The authors describe a method wherein in situ-constrained XCO_2 simulations over North America are used to evaluate retrievals of XCO_2 from OCO-2 v10. The paper could be appropriate for publication after several major concerns are addressed.

Major comments:

On lines 120 and 194 and elsewhere, the authors state their assumption that the influence of recent surface fluxes are zero in the upper troposphere and stratosphere. This is not the case during some forest fires (e.g., Hooghiem et al., 2020 and references therein) and some volcanic eruptions. How does this assumption influence the validity of the results, and in particular, the seasonal cycle? How realistic is the background XCO_2 at those altitudes? The satellites measure through the entire atmosphere, and even if the averaging kernels are smaller in the stratosphere than in the troposphere, they are nonzero – that region of the atmosphere will still influence the retrievals. The authors discuss the errors associated with this around line 295 and dismiss the differences as small, but they do not show profiles comparing the simulated profiles and AirCore above 8 km.

On lines 335-345, the authors evaluate the OCO-2 scaling to the WMO X2007 scale using a single year (2014-2015) and their simulated XCO_2 product. Seasonal scaling does not make sense in this context, as the scaling is largely required because of spectroscopic uncertainties, and it is therefore more likely to be related to the airmass or water vapour interference than to the season. I suggest that these results are checked by binning the C0 values over different airmasses instead of season. The authors also claim that their annual C0,sim (0.9960) is inconsistent with the C0 derived through comparisons with TCCON (0.9959), but given the different timescales and spatial scales used in the analysis, I do not think this is a robust conclusion.

The third and fourth paragraphs of the introduction betray the authors' low opinion of the value of remote sensing relative to in situ measurements. I strongly recommend that the authors completely rewrite those paragraphs as they are misleading. I shall respond to each sentence of these two paragraphs in turn (original text in *red italics*):

In-situ measurements that comprise global networks, such as NOAA's Global Greenhouse Reference Network (https://www.esrl.noaa.gov/gmd/ccgg/ggrn.php), are rigorously evaluated and carefully calibrated relative to the World Meteorological Organization (WMO) calibration scale (data used here are reported on the X2007 scale), thus ensuring the fidelity of these measurements over timescales of seasons to decades (Andrews et al., 2014; Hall et al., 2020).

I agree that in situ measurements like NOAA's Reference Network are rigorously evaluated and calibrated. While the authors do cite the (now published) Hall et al. 2021 paper, it seems relevant to explicitly mention that these calibrations are complex and that periodic and significant changes of scale are possible. From the Hall et al. 2021 paper (italics are my additions): "The new scale *[WMO-CO2-X2019]* is 0.18 μmol mol^{-1} (ppm) greater than the previous scale *[WMO-CO2-X2007]* at 400 ppm CO_2. While this difference is small in relative terms (0.045 %), it is significant in terms of atmospheric monitoring."

The "open-path" nature of space-based XCO_2 measurements however, does not allow for direct calibration.

I agree with this sentence but feel that the authors' use of "however" is unnecessary.

Satellite retrievals of XCO2 require complicated models of atmospheric radiation and are sensitive to a host of assumptions about aerosols, clouds, interference of jointly retrieved parameters, surface properties and details of the instrumentation (Kulawik et al., 2019).

Again, I largely agree with the sentiment, but the language betrays a value judgment ("complicated," "host of assumptions"). There has been significant research into these areas.

Moreover, sensors typically degrade over time, and limited information is available to characterize resulting time-dependent systematic errors.

Sensors can and do degrade over time, but there is sufficient data available to characterize time-dependent systematic errors through radiometric calibrations, lunar calibrations, and comparisons with ground-based data (e.g., Crisp et al., 2017; Bruegge et al., 2019; Yu et al., 2020).

Post-launch data corrections are performed if and when biases (O'Dell et al., 2018) and errors (Kiel et al., 2019) are identified, but are severely limited due to the sparsity of calibrated in-situ vertical profile observations.

Post-launch data corrections are rigorously and systematically evaluated against the available ground-based information provided by multiple sources (TCCON, models, small area analyses, southern hemisphere approximation, etc.). None of this is limited by "the sparsity of calibrated in-situ vertical profile observations." This is not done "if and when biases and errors are identified" – they are actively sought out and significant effort is put into this.

Currently, satellite derived XCO2 retrievals are linked to the WMO scale most directly through a limited set of in-situ profiles obtained over a network of ground-based Fourier Transform Infrared Spectrometers that comprise the Total Carbon Column Observation Network (TCCON; Wunch et al., 2017).

Satellite retrievals of XCO_2 are linked to the WMO scale through thousands of coincident measurements over TCCON stations throughout the lifetime of the mission.

However, TCCON itself provides remotely sensed information about XCO2, and TCCON retrievals undergo a complex validation and bias correction routine (Wunch et al., 2010) that links these retrievals to the WMO scale.

The TCCON data have been rigorously and systematically compared with in situ profiles that are, in turn, calibrated to the WMO X2007 scale. There have been about 80 such CO_2 in situ profiles over TCCON stations – the first collected in 2004, and the most recent from the current year. I do not know what the authors mean by "TCCON retrievals undergo a complex

validation." The paper cited details the method of tying the TCCON retrievals to the WMO X2007 scale.

Moreover, TCCON sites are few.

There are 26 currently operating TCCON stations.

Issues with validation of OCO-2 retrievals via TCCON have been identified- seasonal and site-dependent biases have been reported (Wunch et al., 2017, 2015), raising questions about the adequacy of this network to validate satellite derived XCO2 products (Basu et al., 2013).

I do not see how the 2015 paper (describing the GGG2014 algorithm and dataset) is related to this topic. The 2017 paper identifies biases in OCO-2, and not in the TCCON data – the 2013 paper discusses biases in the RemoTeC GOSAT retrievals relative to TCCON. I do not think these papers question the "adequacy" of the network to provide valuable validation information.

This is especially important since systematic bias corrections of XCO2 from TCCON data are developed over small spatio-temporal scales and extrapolated globally (Wunch et al., 2011) for retrievals over land and ocean respectively (O'Dell et al., 2018).

I do not understand what the authors are referring to in this statement. If the authors are referring to the systematic bias identification related to retrieval parameters like albedo, aerosol, etc., then I do not know what "small spatio-temporal scales" the authors are referring to. The 2011 work did not use TCCON data to identify these systematic biases – it used GOSAT data from the southern hemisphere south of 25S. Comparisons with the TCCON data showed that the bias correction improved the comparisons globally, and identified the scaling required to tie the bias-corrected GOSAT XCO_2 to the WMO scale.

OCO-2 retrievals are additionally corrected for bias by comparing with 4-D CO2 mole fraction fields from global inverse models, and a small-area approximation, but both methods are prone to smoothing across fine-scale variability in XCO2 (O'Dell et al., 2018; Corbin et al., 2008).

I don't understand this statement at all. It should be *comforting* that bias corrections derived from these independent methods are consistent.

While bias correction generally reduces inferred surface flux uncertainty when retrievals are assimilated in atmospheric inversions (Basu et al., 2013), even small retrieval errors can lead to large errors in inferred flux (Takagi et al., 2014; Chevallier et al., 2014).

I agree.

Biases in XCO2 from OCO-2, hereafter XretCO2 , have been identified, and found to be related to surface (e.g. pressure, albedo) and atmospheric (e.g. aerosol loading, sky condition) properties (Kiel et al., 2019).

The Kiel et al., 2019 paper talks about a geolocation error and a meteorological reanalysis sampling error. I'm not sure why the authors attribute aerosol, albedo, cloud cover to this paper.

However, systematic biases not accounted for in the v10 bias correction approach persist and therefore the measurement uncertainty associated with individual retrievals is believed to be at least twice the value currently reported in the OCO-2 data files (Eldering et al., 2017).

I don't see how the authors can justify this claim with the citation provided. Version 10 was released in 2020, and the Eldering paper is dated in 2017.

Thus, a dynamic method to routinely evaluate satellite retrievals is necessary.

This statement does not follow from the previous two paragraphs. Define what is meant by "a dynamic method".

Other comments:

Figure 7 and lines 360-364: There's a 4-ppm latitudinal gradient in the column across North America before and after optimizing biospheric fluxes. This is not "small" from a carbon cycle perspective and should be detectable from space.

References not already cited in the paper:

Hooghiem, J. J. D., et al.: Wildfire smoke in the lower stratosphere identified by in situ CO observations, Atmos. Chem. Phys., 20, 13985–14003, https://doi.org/10.5194/acp-20-13985-2020, 2020.

Hall, B. D., et al.: Revision of the World Meteorological Organization Global Atmosphere Watch (WMO/GAW) CO_2 calibration scale, Atmos. Meas. Tech., 14, 3015–3032, https://doi.org/10.5194/amt-14-3015-2021, 2021.

Crisp, D., et al.: The on-orbit performance of the Orbiting Carbon Observatory-2 (OCO-2) instrument and its radiometrically calibrated products, Atmos. Meas. Tech., 10, 59–81, https://doi.org/10.5194/amt-10-59-2017, 2017.

C. J. Bruegge et al.: Vicarious Calibration of Orbiting Carbon Observatory-2, in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 5135-5145, July 2019, doi: 10.1109/TGRS.2019.2897068.

Yu S, et al.: Stability Assessment of OCO-2 Radiometric Calibration Using Aqua MODIS as a Reference. *Remote Sensing*. 2020; 12(8):1269. https://doi.org/10.3390/rs12081269