Response to Reviewer #1

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.

In my opinion, the underlying machine learning method XGBoost is praised beyond measure (highly interpretable compared to other machine learning algorithms, improved predictive performance compared to Random Forest, highly robust to overfitting, ...). It is not trivial to prove such universal statements, and in the context of this paper it is not even necessary. On the other hand, some of these aspects are not sufficiently dealt with regarding the specific example of bias correction presented here. In this sense, please avoid questionable general statements and elaborate on the available results (the proposed bias correction) instead: How can the actually obtained model be interpreted? What is the most appropriate way to calculate feature importances aiming at maximising interpretability of the specific model in question? What is the ranking of the most important features (for land and ocean data)? Does the model overfit for the chosen parameters?

We fully agree with this statement and have removed the broader claims related to the method. We have added our full responses and manuscript changes to the in-depth points and comments below.

In order to assess (and exclude) potential overfitting tendencies, it might also be useful to define more challenging training and validation data sets or to use completely independent data for validation. The models are trained on data from 2014-2017 and are then evaluated on data from 2018. However, the validation data set is not entirely independent as the biases are likely similar in the statistical sense from year to year. It would be more instructive to leave out whole regions and/or proxy data sets in the training and only use them for validation. It would be interesting whether the figures and tables demonstrating the performance of the correction would change significantly as a result (e.g. Figures 3/4 or Table 6). For example, Figure 5 suggests that the correction does not generalise so well to regions that were rarely considered during training (e.g tropics, parts of South Asia and Canada, compare to Figure 1).

We agree with the need to assess how a machine learning bias correction will generalize to a truly independent data set. An additional section and new Figure 7 have been added to the discussion exploring the ability of the bias correction to generalize to a held-out truth proxy. We would like note that the temporal training/validation split employed in the manuscript is an improvement over the split used by the operational correction which fits on all proxies for a subset of the record and evaluates on data from the same years. Addressing the circularity in the current use of the truth

proxies and identifying new proxies for validation is an ongoing area of study by the OCO science team.

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.

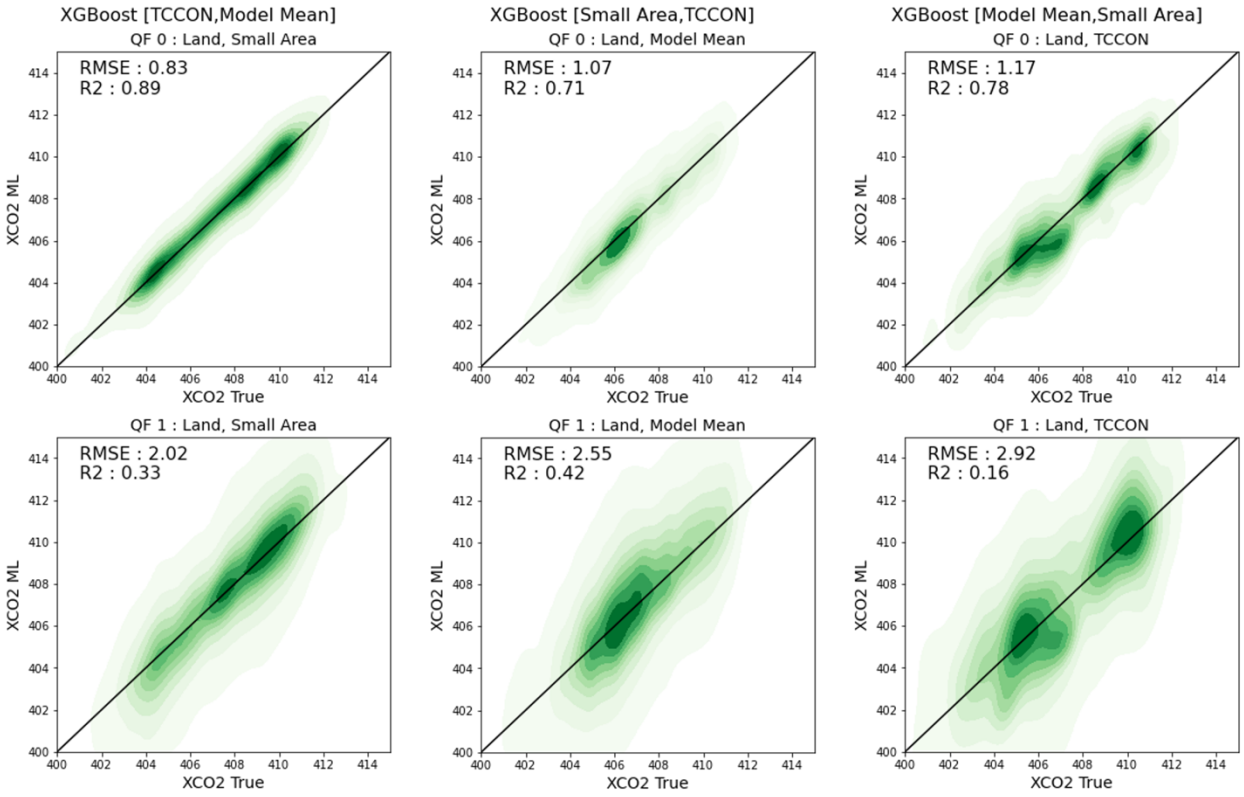


Figure 7. Comparison of XCO_2 derived from truth proxy (XCO_2 True) vs. XCO_2 corrected by XGBoost (XCO_2 ML) for land by hold out proxy set and hold out year (2018). Left-most column displays results of a XGBoost model trained on [TCCON, Model Mean] and evaluated on Small Area. Middle column displays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

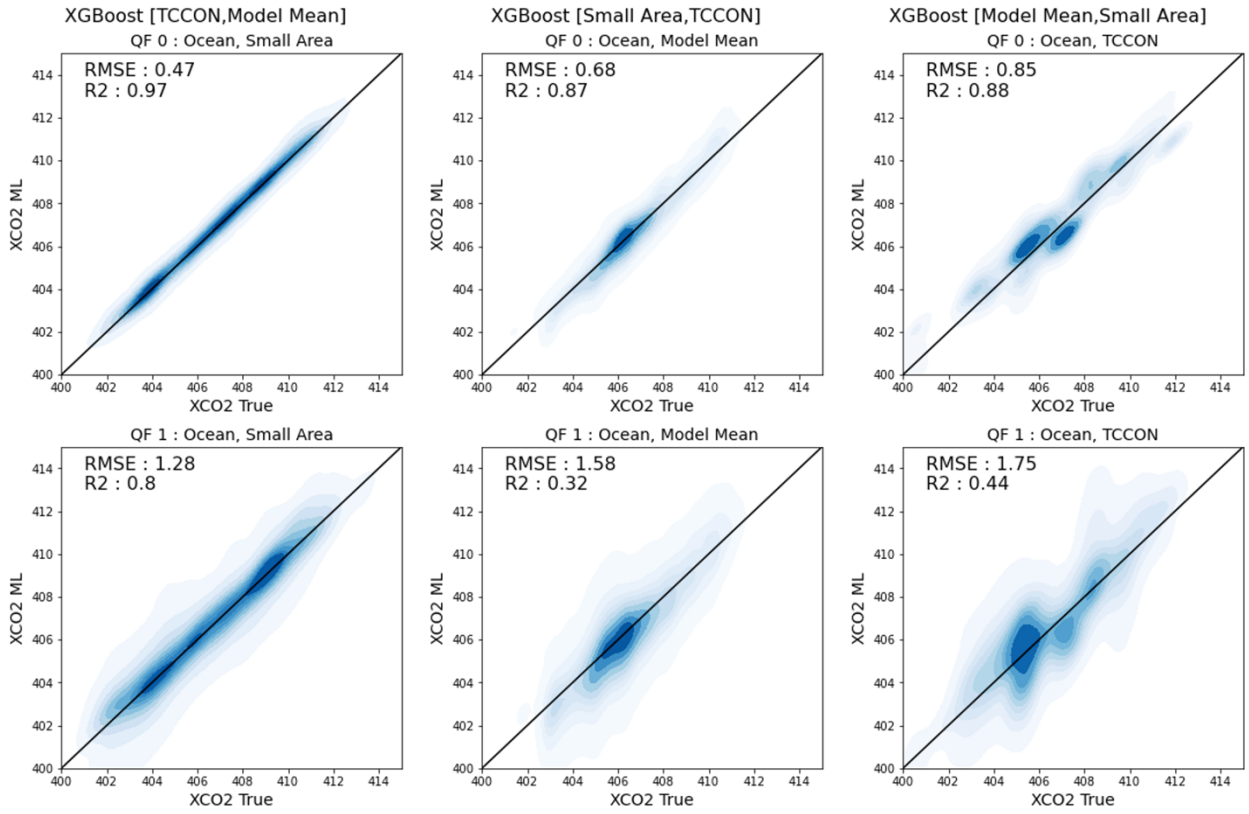


Figure 8. Comparison of XCO_2 derived from truth proxy (XCO_2 True) vs. XCO_2 corrected by XGBoost (XCO_2 ML) for ocean by hold out proxy set and hold out year (2018). Left-most column displays results of a XGBoost model trained on [TCCON,Model Mean] and evaluated on Small Area. Middle column displays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

There are quite a few different XGBoost models, e.g. for testing purposes, in the paper; I am not entirely sure what the proposed bias corrected product is in the end. I suspect it is the data set with the correction (split according to land and ocean) learned on all three proxy datasets for QF=0+1 simultaneously and subsequently restricted to QFNew. Is this correct? Or is it some kind of average of the three models for the different truth proxy data sets? Please make this more clear in the text.

This is correct, we propose a XGBoost model for land data and a XGBoost model for ocean data. We then derive QFNew using the new bias correction. This final proposed bias correction was not clear in the text. We have clarified the language in Section 3.4 and it now reads:

Two methods are compared 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.

Additional wording has been added to the abstract.

In this paper, we demonstrate a clear improvement in the reduction of error variance over the operational correction by using a set of non-linear machine learning models, one for land and one for ocean soundings.

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 XCO_2.

The bias correction (for both land and ocean) includes features measuring the deviation of the retrieved quantities from the prior (surface pressure and vertical CO_2 profile). I assume that these priors are very consistent with the XCO_2 truth used in the supervised learning. Doesn't that run the risk of essentially pushing the XCO_2 results back to the truth/prior in specific cases without noticing in the averaging kernels that hardly any information is gained from the actual measurement? Can you exclude that point sources that are not present or not sufficiently resolved in the truth are artificially attenuated or corrected away in unseen data? Due to these potential pitfalls, the results would be more robust if such parameters (co2_grad_del, dpfrac, dp_sco2) were not used in the bias correction. How important are these features in your machine learning model? Please discuss these issues in the manuscript.

The CO_2 gradient delta, and surface pressure delta terms are used extensively in the operational bias correction and an analysis of the correction using these terms is given in Kulawik et al. 2019. Despite this we agree that there is potentially a risk in over correcting using the more complex machine learning model. We believe that we show that we are robust to this due to the negligible improvement in performance for QF=0 data over the operational correction, where co2_grad_del, dp_sco2, and dpfrac dominate the feature importance. When correcting QF=1 data the feature importance for these terms drops off dramatically as shown in the new Figure 9. Furthermore, we offer an empirical example using a set of plumes not present in the training data (a land example and an ocean example) to illustrate that the method does not remove CO_2 enhancements.

5.3 Preservation of CO_2 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 CO_2 gradient delta, and surface pressure terms (e.g., dpfrac, dp_o2a), 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 XCO_2 to the truth. We compare XGBoost corrected XCO_2 for two known plumes first identified in Nassar et al. 2021. The two example plumes are shown in Figure 10, an ocean glint plume in Taean, South Korea, and a land nadir plume observed over two co-located power plants in Ohio, US. We compare the uncorrected XCO_2 retrieval (B10 Raw), the operationally corrected XCO_2 (B10 Corrected) and the machine learning corrected XCO_2 (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, that also showed fitting a machine learning model for 3D cloud correction, which include similar delta terms, did not correct out CO_2 enhancements.

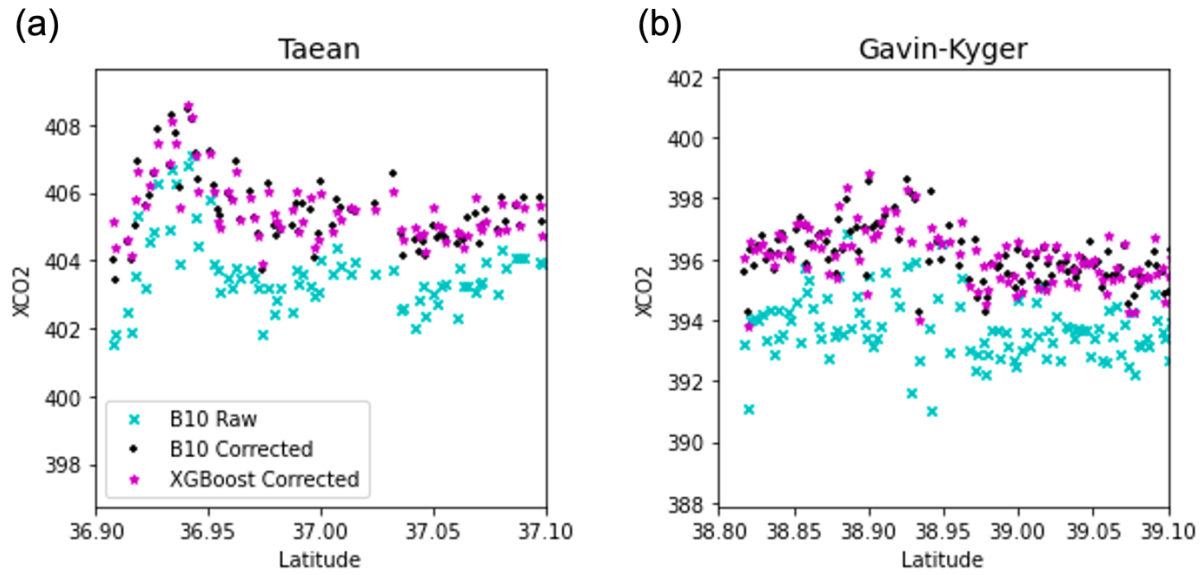


Figure 10. Two CO_2 plumes captured downwind from power plants (Nassar 2021). An ocean glint plume at Taean, South Korea, [lat 36.91°, lon 126.23°] on 2015-04-17 is shown in (a). A land nadir plume near the J. M. Gavin and Kyger Creek power plants in Ohio, USA, [lat 38.93°, lon -82.12°] on 2015-07-30. Regions with the example plumes are not present in the training dataset and consist of QF = 0 + 1 data.

L25-29: There were other satellites measuring CO_2 before GOSAT, e.g. AIRS or SCIAMACHY. GOSAT is considered the first satellite designed specifically for the purpose of measuring atmospheric CO_2 from space. Please be more specific here.

We now correctly acknowledge the contribution of these instruments. The sentence now reads:

Following a long history of critical in situ measurements of CO_2 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 Chartography (SCIAMACHY) in March 2002 (Boevnsmann et al, 1999) and the Atmospheric Infrared Sounder (AIRS) launched in May 2002 (Aumann et al, 2003).

L58-60: Do you mean (Noel et al., 2021) or (Noel et al., 2022)? There are two papers, but only (Noel et al., 2022) is in the References. (Noel et al., 202?) and (Schneising et al., 2019) both use non-linear bias correction techniques but there are no explicit comparisons to linear corrections and the term "operational" does not fit here either. Please revise this sentence.

The citation has been corrected and is no longer used to argue against a linear correction.

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).

L63: Please be more specific what you mean by "interpretable" since XGBoost is a complex black-box model, which is not intrinsically interpretable. Do you mean post-hoc model interpretation methods? Do you refer to global explainability of the model or to local explainability of individual predictions?

We agree that the quality of the post-hoc interpretability provided by the internal splitting criterion was overstated. We have refined the section and clarified that the information gain provides only a global explanation of feature importance. We have clarified this in the text.

This research demonstrates a general non-linear bias correction approach for OCO-2 build 10 (B10, Taylor et al., 2023) via a machine learning method and provide a post-hoc explanation of the overall contribution of the selected state vector features.

L65: "reproducible" is somewhat misleading because it gives the impression that a universal recipe, that can be transferred without any adaptations, is being presented. However, when using other data sets the parameters of the model have to be re-tuned and the used features have to be adapted. Is there a systematic approach, e.g. to feature selection or setting of model parameters, justifying the designation "reproducible"?

We agree that the claim of reproducibility is misleading and have clarified that the general frame work for developing the bias correction can be adapted for future ACOS updates.

The framework presented in this manuscript for identifying informative features for bias correction can be adapted for future OCO-2,3 ACOS algorithm updates.

L66: The GeoCarb mission was cancelled, please remove.

Removed.

L85-92: Please specify which TCCON version you are using (GGG2014 or GGG2020?). It seems to be GGG2014, why didn't you use the most recent version? Does it make any difference in performance when training or validating with one or the other version?

Both the operational correction and our machine learning correction use the same set of soundings co-located with TCCON GGG2014 measurements. A bias correction is currently being developed for GGG2020, however that work is still unpublished. We have clarified our use of GGG2014 data.

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).

L122-L124: I disagree with the statement that XGBoost is highly interpretable compared to other machine learning algorithms; I would even argue that XGBoost is typically the least interpretable well-established machine learning algorithm of all after deep neural networks and also requires post-hoc approaches to understand and explain the model. For example, Support Vector Machines or Random Forest are typically more interpretable than XGBoost. Please revise (or remove) the sentence accordingly.

We have removed the general statements regarding model interpretability comparisons.

L128-129: XGBoost does not average across the ensemble (in contrast to Random Forest). Instead, XGBoost trains each subsequent model in the ensemble to improve upon the errors of the previous models and uses a weighted sum of the individual model predictions to make its final prediction. This is also one of the reasons why XGBoost is usually less interpretable than Random Forest.

This has now been corrected in the sentence.

L130: It depends on the specific task, the available resources, and data set whether XGBoost or Random Forest performs better, as they have different strengths and weaknesses. Moreover, it is hard to do a fair comparison anyway, because you have to tune the corresponding parameters independently, e.g. it makes little sense to fix certain tree structures in a comparison. Therefore, please avoid a general statement comparing the predictive performance of different machine learning algorithms, especially when it comes to XGBoost and Random Forest, which often perform quite similarly after respective optimal parameter tuning (see also general comments). Have you tried other machine learning algorithms for the specific task presented here?

L131-133: The arbitrary application of these strategies does not make XGBoost per se highly robust against overfitting to the training data. Please make clearer that these strategies can avoid overfitting if the parameters and the validation data set are chosen appropriately. Please demonstrate explicitly that there is no overfitting for the non-linear bias correction presented here.

L134-135: How exactly were these parameters determined? By monitoring the performance of the model on a validation dataset during the training process and stopping the training when the performance on the validation dataset does not improve for a given number of consecutive iterations? Are the parameters actually always the same (for all testing models and the final bias correction)?

We agree and have re-written Section 3.1, removing the general comparison on model performance and clarified that regularization alone does not guarantee good generalization to an unseen data set. We have described thoroughly the hyper-parameter tuning process and have further elaborated on the parameters selected for the land and ocean models. The section now reads:

3.1 Gradient boosting

To model systematic error from co-retrieved state vector elements, we employee a machine learning method known as extreme gradient boosting or XGBoost (Chen et al. 2016) which can fit both linear and non-linear relationships. XGBoost is an ensemble model where a set of simple models known as regression trees (Breiman 1984) are sequentially trained, with each new member fit on residuals of the previous trees. During inference, the weighted sum is taken across the ensemble members. Members are grown or fit by selecting features that provide high information gain (Eq. 2). Information gain is calculated by evaluating the sum of the gradients G and hessians H of the loss function at left and right leaf nodes when selecting features during tree fitting (for our experiments we use Mean Squared Error as the loss function shown in Eq. 3. Features that are informative for reducing residual error during tree development yield high gain values. These values can be summed across trees in the ensemble to produce a ranking of feature contribution. This provides a post-hoc method of interpretability yielding a high level or global view of feature importance to correcting

ΔXCO_2. While this method of interpretability is less informative than the regression coefficients provided by a linear model, it is useful for tasks such as feature selection.

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.

$$Information\ Gain = \frac{1}{2}\left[\frac{G_{Left}^2}{H_{Left}+\lambda} + \frac{G_{Right}^2}{H_{Right}+\lambda} - \frac{(G_{Left}+G_{Right})^2}{H_{Left}+H_{Right}+\lambda}\right] - \gamma\ , \quad (2)$$

$$Mean\ Squared\ Error\ loss = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \quad (3)$$

We initially evaluated a Random Forest but found better performance with the gradient booster. Therefore, we have focused only on the XGBoost models for demonstrating a non-linear correction in the manuscript.

L136-145: Have you tried other ways to get feature importances from XGBoost, e.g. permutation based importance or importance computed with SHAP values? Is it possible to improve interpretability by a choice tailored to this specific problem?

We have added comparison to permutation feature importance and note good agreement with information gain. We have added a plot and discussion to Appendix A.

Appendix A: 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.

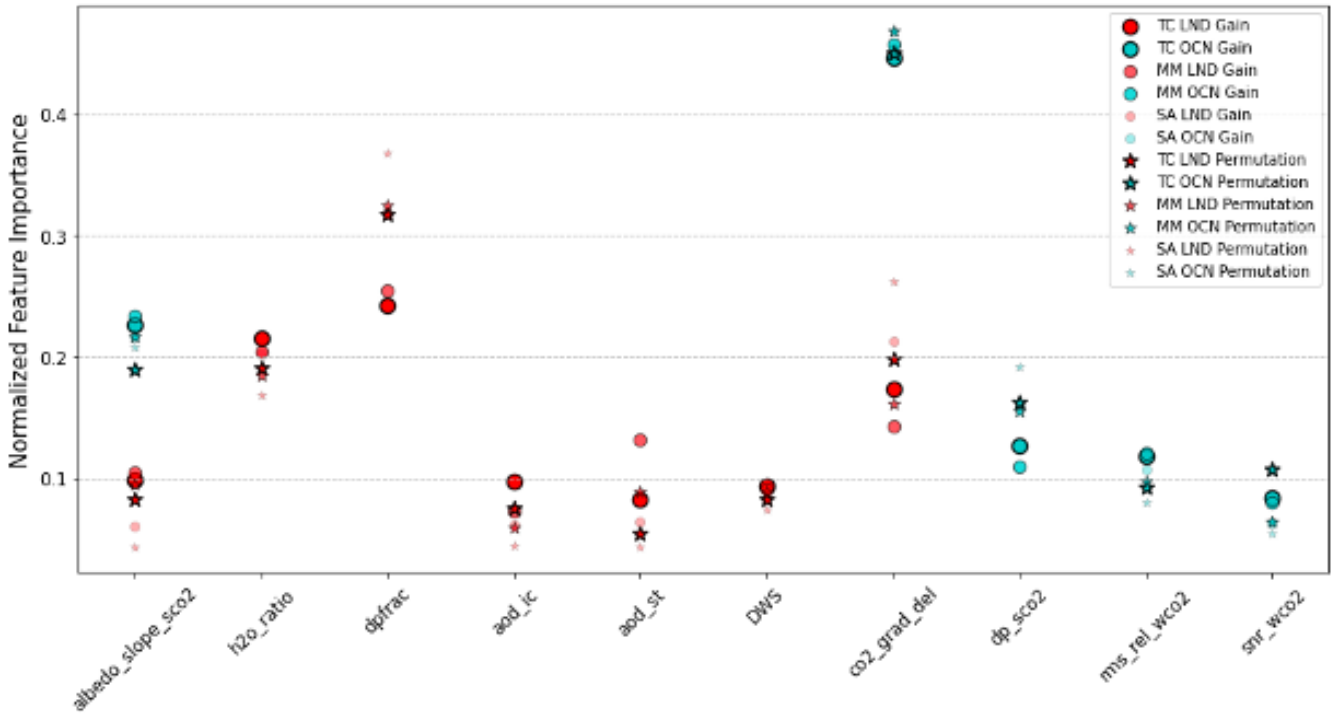


Figure A1. Comparison of feature importance derived from information gain and permutation importance. Normalized importance (permutation importance in stars, and information gain in circles) are shown for land and ocean features, and by truth proxy. Feature importance methods are largely in agreement in ranking and contribution.

L207-210 and Figure 2: Please list the used features and the respective importances (split by land and ocean) for the models based on the different proxy data sets and for the final bias correction separately in a table.

Figure 2 now shows the ranking and names of features used in the final bias correction.

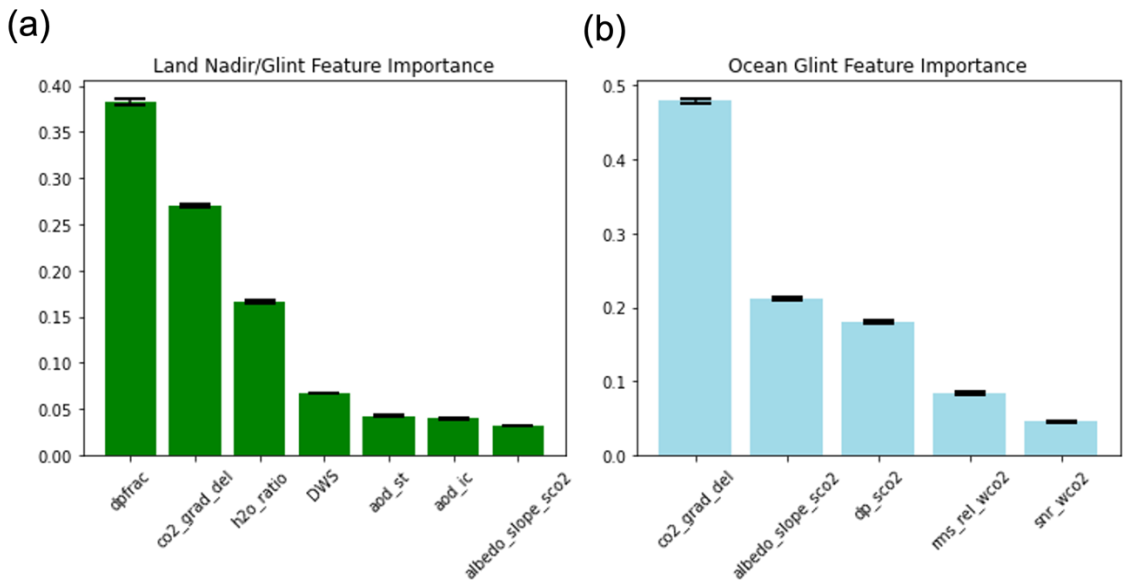


Figure 2. Feature importance for final land model trained with all proxies (a), feature importance for final ocean model trained with all proxies (b). Error bars denote variance in feature importance across 10 runs with different random seeds.

Figure A2 shows the same for the models trained on the individual proxies.

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. 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.

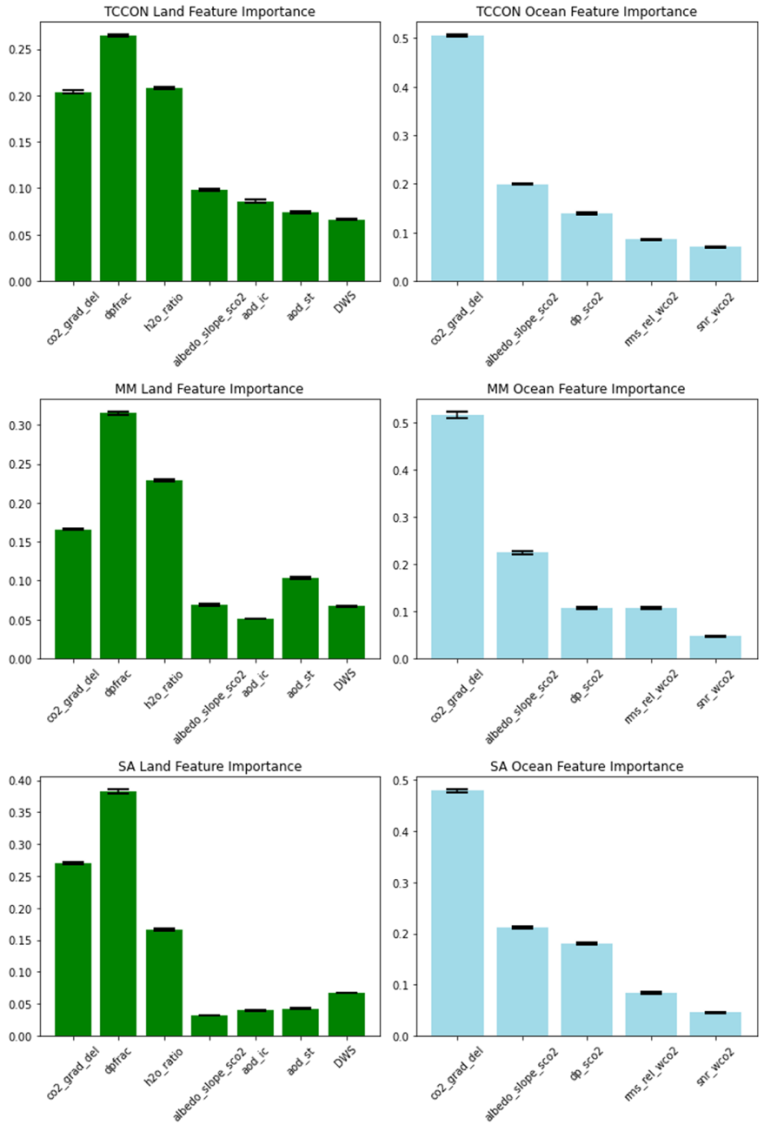


Figure A2. Feature importance for individual truth proxy models. Error bars indicate variance over 10 runs with different random seeds.

L215: Please better explain co2_grad_del.

We have re-written Section 4.1 to improve the overall clarity and description of the features used in the bias corrections including co2_grad_del. Section 4.1 now reads:

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.

L219: Are there different prior surface pressures for the weak and the strong band? If so, why?

There is an alignment offset in the pointing location between the three bands therefore three priors are used.

Table 3: Please remove "large unphysical" in the description of co2_grad_del. Or do you use an additional threshold to rate a deviation from the prior as unphysical?

Removed.

L260: Overfitting of XGBoost cannot be generally excluded. Please prove that there is indeed no overfitting for this specific task and parameter setup. To this end, a more challenging selection of training and validation datasets could also be considered (see also general comments).

We now address this in Section 5.1, shown in the response above to the general comment.

Table 4: Please also list the results for the final proposed bias correction combining all truth proxies.

Added RMSE for combined truth proxies for each model.

Figure 3/4, Table 4/5/6: Only validation data (2018 with truth proxy sampling) is displayed here, right? If this is the case, please be more specific in the description.

All tables and descriptions now thoroughly depict 2018 data.

ADD DESCRIPTION UPDATE HERE

L303-304: Wouldn't it be better to apply the footprint correction before training the bias correction?

This is an excellent point. We now apply the footprint correction first before removing the feature correction. While the effects to performance are negligible, feature contribution to bias is now more accurately depicted. Hyperparameter selection has also been updated.

Is it possible to include the footprint correction in the bias correction by introducing a suitable parameter as feature (e.g. row number)?

This is currently being evaluated for inclusion in a machine learning correction that expands on the approach of this manuscript and scheduled for a L2 litefile update for B11.

Figure 5: Please highlight more (in the text and in the caption) what exactly is shown here: it is all data from 2016-2018 including three types of data (correct me if I'm wrong), namely 1) training data (truth proxy data for 2016-2017), 2) validation data (proxy data for 2018), 3) data beyond (data from

2016-2018 used neither for training nor validation). It would be very enlightening to show (and compare) this kind of figure separately for each of the three data types, because this could provide indications of how well the correction generalises to actually independent data (type 3).

We have now added much needed clarification to the data that is displayed in this plot. In order to increase that data available for plotting we trained three models using with a year of data from 2016 to 2018 used as hold out. The data plotted is the aggregated validation years and would be similarly described as "type 2". Figure 5 text now reads:

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.

Table 6: It would be interesting to add the respective performances for type-3-data (or to introduce more challenging training and validation data sets from the start, see also general comments). Due to the lack of an entirely independent validation dataset, the performance suggested here may be too optimistic.

We now address "type 3" data in the new section 5.1 as described above in the response to the general comment.

Figure 7/8: Please explain all occurring variables.

We have added appendix Table C1 that includes all variables used or considered in either bias correction or filtering.

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

Section 5.1, Figure 9: Please also report the individual information gains and not only the differences in feature importances.

Figure 9 now shows the individual information gain. Wording the section has also been changed to reflect this.

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.

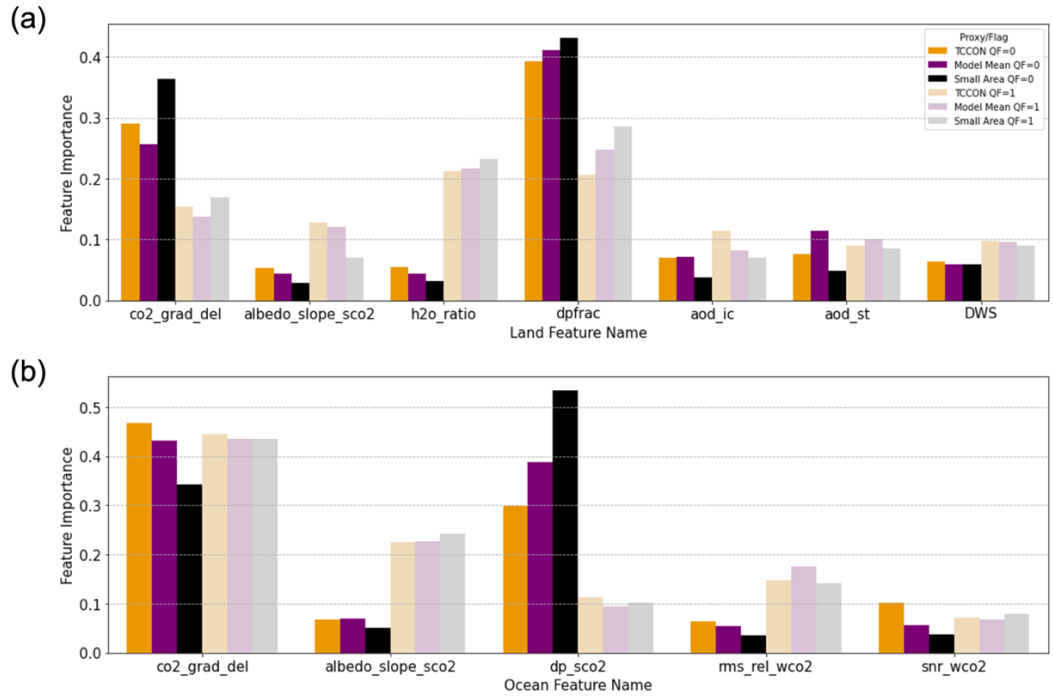


Figure 9: Feature importance for land is shown in (a), feature importance for ocean is shown in (b). Y-axis displays the normalized information gain from XGBoost models with QF=0 shown in darker colours and QF=1 shown in lighter colours.

Technical Corrections

There are incorrect Figure and Table numbers used in the main text. Please check.

Corrected.

Citations in the text do not always match the ones in the References section. Please check.

Thank you for catching these, we have corrected several citation errors.

L12: "Obersvatory" → "Observatory"

Corrected.

L15: "correlate" → "correlated"

Corrected.

L35: Please remove the closing bracket.

Corrected.

L36: It is (Rodgers, 2000).

Corrected.

L47: "epmerically". Do you mean "empirically"?

Corrected.

L49: "inturn" → "in turn"

Corrected.

L53: "reduce" → "reduces"

Corrected.

L55: "applying quality filter" → "applying the quality filter"

Corrected.

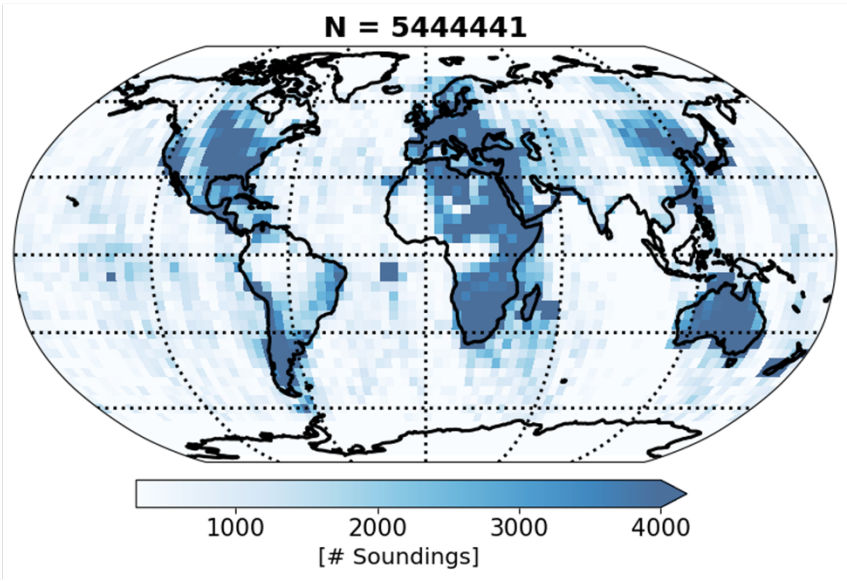
L55: Please remove the full stop between "correction" and "to" or rephrase both sentences.

Corrected

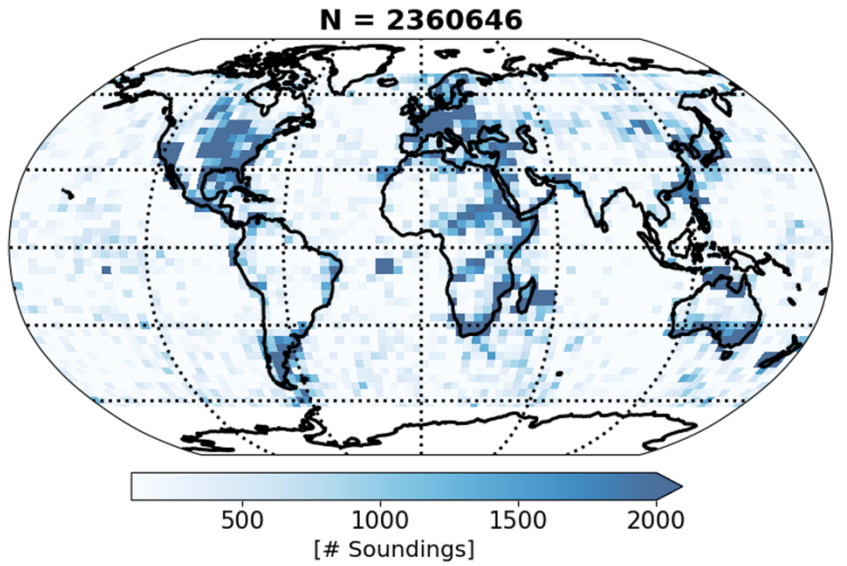
Figure 1: The number of soundings N is wrong in panel (a). The colour scale is oversaturated. Please extend the value range (0-2000 does not seen to be optimal).

Corrected

(a)


N = 5444441
1000
2000
3000
4000
[# Soundings]

(b)


N = 2360646
500
1000
1500
2000
[# Soundings]

(c)

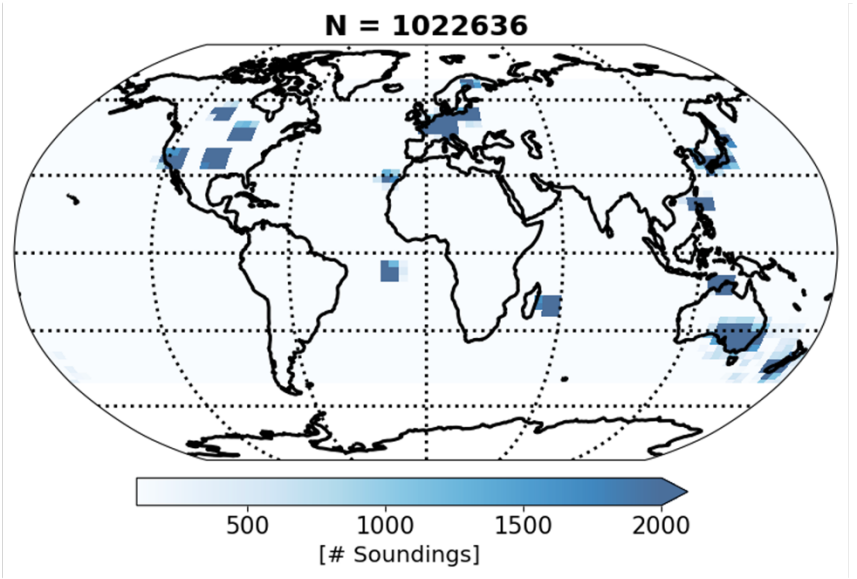

N = 1022636
500
1000
1500
2000
[# Soundings]

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).

L106: "overs" → "offers"

Corrected

L178: "trainined" → "trained"

Corrected

L186: "This, allows" → "This allows"

Corrected

L223: "retrivals" → "retrievals"

Corrected

L252: The section numbering is inconsistent. This number already exists.

Corrected

This list of technical corrections is (likely) not exhaustive. Please check the text for further typing errors

Reference:

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.