Response to Reviewer #2

A non-linear data driven approach to bias correction of XCO2 for OCO-2 NASA ACOS version 10

William Keely et al.

We sincerely thank the reviewer for their time in giving thorough feedback and suggestions. The points raised have led to significant changes to the manuscript. The points from the reviewer are shown in black, with our response in blue, and changes or additions to the manuscript in red.

Could the authors clarify if this understanding is correct? There are two XGBoost models (one for ocean and one for land) trained on data from all three proxy datasets that are the main contribution of this paper and these two models are used for Table 4, Table 5, Figure 3, Figure 4, Table 6, Figure 5, Table 7, Figure 7, Figure 8, and Figure 10. An additional six XGBoost models (one for each of the two surface types and three proxy datasets) are trained for Figures 2 and 9 to understand feature importance, but these models are not applied elsewhere. If this is correct (or incorrect), I believe Section 3.4 on Experiment Design could make this more clear.

This is correct, our final proposed bias correction is a land model and an ocean model trained on all three truth proxies. To clarify this, we have re-written and streamlined the Experiment Design section and throughout the manuscript.

3.4 Experiment Design

First, the footprint correction as described in O'Dell et al., 2018 is applied to the training and evaluation datasets. We then evaluate two methods for bias correcting retrieved XCO_2: a non-linear machine learning model called XGBoost; and as a baseline, we also train a MLR model 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 non-linear model and baseline linear method to properly assess that improved fit is coming from the captured non-linearity 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 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 as 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 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 labelled QF = 1 where non-linear relationships between ΔXCO_2 and predictors become more pronounced. Finally, we derive a new quality flag (QFNew) used in conjunction with the non-linear correction to that increases the throughput of well corrected data while maintaining similar error metrics as the operational filter and correction.

The authors seem to go through a lot of trouble to produce the XGBoost models for different proxy types for Figures 2 and 9, but there is little discussion other than Lines 209-213, Lines 399-401, and Lines 405-406 which take the form of "they are different for different proxy types." I suggest the authors either simplify the feature importance discussion to the XGBoost models used for the rest of the paper (trained on all proxy data and just divided by land/water) or improve the discussion related to information from different proxy datasets.

In the Feature Selection section we have simplified the wording to a more succinct explanation of the feature selection process, including a Figure 2 that now only shows the final selected variables for the proposed land and ocean models.

4.1 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 ΔXCO_2 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 we (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 2. The resulting features used in the final models and a brief description is shown in Table 3.

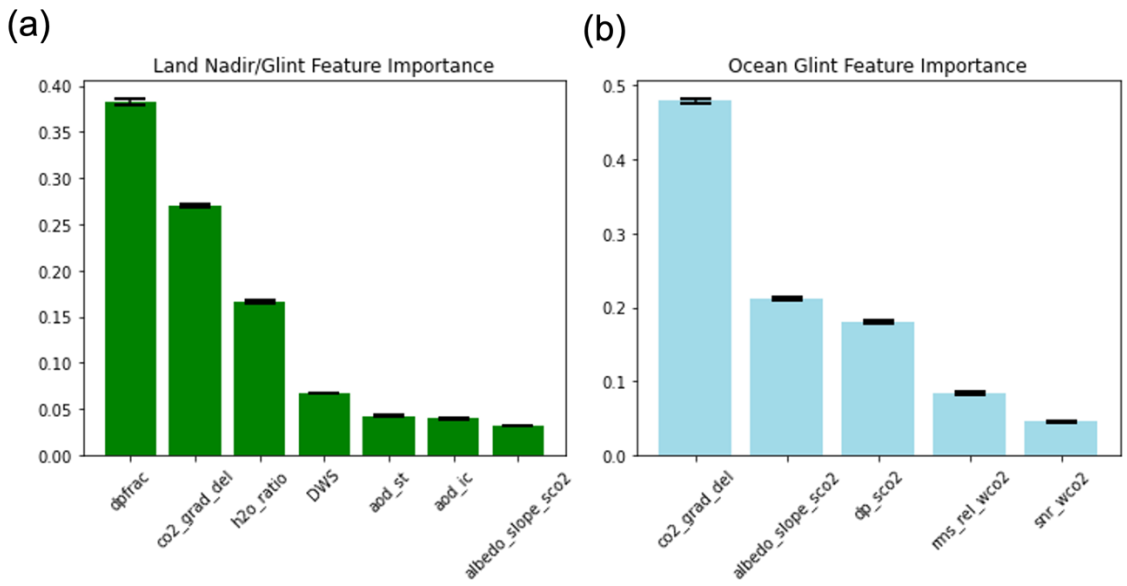

Figure 2. Feature importance for final land model trained with all proxies (a), feature importance for final ocean model trained with all proxies (b). Error bars denote variance in feature importance across 10 runs with different random seeds.

Section 5.1 (now 5.2) more thoroughly explores feature importance between filtering regimes and truth proxies.

5.2 Evaluating feature importance between filter regimes

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 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 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 B10) to date. These variables are responsible for the largest reduction in unexplained variance in the filtered regime (Payne et al. 2022; Osterman et al. 2020; O'Dell et al. 2018)

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 ΔXCO_2. In the QF=1 regime, h2o_ratio corresponds to a significant negative bias. Larger values of h2o_ratio are explained in Taylor et al. 2016, where it was shown that retrieved surface albedo from the strong CO_2 band is generally lower than the weak CO_2 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 CO_2 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 ΔXCO_2. Similarly, to land, the albedo term for the strong CO_2 band more informative for correcting outside the filtered regime along with the residual error between forward modelled radiances and measurements in the weak CO_2 band.

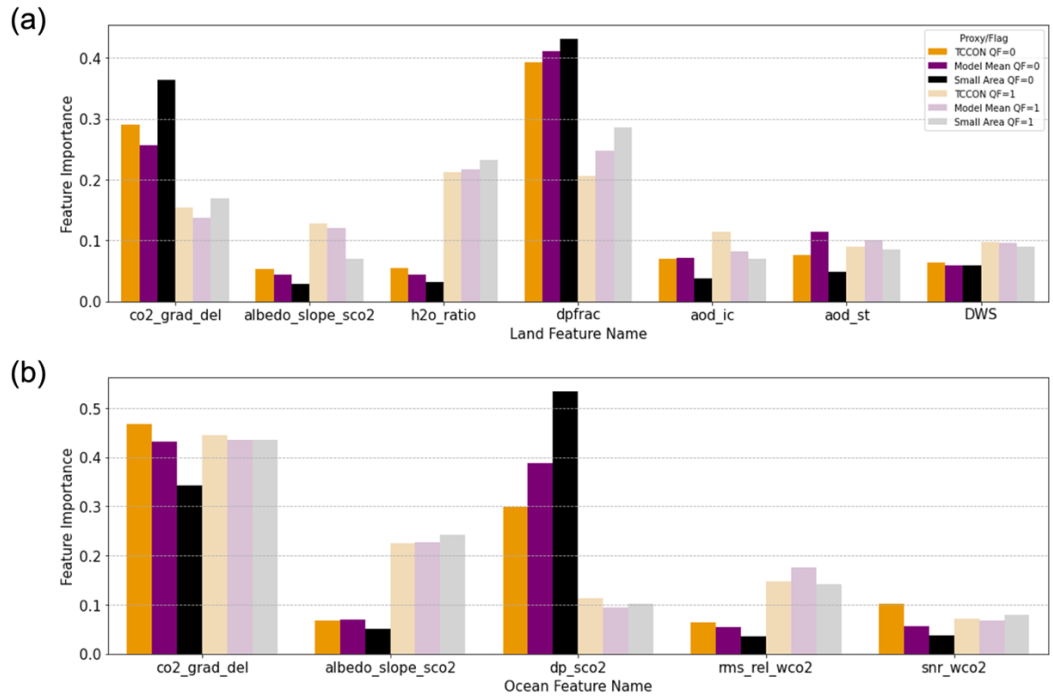

Figure 9: Feature importance for land is shown in (a), feature importance for ocean is shown in (b). Y-axis displays the normalized information gain from XGBoost models with QF=0 shown in darker colours and QF=1 shown in lighter colours.

It would be very helpful to have a table (could be in the main text or an Appendix) defining all of the variables discussed in this paper. Ideally this table would contain all of the variables considered for all of your models (I believe the "subset of 27 co-retrieved state vector variables" stated in Line 202). This table would be useful for a few reasons. (1) While most variables are already defined in Table 3, some of the variables used for QFNew in Figures 7 and 8 are never defined (e.g., max_declocking_3). (2) It would be useful to have a little more information about the variables, such as how co2_grad_del is calculated (Equation 5 from O'Dell2018).

We have added Table C1 to the appendix that includes all variables considered or used in the operational and proposed bias correction and quality filters.

Table C1. Features used or considered for the operational and proposed bias correction and filtering.

State Variable	Description	Used for: [B10 BC, ML BC, QF, QFNew]

dpfrac	Surface pressure difference that considers smaller dry air columns over higher elevations (Kiel et al. 2019).	B10 BC, ML BC, QF, QFNew
h2o_ratio	Ratio of retrieved H_2O column in weak and strong CO_2 bands by IMAP-DOAS.	ML BC, QF, QFNew*
DWS	Additive combination of retrieved dust, water, and sea salt aerosol optical depth.	B10 BC, ML BC, QF, QFNew
aod_stratear	Retrieved upper tropo+stratospheric aerosol optical depth at 0.755 microns.	ML BC, QF, QFNew
aod_ice	Retrieved ice cloud optical depth at 0.755 microns.	ML BC, QF, QFNew*
co2_grad_del	Large unphysical variation between the retrieved vertical CO_2 profile and prior.	B10 BC, ML BC, QF, QFNew
dp_sco2	Surface pressure difference between the retrieved and prior, evaluated for the strong CO_2 band location on the ground.	B10 BC, ML BC, QF, QFNew*
snr_wco2	The estimated signal-to-noise ratio in the continuum of the weak CO_2 band.	ML BC, QF, QFNew
co2_ratio	Ratio of retrieved CO_2 column in the weak and strong CO_2 bands by IMAP-DOAS	QF, QFNew*
altitude_stddev	The standard deviation of the surface elevation in the target field of view. Unit is in meters.	QF, QFNew*
max_declocking_wco2	An estimate of the absolute value of the clocking error in the weak CO_2 band expressed as a percent.	QF, QFNew*
max_declocking_sco2	An estimate of the absolute value of the clocking error in the strong CO_2 band expressed as a percent.	QF, QFNew*
dp_o2a	The difference in retrieved surface pressure to O_2A surface pressure prior.	QF, QFNew
dp_abp	The difference in the retrieved surface pressure to the fast O_2A band pre-processor retrieval.	QF, QFNew*
albedo_slope_sco2	Retrieved strong band reflectance slope(land) or slope of Lambertian albedo component of BRDF (ocean).	ML BC, QF, QFNew
albedo_slope_wco2	Slope of the weak CO_2 band albedo with respect to wavenumber.	QF, QFNew*
albedo_sco2	Surface reflectance at a reference wavelength in the strong CO_2 band in the primary scattering geometry from the retrieved BRDF (land). Retrieved Lambertian albedo (ocean).	QF, QFNew*
albedo_quad_sco2	Quadratic coefficient of the albedo_sco2 term with respect to wavenumber (land only).	QF, QFNew
albedo_quad_wco2	Quadratic coefficient of the albedo_wco2 term with respect to wavenumber (land only).	QF, QFNew
aod_total	Retrieved aerosol optical depth of cloud and aerosol at 0.755 microns.	QF, QFNew*
rms_rel_sco2	RMSE of the L2 fit residuals in the strong CO_2 band relative to the signal.	QF, QFNew

rms_rel_wco2	RMSE of the L2 fit residuals in the weak CO_2 band relative to the signal.	ML BC, QF, QFNew
detlaT	Retrieved offset to prior temperature profile in Kelvin.	QF, QFNew
aod_sulfate	Retrieved aerosol optical depth of sulfate aerosol at 0.755 microns.	B10 BC, QF, QFNew
aod_oc	Retrieved aerosol optical depth of organic carbon aerosol at 0.755 microns.	B10 BC, QF, QFNew
aod_water	Retrieved aerosol optical depth of water aerosol at 0.755 microns.	QF, QFNew
dust_height	Retrieved central pressure of the dust aerosol layer, relative to the retrieved surface pressure.	QF, QFNew
aod_seasalt	Retrieved aerosol optical depth of sea salt aerosol at 0.755 microns.	QF, QFNew
Fs_rel	Retrieved fluorescence relative to the O_2A band continuum signal.	QF, QFNew
chi2_wco2	Reduced chi-squared value of the L2 fit residuals for the weak CO_2 band.	QF, QFNew
windspeed	Retrieved surface wind speed over water surfaces.	QF, QFNew
water_height	Retrieved central pressure of the cloud water layer, relative to the retrieved surface pressure.	QF, QFNew

We have also provided more information for co2_grad_del. We have also added clarification to section 4.1 as shown in a response farther below.

$$co2\_grad\_del = [CO_{2,ret}(1) - CO_{2,ret}(0.6316)] - [CO_{2,prior}(1) - CO_{2,prior}(0.6316)] \quad (4)$$

For the machine learning model evaluation, it is my impression only 2018 should be used (the testing dataset). It is not clear in Figures 3/4/7/8/10 what date range is being used, but it should in theory be only 2018 since the model has been trained with the data for other years and the goal is to see how generalizable the model is to data it has never seen. This is concerning for Figure 5, which tries to evaluate the model using (in part) data from 2016 and 2017 that the model was already trained on. This could suggest corrections that are overly optimistic.

We have clarified the date ranges used in figures throughout the manuscript. For Figure 5 - to increase the amount of data for plotting we use three models. Each model is used to infer bias for a validation year that is held out during model training.

Figure 5. Remaining XCO_2 biases (ΔXCO_2) after correction for 2016-2018 and model mean proxy, binned to a 2°x2° resolution. ΔXCO_2 after the XGBoost correction for QF=0 is shown in (a), ΔXCO_2 after the B10 correction for QF=0 is shown in (b), ΔXCO_2 after the XGBoost correction for QF=1 is shown in (c), ΔXCO_2 after the B10 correction for QF=1 is shown in (d), and difference (B10 –

XGB) for QF=0 is shown in (e). Three models are trained each with one year in [2016,2017,2018] used as holdout. The results on the holdout sets are then used for plotting.

To go further with validation, have the authors considered leaving one of these proxies (like TCCON) completely out of their bias correction (or bringing in an independent dataset)? It seems to me that at the end of this, you have no independent datasets to evaluate your bias-corrected retrievals with, as they have all been used for training the model.

We now address the potential circularity induced by the lack of a truly independent proxy in section 5.1. The issue of lack of independent truth proxy in the operational correction is discussed in Taylor et al., 2023 and is still a current research focus by the OCO science team. We evaluate bias correction models trained on two of the proxies and evaluate on the third withheld proxy.

5.1 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 Taylor et al. 2023 and comparison and selection independent validation data sets is still an open area of study. The risk of overfitting due to circularity become 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 7, for land, and Figure 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.

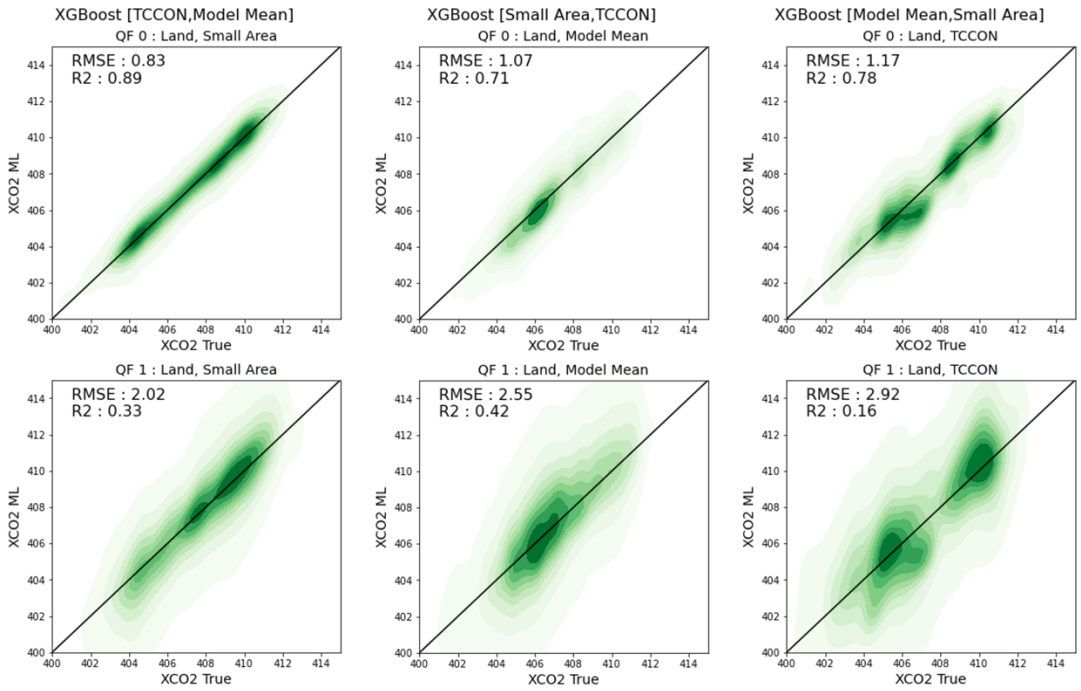

Figure 7. Comparison of XCO_2 derived from truth proxy (XCO_2 True) vs. XCO_2 corrected by XGBoost (XCO_2 ML) for land by hold out proxy set and hold out year (2018). Left-most column displays results of a XGBoost model trained on [TCCON, Model Mean] and evaluated on Small Area. Middle column displays 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 proxy and QF=0 is shown in the top row and QF=1 in the bottom.

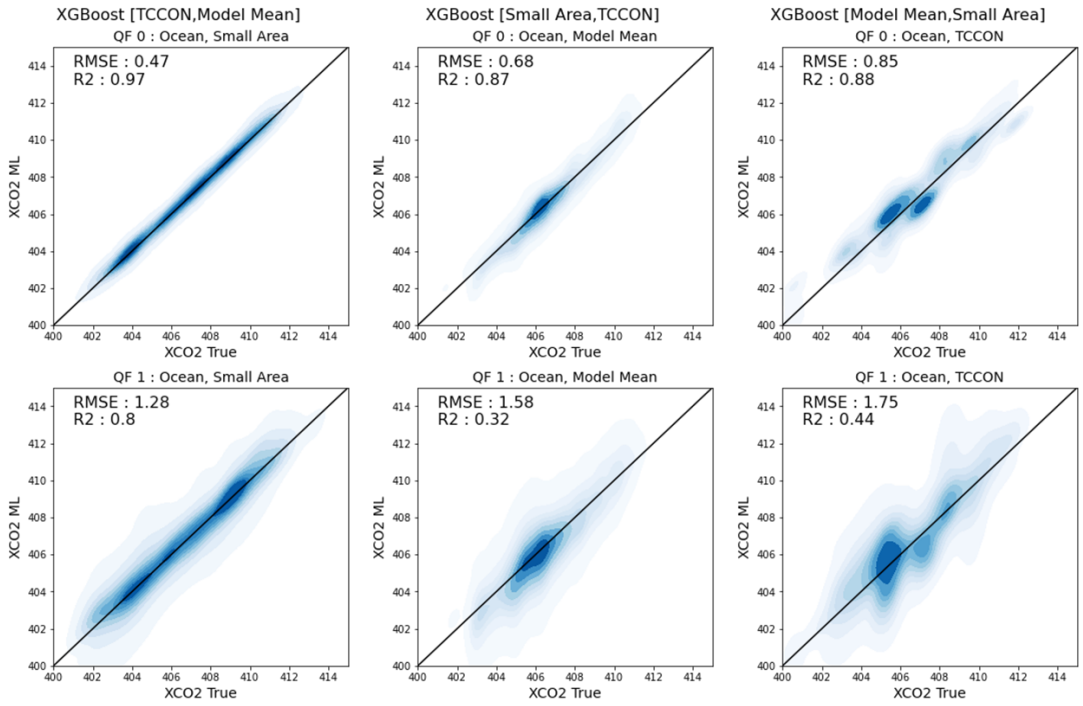

Figure 8. Comparison of XCO_2 derived from truth proxy (XCO_2 True) vs. XCO_2 corrected by XGBoost (XCO_2 ML) for ocean by hold out proxy set and hold out year (2018). Left-most column displays results of a XGBoost model trained on [TCCON,Model Mean] and evaluated on Small Area. Middle column displays 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 proxy and QF=0 is shown in the top row and QF=1 in the bottom.

Reference added: ADD TO BOTH RESPONSES

For Table 1 and the rest of the paper, what version of TCCON data is used here? Presumably GGG2014 given the reference list. Is there a reason to not use GGG2020 at this point? It is my understanding that OCO-2 B10 uses the same prior as GGG2020, so this would be more appropriate. If not, it might be necessary to account for the difference in priors when comparing GGG2014 and OCO-2 data in calculating deltaXCO2 (if this effect is large).

The operational correction for B10 is fit using GGG2014. In order to provide a faithful comparison to the operational correction we also use GGG2014. We now adequately clarify this in the manuscript.

2.1 TCCON truth proxy

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 XCO_2 (Wunch et al., 2010). 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 offering non-uniform point comparisons. We use the same dataset as the operational correction consisting of OCO-2 soundings co-located TCCON GGG2014 measurements (Wunch et al., 2017; Wunch et al., 2011) in space (2.5º lat, 5º lon) and time (2h).

I am suspect of the authors' claim (Lines 65, 438) that this method is "reproducible." There is still some hand-tuning in these methods, including picking which variables to include for the regression task (How do you reconcile the different proxies saying different variables are important in Figure 2? How do you pick which redundant variables to drop based on correlation before doing the analysis in Figure 2?) and how to adjust the filters for QFNew. This is fine, but with no code published alongside the paper, this could be difficult to reproduce.

We fully agree with this statement. Subjectivity is still largely present in the selection of variables and hand tuning of filter thresholds. We therefore only claim that the framework can be adapted to future algorithm updates.

Is there a plan to incorporate this into future versions of the OCO-2 data? Regardless, will the authors be making available the bias-corrected data produced in this paper?

A machine learning bias correction is planned for a future lite file update for B11 that expands on the approach discussed in this work. While there are currently no plans for a machine learning correction for B10 we plan on providing the data used in this paper.

Specific Comments

Line 59: I am not sure if "relative to the operational linear correction" is accurate with respect to the TROPOMI methane retrieval in Schneising et al. (2019).

Line 61: a slightly longer discussion of how this work differentiates from Mauceri2023 could be appropriate.

Line 59-61: We now address both comments in the paragraph.

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 ΔXCO_2. Due to loss of data, the bias correction and quality filter are often disregarded for local studies (Nassar et al., 2017; Mendonca et al., 2021) or too limiting for certain regions (Jacobs et. al., 2020). Applying non-linear machine learning techniques have shown great promise for the task of bias correction for GOSAT/GOSAT-2 (Noël et al., 2022) and TROPOMI (Schneising et al., 2019). Specific correction of 3D cloud biases for OCO-2 retrieved XCO_2 (Massie et al. 2016) using a non-linear method fit on a small set of features correlated with 3D cloud effects in addition to the linear operational correction, is demonstrated in Mauceri et al. (2022).

Line 70: are the averaging kernels taken into account when comparing OCO-2 to TCCON or the model atmospheres?

The averaging kernels are used for both the TCCON truth proxy and concentration fields from the flux models proxy. We have added an additional citation to the paragraph that further explains the application of the OCO-2 averaging kernel in the derivation of the truth proxies.

To develop a bias correction, we define three truth proxy data sets for the true atmospheric column mode fraction. ΔXCO_2 is then set as the difference between the raw ACOS retrieval of XCO_2 and the truth proxy estimate of XCO_2 as shown in Eq. 1. For the TCCON and model mean truth proxies, the OCO-2 kernel is also applied as described in Taylor et al., 2023.

Figure 1: The "N = 1022636" title for Figure 1a seems to be a typo. I would expect the number of soundings here to be close to the total number of OCO-2 soundings since model grids are continuous in space and time (depending on how often the 1.5 ppm threshold is passed) and thus N for the model proxy should be > N for TCCON, not equal. On a related note, what is the total number of OCO-2 soundings considered to give a sense of the percentage used for each proxy dataset?

We have corrected the OCO-2 sounding count in the figure. The soundings come a quick test set (QTS) which is approximately 5% of the full OCO-2 record as described in Taylor et al., 2023.

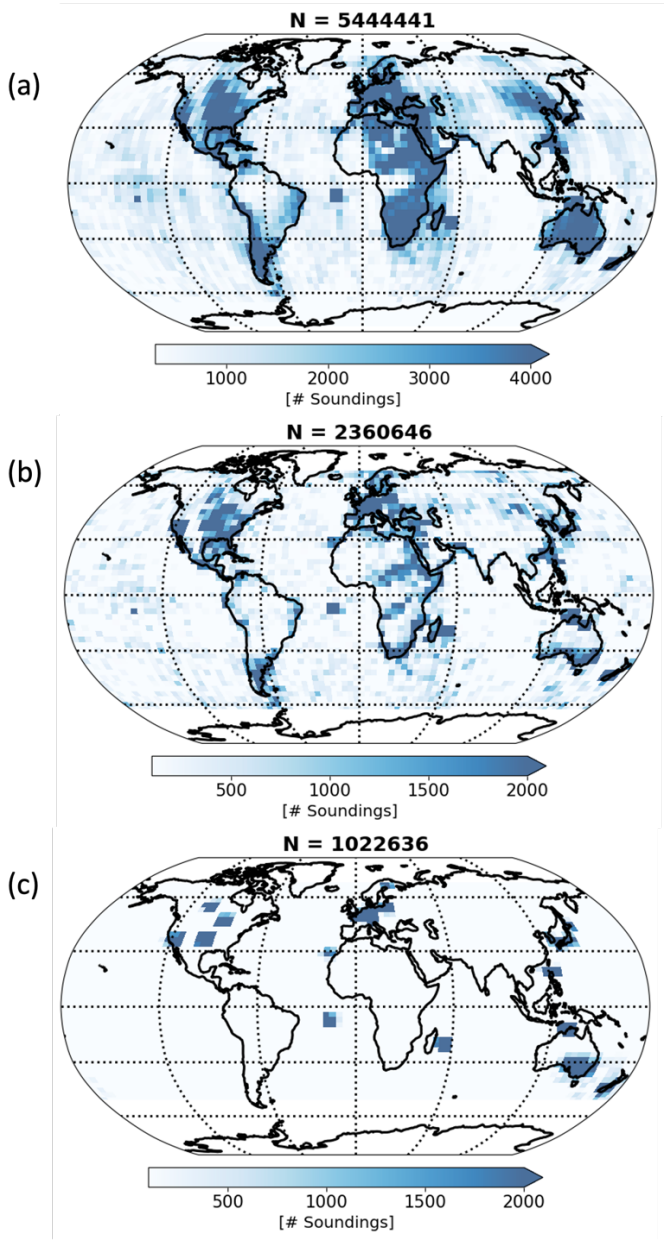

Figure 1: Spatial coverage for each truth proxy. The mean of a set of flux models is shown in (a), small area approximation is shown in (b), and TCCON is shown in (c).

The soundings considered come a quick test set (QTS) which is approximately 5% of the full OCO-2 record as described in Taylor et al., 2023.

Table 1 contains errors for the Tsukuba altitude, Sodankylä altitude, Izaña altitude, Wollongong latitude, Réunion latitude, Lamont latitude, and Karlsruhe latitude (and maybe others).

Table 1 has now been corrected.

Table 2: could you specify the resolutions of these models?

Table 2 now includes an additional column that with spatial and temporal resolution.

Table 2. Flux models used for the model mean truth proxy. TM5 – Transport model 5, TM3 – Transport model 3, LMDZ – Laboratoire de Meteorology, EnKF – Ensemble Kalman Filter, 4D-Var – 4 Dimensional Variation.

Model name	Institute	Transport model	Resolution [latxlonxtime]	Inverse method	Citation
CarbonTracker	NOAA Global Monitoring Laboratory	TM5	2ºx3ºx3h	EnKF	Peters et al. (2007) CarbonTracker (2021)
CarboScope	Max Planck Institute for Biogeocehmistry	TM3	4ºx5ºx6h	4D-Var	Rödenbeck (2005); Rödenbeck et al. (2018) CarboScope (2021)
CAMS	Copernicus Amosphere Monitoring Service	LMDZ	1.9ºx3.75ºx3h	4D-Var	Chevallier et al. (2010) CAMS (2021)

Line 128-129: it is my understanding (and you state as such in line 130) that XGBoost is not the average across an ensemble, but rather the sum across the ensemble.

Yes, thank you for catching this error. We have corrected this in the manuscript.

Line 135: how do you search for these hyperparameters? Are these the same for all of the XGBoost models discussed in this paper?

We now thoroughly explain the hyperparameter search and regularization values for the final proposed bias correction models in the paragraph.

XGBoost employs L_1 and L_2 norm regularization to reduce overfitting to outliers present in the training dataset. The effect of the regularization is governed by the hyper-parameters λ and γ, and must be carefully selected or tuned. To find these hyper-parameters we use a k-fold cross validation strategy in which the training dataset is divided into k subsets (we use k=10) and each subset is sequentially held out for evaluation for a model trained on the rest of the data. Performance across the k-folds is averaged and the process is repeated for each potential selection of hyper-parameters. We found a λ_{LAND}=2.5 and γ_{LAND}=3.75 for the land correction, and λ_{OCEAN}=2.0 and γ_{OCEAN}=10.0.

Line 179: Is data from 2014-2018 used for all three proxy datasets?

Yes, this is correct. Data range for all three proxies is from 2014-2018.

Line 184-185: In Section 3.3 and elsewhere, it is stated that the models are trained for Ocean G and Land NG (2 models). In these lines, it is suggested that there are three models.

We agree the final proposed correction is not clear and the section has been reworded. We have also worked to clarify this throughout the manuscript.

3.3 Training and test split

For training and evaluating the non-linear correction, we subset each of our truth proxy datasets into a training and testing datasets. First, datasets are split by the two surface types: ocean and land. In the B10, both operation modes (nadir and glint) are combined for the land bias correction due to low variance in feature importance between nadir and glint (O'Dell et al. 2018). 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.

Line 206: Is there a threshold you used for the correlation coefficient?

We now clarify this value in the revised paragraph.

To ensure robustness to correlation among features we (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.

Line 219: Is there a reason to specify that the prior pressure is from the strong band? Is the prior pressure different for the weak band?

Yes, this due to an alignment offset in the pointing location between the three bands and there for the priors are used.

Line 233: Why is the variable dp_sco2 considered? Lines 220-221 dp_frac discussed the disadvantages of this kind of pressure difference term.

dpfrac takes into account the surface elevation and is only available over land, while dp_sco2 is used for ocean glint scenes. We have clarified this in the Section 4.1:

4.1 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 ΔXCO_2 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 we (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 2. The resulting features used in the final models and a brief description is shown in Table 3.

Features used for operational correction are also highly informative for the proposed non-linear corrections and include the difference between the retrieved CO_2 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 (Kiel et al., 2019). The co2_grad_del is the change in profile shape and prior and is calculated as the difference dry air mole fraction at the surface, denoted as $CO_2(1)$, 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 $X_{CO2,raw}$ is the uncorrected retrieval of the column average and $P_{ap,SCO2}$ 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 Kulawik et al., 2019.

$$co2\_grad\_del = [CO_{2,ret}(1) - CO_{2,ret}(0.6316)] - [CO_{2,prior}(1) - CO_{2,prior}(0.6316)] \quad (4)$$

$$dpfrac = X_{CO2,raw}\left(1 - {P_{ap,SCO2}}/{P_{ret}}\right) \quad (5)$$

For land, the *h2o_ratio* is used and is the ratio of XH_2O estimated by single band retrievals from the strong and weak CO_2 bands separately using the IMAP-DOAS algorithm, which can differ from unity in the presence of atmospheric scattering (Taylor et al. 2016). 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_stratear*). The last feature used for land, as well as for ocean, is the albedo slope for the strong CO_2 band termed *albedo_slope_sco2*. This variable represents the slope of the reflectance across the strong CO_2 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 (Cox and Munk, 1954). 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 modelled radiance for the weak CO_2 to the measured radiance.

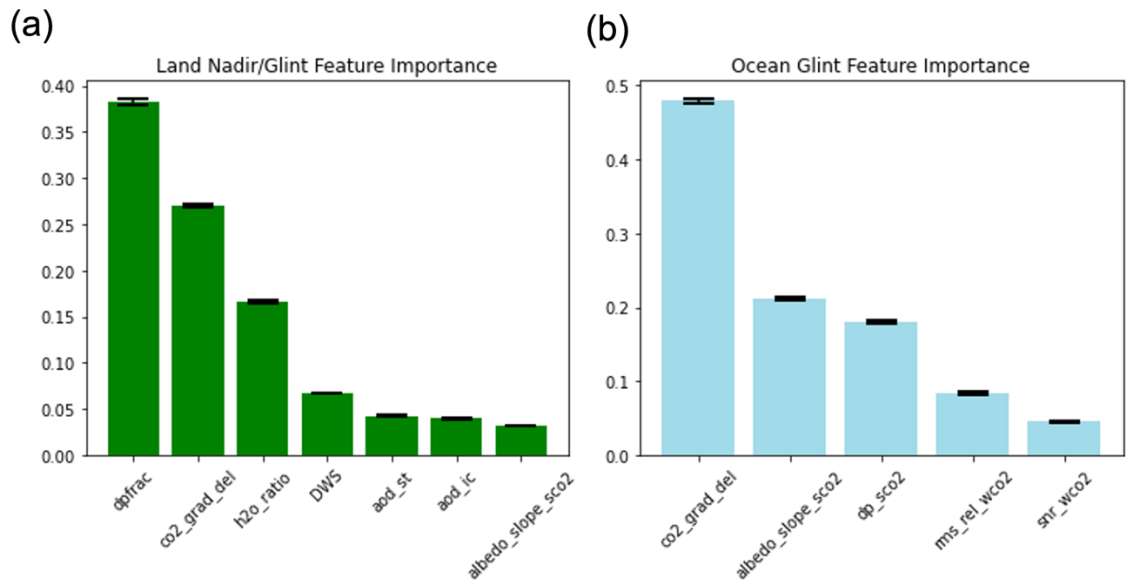

Figure 2. Feature importance for final land model trained with all proxies (a), feature importance for final ocean model trained with all proxies (b). Error bars denote variance in feature importance across 10 runs with different random seeds.

Figure 2: Why are there a different number of considered features for land and ocean? Were the same 27 variables (Line 202) started with and a different subset dropped for LandNG versus OceanG because of different correlations (Line 206) for the different operation modes?

Some features are not available or held constant over the ocean. We have simplified Figure 2 to show only the final feature importance of the proposed correction models, as seen above.

Line 276: what percentage of retrievals are filtered because of this?

h2o_ratio filters out ~10% of the soundings.

Figures 3 and 4: Is this just 2018 data? Could you add arrows to indicate the direction of the filters (or write the ranges like in Figures 7 and 8)?

We have added arrows to indicate the region assigned QF=0 or "good" quality data.

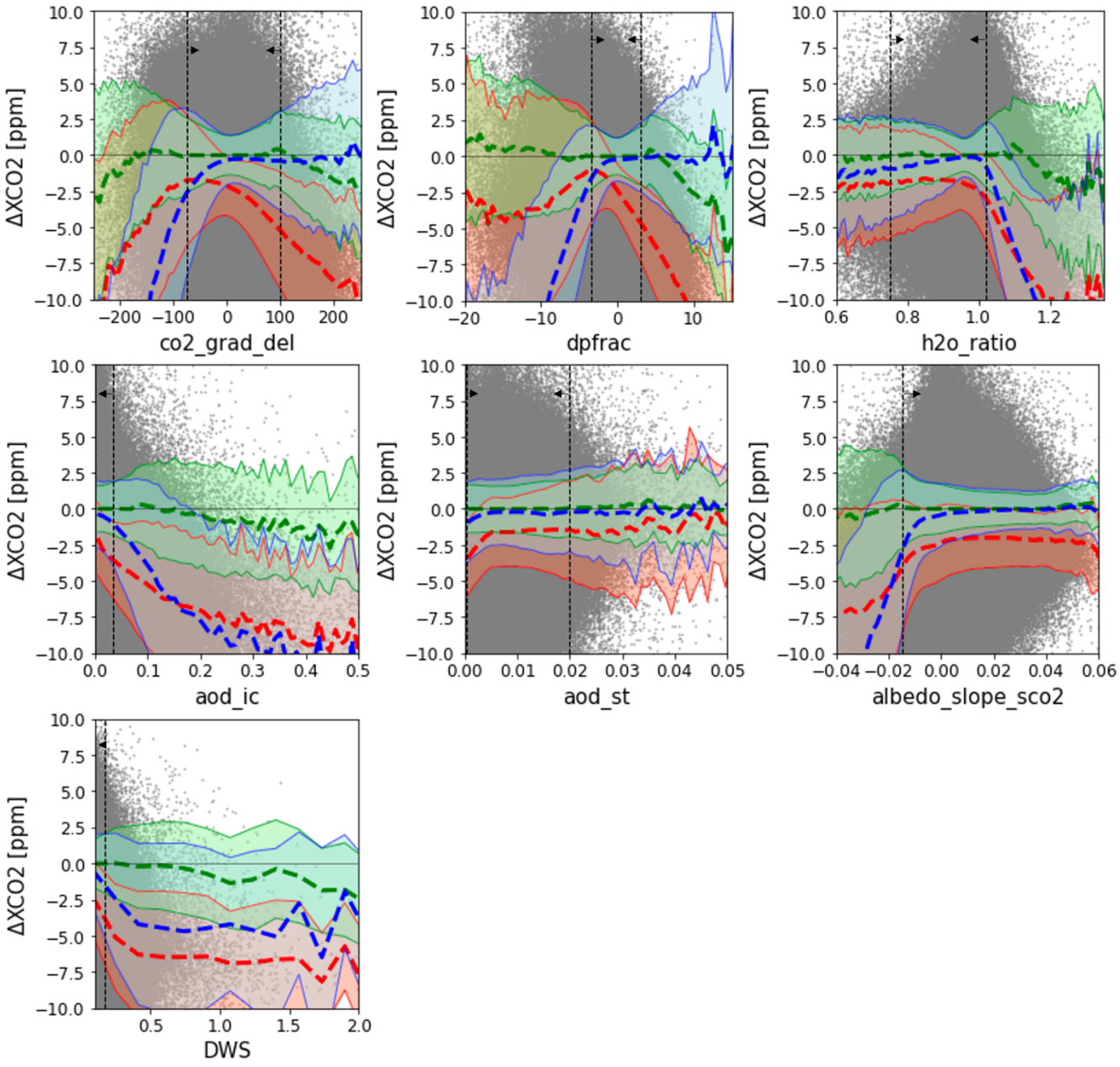

Figure 3. ΔXCO_2 vs land features for 2018. Mean interaction and 2σ Stddev for uncorrected ΔXCO_2 plotted in red, XGBoost corrected in green and B10 corrected in blue. The vertical black dotted lines indicate B10 QF filters and arrows point towards the region assigned QF=0. Individual soundings are shown with grey scatter.

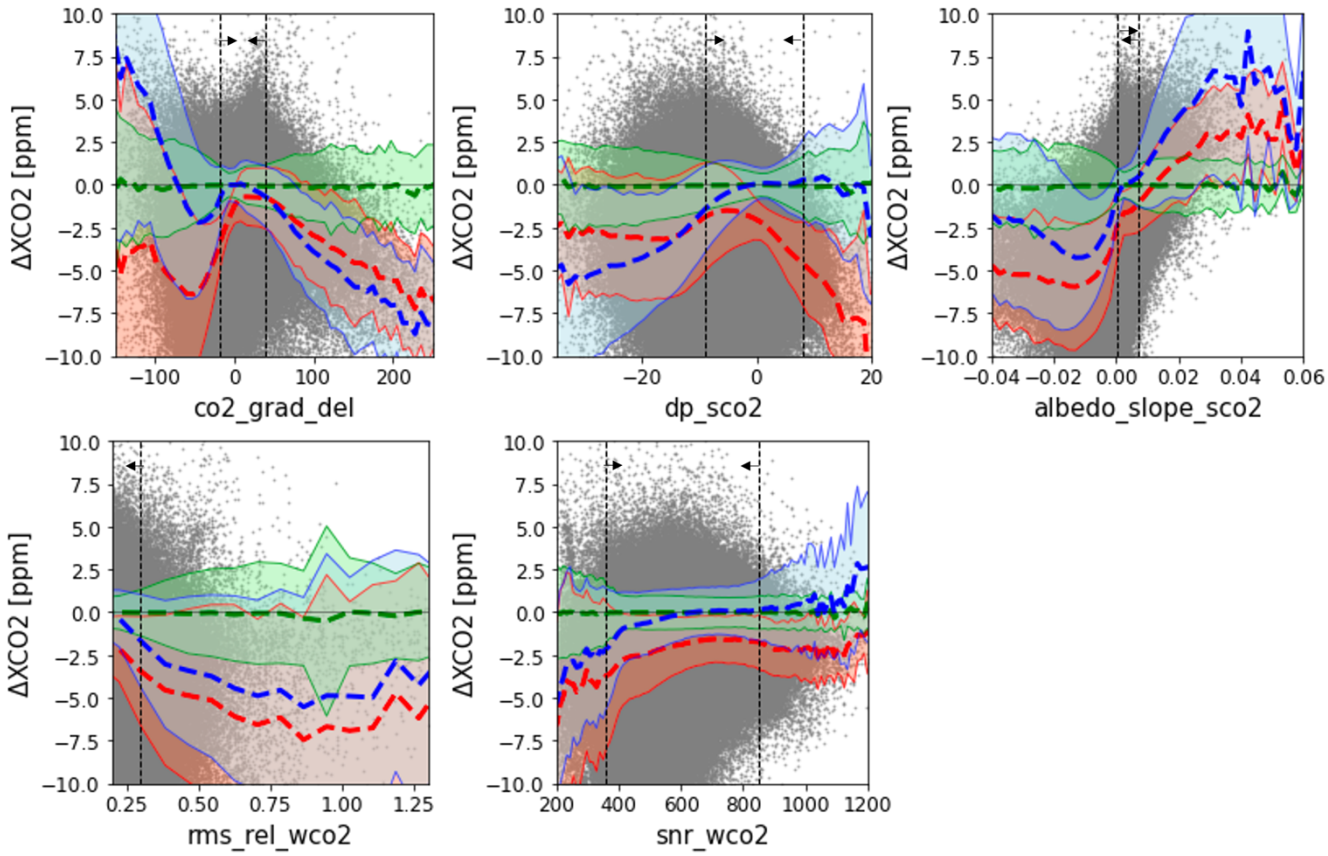

Figure 4. ΔXCO_2 vs ocean features for 2018. Mean interaction and 2σ Stddev for uncorrected ΔXCO_2 plotted in red, XGBoost corrected in green and B10 corrected in blue. The vertical black dotted lines indicate B10 QF filters and arrows point towards the region assigned QF=0. Individual soundings are shown with grey scatter.

Figure 5: It seems to me you cannot properly evaluate the model with the training data (2016-2017) and this plot should only show 2018 data. Additionally, I am interested to know if this is data from all proxy datasets? This would help answer the question of if the remaining differences are due to shortcomings in the bias correction method or in the proxy datasets.

Thank you for raising these points, this plot needed much clarification. The biases for each year in 2016-2017 are estimated by a model (3 in total) trained on all years except for one year within the range which was used for inference. We used this process to increase the amount of available data for plotting – only the model mean is plotted but is included in for training. To address generalization to a held proxy we now have added a new Section 5.1 as shown above.

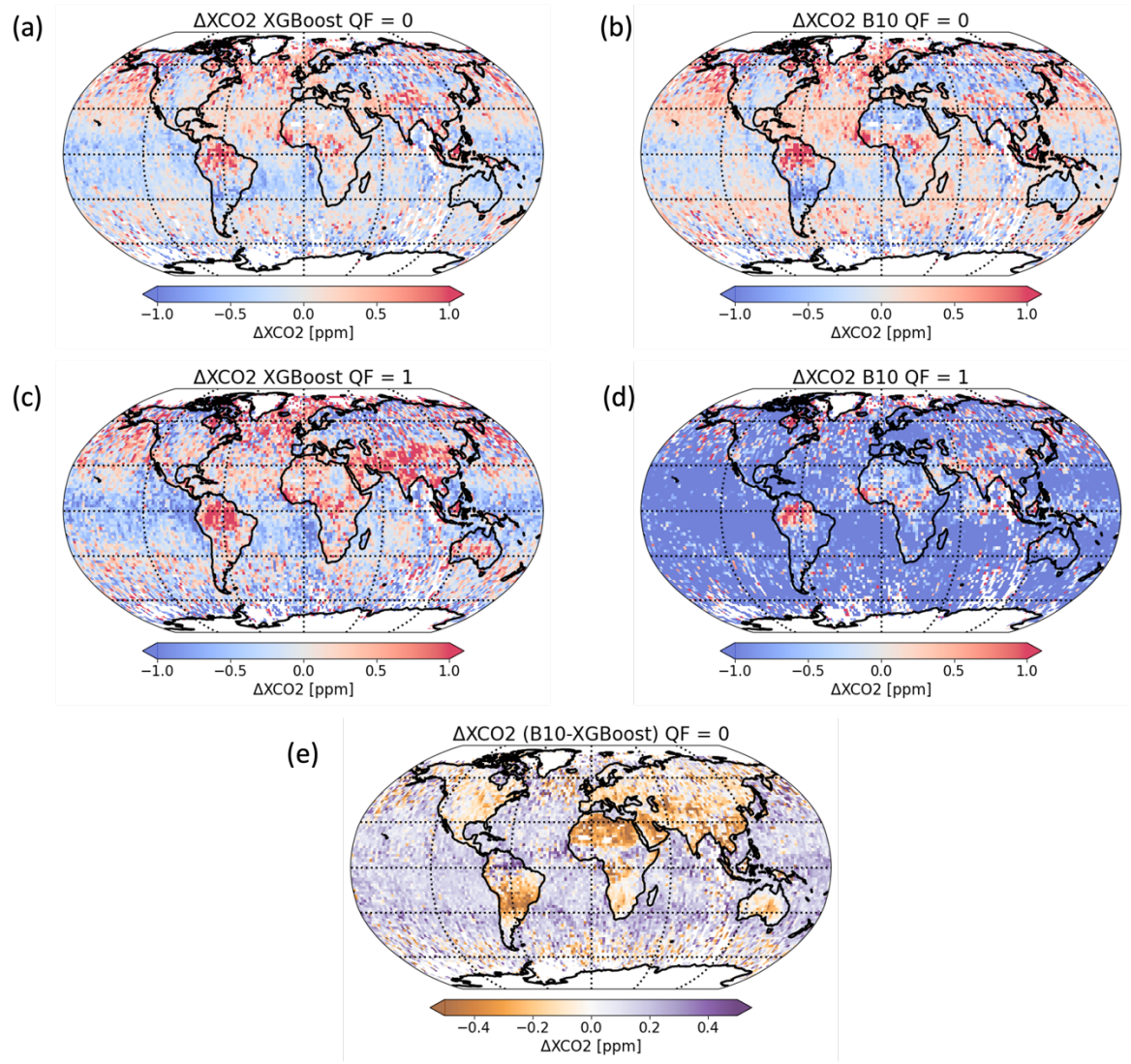

Figure 5. Remaining XCO_2 biases (ΔXCO_2) after correction for 2016-2018 and model mean proxy, binned to a 2°x2° resolution. ΔXCO_2 after the XGBoost correction for QF=0 is shown in (a), ΔXCO_2 after the B10 correction for QF=0 is shown in (b), ΔXCO_2 after the XGBoost correction for QF=1 is shown in (c), ΔXCO_2 after the B10 correction for QF=1 is shown in (d), and difference (B10 – XGB) for QF=0 is shown in (e). Three models are trained each with one year in [2016,2017,2018] used as holdout. The results on the holdout sets are then used for plotting.

Line 338: Is this trained on QF = 0 + 1 data and all three of the truth proxies? And two different models (one for land and one for ocean)?

Yes, this is correct. We have clarified that in the manuscript.

4.5 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 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 XCO_2 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 **16%** over the B10 QF while matching the RMSE of the B10 correction as shown in Table 5.

Figures 7/8: It is not clear to me what each of the colors represent. The black line is deltaXCO2 for the raw XCO2 retrievals. But between Line 348 and the figure captions, I can't figure out what the difference between the light and dark green/blue lines are (maybe dark is XGBoost bias-corrected and light is operationally bias-corrected?).

We have added clarification to the figure text. Figure 7 & 8 have now been moved to the appendix in order to stream line the section.

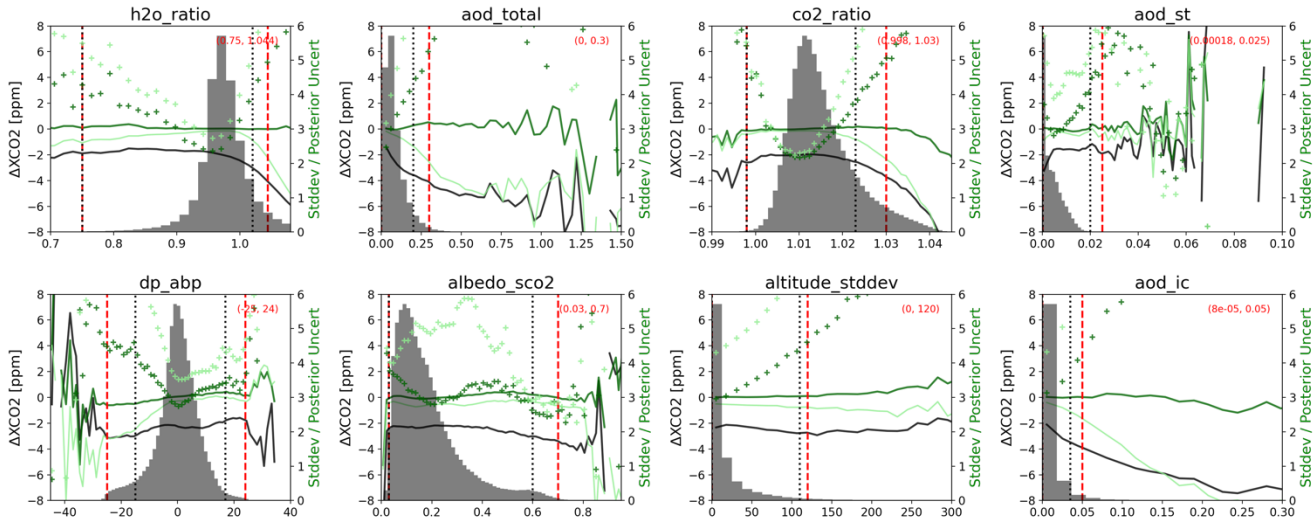

Figure B1: Variables selected for land QFNew: the difference between the uncorrected retrieval and the model mean truth proxy is shown with the black curve. The difference between the operational correction and the model mean truth proxy is shown in the light green curve. The difference after the non-linear correction is show by the dark green curve. The binned Std error divided by the posterior uncertainty of XCO2 is show by the green pluses and right y-axis. B10 QF filters are indicated by the black vertical dashed lines and QFNew is shown by the red dashed lines. Region of data denoted as QF=0 is contained within the red values in the paratheses.

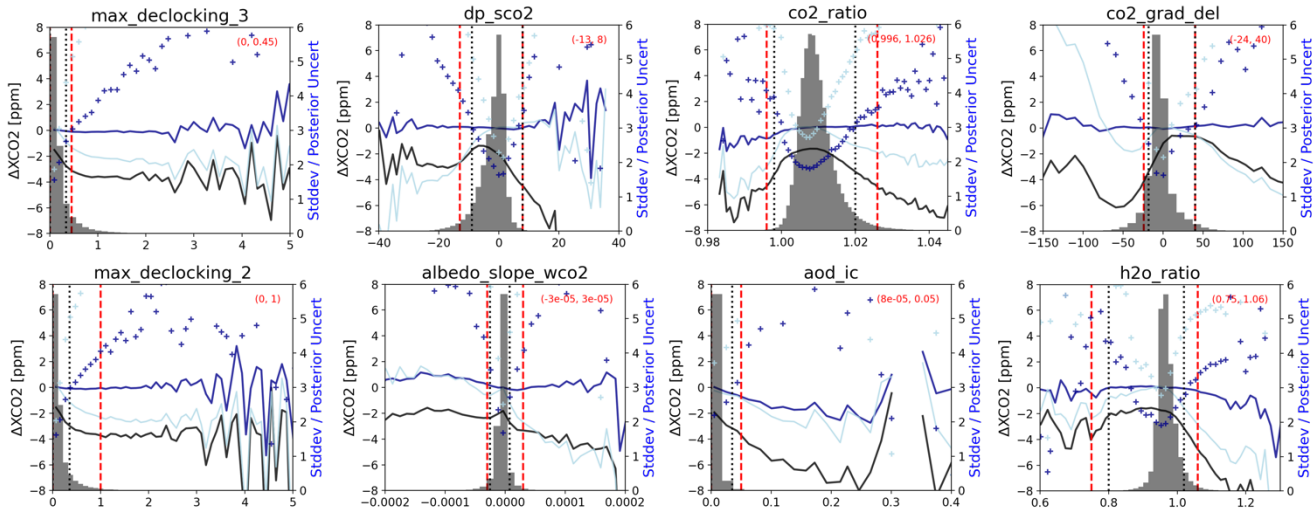

Figure B2: Variables selected for ocean QFNew: the difference between the raw retrieval uncorrected retrieval and the model mean truth proxy is shown with the black curve. The difference between the operational correction and the model mean truth proxy is shown in the light blue curve. The difference after the non-linear correction is show by the dark blue curve. The binned Std error divided by the posterior uncertainty of XCO2 is show by the blue diamonds and right y-axis. B10 filters are indicated by the black vertical dashed lines and a potential filter is shown by the red dashed lines. Region of data denoted as QF=0 is contained within the red values in the paratheses.

Figures 5/10: titles or different labels on the colormaps might make it more clear which plots are for operationally bias-corrected data and which are for XGBoost bias-corrected data (but this is clear in the caption).

We agree and have added titles as seen above.

Figure 10: Is this just for 2018? Is it for all proxy datasets?

The figure shows 2018 data, this is now clarified in the figure text. Note, Figure 10 is now Figure 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 X_{CO2} 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.

There are in-text references for Kuze2009, Palmer2019, Crowell2019, Peiro2021, Mendonca2021, Jacobs2020, Osterman2020, Worden2017, Taylor2012, Morino2018a, Morino2018b, and Hase2015 (and maybe others) in the text and tables that are missing in the Reference section.

Thank you for thoroughly checking for and catching these errors. We have now corrected these in the revised manuscript.

Technical Corrections

There are many typos throughout this paper. It needs to be significantly cleaned up before publication. I tried to catch as many as I could here.

Line 12: Obersvatory => Observatory

Corrected.

Line 15: correlate => correlated

Corrected.

Line 35: m). => m.

Corrected.

Line 36: Rogers => Rodgers

Corrected.

Line 42: Missing subscript on XCO2

Corrected.

Line 47: emperically => empirically

Corrected.

Line 53: reduce => reduces

Corrected.

Line 55: filter => filters

Corrected.

Line 55: correction. to which => correction to which

Corrected.

Line 57: missing word before “or too limiting”

Or is too limiting

Line 59: 2021 => 2022

Corrected.

Line 65: reproduceable => reproducible (or be consistent with Line 438)

Claims of reproducibility have been removed.

Line 66: remove “upcoming missions such as GeoCarb”

This has now been removed.

Line 68: mode => mole

Corrected.

Line 78: Lever => Level

Corrected.

Line 86: remove comma

Removed

Line 91: offering => offers

Corrected.

Table 1 caption: rephrase/missing words in “proxy the TCCON”

Table 1. TCCON sites used in bias correction and filtering for B10 ACOS.

Line 104: missing subscript on XCO2

Corrected.

Line 106: XCO2, is => XCO2 is

Removed.

Line 106: overs => offers

Corrected.

Line 110: Table 1 => Table 2

Corrected.

Line 121: employee => employ

Corrected.

Line 132: into => in

Removed.

Line 134: we hold out small subset => we hold out a small subset

Removed.

Line 145: variables not defined (though common notation is used)

The variables in Eq 3 are now explained in the text.

Line 172: remove “a”

Removed.

Line 176: mode; => mode,

Corrected.

Line 185: or features are => or features, are

Removed.

Line 186: This, allows =? This allows

Removed.

Line 192: data then, => data and then

Removed.

Line 205: delete “we”

Removed.

Line 213: Table 1 => Table 3

Corrected.

Line 223: retrivals => retrievals

Corrected.

Line 227 (and Table 3): aod_stratear => aod_strataer

Corrected.

Line 232: In addition to the albedo_slope_sco2, four => In addition to albedo_slope_sco2 and co2_grad_del, three

Removed.

Table 3 caption: features for in => features for use in

Corrected.

Table 3: dpfrac versus dp_frac is inconsistent between the text (Line 216) and here/other figures.

Corrected throughout to dpfrac to stay consistent with lite file variable naming.

Section 4.1 is repeated (4.1 Feature Selection, 4.1 Model evaluation for QF = 0).

Corrected.

Line 261: Table 2 => Table 4

Corrected.

Line 275: Sentence beginning with “However,” could be rephrased. The portion inside parentheses is confusing.

Sentence now reads:

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.

Line 280: Table 3 => Table 5

Corrected.

Figures 3/4: inconsistent x-axis labels with the text/Table 3 (e.g., aod_st versus aod_strataer). Arrows designating the direction of the filters (e.g., on the aod_st plot) could be helpful but are not necessary.

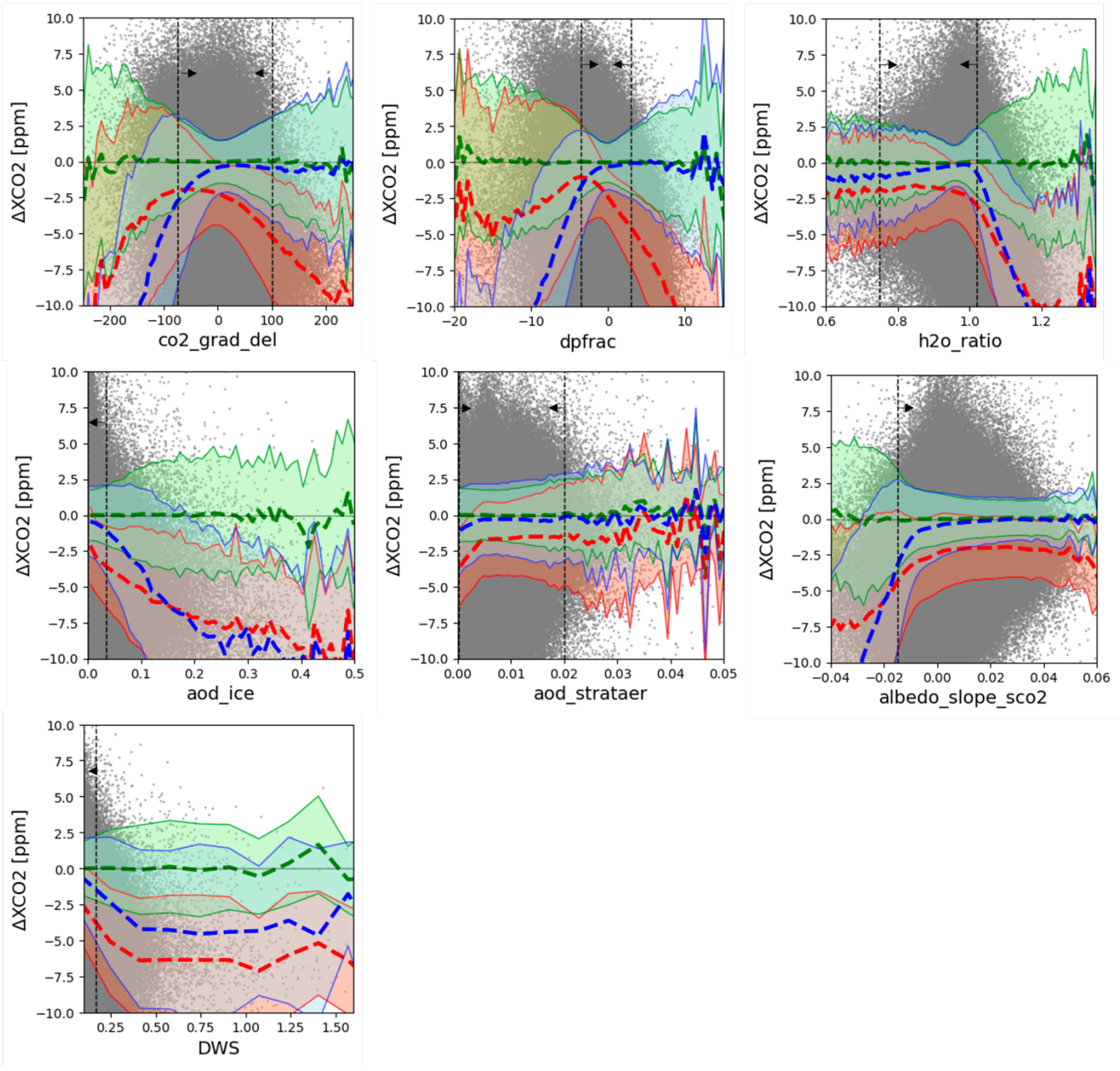

Figure 3. ΔXCO_2 vs land features for 2018. Mean interaction and 2σ Stddev for uncorrected ΔXCO_2 plotted in red, XGBoost corrected in green and B10 corrected in blue. The vertical black

dotted lines indicate B10 QF filters and arrows point towards the region assigned QF=0. Individual soundings are shown with grey scatter.

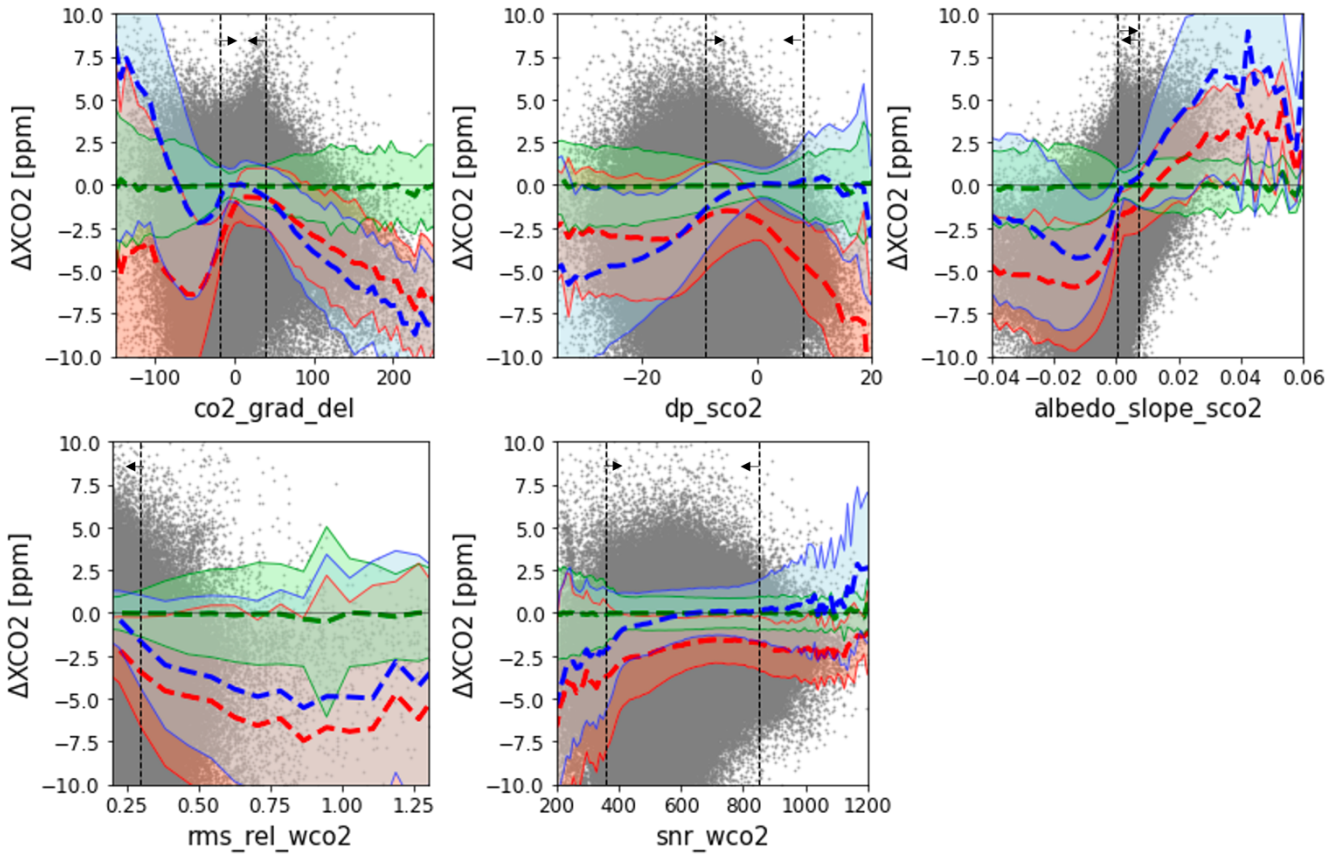

Figure 4. ΔXCO_2 vs ocean features for 2018. Mean interaction and 2σ Stddev for uncorrected ΔXCO_2 plotted in red, XGBoost corrected in green and B10 corrected in blue. The vertical black dotted lines indicate B10 QF filters and arrows point towards the region assigned QF=0. Individual soundings are shown with grey scatter.

Line 304: Table 4 => Table 6

Corrected.

Line 307: double-check percentages. For example, for ocean QF=0, ((0.67^2)-(0.61^2))/(0.67^2) = 17%, not 4% (if I correctly understand your methodology).

Thank you for catching this. We have updated these values.

Table 6: stddev => standard deviation; caption mentions raw XCO2 data, but this is not present in the table.

Raw retrieval values are now correctly included in Table 6.

Line 342: Table 5 => Table 7

Corrected.

Line 348: Figure 8 and 9 => Figures 7 and 8

Corrected to Figure B1 and Figure B2.

Line 354: benefit of quality => benefit of a quality

Corrected.

Table 7: Region/Truth Proxy => “Surface/Mode” (?); through put => throughput

Corrected.

Line 383: Qf => QF

Corrected.

Line 384: Figure 9 => Figure 10; Feaures => Features

Corrected.

Line 394: filter bound h2o_ratio => filter bounds, h2o_ratio

Corrected.

Line 400: noteably => notably

Corrected

Lines 415-416: Figure 11 => Figure 10

Figure 10 has now become Figure 11

Line 442: remove ; and rephrase

Corrected.

Line 449: ISS not defined in text

ISS is now defined correctly before abbreviation.

Title: data driven => data-driven. Hyphen usage should be reviewed throughout (Lines 12, 13, 19, 32, 52, 58, 87, 89, 103, 157, 164, 273, 336, 355, 421, 438, 447, etc.).

Corrected throughout the manuscript.

References:

Kulawik, S. S., O'Dell, C., Nelson, R. R., and Taylor, T. E.: Validation of OCO-2 error analysis using simulated retrievals, Atmos. Meas. Tech., 12, 5317–5334, https://doi.org/10.5194/amt-12-5317-2019, 2019.

Taylor, T. E., O'Dell, C. W., Baker, D., Bruegge, C., Chang, A., Chapsky, L., Chatterjee, A., Cheng, C., Chevallier, F., Crisp, D., Dang, L., Drouin, B., Eldering, A., Feng, L., Fisher, B., Fu, D., Gunson, M., Haemmerle, V., Keller, G. R., Kiel, M., Kuai, L., Kurosu, T., Lambert, A., Laughner, J., Lee, R., Liu, J., Mandrake, L., Marchetti, Y., McGarragh, G., Merrelli, A., Nelson, R. R., Osterman, G., Oyafuso, F., Palmer, P. I., Payne, V. H., Rosenberg, R., Somkuti, P., Spiers, G., To, C., Weir, B., Wennberg, P. O., Yu, S., and Zong, J.: Evaluating the consistency between OCO-2 and OCO-3 XCO_2 estimates derived from the NASA ACOS version 10 retrieval algorithm, Atmos. Meas. Tech., 16, 3173–3209, https://doi.org/10.5194/amt-16-3173-2023 , 2023.