Response to referee #1

Major comments:

1. The paper is focused on model evaluations. But no necessary detail on these models such as the surface CH4 fluxes, meteorological fields, model resolution, and chemistry scheme are presented (although references are provided), or are used to explain their different performances (for example Figure 3).

All the three models optimize CH_4 field against in situ measurements at the surface through inversions of the CH_4 emissions. The chemical reactions considered in the models are the oxidation by OH in the troposphere, and by Cl, OH and $O(^1D)$ in the stratosphere. The fields of the radicals are prescribed monthly with no interannual changes. A description added in the first paragraph in page 5 to illustrate those. The meteorological fields, model resolution are included in Table 2.

2. Biases in GOSAT retrievals, and their implications on the model evaluations have not been discussed by the authors. GOSAT XCH4 has not been fully validated (particularly)over tropical regions, and could itself have latitude-dependent bias as well. I suggest the authors use more recent version of GOSAT XCH4 retrievals (such as OCPR v7) as well.

We follow the advise of the referee, and the version OCPR v7 of GOSAT data is used in the Figure 2.

Minor comments:

1. Line 37, Page 1: '... 6.2+/-11.2' ppb in the stratosphere ... ' The notation of +/-11.2 ppb may be mis-leading, as in this case the 'amplitude' as defined by the authors could not be negative.

The symbol '+/-' has been changed to '±' in the text.

2. Figure 1: Caption and main text does not provide necessary information, for example, the information about IMECC, and aircore data etc.

The FTS data are averaged for the in situ measurement periods. The IMECC is an aircraft campaign over Europe (Geibel et al., 2012). The Lamont-AirCore measurements are from Greenhouse Gas Group Aircraft Program (http://www.esrl.noaa.gov/gmd/ccgg/aircraft/). The AirCore data at Sodankylä is from the FTS group there. Following this advise, the information has been added to the caption of Fig. 1.

3. Line 6, Page 4: '... infers dry air columns from the CO2 columns retrieved from the same spectra as used in the CH4 retrieval' The sentence is not clear, and no mention of model CO2 concentrations, which is one of the possible sources for biases in GOSAT proxy XCH4 data.

The sentence has been changed to ' .... spectra as used in the CH_4 retrieval. This method assumes the CO2 concentrations are known and provided by model simulations. ' at line 16 in page 4.

4. Line 14, page 4: '... F07_10 data are applied and measurements with less than 1.4 DOFS are filtered out...', More detailed information such as the observation coverage and errors will be helpful.

Validation of F07_10 data against to HIPPO measurements shows a bias of -8 ~ 5 ppb with standard deviations of 25 ~ 50 ppb below 100 hPa (Herman and Osterman, 2014). According to the advise of the referee the information has been added to the line 24~25 in page 4.

5. Line 30, Page 5: '... Figure 3 shows yearly and seasonal median model biases scaled by the fraction of the air column in the troposphere and stratosphere ...'. I suggest adding the number of the TCCON observations at different months to the plot. Also it is interesting to know whether TCCON retrievals have biases depending on the solar zenith angles.

There are 10 sites used in that plot, for all seasons, except for measurements at ZEP is absent during the season DJF and SON. This is clearly seen in Fig. 3. The TCCON products has been corrected for solar zenith angle dependence. So this bias should be minor.

6. Line 5, Page 6: '... one can see that the latitudinal pattern of model biases in total column-averaged CH4 results from both the stratosphere and troposphere for ...' Some explanation of different performances of the three models shown in Figure 3 in terms of surface fluxes, transport or chemistry scheme will be helpful.

All the models are optimized with respect to surface measurements already. So the surface emission might not been the main reason to the different performances. The tropospheric oxidation also directly influence the surface CH4 concentrations, and the optimization process should give emissions consistent with the prescribed OH field in each model. The only significant difference among the models could come from convection, the North-South transport, and the transport from the troposphere to the stratosphere. However, it is difficult to give some useful discussion since only column measurements are used to evaluate the models.

7. Line 20, Page 6: TCCON and in situ sites are selected to be located close to one another so that both instruments measure similar airmasses ::: The TCCON and in-situ measurements have different measurement frequencies. For example, availability of TCCON data usually has strong seasonal variations. How will these differences affect the results presented in Figure 4?

Yes, the TCCON measurement has a different sampling frequencies compared to the in situ measurement. In our analysis the measurement series of TCCON and in situ has been filtered to extract variations with temporal scale longer than 1.4 years and only multi-year (longer than 3 years) averaged results are used. Besides, the models are matched to TCCON and in situ measurements in time, respectively, and undergo the same analysis with the measurements. So the model performance against to the measurements should not be affected.

8. Table 3: typo: The latitude of the Lauder TCCON site should be -45.038.

We follow the advise, and this error has been corrected.

Response to referee #2

The comments by refree #2 mainly include two parts: 1) The tropopause applied to integrate the model CH_4 to obtain tropospheric and stratospheric column-averaged CH_4 should be check and the results could depend on the definition of the tropopause. 2) The comparison of the model with TES measurements does not support our conclusions.

1) For the first comment some sensitive tests have been conducted, a different definitions for the tropopause are applied to integrate the model outputs. That include, thermal tropopause according to WMO definition, the dynamical tropospause defined as 1.5, 2.5, 3.5 and 4.0 PUV surface in the extratropics and 380 K potential temperature surface in the tropics. The ECMWF-interim reanalysis data is used to calculate the tropopause.

The sensitive test is applied to TM3 and LMDz-PYVAR (TM5 has a similar configuration with TM3 and then not been tested). These sensitive tests show (see following plots) that, the tropospheric model bias almost is not affected by the selection of the tropopause, even for the unrealistically low tropopause of 1.5 PVU. The amplitude of the stratospheric mode bias changes between the thermal tropoause and the dynamical tropopause of 2.5~4.0 PVU. However, there is still not a consistent latitudinal gradient existing during whole year for the stratospheric model biases. So the conclusion that the latitudinal gradient in the model bias of total column-averaged CH_4 come from the troposphere is valid. The dynamical tropopause of 1.5 PUV give some latitudinal patten in the stratospheric model biases, but, that tropoause is unrealistically low and frequently reach 170 hPb (below 380 K potential temperature surface) in the tropics.

2) For the second comment, the results from TES actually support the conclusion that the inconsistence between the HIPPO and TCCON comparisons with the models come from the longitudinal dependence of latitudinal gradient in the tropospheric model bias. But there are writing errors in the figure caption of the Figure 6 in the manuscript. In the third panel of Figure 6, the black points correspond to HIPPO sampling area (110°W~150°E) and the red points to the region beyond it.

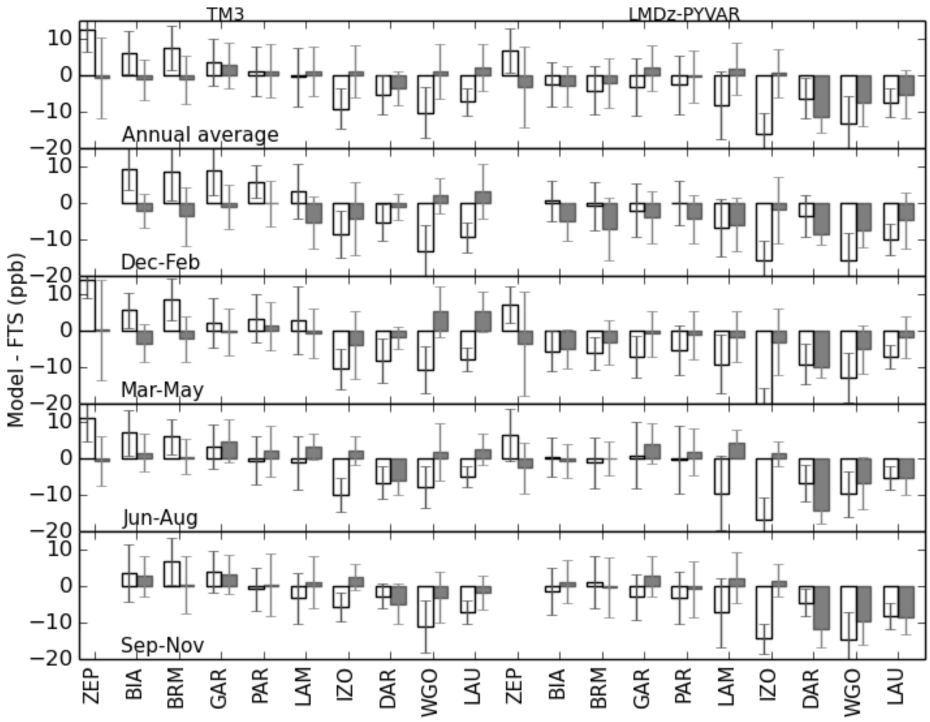

Figure 1. Yearly and seasonal medians of the scaled stratospheric and tropospheric contributions in modeled total column biases at TCCON sites. The sites from left to right is North to South. The white bar denotes the tropospheric bias, the grey bar for the stratospheric bias. The scale factor for the model bias are the air column fractions $P_t/1000$ (stratosphere) and $(1-P_t/1000)$ (troposphere), where P_t is the thermal tropopause pressure. The error bar are the standard deviations of the model biases. The results are averaged for 2007-2011 when FTS measurements are available.

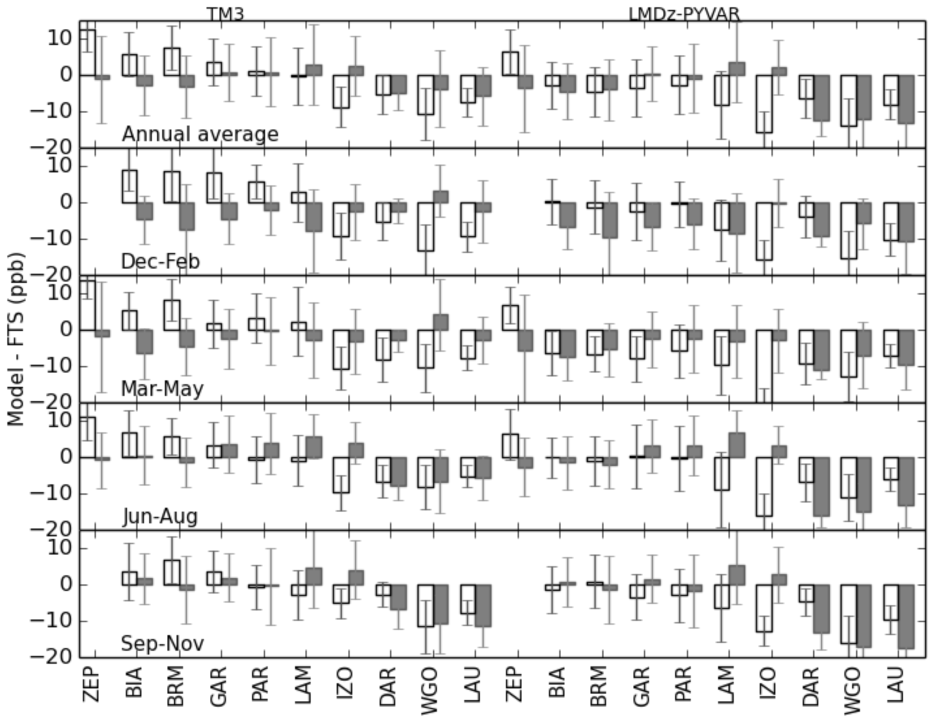

Figure 2. Same as Fig. 1 except for 4.0 PUV dynamical tropopause is applied.

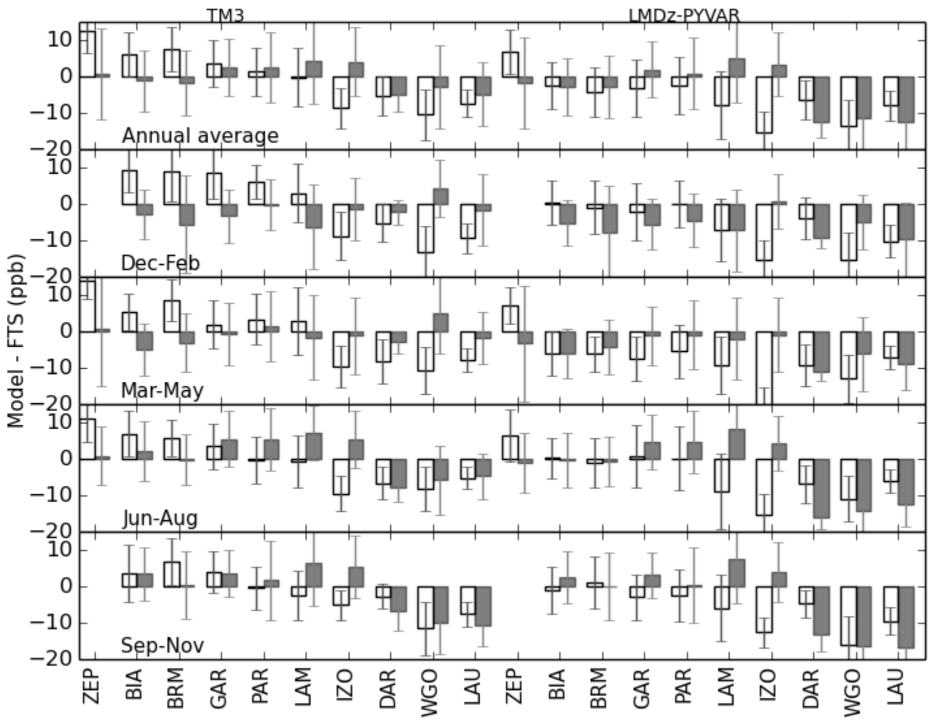

Figure 3. Same as Fig. 1 except for 3.5 PUV dynamical tropopause is applied.

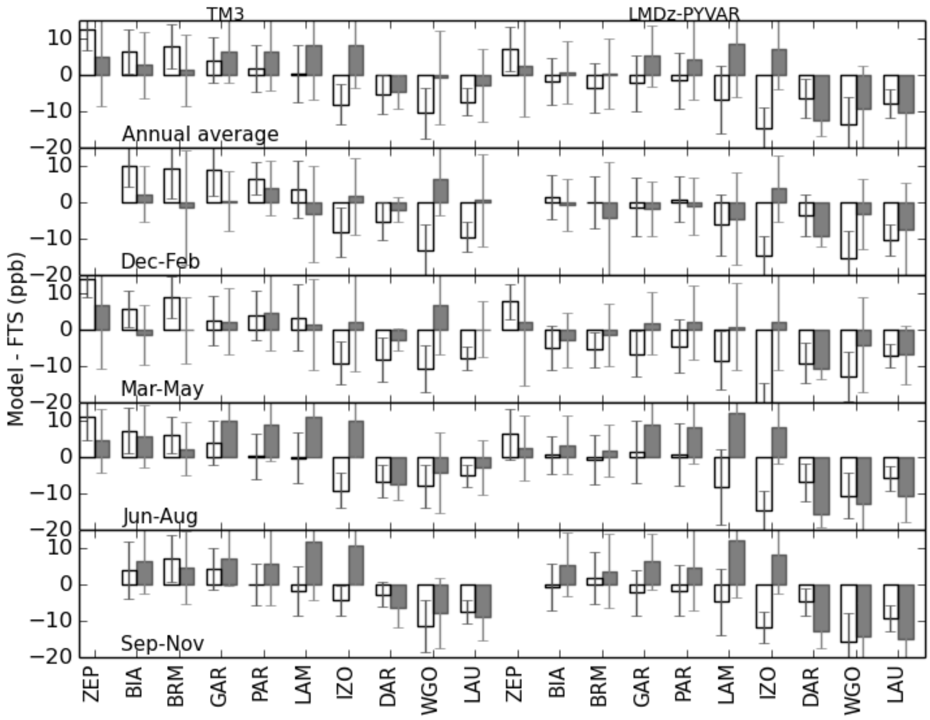

Figure 4. Same as Fig. 1 except for 2.5 PUV dynamical tropopause is applied.

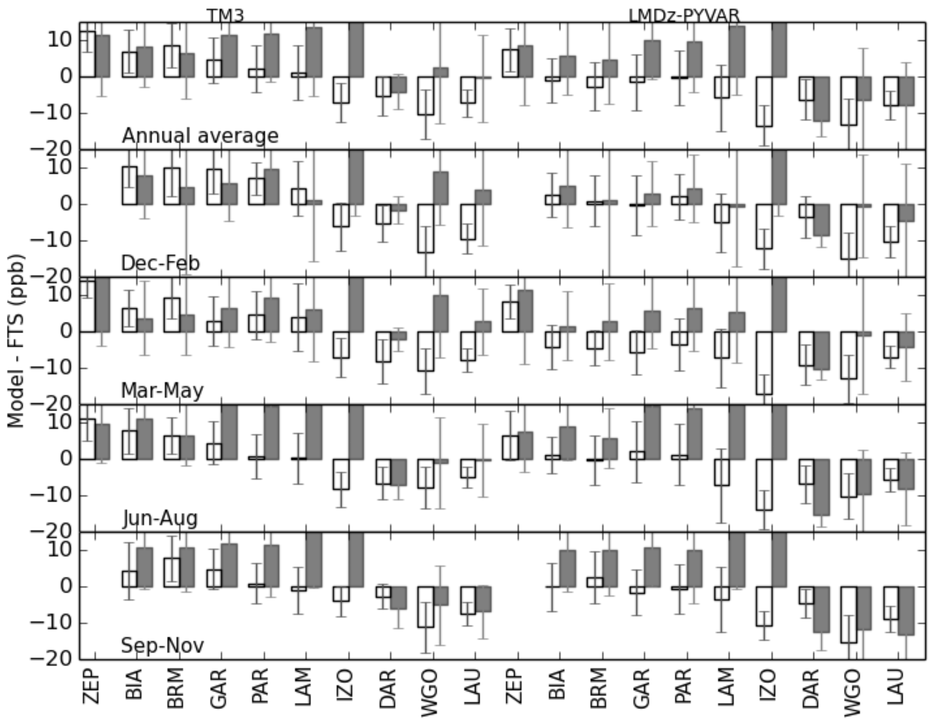

Figure 5. Same as Fig. 1 except for 1.5 PUV dynamical tropopause is applied.