We thank the reviewers for their constructive and helpful suggestions. We have provided our responses to the reviewers' comments and believe that our manuscript is much improved as a result.

The main paper improvements are:

- Section 2. Method was revised;
- More details regarding the NIES TM and FLEXPART modelling;
- Figures 7 and 10 are updated;

The reviewer's specific comments (shown in blue) are addressed below.

Anonymous Referee #2

Received and published: 2 June 2016

This manuscript presents a new method for selecting satellite data for validation against the ground-based TCCON sites. This method is based on calculated footprints that represent the sensitivity of the site to surrounding CO2 concentrations. This new method is compared to simpler geographic-based data selection methods. This paper is methodologically sound but needs a few improvements in the description of the method, therefore, I recommend this manuscript for publication after minor changes have been made.

General comments

Method: more explanation of the FLEXPART modelling should be given. Specifically, which meteorological data were used, how the footprints are calculated and at what temporal and spatial resolution. Also, how the NIES TM model was used to initialize the FLEXPART simulations.

Whole section 2. Method was revised in order to provide more clear description.

The authors state that the TCCON observations are mostly sensitive to the lower troposphere, and that is why virtual particles were released at 1000 m. The authors however, do not take into account the vertical sensitivity of the TCCON measurements. What is the influence of assuming that the measurements are only sensitive to heights of circa 1000 m?

In this work we focused on study of the footprints of short-term (around one week) variation in XCO2 observed by TCCON and GOSAT. This variation is mainly managed by change of CO2 concentration in boundary layer (Keppel-Aleks et al. 2012).

Could the authors include a sensitivity test using multiple releases in FLEXPART to represent the averaging kernel?

Sensitivity test using multiple releases in FLEXPART to represent the averaging kernel is very expensive task, as it requires running trajectories for several months for the free troposphere and several years for the stratosphere. We have no available computational system to perform this task.

The colocation methods are compared for different GOSAT retrieval products. If the footprint-based method is considered the most comprehensive colocation method, would it be useful for also validating/assessing the different retrieval products by comparing these against the TCCON data. It would be interesting to include this comparison between

retrieval products in the manuscript. This would also make the manuscript of greater interest to the community.

Validating/assessing of the different retrieval products is a quit complicated and self-sufficient task, though we think such kind analysis is out of the scope of this manuscript. Moreover, recent papers cover this problem quit well: Oshchepkov, S., Bril, A., Yokota, T., Morino, I., Yoshida, Y., Matsunaga, T., Belikov, D. A., Wunch, D., Wennberg, P., Toon, G., O'Dell, C., Butz, A., Guerlet, S., Cogan, A., Bosch, H., Eguchi, N., Deutscher, N., Griffith, D., Macatangay, R., Notholt, J., Sussmann, R., Rettinger, M., Sherlock, V., Robinson, J., Kyro, E., Heikkinen, P., Feist, D. G., Nagahama, T., Kadygrov, N., Maksyutov, S., Uchino, O., and Watanabe, H.: Effects of atmospheric light scattering on spectroscopic observations of greenhouse gases from space. Part 2: Algorithm intercomparison in the GOSAT data processing for CO2 retrievals over TCCON sites, J. Geophys. Res., 118, doi: 10.1029/2012JD018782, 2013.

It is more reasonable to assess the different retrieval of OCO-2 products, as this instrument provides better coverage and larger number of observation points.

Specific comments

P3, L4-8: Suggest adding the years when data are available from each satellite, e.g., SCIAMACHY was discontinued in 2012 and OCO-2 is only available since mid-2014.

Done

P3, L28-32: I suggest that the authors make it clear here that the region and time period selected is for selecting the satellite data, just to make it unambiguous.

Done

P4, L10-11: This sentence needs a bit more explanation. It should be stated that Bremen, Garmisch etc. are TCCON sites, and the acronym JPL should be explained. Also, it is not clear in which averages these sites are not included – is this a different method again, if so it needs explanation.

P4, L11-12: It looks here as though the authors forgot to remove their own comment?

Yes, it is true. The paragraph "Bremen, Garmisch, Four Corners, JPL, and Izaña are influenced by local effects or complex terrain and are not included in averages (Kulawik ATM 2016). Limitations of the techinks !!!" was deleted.

P5, L4-5: It is unclear how the CO2 concentration fields from NIES TM are used to initialize the backward simulations with the LPDM. Also, the LPDM, FLEXPART, needs wind fields from e.g. meteorological reanalysis, which wind fields were used?

P5, L14-17: Firstly we run NIES TM for the target period (January 2010 to February 2011) using ten year's spin-up to ensure reduction of initialization errors. Then NIES TM CO_2 concentrations sampled at the location of TCCON sites at the level of 1 km above ground at 13:00 local time were used to initialize backward tracer simulations with the FLEXPART model.

P6, L7-8: Both models are driven by the Japanese Meteorological Agency Climate Data Assimilation System (JCDAS) datasets (Onogi et al., 2007).

P6, L11-14: It needs to be explained more clearly what the FLEXPART calculated footprints represent. Only from the figures is shown that the footprints are sensitivities to CO2 concentrations (at 1 km?) and have units of ppm per umol/m2/s but this is not explained in the methods.

The FLEXPART calculated footprints represent the source-receptor relationship which is an important concept in air quality modelling. It describes the sensitivity of a "receptor" element y to a "source" x. The receptor could be, for example, the average concentration of a certain atmospheric trace substance in a given grid cell during a given time interval (P. Seibert and A. Frank, 2004). The traditional way of evaluating the footprint is to overlay the area with a grid and sum the footprint value of the particle touchdowns within each grid cell. Using this method, the footprint strongly depends on the selected grid spacing and on the number of particles released in the simulation (Kljun et al., 2002).

For more details please check the following papers:

> Kljun, N., Rotach, M. W., and Schmid, H. P.: A 3D backward Lagrangian footprint model for a wide range of boundary layer stratifications, Bound.-Layer Meteorol. 103, 205–226, 2002.
>
> Seibert P, Frank A.: Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mode, Atmospheric Chemistry and Physics, 23, 4(1), 51-63, 2004.

Is it true that the footprints can be understood to represent a type of averaging kernel of the CO2 concentrations surrounding the site?

No. Footprint represents area to which observation site is sensitive.

Averaging kernels (AK) are a linear representation of the weighting of information content of retrieval parameters. AK related to the partition of information determined from the radiances and the prior information (usually a first guess) and usually employed for assessment of vertical information.

P8, L28: I think the authors mean "additional use of CO2 observations"?

Disagree. CO observations are necessary to distinguish biosphere and anthropogenic fluxes.

P9, L24-29: Some clarification is perhaps needed here; cases C1-4 were using a cutoff limit of the footprint and cases C5-8 were standard geographical comparisons?

Yes, exactly.

P10, L12-13: While it is true that the bias is smallest with one observation, this could just be by chance? How significant are the differences between the biases?

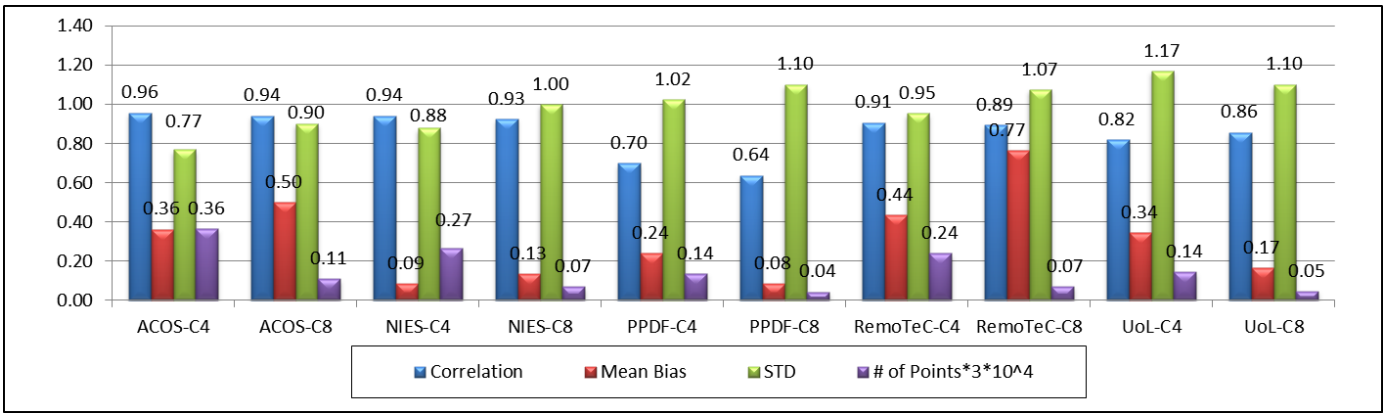
The difference between the two considered colocation methods is determined by the approaches to select spatial region around a TCCON site for selecting the satellite data. Case studies show how this spatial region may differ for cases 1-4 and 5-8. However, limited and uneven distribution of GOSAT data does not allow performing a more detailed analysis. Nevertheless, even a small reduction in a bias looks promising, as even a small bias can significantly affect the results of inverse modeling.

P11, L2: It's not clear what the authors mean by "colocation efficiency" if they mean the method, then there is a quite strong influence on the number of observations include in the comparison. Or do the author's rather mean that there is no dependence of the TCCON-GOSAT agreement on the colocation method?

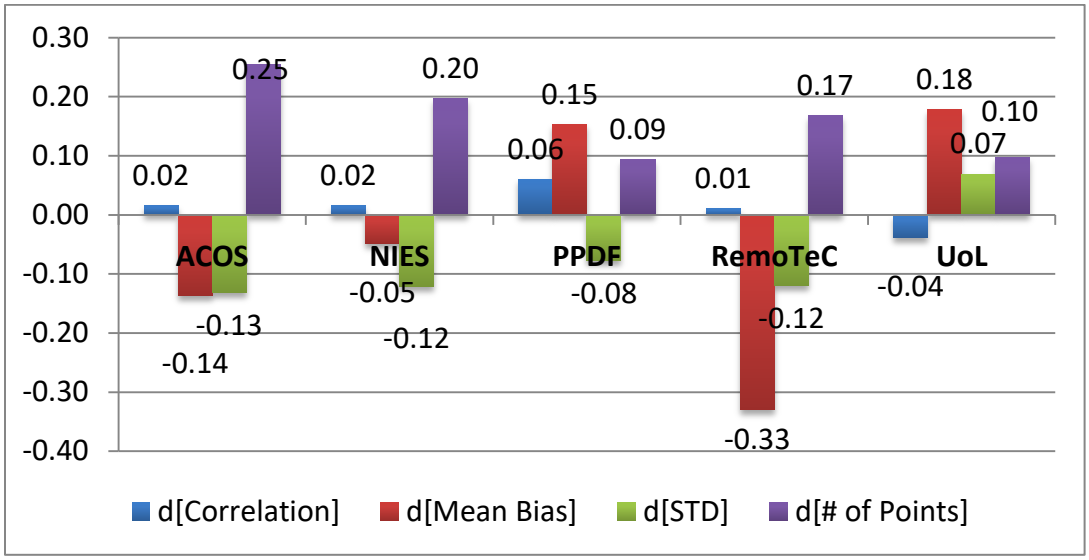
Here we stated that colocation efficiency doesn't depend on the number of observations include in the comparison.

Figure 7 and 10. By showing the difference between the two methods it is not clear which performs better. Instead it would be clearer and more meaningful to show these parameters (correlation etc.) for both methods.

Disagree. It is quite messy to show all these parameters (correlation etc.) for both methods. See below.



To make the figures more clear we added data labels:



It would also be helpful to briefly state again what each method is in the caption.

Done

Technical comments

P11, L1: please change to “differs by approximately a factor of 5”

Done