Response to reviewer #2

We thank reviewer #2 for his or her insightful comments on the paper, in particular on improving the evaluation section. We were very impressed by the reviewer's attention to detail in spotting inconsistencies in the listing of the chemical mechanism in the supplement.

We would like to respond to the review as follows. The reviewer's comments are put in italics. Our suggestions for changes to the text are given in quotation marks.

p.7736, l.26: WRF-Chem is a regional model, but it is implied to be global in this sentence.

> We agree with reviewer #2 that WRF-Chem is mainly a regional model but there are also global applications of the model, e.g. Zhang et al. (2012).
>
> We will replace in the manuscript "WRF/chem (Grell et al. 2005)" with "GU-WRF/chem (Zhang et al., 2012)"
>
> Zhang, Y., P. Karamchandani, T. Glotfelty, D. G. Streets, G. Grell, A. Nenes, F. Yu, and R. Bennartz (2012), Development and initial application of the global-through-urban weather research and forecasting model with chemistry (GU-WRF/Chem), J. Geophys. Res., 117, D20206, doi:10.1029/2012JD017966.

p. 7755 and Table 2: North America ozone average of MOZAIC profiles and ozonesondes – It does not seem valid to average together all the stations of US and Canada. I would not consider Atlanta and Vancouver as having similar conditions at all.

> We agree with reviewer #2 that spatially averaging the observations and model results over a larger area needs to be done with caution. For the sake of generalisation of the global results we choose rather large areas. As we show monthly mean values averaged over layers of 200-300 hPa, we concluded that our approach is scientifically sound. The averages are calculated in such a way that stations/airports with more observations get a greater weight in the regional average.
>
> The varying data availability was a major motivation to average the MOZAIC profiles over North America. For example Vancouver and Toronto had no observation from April to September whereas Dallas had most of the observation in this period. We therefore averaged over a larger number of airports to obtain a more complete monthly time series for 2008. Because of the airport location and the number of available profiles, the plots are dominated by observation in the Eastern US. Please find for your information below the time series plots separately for the eastern and western North-America and for the whole sub-continent (Fig R1). We argue that the differences in the CO bias of different regions compared to the North-American biases is not worth including the region-specific plots in the paper.
>
> We will add in section 3.2.1
> "Because of the varying data availability the North-American mean is dominated by the airports in the Eastern United States. "

Tilmes et al. (ACP,2012, doi:10.5194/acp-12-7475-2012) shows significant differences among 4 ozonesonde sites spread across N. America, recommending against averaging them together for model evaluation. Please explain in more detail how this comparison was done. Was the model extracted for each site and then averaged? It seems it would be better to determine a model-measurement bias for each site, and then perhaps it is ok to average the biases.

We were happy to follow the reviewer suggestion to consider sub-region in North-America for the evaluation with ozone sondes as suggested by Tilmes et al. (2012). We also divided the Tropics in three sub-regions as suggested by Tilmes et al. (2012). We found that more detailed information can be gained but also that the smaller averaging regions were more susceptible to data gaps and distortion by outlier observations. (The Tilmes et al. (2012) data set is probably less affected by these problems since they consider a 17 year period whereas we only study one year.) The structure of the biases in the sub-regions did confirm the conclusions drawn from the larger averaging area. We therefore suggest to discuss the biases for three sub-region in North-America as well as for three sub-region in the Tropics but to include the corresponding pictures in the supplement (see below Fig R2 and R3).

We will add in section 3.2.1

"Tilmes et al. (2012) suggest a further refinement of the North-America region into Canada, Eastern and Western United States as well of the Tropics into Atlantic/Africa, equatorial Americas and Eastern Indian Ocean/Western Pacific based on the inter-comparison of ozone sonde observation for the 1994-2010 period. The results will be discussed also for these sub-regions and corresponding figures will be presented in the supplement. "

in section 3.3
"A more detailed breakdown of North America (Canada, Eastern and Western United States) and the Tropics (Atlantic/Africa, equatorial Americas and Eastern Indian Ocean/Western Pacific) following Tilmes et al. (2012) is presented in the supplement."

in section 3.3
"The LT underestimation occurred in all regions but was largest in early spring over Canada. C-IFS also underestimated over North America in MT. LT summer time ozone is overestimated in North-America by all models, in particular over the Eastern United States. The bias of C-IFS was the smallest in LT but in contrast to MOZ and REAN C-IFS underestimates summer time ozone in MT over the Eastern United States. The overestimation of UT ozone by MOZ was most pronounced in Canada."

in section 3.3
"A more detailed analysis for different tropical regions shows that the seasonality is mostly well captured by all models over Atlantic-Africa, equatorial America and eastern Indian Ocean/Western Pacific in all three levels. Only the maxima occurring in equatorial America in September were underestimated by up to 15 ppb in MT and UT."

p. 7757: The description of the MOPITT data set is not written very clearly. Was the Level 3 product used, or did the authors perform their own gridding to 1x1 degree? Presumably the model profiles were transformed, taking into account the a priori profile as well as the averaging kernel (this should be stated more clearly -
l.21-23 seems a little confused - it is 2 operations).
l.20: The increased sensitivity at the surface of the joint (NIR+TIR) retrieval is due to the inclusion of the NIR channel.
l.24-26: I don't understand the point of this sentence.

We used level 2 data and individual MOPITT pixels were binned onto a common 1x1 degree grid. The averaging kernels in combination with the a-priori profile were applied to the model profiles of CO.

The respective section (section 3.2.2) has been re-written as follows:
"MOPITT is a multispectral thermal infrared (TIR) / near infrared (NIR) instrument onboard the TERRA satellite with a pixel resolution of 22 km. TERRA's local equatorial crossing time is approximately 10:30 a.m. The MOPITT CO level 2 pixels were binned within 1x1° within each month. Deeter et al. (2013a) report a bias of about +0.08e18 molec/cm2 and a standard deviation (SD) of the error of 0.19e18 molec/cm2 for the TIR/NIR product version 5. This is equivalent to a bias of about 4 % and a SD of 10% respectively assuming typical observations of 2.0 e18 molec/cm2. For the calculation of the simulated CO total column the a-priori profile in combination with the averaging kernels (AK) of the retrievals were applied. They have the largest values between 300 and 800 hPa. The AK have been applied to ensure that the difference between retrieval and AK-weighted model column is independent of the a-priori CO profiles used in the retrieval. One should note however, that the AK-weighted column is not equivalent to the modelled atmospheric CO burden anymore."

p.7758: As with the MOPITT description, it is not clear if the authors performed some of the processing of the GOME-2 retrievals or if they are describing the product they used. Please clarify. Was any transformation of model profiles performed to account for the sensitivity of the GOME-2 columns to the true profile (i.e., averaging kernels or airmass factors)?

The modelled tropospheric columns have only been interpolated to the times and location of the observations. As the uncertainty in the NO_2 and HCHO retrievals are considerable, AK have not been applied to the modelled tropospheric columns. The retrieved and modelled tropospheric columns have been compared at time of and location of the satellite observations. Air mass factors were used for the retrievals.

We will in section 3.2.2

" For comparison to GOME-2 data, model data are vertically integrated without applying AK to …"

We also added a reference for the HCHO retrievals
"(Vrekoussis et al., 2010)."

Vrekoussis, M., Wittrock, F., Richter, A., and Burrows, J. P.: GOME-2 observations of oxygenated VOCs: what can we learn from the ratio glyoxal to formaldehyde on a global scale?, Atmos. Chem. Phys., 10, 10145-10160, doi:10.5194/acp-10-10145-2010, 2010.

p. 7760: The bias in surface ozone in MOZART could be at least partially due to a recently documented error in the dry deposition calculation for all versions of MOZART, which led to reduced deposition velocities than intended, and thus over-estimate of surface ozone, as described in Val Martin et al. (GRL, 2014, doi:10.1002/2014GL059651).

We agree that differences in dry deposition velocities are also an important factor and will add in section 3.3:

"The recently reported (ValMartin at al. 2014) missing coupling of the leaf area index to the leaf and stomatal vegetation resistance in the calculation of dry deposition velocities could be an explanation for the MOZ bias."

Val Martin, M., Heald, C. L. and Arnold, S. R.: Coupling dry deposition to vegetation phenology in the Community Earth System Model: Implications for the simulation of surface O3, Geophys. Res. Lett., 41, 2988–2996, doi:10.1002/2014GL059651., 2014.

p. 7761: It would be much easier to follow the arguments about the size of biases if the actual bias were plotted for each model. For example, the argument that "the bias of MOZ seems stronger over land" is hard to verify from these plots.

We decided to show the simulated TC (AK applied) rather than the biases because they give a better impression of the actual fields. Please find below (Fig R4 and Fig R5) the biases corresponding to Fig 6 and Fig 7. We find that the bias in MOZ follows the land-sea patterns more than in the other models. However, it is difficult to exclude the possibility that the contrast is caused by satellite data retrieval rather than the model. The retrieval is sensitive to changes surface temperature and albedo

.

Fig. 10 is only mentioned in passing in between discussion of Figs 8 and 9. It should be put in order and discussed more completely.

We will describe Fig. 10 in more detail:

"The outcome of the comparison with LT CO from MOZAIC is consistent with the model bias with respect to the GAW surface observations in Europe (Figure 10). The winter biases were larger than summer biases and MOZ showed the largest underestimation. The GAW stations measuring CO are mostly located on mountains in the Alpine region and typical annual biases were about -15, - 20 and -35 ppb for REAN, C-IFS and MOZ respectively. Biases of stations in flatter terrain such as Kollumerwaard tended to be larger."

p. 7762, l.18-19: The altitude levels that have the highest sensitivity for MOPITT should not have any bearing on the performance of the model, if the averaging kernels and a priori have been taken into account. I would remove this sentence.

In this sentence we refer to REAN, which assimilated CO from MOPITT. We think it is a valid statement.

All technical errors have been corrected.

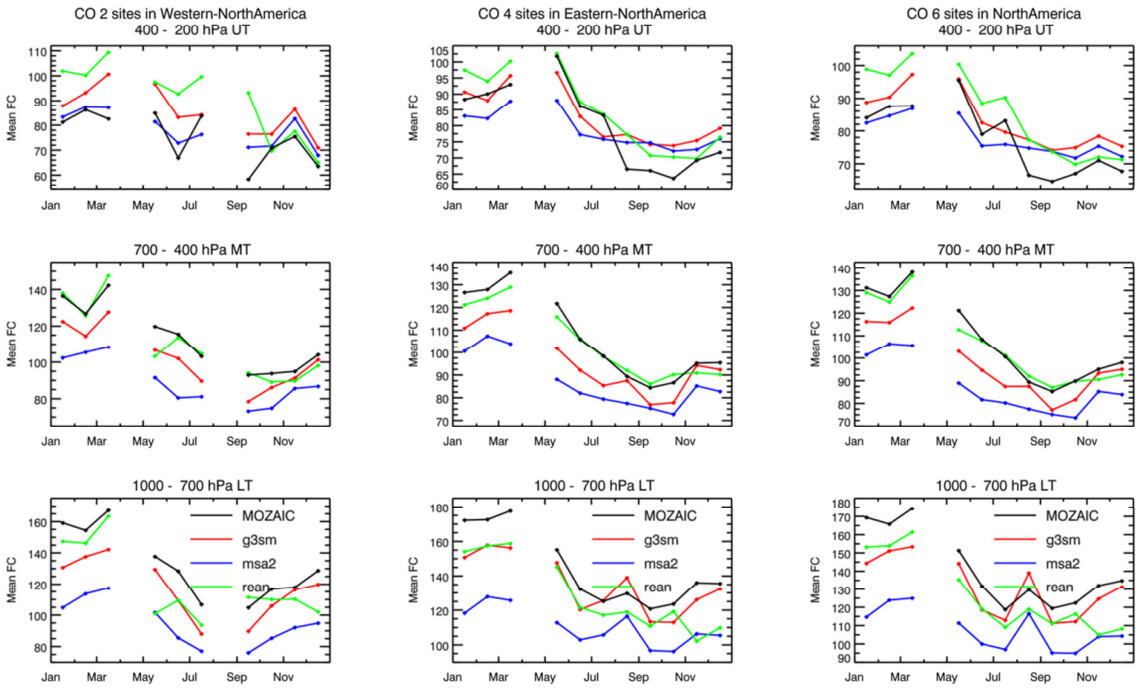

Figure R1 CO volume mixing ratios (ppb) over Western North America (left, 2 sites), Eastern North America (middle, 4 sites) and whole North America (right, 6 sites) averaged in the pressure bands 1000-700 hPa (bottom), 700-400 hPa (middle) and 400-200 hPa (top) observed by MOZAIC and simulated by C-IFS (red), MOZ (blue) and REAN (green) in 2008.

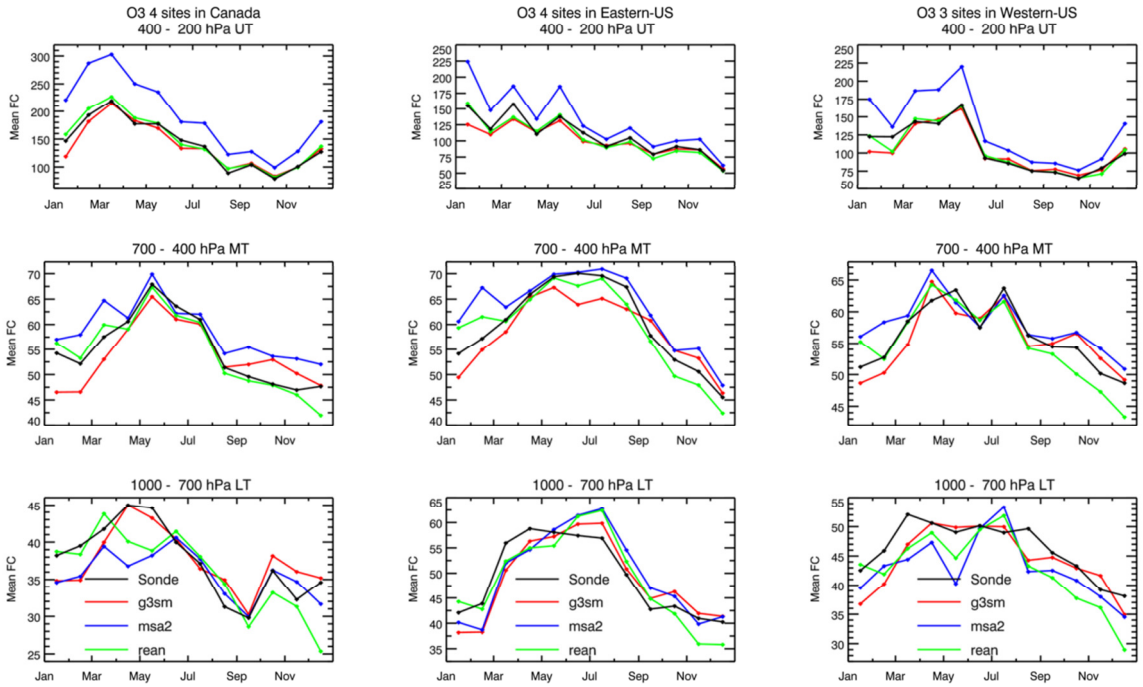

Figure R2 Tropospheric ozone volume mixing ratios (ppb) over the Western-US (right) and Eastern-US (middle) and Canada (left) averaged in the pressure range 1000-700 hPa (bottom), 700-400 hPa (middle) and 400-200 hPa (top) observed by ozonesondes (black) and simulated by C-IFS (red), MOZ (blue) and REAN (green) in 2008

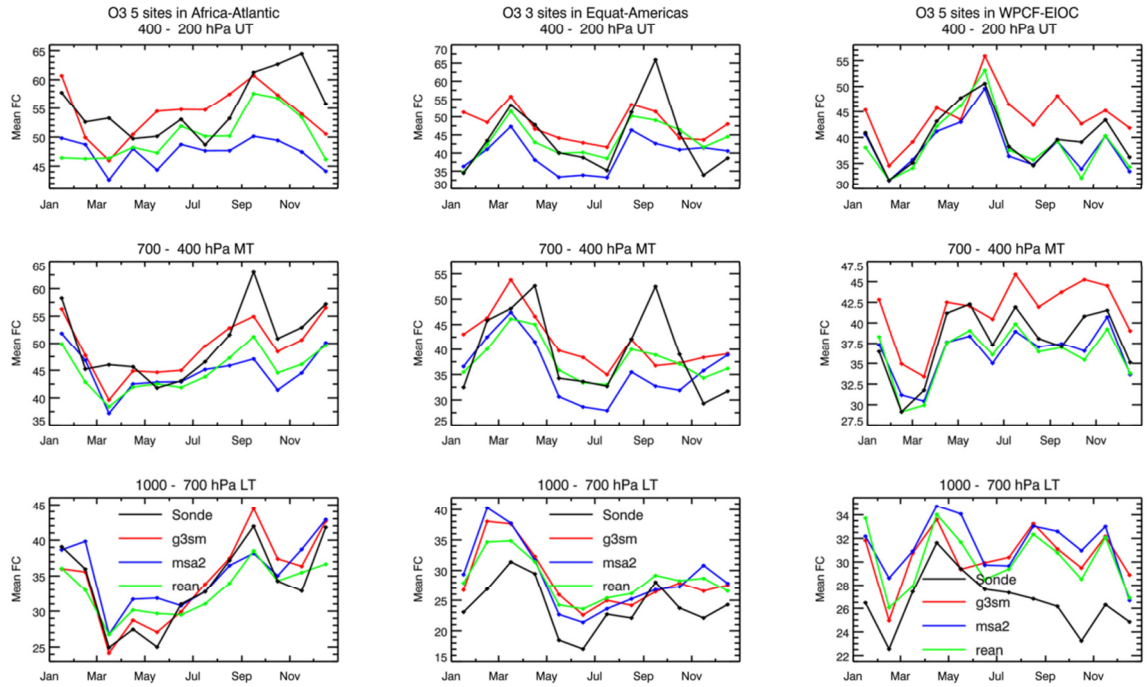

Figure R3 ***Tropospheric ozone volume mixing ratios (ppb) in the Tropics over Atlantic-Africa region (left) and Eastern Pacific and Indian Ocean (right) and equatorial Americas (middle) averaged in the pressure range 1000-700 hPa (bottom), 700-400 hPa (middle) and 400-200 hPa (top) observed by ozonesondes (black) and simulated by C-IFS (red), MOZ (blue) and REAN (green) in 2008***

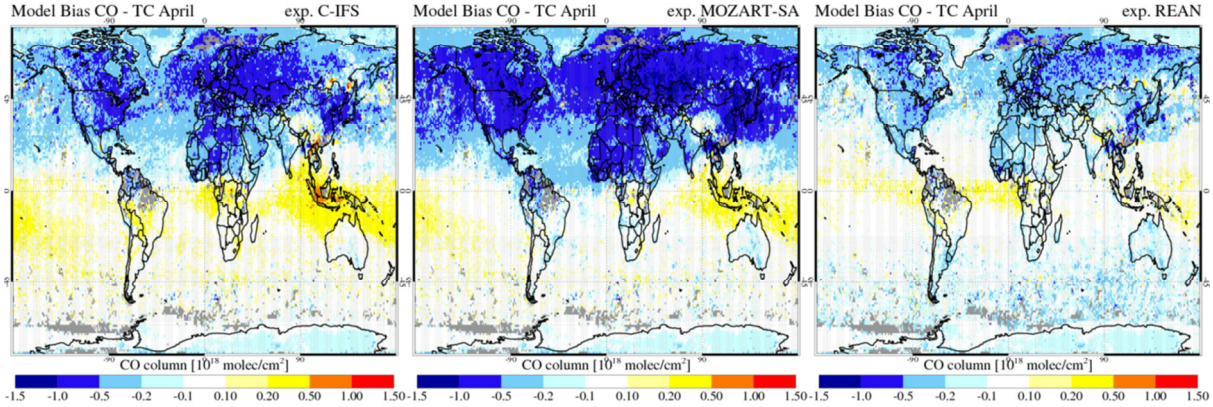

Figure R4 Bias of CO total column with respect to retrieval MOPITT V6 for April 2008 of C-IFS (left), MOZ (middle) and REAN (right).

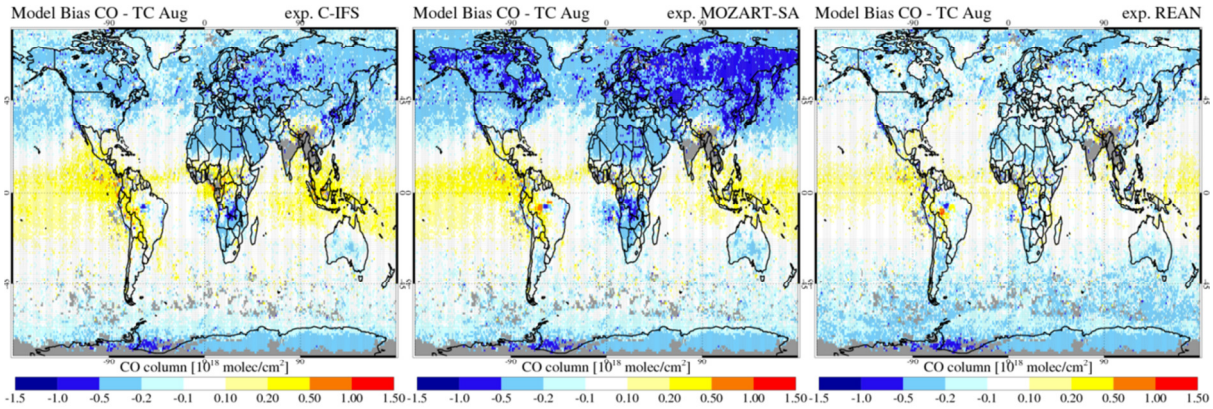

Figure R5 Bias of CO total column with respect to retrieval MOPITT V6 for August 2008 of C-IFS (left), MOZ (middle) and REAN (right)