

[67] Jake Lee, Brian Bue, Philip G. Brodrick, Andrew K. Thorpe, William Keely, Michael Kiper, and Jay Fahlen. Spectral Transformers for EMIT Methane Retrieval. In *AGU Fall Meeting Abstracts*, volume 2024, pages A53I–2209, December 2024.

[68] Jake H Lee, Michael Kiper, David R Thompson, and Philip G Brodrick. Spectf: Transformers enable data-driven imaging spectroscopy cloud detection. *arXiv preprint arXiv:2501.04916*, 2025.

[69] A. Liang, W. Gong, G. Han, and C. Xiang. Comparison of satellite-observed xco_2 from gosat, oco-2, and ground-based tccon. *Remote Sensing*, 9(10):1033, 2017.

[70] J. Liu, K. W. Bowman, D. S. Schimel, N. C. Parazoo, Z. Jiang, M. Lee, A. A. Bloom, D. Wunch, C. Frankenberg, Y. Sun, and *et al.* Contrasting carbon-cycle responses of the tropical continents to the 2015–2016 El Niño. *Science*, 358:eaam5690, 2017.

[71] R Lucchesi. File specification for geos-5 fp-it (forward processing for instrument teams). Technical report, NASA, 2013.

[72] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[73] Tristan S L'Ecuyer, Yun Hang, Alexander V Matus, and Zhien Wang. Reassessing the effect of cloud type on earth's energy balance in the age of active spaceborne observations. part i: Top of atmosphere and surface. *Journal of Climate*, 32(19):6197–6217, 2019.

[74] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

[75] L. Mandrake, C. W. O'Dell, D. Wunch, P. O. Wennberg, B. Fisher, G. B. Osterman, and A. Eldering. Orbiting carbon observatory-2 (oco-2) warn level, bias correction, and lite file product description. Technical report, Jet Propulsion Laboratory, California Institute of Technology, 2015. `https://disc.sci.gsfc.nasa.gov/OCO-2/documentation/oco-2-v7/OCO2_XCO2_Lite_Files_and_Bias_Correction_0915_sm.pdf`.

[76] K. A. Masarie, W. Peters, A. R. Jacobson, and P. P. Tans. Obspack: A framework for the preparation, delivery, and attribution of atmospheric greenhouse gas measurements. *Earth System Science Data*, 6:375–384, 2014.

[77] S. Massie, H. Cronk, A. Merrelli, K. S. Schmidt, H. Chen, and D. Baker. 3d cloud metrics for oco-2 observations, 2020. last accessed: 2025-01.

[78] S. Massie, S. Sebastian, A. Eldering, and D. Crisp. Observational evidence of 3-d cloud effects in oco-2 co_2 retrievals. *Journal of Geophysical Research: Atmospheres*, 122(13):7064–7085, 2017.

[79] S. T. Massie, H. Cronk, A. Merrelli, C. O'Dell, K. S. Schmidt, H. Chen, and D. Baker. Analysis of 3d cloud effects in oco-2 xco_2 retrievals. *Atmospheric Measurement Techniques*, 14:1475–1499, 2021.

[80] S. T. Massie, H. Cronk, A. Merrelli, S. Schmidt, and S. Mauceri. Insights into 3d cloud radiative-transfer effects for the orbiting carbon observatory. *Atmospheric Measurement Techniques*, 16:2145–2166, 2023.

[81] V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the IPCC*. Cambridge University Press, 2021.

[82] S. Mauceri and W. Keely. Uncertainty-aware machine learning bias correction and filtering for oco-2 | 2014–2024 (version v1), 2025. Dataset.

[83] S. Mauceri, S. Massie, and S. Schmidt. Correcting 3d cloud effects in xco_2 retrievals from the orbiting carbon observatory-2 (oco-2). *Atmospheric Measurement Techniques*, 16:1461–1476, 2023.

[84] Steffen Mauceri, William Keely, Joshua Laughner, Christopher O'Dell, Steven T Massie, Robert Roland Nelson, David F Baker, Matthäus Kiel, Otto Lamminpaeae, Jonathan M Hobbs, et al. Uncertainty-aware machine learning bias correction and filtering for oco-2: Part 1. *Earth and Space Science*, 2025. in press.

[85] Max Planck Institute for Biogeochemistry. Carboscope. `https://www.bgc-jena.mpg.de/CarboScope`, 2021. accessed 2022-01-10.

[86] N. Meinshausen and G. Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[87] J. Mendonca, R. Nassar, C. W. O'Dell, R. Kivi, I. Morino, J. Notholt, C. Petri, K. Strong, and D. Wunch. Assessing the feasibility of using a neural network to filter orbiting carbon observatory-2 (oco-2) retrievals at northern high latitudes. *Atmospheric Measurement Techniques*, 14:7511–7524, 2021.

[88] C. E. Miller, D. Crisp, P. L. DeCola, S. C. Olsen, J. T. Randerson, A. M. Michalak, and *et al.* Precision requirements for space-based column co_2 data. *Journal of Geophysical Research: Atmospheres*, 112:D10314, 2007.

[89] I. Morino, T. Matsuzaki, and M. Horikawa. TCCON data from tsukuba (jp), 125hr, release ggg2014.r2, 2018. Dataset.

[90] I. Morino, V. A. Velazco, A. Hori, O. Uchino, and D. W. T. Griffith. TCCON data from burgos, ilocos norte (ph), release ggg2014.r0, 2018. Dataset.

[91] I. Morino, N. Yokozeki, A. Matzuzaki, and A. Shishime. TCCON data from rikubetsu (jp), release ggg2014.r2, 2018. Dataset.

[92] G. Myhre, C. R. Hoyle, T. F. Berglen, B. T. Johnson, and J. M. Haywood. Modeling of the solar radiative impact of biomass-burning aerosols during the dust and biomass-burning experiment (dabex). *Journal of Geophysical Research: Atmospheres*, 113:D00C16, 2008.

[93] R. Nassar, T. G. Hill, C. A. McLinden, D. Wunch, D. B. A. Jones, and D. Crisp. Quantifying co_2 emissions from individual power plants from space. *Geophysical Research Letters*, 44:10045–10053, 2017.

[94] R. Nassar, J.-P. Mastrogiacomo, W. Bateman-Hemphill, C. McCracken, C. G. MacDonald, T. Hill, C. W. O'Dell, M. Kiel, and D. Crisp. Advances in quantifying power plant co_2 emissions with oco-2. *Remote Sensing of Environment*, 264:112579, 2021.

[95] R. Nassar, O. Moeini, J.-P. Mastrogiacomo, C. W. O'Dell, R. R. Nelson, M. Kiel, A. Chatterjee, A. Eldering, and D. Crisp. Tracking co_2 emission reductions from space: A case study at europe's largest fossil-fuel power plant. *Frontiers in Remote Sensing*, 3:1028240, 2022.

[96] B. A. Nault, P. Campuzano-Jost, D. A. Day, D. S. Jo, J. C. Schroder, H. M. Allen, and *et al.* Airborne observations and modeling comparison of global inorganic aerosol acidity, 2022. Dataset, ORNL DAAC, DOI assigned by ORNL.

[97] H. Nguyen, G. Osterman, D. Wunch, C. O'Dell, L. Mandrake, P. Wennberg, B. Fisher, and R. Castano. A method for colocating satellite xco_2 data to ground-based data and its application to acos-gosat and tccon. *Atmospheric Measurement Techniques*, 7(8):2631–2644, 2014.

[98] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion prob-

abilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[99] NOAA Earth System Research Laboratory. Carbontracker. `https://carbontracker.noaa.gov`, 2021. accessed 2022-01-10.

[100] S. Noël, M. Reuter, M. Buchwitz, J. Borchardt, M. Hilker, O. Schneising, H. Bovensmann, J. P. Burrows, A. Di Noia, R. J. Parker, and *et al.* Retrieval of greenhouse gases from gosat and gosat-2 using the focal algorithm. *Atmospheric Measurement Techniques*, 15:3401–3437, 2022.

[101] J. Notholt, C. Petri, T. Warneke, N. M. Deutscher, M. Palm, M. Buschmann, C. Weinzierl, R. C. Macatangay, and P. Grupe. TCCON data from bremen (de), release ggg2014.r1, 2019. Dataset.

[102] Christopher W O'dell, Annmarie Eldering, Paul O Wennberg, David Crisp, Michael R Gunson, Brendan Fisher, Christian Frankenberg, Matthäus Kiel, Hannakaisa Lindqvist, Lukas Mandrake, et al. Improved retrievals of carbon dioxide from orbiting carbon observatory-2 with the version 8 acos algorithm. *Atmospheric Measurement Techniques*, 11(12):6539–6576, 2018.

[103] Christopher W. O'Dell and et al. Improved retrievals of carbon dioxide from oco-2 with version 9 of the imap–doas algorithm. *Atmospheric Measurement Techniques*, 11(12):6539–6576, 2018.

[104] CW O'Dell, B Connor, H Bösch, D O'Brien, C Frankenberg, R Castano, M Christi, D Eldering, B Fisher, M Gunson, et al. The acos co2 retrieval algorithm–part 1: Description and validation against synthetic observations. *Atmospheric Measurement Techniques*, 5(1):99–121, 2012.

[105] S. C. Olsen and J. T. Randerson. Differences between surface and column atmospheric

co$_2$ and implications for carbon-cycle research. *Journal of Geophysical Research: Atmospheres*, 109:D02301, 2004.

[106] G. B. Osterman, C. W. O'Dell, A. Eldering, B. Fisher, D. Crisp, C. Cheng, C. Frankenberg, A. Lambert, M. R. Gunson, L. Mandrake, and D. Wunch. Orbiting carbon observatory-2 & 3 (oco-2 & oco-3) data product user's guide: Operational level 2 data v10 and lite v10. Technical report, Jet Propulsion Laboratory, California Institute of Technology, 2020.

[107] Y. Ozaki, Y. Tanigaki, S. Watanabe, and M. Onishi. Multi-objective tree-structured parzen estimator for computationally expensive optimization problems. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference (GECCO '20)*, pages 533–541, New York, NY, USA, 2020. Association for Computing Machinery.

[108] P. I. Palmer, L. Feng, D. Baker, F. Chevallier, B. Hartman, and P. Somkuti. Net carbon emissions from african biosphere dominate pan-tropical atmospheric co$_2$ signal. *Nature Communications*, 10:3344, 2019.

[109] V. H. Payne, A. Chatterjee, R. Rosenberg, M. Kiel, B. Fisher, and L. Dang. Carbon observatory-2 & 3 (oco-2 & oco-3) data product user's guide, operational level 2 lite files. Technical report, 2023.

[110] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and *et al.* Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[111] Hélène Peiro, Sean Crowell, Andrew Schuh, David F Baker, Chris O'Dell, Andrew R Jacobson, Frédéric Chevallier, Junjie Liu, Annmarie Eldering, David Crisp, et al. Four years of global carbon cycle observed from the orbiting carbon observatory 2 (oco-2) version 9 and in situ data and comparison to oco-2 version 7. *Atmospheric Chemistry and Physics*, 22(2):1097–1130, 2022.

[112] W. Peters, A. R. Jacobson, C. Sweeney, A. E. Andrews, T. J. Conway, K. Masarie, J. B. Miller, L. M. P. Bruhwiler, G. Pétron, A. I. Hirsch, D. E. J. Worthy, G. R. van der Werf, J. T. Randerson, P. O. Wennberg, M. C. Krol, and P. P. Tans. An atmospheric perspective on north american carbon dioxide exchange: Carbontracker. *Proceedings of the National Academy of Sciences of the United States of America*, 104:18925–18930, 2007.

[113] D. F. Pollard, J. Robinson, and H. Shiona. TCCON data from lauder (nz), release ggg2014.r0, 2019. Dataset.

[114] M. Reuter, M. Hilker, S. Noël, A. Di Noia, M. Weimer, O. Schneising, and *et al.* Retrieving the atmospheric concentrations of co_2 and ch_4 from the european copernicus co2m satellite mission using artificial neural networks. *EGUsphere*, pages 1–32, 2024.

[115] C. Rödenbeck. Estimating co_2 sources and sinks from atmospheric mixing ratios using a global inversion of atmospheric transport. Technical report, Max Planck Institute for Biogeochemistry, Jena, Germany, 2005.

[116] C. Rödenbeck, S. Zaehle, R. Keeling, and M. Heimann. How does the terrestrial carbon exchange respond to inter-annual climatic variations? a quantification based on atmospheric co_2 data. *Biogeosciences*, 15:2481–2498, 2018.

[117] Clive D. Rodgers. *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific, Singapore, 2000.

[118] Clive D Rodgers. *Inverse methods for atmospheric sounding: theory and practice*, volume 2. World scientific, 2000.

[119] Esther Rolf, Sara Beery, Nicolas Audebert, et al. Satellite data is a distinct modality in machine learning. *arXiv preprint arXiv:2402.01444*, 2024.

[120] Jay Rothenberger, Sean Crowell, William Keely, Daniel Cusworth, and Katherine Howell. The Promise of Foundation Models for Improving Automated Detection of Methane Sources in Hyperspectral Imagers: A Case Study for the EMIT Instrument. In *AGU Fall Meeting Abstracts*, volume 2024, pages H01–08, December 2024.

[121] O. Schneising, M. Buchwitz, M. Reuter, H. Bovensmann, J. P. Burrows, T. Borsdorff, N. M. Deutscher, D. G. Feist, D. W. T. Griffith, F. Hase, C. Hermans, L. T. Iraci, R. Kivi, J. Landgraf, I. Morino, J. Notholt, C. Petri, D. F. Pollard, S. Roche, K. Shiomi, K. Strong, R. Sussmann, V. A. Velazco, T. Warneke, and D. Wunch. A scientific algorithm to simultaneously retrieve carbon monoxide and methane from tropomi onboard sentinel-5 precursor. *Atmospheric Measurement Techniques*, 12:6771–6802, 2019.

[122] A. E. Schuh, A. R. Jacobson, S. Basu, B. Weir, D. Baker, K. Bowman, F. Chevallier, S. Crowell, and *et al.* Quantifying the impact of atmospheric transport uncertainty on co_2 surface-flux estimates. *Global Biogeochemical Cycles*, 33(4):484–500, 2019.

[123] K. Shiomi, S. Kawakami, H. Ohyama, K. Arai, H. Okumura, C. Taura, T. Fukamachi, and M. Sakashita. TCCON data from saga (jp), release ggg2014.r0, 2014. Dataset.

[124] Bernd Sierk, Valerie Fernandez, J-L Bézy, Y Meijer, Y Durand, G Bazalgette Courrèges-Lacoste, C Pachot, A Löscher, H Nett, K Minoglou, et al. The copernicus co2m mission for monitoring anthropogenic carbon dioxide emissions from space. In *International Conference on Space Optics—ICSO 2020*, volume 11852, pages 1563–1580. SPIE, 2021.

[125] R. Sussmann and M. Rettinger. TCCON data from garmisch (de), release ggg2014.r2, 2018. Dataset.

[126] T. E. Taylor, A. Eldering, A. Merrelli, M. Kiel, P. Somkuti, C. Cheng, R. Rosenberg, B. Fisher, D. Crisp, R. Basilio, and M. Bennett. Oco-3 early mission operations and

initial (vearly) xco_2 and sif retrievals. *Remote Sensing of Environment*, 251:112032, 2020.

[127] T. E. Taylor, C. W. O'Dell, D. Baker, C. Bruegge, A. Chang, L. Chapsky, A. Chatterjee, C. Cheng, F. Chevallier, D. Crisp, L. Dang, B. Drouin, A. Eldering, L. Feng, B. Fisher, D. Fu, M. Gunson, V. Haemmerle, G. R. Keller, M. Kiel, L. Kuai, T. Kurosu, A. Lambert, J. Laughner, R. Lee, J. Liu, L. Mandrake, Y. Marchetti, G. McGarragh, A. Merrelli, R. R. Nelson, G. Osterman, F. Oyafuso, P. I. Palmer, V. H. Payne, R. Rosenberg, P. Somkuti, G. Spiers, C. To, B. Weir, P. O. Wennberg, S. Yu, and J. Zong. Evaluating the consistency between oco-2 and oco-3 xco_2 estimates derived from the nasa acos version 10 retrieval algorithm. *Atmospheric Measurement Techniques*, 16:3173–3209, 2023.

[128] T. E. Taylor, C. W. O'Dell, C. Frankenberg, P. T. Partain, H. Q. Cronk, A. Savtchenko, R. R. Nelson, E. J. Rosenthal, A. Y. Chang, B. Fisher, G. B. Osterman, R. H. Pollock, D. Crisp, A. Eldering, and M. R. Gunson. Orbiting carbon observatory-2 (oco-2) cloud screening algorithms: validation against collocated modis and caliop data. *Atmospheric Measurement Techniques*, 9:973–989, 2016.

[129] Thomas E Taylor, Christopher W O'Dell, David Baker, Carol Bruegge, Albert Chang, Lars Chapsky, Abhishek Chatterjee, Cecilia Cheng, Frédéric Chevallier, David Crisp, et al. Evaluating the consistency between oco-2 and oco-3 xco2 estimates derived from the nasa acos version 10 retrieval algorithm. *Atmospheric Measurement Techniques Discussions*, 2023:1–61, 2023.

[130] Y. Té, P. Jeseck, and C. Janssen. TCCON data from paris (fr), release ggg2014.r0, 2014. Dataset.

[131] D. R. Thompson, I. Leifer, H. Bovensmann, M. Eastwood, M. Fladeland, C. Frankenberg, K. Gerilowski, R. O. Green, S. Kratwurst, T. Krings, B. Luna, and A. K. Thorpe.

Real-time remote detection and measurement for airborne imaging spectroscopy: A case study with methane. *Atmospheric Measurement Techniques*, 8:4383–4397, 2015.

[132] A. K. Thorpe, D. A. Roberts, E. S. Bradley, C. C. Funk, P. E. Dennison, and I. Leifer. High-resolution mapping of methane emissions from marine and terrestrial sources using a cluster-tuned matched filter technique and imaging spectrometry. *Remote Sensing of Environment*, 134:305–318, 2013.

[133] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[134] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. *arXiv preprint arXiv:2502.09356*, 2025.

[135] Michael Turmon and Amy Braverman. Uncertainty quantification for JPL retrievals. Technical report, 2019.

[136] T. Warneke, J. Messerschmidt, J. Notholt, C. Weinzierl, N. M. Deutscher, C. Petri, and P. Grupe. TCCON data from orléans (fr), release ggg2014.r1, 2019. Dataset.

[137] B. Weir, L. E. Ott, G. J. Collatz, S. R. Kawa, B. Poulter, A. Chatterjee, T. Oda, and S. Pawson. Bias-correcting carbon fluxes derived from land-surface satellite data for retrospective and near-real-time assimilation systems. *Atmospheric Chemistry and Physics*, 21:9609–9628, 2021.

[138] P. O. Wennberg, C. Roehl, D. Wunch, G. C. Toon, J.-F. Blavier, R. A. Washenfelder, G. Keppel-Aleks, N. Allen, and J. Ayers. TCCON data from park falls (us), release ggg2014.r1, 2017. Dataset.

[139] P. O. Wennberg, C. M. Roehl, J.-F. Blavier, D. Wunch, and N. T. Allen. TCCON data from jet propulsion laboratory (us), 2011, release ggg2014.r1, 2017. Dataset.

[140] P. O. Wennberg, D. Wunch, C. Roehl, J.-F. Blavier, G. C. Toon, N. Allen, P. Dowell, K. Teske, C. Martin, and J. Martin. TCCON data from lamont (us), release ggg2014.r1, 2016. Dataset.

[141] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

[142] J. R. Worden, G. Doran, S. Kulawik, A. Eldering, D. Crisp, C. Frankenberg, C. O'Dell, and K. Bowman. Evaluation and attribution of oco-2 xco_2 uncertainties. *Atmospheric Measurement Techniques*, 10:2759–2771, 2017.

[143] D. Wunch, G. C. Toon, J.-F. L. Blavier, R. A. Washenfelder, J. Notholt, B. J. Connor, D. W. T. Griffith, V. Sherlock, and P. O. Wennberg. The total carbon column observing network. *Philosophical Transactions of the Royal Society A*, 369(1943):2087–2112, 2011.

[144] D. Wunch, G. C. Toon, P. O. Wennberg, S. C. Wofsy, B. B. Stephens, M. L. Fischer, O. Uchino, J. B. Abshire, and *et al.* Calibration of the total carbon column observing network using aircraft profile data. *Atmospheric Measurement Techniques*, 3:1351–1362, 2010.

[145] D. Wunch, P. O. Wennberg, G. C. Toon, B. J. Connor, B. Fisher, G. B. Osterman, C. Frankenberg, L. Mandrake, C. O'Dell, P. Ahonen, S. C. Biraud, R. Castano, N. Cressie, D. Crisp, N. M. Deutscher, A. Eldering, M. L. Fisher, D. W. T. Griffith, M. Gunson, P. Heikkinen, G. Keppel-Aleks, E. Kyrö, R. Lindenmaier, R. Macatangay, J. Mendonca, J. Messerschmidt, C. E. Miller, I. Morino, J. Notholt, F. A. Oyafuso, M. Rettinger, J. Robinson, C. M. Roehl, R. J. Salawitch, V. Sherlock, K. Strong, R. Sussmann, T. Tanaka, D. R. Thompson, O. Uchino, T. Warneke, and S. C. Wofsy.

A method for evaluating bias in global measurements of co_2 total columns from space. *Atmospheric Chemistry and Physics*, 11:12317–12337, 2011.

[146] Debra Wunch, Geoffrey C Toon, Jean-François L Blavier, Rebecca A Washenfelder, Justus Notholt, Brian J Connor, David WT Griffith, Vanessa Sherlock, and Paul O Wennberg. The total carbon column observing network. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1943):2087–2112, 2011.

[147] Debra Wunch, Paul O Wennberg, Gregory Osterman, Brendan Fisher, Bret Naylor, Coleen M Roehl, Christopher O'Dell, Lukas Mandrake, Camille Viatte, Matthäus Kiel, et al. Comparisons of the orbiting carbon observatory-2 (oco-2) xco2 measurements with tccon. *Atmospheric Measurement Techniques*, 10(6):2209–2238, 2017.

[148] Fengxin Xie, Tao Ren, Changying Zhao, Yuan Wen, Yilei Gu, Minqiang Zhou, Pucai Wang, Kei Shiomi, and Isamu Morino. Fast retrieval of xco2 over east asia based on orbiting carbon observatory-2 (oco-2) spectral measurements. *Atmospheric Measurement Techniques*, 17(13):3949–3967, 2024.

[149] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR, Jul 2020.

[150] T Yokota, Y Yoshida, N Eguchi, Y Ota, T Tanaka, H Watanabe, and S Maksyutov. Global concentrations of co2 and ch4 retrieved from gosat: First preliminary results. *Sola*, 5:160–163, 2009.

[151] L. Zhang, K. J. Davis, A. E. Schuh, A. R. Jacobson, S. Pal, Y. Y. Cui, D. Baker, S. Crowell, F. Chevallier, M. Remaud, J. Liu, B. Weir, S. Philip, M. S. Johnson,

F. Deng, and S. Basu. Multi-season evaluation of co_2 weather in oco-2 mip models. *Journal of Geophysical Research: Atmospheres*, 127(2):e2021JD035457, 2022.