Review of *"On the statistical optimality of CO2 atmospheric inversions assimilating CO2 column retrievals"* (F. Chevallier)

This work presents essentially two different, though related, pieces of research. The first argues that the current pipeline used to derive optimal flux estimates from satellite measurements of column CO2 (XCO2) are fundamentally flawed. They are flawed because different prior assumptions are used in the retrieval as compared to the inversion, and the author argues that this inconsistency could bias inversion results. The author then argues that using a strong prior constraint (as most inverse models would suggest) in the GOSAT retrieval algorithm seems to yield better agreement between the XCO2 in his MACC (v13.1) model, than do comparisons with the standard ACOS (v3.5) XCO2 retrievals. He also states that ACOS – MACC XCO2 differences appear to be correlated with surface albedo, though only upon visual inspection of difference maps.

This paper, while certainly thought-provoking, suffers from a severe logical deficiency that must be addressed before publication.

Regarding the first point, of the basic inconsistency between the GOSAT retrieval's prior CO2 covariance assumption and that of the model, it is worth stating that retrieval groups use a loose prior primarily **because** they want to be maximally consistent with any model prior covariance. A sufficiently loose covariance is always consistent with a tighter one, but not necessarily the other way around. Therefore, it is not clear to me that using a tighter covariance is required to yield formal mathematical consistency upon assimilation of the satellite-retrieved XCO2, assuming the averaging kernels are fairly applied.

My strongest concern, however, regards the author's evaluation of the GOSAT XCO2 retrieval quality via the comparison to a single model. Disagreement does not necessarily mean the GOSAT retrievals are biased. Models have many sources of error: transport model error, imperfect prior fluxes, and the assimilation of datasets that are sparse in many regions of the world. The author's only serious argument is that the difference map between the model-predicted and satellite-retrieved XCO2 should not have sharp spatial gradients because these should be smoothed out by transport effects (page 1900, line 8). But this argument problematic for at least two reasons:

- He does not specifically demonstrate that there is no way such a spatial gradient can be supported by transport, even if the underlying flux was large and itself contained a strong spatial boundary, as of course happens in some ecotones as well as at land/ocean interfaces; and
- One certainly cannot make this argument on maps that contain variable spatio-temporal sampling all plotted on the same map. For example, in the seasonally dry African Sahel region, the satellite has strong seasonality in its ability to monitor this region (namely due to wet vs. dry seasons), and this in

and of itself could cause *apparent* spatial gradients because, in fact, different times are plotted on the same map.

Secondly, the author states that in certain regions of large (1-2 ppm) model-satellite disagreement, the fault likely lies in the satellite data. While this is certainly possible, the reverse is of course also possible in the lack of additional information. Even though the author admits a few times in the text that the model may be imperfect, he does not comment about the general agreement (or disagreement) between the XCO2 of different carbon inverse model systems. These differences exist and they have been shown to be notable especially in regions where the models are not well constrained by in-situ data. For example, Kulawik et al. (AMTD, 2015) and Lindqvist et al. (ACPD, 2015) have recently shown that inversion models can have major differences in the seasonal magnitude of their optimized XCO2 values both latitudinally and longitudinally. Most of the regions with large retrieval-to-model differences in Fig. 2a are, interestingly, the same regions where also model-to-model differences in XCO2 can be notable: for example, in the African savannas, in seasonally dry forest/grassland regions in South America, in India, and in the high northern latitudes there can be up to 1-3 ppm differences in monthly averages *between different inverse models constrained by in-situ measurements*. Ultimately, of course, we would like to know what is driving these persistent model differences. Nevertheless, the author's conclusions would be on much more solid ground if independent model data sets were shown to support the author's arguments both about the surface albedo effect on retrievals and over-fitting of the radiances.

The author argues that the differences between the model and the retrieval over land at high latitudes are likely due to retrieval errors over dark surfaces. While this argument might have some truth to it (as retrieved XCO2 is indeed sensitive to the surface albedo in all three bands, and to its changes within each band), it is not entirely supported by the figures shown: the map of the mean surface albedo (Fig. 4) shows that the darkest land regions are in Scandinavia and the westernmost Russia while the largest positive differences are most continuous and consistent in central and eastern Russia. Moreover, the author says that the regions with the largest positive differences correspond to the evergreen needle leaf forest biome type, which is not true especially for central Russia where differences in June vary from -1.5 to 1.5 ppm inconsistently (Fig. 3b) and parts of Alaska.

The author finds substantial model-to-retrieval differences in the African savanna/Sahel region, and attributes these differences to "systematic errors in the retrievals", speculating about averaging kernels not peaking low enough in the atmosphere due to too loose retrieval prior error variances. However, the author does not speculate more about the reason for such regionally constrained errors: why would the prior error variances have more impact in that particular region compared to elsewhere? He suggests that CO2 from fires inaccurately represented in the MACC model might be another cause for the differences but considers this unlikely. However, a look at this particular region's optimized, natural CO2 fluxes inverted by different models reveals extremely large differences in the fluxes, and also that similar differences are reflected in that region's XCO2. As long as the model differences in this region are unexplainably large, one of the models cannot be fairly used to speculate about biases in the satellite retrievals in that region without some kind of additional information.

The author presents in Figs. 5-8 an interesting metric for evaluating overfitting in the retrievals (i.e., too tight a prior), and shows that increasing the weight of the prior XCO2 could make the retrievals statistically more consistent with the model. However, he does not show any spatial patterns of this metric; therefore it remains unclear if the suggested change in the retrieval prior errors would lead to worse misfits in some currently well-matched regions in addition to the likely improvements in the model-retrieval misfits in the regions where the differences are large. And even if he did, it would still suffer from the problem of comparing to a single model, and the fact that it couldn't be accounted for by faithfully using the column averaging kernel in the assimilation. Overall, by counting too much on the results obtained by this metric, we risk the possibility of both the model and the prior XCO2 being wrong and the satellite observations the truth. The satellite retrievals are certainly not (yet) completely free of retrieval biases, but it is fruitful to remind oneself *why* they are being carried out: because neither our prior knowledge nor our models are perfect. Even if similar results were obtained based on comparisons to other models, this philosophical dilemma would still remain in the background but the reasons that support to change the current retrieval procedure would be stronger.

Detailed comments:

- Page 11893, line 21. The author should state that the use of a rather loose prior CO2 covariance is not specific to ACOS, with some examples. For instance: 1) The RemoTeC retrieval has a formally unconstrained XCO2 (Butz et al., Applied Optics, 2009), and 2) the BESD retrieval uses a prior error on XCO2 of 15.6 ppm (Reuter et al, AMT, 2010). etc.
- Page 11896, line 3: "H a linearized" → H is a linearized
- Page 11896, line 12: "inversion window for the inversion" → inversion window
- Page 11897, Eq. (4): might be more informative to simply show the derivation of Eq. (4) instead of describing it in the previous paragraph.
- Page 11899, line 16: "long-tern" → long-term
- Page 11899, lines 17-18: variability in the XCO2 field is ~8 ppm in Fig. 1, retrieval-to-model differences are most typically less than 1 ppm (Fig. 2a). Therefore, the retrieval-model difference *is* "much less" than the variability within the modeled or retrieved XCO2 field.

- Page 11900, lines 10-11: it is incorrect to say that the local spatial gradients mostly reflect the retrieval gradients. For example, the gradients in Fig. 3a for South Africa, South America, and the latitudinal gradients in the oceans are not obviously wrong in the retrievals (Fig. 1a).
- Page 11900, lines 23-25: The surprising discontinuity in XCO2 on the NW coast of the U.S. compared to the adjacent ocean data is more clearly seen in the model (Fig. 1b) than in the retrieval.
- The benefits of showing Tables 1 & 2 are not clear. Because the paper otherwise concentrates on the GOSAT data years 2009-2013, it might be more helpful to the reader to see a map of where the in-situ data were collected during these years.
- Figures 5-8 need a more informative y-axis label. For example "XCO2^a – XCO2^model (ppm), mean(---) or \sigma (____)", or something similar.
- Figure 6: the two blue shades look very similar in the printed version. Consider colors with a larger contrast.