Response to anonymous referee 2

S. Basu, et al.

May 24, 2013

We thank the referee for the constructive feedback. Please find below our responses.

The comparison of GOSAT, OCO2, etc., to satellites not primarily designed to measure CO_2 seems a bit unfair; I'm not sure the point.

Our point was to make clear that GOSAT was not the first instrument to ever measure XCO_2. We did mention that they were not geared to study sources and sinks of CO_2, due to their lack of near-surface sensitivity. However, we realize that now it reads as if we are comparing (unfairly) TES, IASI, TOVS and AIRS to GOSAT, so we have changed the language to say that the earlier instruments were not designed to study sources and sinks of CO_2. Nonetheless, we do not think that the comparison is completely irrelevant, since inverse modelers such as Chevallier et al. [2005], Chevallier et al. [2009] and Nassar et al. [2011] did try to extract CO_2 source-sink signals from those instruments.

The introduction mentions in many places that the remote sensing observations based on instruments whose sensitivities peak at various altitudes in the troposphere have correspondingly variable sensitivity to surface emissions. While intuitively this makes sense, the extent to which this really matters hasn't been shown. Given their adjoint of TM5, the authors could easily quantify this through sensitivity simulations.

The relative merits of satellite instruments with peak sensitivities at different altitudes for estimating surface fluxes has been assessed by past studies such as [Houweling et al., 2004], and therefore were not discussed in our manuscript.

Can the authors explain why assimilating the full set of hourly observations leads to biased inversions? This reason doesn't seem obvious. Rather, it seems there is a sampling bias by selecting observations only at particular times. Further does the selection criteria introduce correlations in the observations by considering e.g., only nighttime measurements from high altitude sites?

The basic philosphy behind choosing which observations to assimilate is that one should only assimilate observations that can be reliably modeled. This is why only daytime averaged observations are assimilated on the plains, and also why this selection shifts to nighttime averages in the mountains.

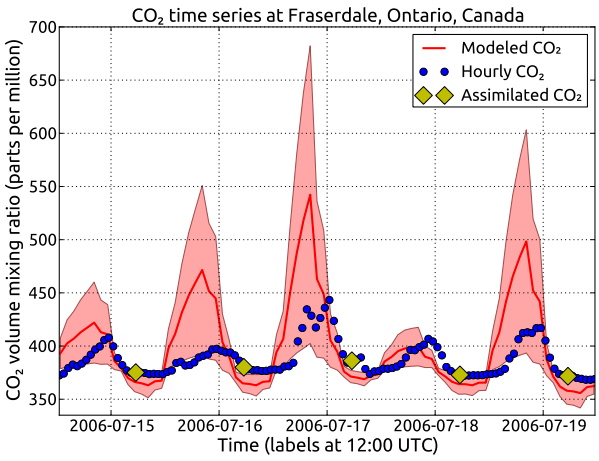

FIGURE 1: The hourly time series of CO_2 at Fraserdale, Ontario (Canada) in blue, compared to the modeled CO_2 concentration in red. The red shaded region is the estimated error in the simulated CO_2. The model was optimized by assimilating mid-afternoon averaged CO_2, in yellow rhombi.

The CO_2 concentration near the surface has a daily cycle, varying from a low in the late afternoon due to photosynthesis to a high in the early morning due to respiration. This daily cycle is amplified by the elevation of the planetary boundary layer (PBL) during the day (which dilutes the already low CO_2) and its subsidence at night (which concentrates the already high CO_2). Transport models in general – and certainly TM5 – are not equally good at simulating the daytime and nighttime PBL heights (BLH). In particular, the nighttime BLH is underestimated, resulting in spuriously high modeled CO_2 mixing ratios at night. Figure 1 illustrates this phenomenon by comparing the simulated CO_2 time series at Fraserdale, Ontario with the observed hourly time series. The modeled CO_2 has a much higher daily cycle amplitude compared to the observations. Therefore we cannot match the entire hourly observational time series, and must choose whether to assimilate daytime or nighttime observations. We know that the TM5 BLH is incorrect at night [Peters et al., 2004]. Therefore it makes sense to only assimilate late afternoon observations (yellow rhombi in figure 1), when we know that the TM5 BLH is most accurate. If we tried to match the entire time series, the model would try to fit both daytime and nighttime observations by lowering the overall flux. This would end up fitting neither, but result in a low-biased flux estimate.

High altitude sites such as Mauna Loa, on the other hand, are close to the PBL during late afternoon, but well above the PBL and in the free troposphere in the early morning. It is difficult for a coarse resolution transport model to accurately simulate the mixing near a complex, mountainous topography, which is essential to determine whether, during the daytime, the site is above or below the PBL. On the other hand, at night, due to PBL subsidence, the site is in the free troposphere above the PBL, which is much easier to model. Since the CO_2 concentration at high altitude sites are better modeled at night than during the day, we assimilate late nigh/early morning observations from such sites.

hint on → hint at

Corrected.

Noting that there is 4.5 times as many remote sensing observations as there are surface observations is a simple straw man argument and could be dropped. Just present that actual sensible approach, rather than suggesting a transparently flawed view first (i.e., "going by the numbers").

Good point. We have changed the sentence to simply say "The 77,769 assimilated soundings do not necessarily contribute 77,769 independent observations."

What are the values of R and T ? I don't see that included here. If they are mentioned elsewhere, it would be useful to repeat that here where the variables are introduced?

That was an omission. The values of R (500 km) and T (1 hour) have been included in the text now.

Eq 3: what is the meaning of the "hor" subscript on i?

The "hor" stands for horizontal, since i is an index for surface grid cells. So for a $6° \times 4°$ model, i_{hor} would run from 1 to $60 \times 45 = 2700$.

what is Q_{10} ?

It was a notation borrowed from Olsen and Randerson [2004], and the referee is right that it does not make sense unless explained. We have replaced it with "are added to NEE every three hours using a relation involving the two-meter air temperature and incident solar radiation as described in Olsen and Randerson [2004]".

Given the off diagonal elements and size of B, were there any numerical issues involved with calculating B^{-1} , particularly for the high resolution inversion tests?

B was never inverted; the worst we had to do was calculate the eigenvector decomposition of B for preconditioning the flux state vector, which, for a 10800×10800 matrix, can be time consuming but is still possible. One issue with diagonalizing such large matrices was that even though B was positive definite by construction, due to numerical artifacts we always obtained some small negative eigenvalues, which had to be truncated to a small positive value and the matrix B reconstituted.

Did the authors find a tuning that was unique, or are multiple possible values possible?

To obtain a given global uncertainty, multiple parameter combinations are possible. For each category there were three parameters (L, T and ξ), and choosing any two automatically fixes the third. In practice we selected realistic L and T and tuned ξ to fit the desired uncertainty on the global total. Interestingly, the first referee raised a question about how sensitive our flux estimates were to the choice of the parameters in Table 1 (not a whole lot), and our response now contains a sensitivity test where we drastically vary L and T (and therefore ξ) and repeat our inversions.

I applaud the authors attempt to quantify the model representativeness error. It's clear that their understanding of what this term is suppose to represent is correct. Too often (i.e., one of the other reviewers for this manuscript) this term is interpreted as "model error", which it is not. The approach taken here to estimate the representativeness error, however, doesn't seem the best. The goal is to quantify the degree to which sub-grid scale variability will make accurate matching of the observations impossible, even if x is perfect. I doubt however that the gradients across the coarse $6° \times 4°$ grid cells give a good sense of the subgrid variability. Instead, the authors should make use of their $3° \times 2°$ model run to evaluate where their are strong gradients near observations.

This is a good point. Our idea was that if a certain grid cell shows a high tracer gradient to its neighbors, there is most likely a lot of sub-grid scale spatial variability in the tracer mass there, which means a high representation error. This puts more weight on observations that are taken in areas of lower CO_2 variability. However, the referee is right that a higher resolution simulation is also a good candidate for assessing the sub-grid scale variability. Although we have not done it for this work, we plan to look at it in the future. One advantage of calculating the representation error online is that it does not require a separate simulation, and can adapt to changing x, for example if there is a lot of emission from a cell at a particular time step, we expect the representation error to be higher, which in this case will be ensured by the higher tracer mass gradient due to the emission.

I don't understand what the authors mean by σ_{mod} of the satellite observations, since this is the model representativeness error. Were they using gradients in GOSAT observations as an estimate of sub-grid scale variability to calculate the model representativeness error? This would make more sense than gradients in the $6° \times 4°$ grid cells... Sorry if I'm confused here, but perhaps it could just be explained a bit more clearly.

What is meant is the following. To calculate the model representativeness error of the total column, assume that there is a station in each model layer at the (latitude,longitude) of a sounding. Then calculate the representativeness

error in the horizontal direction only (since the vertical errors, i.e., deviations from the gridbox mean, will cancel out once we take the column average) for each of these stations. Finally, sum the representativeness error from all the layers after weighing them with the averaging kernel and the pressure thickness of the layers. Again, this is not a perfect measure, but a workable measure that weighs observations less if they occur in areas of high CO_2 variability.

I understand what the authors mean, but the way it is currently written sounds as if less averaging leads to less noise, where the opposite is true, so perhaps they could address this more carefully as well.

Good point. We have tried to explain it better in the revised text.

I love this book as much as any, but Tarantola (2005) is an odd reference to provide for gradient-based approaches to minimizing the cost function.

Tarantola does discuss adjoint operators for solving inverse problems, albeit in the appendices. We've changed the citation to [Kaminski et al., 1999].

Meirinik et al. (2008)

Corrected (parentheses removed)

One of the few areas where the methodology is a bit weak is the convergence criteria, which seems a bit arbitrary. Have the authors considered other standard evaluations of convergence such as χ^2 test, or comparing the magnitude of J to the number of observations, etc.?

Our convergence criterion of a gradient norm reduction of 10^8 – which is stricter than what has been used in previous studies involving 4DVAR, for example [Chevallier et al., 2005] – is based on the fact that the information content of the observations is the sum of the logarithms of the eigenvalues of the Hessian [Rodgers, 2000]. At a gradient norm reduction of 10^8, which is typically achieved after 60 - 100 iterations, the lowest of the leading eigenvalues was seen to be $\leq$ 1.03, and further iterations would have yielded progressively smaller eigenvalues. Therefore, the observational information that could be harvested with more iterations was going to be minimal. Figure 2 shows – at our convergence criterion – the leading eigenvalues of the Hessian, and the total information gleaned from the observations. We can see the total information saturating as the eigenvalues approach one. Further iterations would have yielded eigenvalues closer and closer to one, and therefore less and less information from the observational constraints.

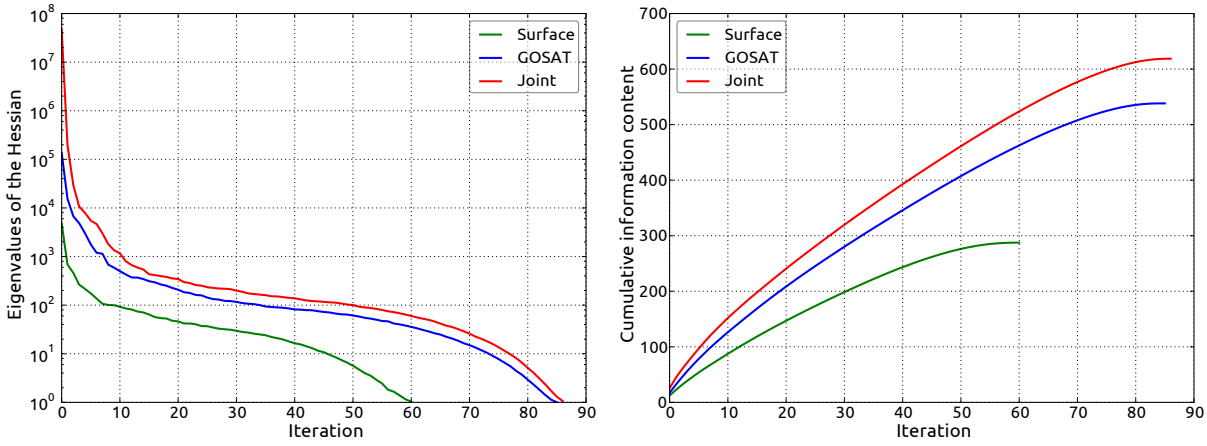

FIGURE 2: [**left**] The leading eigenvalues of the Hessian after the gradient norm has been reduced by 10^8, for three inversions. [**right**] The cumulative observational information content (in bits) gleaned after each iteration. As the eigenvalues dropped to one, the additional information obtained dropped to zero, and the total information obtained levelled off.

I think that ideally an inversion is judged by how well it constrains the sources, which we care more about than the actual distribution of CO_2.

From the point of view of setting up an inversion that is mathematically sound and self-consistent, this is entirely true. Indeed, this is the metric that is used to judge, for example, the usefulness of a new observational instrument. Having said that, we would not consider an atmospheric inversion to be useful for carbon cycle science if our mathematically sound machinery estimated a net CO_2 source in summer and a sink in winter over North America. When we speak of validating our flux estimates, we want to check how realistic those flux estimates are, keeping in mind that even for very sophisticated inversion techniques, observational biases and erroneous flux priors can lead to flawed estimates.

"spot on" is a bit casual for a journal article. I suggest the authors be more quantitative here. Specifically, statistics such as error, bias, correlation, etc., should be provided for the different inversion results. The can be included in the white space on the plots, or consolidated into a table.

We agree that "spot on" has no place in a journal article. We were specifically talking about the phasing of the

seasonal cycle here, which is hard to quantify for time series such as those in Figure 3. Therefore, we have changed the phrasing to "The simulated phasing of the seasonal cycle also matches the observations".

The selection of these 4 stations still seems arbitrary. A table presenting overall performance statistics to the observations should be presented.

We chose those stations as typical of each latitude band. We would rather not present a table of station-by-station statistics in the main text, since in our experience it is hard for a reader to parse such a table and draw any useful conclusion. Statistics on total mismatches also hide a lot, such as whether the "goodness" of fit at a station has a seasonal dependence, or if the model-observation bias at a station has a seasonal variation. However, we will –if the referee wants – happily upload our posterior time series at all the stations we've assimilated, for all of our inversions, as supplementary material.

Is is really necessary to show the results at such high frequency? There is no discussion of the high frequency aspects of the data, so I suggest smoothing the data would make it much easier to distinguish the different results, particularly for Park Falls.

The referee is correct that we do not discuss the high frequency aspect of the data, and therefore, for visual clarity, it would be useful to present smoothe/averaged model time series, especially for Park Falls. Therefore, in that figure, we have replace the time series at Park Falls with one where both the observations and the co-sampled modeled CO_2 have been averaged over three days (figure 3 below). We have mentioned this in the caption as well.

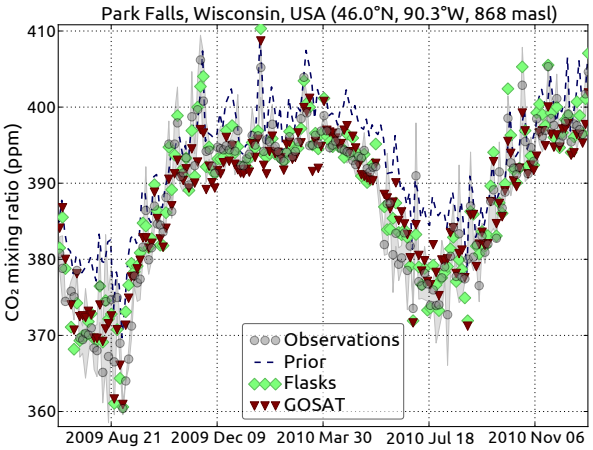

FIGURE 3: The CO_2 time series at the tall tower site in Park Falls, Wisconsin, from a surface-only inversion and a GOSAT-only inversion, compared with the prior time series. The model fields from optimized models have been co-sampled with the observations, then averaged over three day windows.

It seems like this excessive drawdown is almost as prevalent in the norther data on the left of Fig 5 as well, so it's not clear why it is only mentioned with respect to the tropics.

We are not sure if we have understood this particular comment. The comparison between CONTRAIL and optimized CO_2 fields was presented on page 4555, where we showed that a surface-only inversion agreed with CONTRAIL, whereas including GOSAT resulted in a higher drawdown. Therefore, we could not exclude the possibility that the higher drawdown was an artifact of our particular GOSAT retrieval. In contrast, in the tropics the GOSAT-based inversions agreed better with CONTRAIL compared to the surface-only inversion. Perhaps this is what the referee was looking for?

Please include in all figure captions an explanation of the legends to clarify, for example, the difference between GOSAT assimilation results and GOSAT observations.

Only two figures (2 and 6 in the manuscript) contained GOSAT observational data. That information has been added to the figure captions.

I would think that the GOSAT assimilation would lie between the prior and the GOSAT observations, but in many places (e.g., Sep 2009) this isn't the case. Please explain.

In figure 6 in the manuscript the posterior modeled fields are averaged over 7 days, whereas the GOSAT XCO_2 data are averaged over 15 days due to the higher scatter (the spread in GOSAT XCO_2 is the light green zone). This made the curve with GOSAT observations smoother, suppressing extreme values. In reality, in September 2009 there were GOSAT observations even lower than the GOSAT-optimized TCCON-cosampled model XCO_2, evidenced by the fact that the green shaded region extends below the lowest brown rhombus.

One of the main novelties of this work, as expounded upon in the introduction, is the use of GOSAT and surface flask measurements individually and in tandem for the assimilation. However, while section 4 includes a detailed description of the GOSAT assimilation, it doesn't directly explain the flask-only assimilation or the joint

assimilation. Results from these are apparent in the figures and appear scattered throughout the text, but there isn't any section dedicated to them. I suggest these two other assimilation results be specifically discussed in Section 4.

We did not discuss the flask-only inversion since it is the more "traditional" inversion, which we use as the baseline. Both the methodology and typical conclusions have been published before, for example by Chevallier et al. [2010]. In § 4.5, most of the results we discuss – for example those on pages 4563 and 4564 – are from the joint inversion.

It is interesting that the higher resolution source regions contain less correlations. Can the authors comment on this with regards to inversion techniques using grid-scale vs aggregated source regions?

In general smaller regions are expected to be less correlated than larger regions. To take an example, we expect global land and global ocean to be highly correlated, since the inversion will try to maintain a carbon budget consistent with the observed growth rate, and anything that is not coming from the land must come from the water and vice-versa. However, we expect less correlation between tropical land and northern extra-tropical ocean, since what is not coming from tropical land does not necessarily have to come from northern extra-tropical ocean. Furthermore, many of the larger regions to the right of figure 9 are overlapping, so for example global land includes northern extra-tropical land, and since northern extra-tropical land is one of the major land-based carbon sinks, the two are highly positively correlated.

Did the authors consider that timing of the GOSAT observations (i.e., only at 13:00 local time) could potentially bias the inversion? Is the mean CO_2 at that time an unbiased estimate of mean CO_2 , and would that make a difference?

This is not a significant issue for assimilating satellite observations, since the daily cycle in the total column is very small, and the problematic dynamics of the PBL discussed before has no effect on the model's ability to simulate the total column.

It seems the discussion of σ_b here is out of place. Wouldn't it fit better in section 3.1.4, which describes R?

We do not use σ_b in the data assimilation. It is a temporary variable introduced to explain our error inflation scheme (equation 1 in the manuscript), which is why it is in § 2.2.

enough to detect

Corrected

Chevallier et al. (2011)

Corrected (parentheses removed)

References

Chevallier, F., Fisher, M., Peylin, P., Serrar, S., Bousquet, P., Bréon, F.-M., Chédin, A., and Ciais, P.: Inferring CO2 sources and sinks from satellite observations: Method and application to TOVS data, J. Geophys. Res., 110, D24 309, doi:10.1029/2005JD006390, URL http://dx.doi.org/10.1029/2005JD006390, 2005.

Chevallier, F., Engelen, R. J., Carouge, C., Conway, T. J., Peylin, P., Pickett-Heaps, C., Ramonet, M., Rayner, P. J., and Xueref-Remy, I.: AIRS-based versus flask-based estimation of carbon surface fluxes, J. Geophys. Res., 114, D20 303–D20 303, doi:10.1029/2009JD012311, URL http://dx.doi.org/10.1029/2009JD012311, 2009.

Chevallier, F., Ciais, P., Conway, T. J., Aalto, T., Anderson, B. E., Bousquet, P., Brunke, E. G., Ciattaglia, L., Esaki, Y., Fröhlich, M., Gomez, A., Gomez-Pelaez, A. J., Haszpra, L., Krummel, P. B., Langenfelds, R. L., Leuenberger, M., Machida, T., Maignan, F., Matsueda, H., Morguí, J. A., Mukai, H., Nakazawa, T., Peylin, P., Ramonet, M., Rivier, L., Sawa, Y., Schmidt, M., Steele, L. P., Vay, S. A., Vermeulen, A. T., Wofsy, S., and Worthy, D.: CO2 surface fluxes at grid point scale estimated from a global 21 year reanalysis of atmospheric measurements, J. Geophys. Res., 115, D21 307–D21 307, URL http://dx.doi.org/10.1029/2010JD013887DO-10.1029/2010JD013887, 2010.

Houweling, S., Breon, F.-M., Aben, I., Rödenbeck, C., Gloor, M., Heimann, M., and Ciais, P.: Inverse modeling of CO_2 sources and sinks using satellite data: a synthetic inter-comparison of measurement techniques and their performance as a function of space and time, Atmospheric Chemistry and Physics, 4, 523–538, doi:10.5194/acp-4-523-2004, URL http://www.atmos-chem-phys.net/4/523/2004/, 2004.

Kaminski, T., Heimann, M., and Giering, R.: A coarse grid three-dimensional global inverse model of the atmospheric transport 1. Adjoint model and Jacobian matrix, Journal of Geophysical Research, 104, 18 535–18 553, doi:10.1029/1999JD900147, URL http://www.agu.org/pubs/crossref/1999/1999JD900147.shtml, 1999.

Nassar, R., Jones, D. B. A., Kulawik, S. S., Worden, J. R., Bowman, K. W., Andres, R. J., Suntharalingam, P., Chen, J. M., Brenninkmeijer, C. A. M., Schuck, T. J., Conway, T. J., and Worthy, D. E.: Inverse modeling of CO_2 sources and sinks using satellite observations of CO_2 from TES and surface flask measurements, Atmospheric Chemistry and Physics, 11, 6029–6047, doi:10.5194/acp-11-6029-2011, URL http://www.atmos-chem-phys.net/11/6029/2011/, 2011.

Olsen, S. C. and Randerson, J. T.: Differences between surface and column atmospheric CO2 and implications for carbon cycle research, J. Geophys. Res., 109, D02 301, doi:10.1029/2003JD003968, URL http://dx.doi.org/10.1029/2003JD003968, 2004.

Peters, W., Krol, M. C., Dlugokencky, E. J., Dentener, F. J., Bergamaschi, P., Dutton, G., Velthoven, P. v., Miller, J. B., Bruhwiler, L., and Tans, P. P.: Toward regional-scale modeling using the two-way nested global model TM5: Characterization of transport using SF6, Journal of Geophysical Research: Atmospheres, 109, n/a—-n/a, doi:10.1029/2004JD005020, URL http://dx.doi.org/10.1029/2004JD005020, 2004.

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding : Theory and Practice, World Scientific Publishing Co Inc, URL http://www.amazon.com/Inverse-Methods-Atmospheric-Sounding-Planetary/dp/981022740X, 2000.