We thank the two referees for their thoughtful and useful comments, and we have incorporated their suggestions into the revised manuscript. The comments from each referee are copied below, with our responses in blue and new text in **bold**.

Anonymous Referee #1:

Overall, this is a well written manuscript with clear goals. While I'm typically left disappointed by model intercomparison (MIP) papers, I appreciate the authors' efforts to disentangle the influence of differences in model chemistry and transport on the simulated vertical structure of CO in the southern hemisphere. I have no major concerns, only minor comments.

Minor Comments

I found it frustrating that Zeng et al. (2014) was referenced repeatedly, but it wasn't available to me: Zeng, G.,: Multimodel assessment of the influence of uncertainties in biogenic emission estimates on the distribution of CO and HCHO in the Southern Hemisphere, in preparation, Atmos. Chem. Phys. Discuss., 2014. Ideally, the two articles should have been submitted simultaneously. I really would have liked to seen the model evaluation with observations, including satellite observations.

We apologize for the inconvenience. Zeng et al. (2015) is now available in ACPD, and can be reviewed at http://www.atmos-chem-phys-discuss.net/15/2615/2015/acpd-15-2615-2015.html. Model evaluation in that work is focused on surface and ground-based total column observations rather than satellite observations.

The abstract reads a bit like a laundry list. I suggest that the authors clearly articulate the importance of this paper. What is new and exciting? Why should anyone want to read this paper? I wasn't too intrigued by the abstract.

Thanks for the suggestion. We have now rewritten the abstract:

"**The combination of low anthropogenic emissions and large biogenic sources that characterizes the Southern Hemisphere (SH) leads to significant differences in atmospheric composition relative to the better studied Northern Hemisphere. This unique balance of sources poses significant challenges for global models. Carbon monoxide (CO) in particular is difficult to simulate in the SH due to the increased importance of secondary chemical production associated with the much more limited primary emissions. Here, we use aircraft observations from the 1991–2000 Cape Grim Overflight Program and the 2009–2011 HIAPER Pole-to-Pole Observations, together with model output from the SH Model Intercomparison Project, to elucidate the drivers of CO vertical structure in the remote SH. Observed CO vertical profiles from Cape Grim are remarkably consistent with those observed over the southern mid-latitudes Pacific 10-20 years later, despite major differences in time periods, flight locations, and sampling strategies between the two datasets. These similarities suggest the processes driving observed vertical gradients are coherent across much of the remote SH and have not changed significantly over the past two decades. Model ability to simulate CO profiles reflects the interplay between biogenic emission sources, the chemical mechanisms that drive CO production from these sources, and the transport that redistributes this CO throughout the SH. The four chemistry-climate and chemical transport models included in the intercomparison show large variability in their abilities to reproduce the observed CO profiles. In particular, two of the four models significantly underestimate vertical gradients in austral summer and autumn, which we find are driven by long-range transport of CO produced from oxidation of biogenic**

compounds. Comparisons between the models show that more complex chemical mechanisms do not necessarily provide more accurate simulation of CO vertical gradients due to the convolved impacts of emissions, chemistry, and transport. Our results imply a large sensitivity of the remote SH troposphere to biogenic emissions and chemistry, both of which remain key uncertainties in global modeling. We suggest that the CO vertical gradient can be used as a metric for future model evaluation as it provides a sensitive test of the processes that define the chemical state of the background atmosphere."

First paragraph of Introduction: Line 12. CO levels in cleaner areas of the remote SH are often simply determined by the methane-CO-OH cycle. Methane oxidation is an important source of atmospheric CO. Possibly you mean the variations in CO levels are primarily caused by transport.
Thanks for catching this omission. We have revised the sentence to read:
"Much of the SH is characterized by very low CO emissions, and in these remote regions CO is largely controlled by the balance between long range transport, **production from methane oxidation**, and chemical removal via reaction with OH."

Fourth paragraph of Introduction: Are there any Japan Airlines data and MOZAIC aircraft data of use? These are longer records, but mainly in the UT. However they do provide gradients across the equator.
We have added the following:
"**Ongoing programs such as IAGOS/MOZAIC that conduct measurements from aboard commercial aircraft have been limited in the SH, with most concentrated over the African outflow region of the equatorial Atlantic. Neither of these programs included flights over the Pacific or Indian Oceans; however, IAGOS flights to Australia began in late 2013 and will likely provide a valuable additional SH dataset in future.**"
We contacted the PIs of CONTRAIL (Japan Airlines) in mid-2014, but at that point they did not have quality-controlled CO data that we could use in this project.

Why don't you show the vertical structure of OH between the four models over the SH?
We have now added OH to the longitude-altitude cross sections in Figures 9 and S4. We also now refer to these plots in Section 5:
"**The effects of the enhanced isoprene uplift in NIWA-UKCA are compounded by lower OH in the UT in this region (Fig. 9b).** The net result for both CH2O (Fig. 9c) and CO (Fig. 9d) is more UT production, **less UT destruction**, and therefore higher UT mixing ratios in NIWA-UKCA than GEOS-Chem."

In addition, vertical profiles of OH averaged over three SH regions (0-30, 30-60, and 60-90°S) from the four models are now available in Zeng et al. (2015), Figure 14.

I think a global and SH CO budgets for each model would be most helpful.
We have added a new Table 2 with global and SH CO budgets. GEOS-Chem rates were only archived for 2004, so we use 2004 as an indicative year in the budget.

We also now refer to the budgets in Sections 2.3 and 5.
Section 2.3: "Additional details of the model configurations and major differences between models are given in Table 1 and described in more detail in Zeng et al. (2015). **Indicative global and SH budgets for 2004 are shown in Table 2.** "

Section 5: "This is seen in Fig. 8f, which shows that the net balance between CO chemical production (P_{CO}) and CO chemical loss (L_{CO}) is more strongly weighted towards production in NIWA-UKCA, leading to slight enhancements in boundary layer CO over biogenic source regions (e.g., South America, Fig. 8a). **While differences in CO loss rates are likely partially responsible, we expect that CO production contributes more to the P_{CO}-L_{CO} differences** given the similarity of surface OH between models, particularly over South America where all models show OH titration (Fig. 8d). **The near-source surface differences between the two models are consistent with the whole troposphere budgets for the SH given in Table 2, which show total CO production is about 10% higher in NIWA-UKCA than GEOS-Chem, while total loss is about 5% lower.**"

Section 4.3: It seems that a better or additional experiment would be to use one model (e.g., GEOS-Chem) to run 4 tracer simulations (i.e., tagged CO), each one with one of the models' OH fields. This would remove each model's transport as a complicating factor.
We agree that this would be an interesting experiment that may provide additional insight into the relative importance of OH vs. transport in driving inter-model variability. However, setting up and interpreting the results of such an experiment would be non-trivial. For example, given the resolution differences between models, regridding the OH from one grid to another may result in vertical or horizontal offsets in the locations of major features and/or smearing of OH influences (e.g. going from coarser to finer resolution). Perhaps more concerning, this would decouple OH from the rest of the VOC chemistry – complicating interpretation of the role of chemical production (the primary goal of Section 4.3). Further, if we were to use the current version of the GEOS-Chem tagged CO simulation for such an experiment, it would only provide additional information on *primary* (rather than secondary) contributions, as CO chemical production is treated differently in the tagged CO model than in the full model used for the intercomparison. We therefore feel that such an experiment should be investigated in the context of future studies but is beyond the scope of this work.

D. Parrish, Referee #2:

Summary:
This paper presents an interesting discussion of measurement-model comparisons of CO in the southern hemisphere, and presents insightful analysis that is quite valuable. It is worthy of publication as it now stands. However, below are some specific suggestions for additional discussion to include in the paper that would further increase its value. Also, some relatively minor issues discussed below should be addressed before publication.

Suggestions for additional discussion:
This paper, in common with many such papers, presents a very useful comparison between results from different models and with measurements. Here the comparison was organized and presented in a manner to test important aspects of the models. Conclusions are reached regarding the importance of secondary CO production from biogenic hydrocarbons; these have important implications for the use of model inversion studies to correct emission estimates.

However, in the end this and other such model-measurement comparison studies do not have as much impact on model development as would be desirable. I think that this lack of impact arises from two problems. First, completing the model calculations, conducting the comparisons and publishing the results requires a significant period of time, likely years; over that time models evolve, and so it is not clear that the published comparison results are relevant for present-day models. Second, model-measurement differences point to shortcomings in the models, but exactly how to remove those short-comings and improve the models is not identified.

The following two suggestions for additional discussion may help to increase the impact of the work presented in this paper.
1) Future comparisons of model results with observations following the approaches presented in this paper will involve results from different (hopefully improved) models, but the observational data set will not change. The comparisons developed in this paper using these observations can lead to quantitative metrics for comparison with the new model results that can be conducted immediately after completion of the model simulations. If these metrics are explicitly tabulated in this paper, then their use in future model-measurement comparisons will be greatly facilitated, and thus more likely to effectively guide model evolution. Suggested below are two observationally derived metrics that could be tabulated in this paper.

- Seasonal cycles of median monthly CO observed in three altitude ranges as shown in Figure 3. Rather than tabulate the 12 monthly medians (with corresponding median deviations), it likely will be more effective to tabulate the few harmonic terms that define the seasonal cycle (Francey et al., 1999 effectively did this in their Figure 5). The derived quantitative metrics would then include the annual mean plus amplitude and phase (with confidence limits) of all statistically significant harmonic terms in the seasonal cycle in each altitude range. I expect that no more than two or three harmonic terms would be significant in each altitude range. Extracting these harmonics is equivalent to performing a Fourier analysis of the monthly median data. A similar analysis of any present or future model results followed by comparison of the measurement and model derived parameters would allow a prompt evaluation of the model calculated seasonal cycle.

– Median CO vertical profiles for each season as shown in Figure 4, much as was partially included in the top line of Table 2. Here the absolute concentration, which was normalized out in Figure 4, could be included along with the slope and other statistically significant terms (quadratic and perhaps even cubic, all with confidence limits) to fully describe the vertical profiles. These are only poorly described by the slopes (i.e., average vertical gradients) included in Table 2. Extracting these polynomial terms is equivalent to performing a power series expansion of the vertical profile data. A similar analysis of any present or future model results would allow a prompt evaluation of the model calculated vertical profiles. Here particular attention should be paid to properly describing the profile through the marine boundary layer, which as expected, appears uniform below 2 km in Figure 4.

We appreciate the suggestions. For the seasonal cycles, the first harmonic in each altitude range is now given in Table 5. Only these first harmonic terms were statistically significant. For the vertical profiles, the polynomial terms in each season are now given in Table 6. We have added discussion of these metrics and their potential value to the Conclusions section: "… and have not changed significantly over the past two decades. **The consistency between the two datasets further suggests that quantitative metrics derived from the CGOP observations can be used to diagnose model performance, both for the SHMIP models used here and more generally for future revisions of these and other models. Tables 5 and 6 provide tabulated observation-based metrics for the two salient features of the CGOP data: the seasonal cycle at different altitudes (represented in Table 5 by a harmonic fit), and the vertical profile in different seasons (represented in Table 6 by a polynomial fit). Tables S2 and S3 in the Supplement provide the equivalent parameters for the SHMIP models as a baseline against which to test future improvements to these models. The fitting methodologies are described in detail in the Supplement and can be easily applied to any atmospheric chemistry model for quick-look diagnosis of the ability to represent the SH free tropospheric CO background.**"

In the Supplement, we have added a description of the fitting methodologies in Section S1, figures showing the original and reconstructed time series (Fig. S5) and vertical profiles (Fig. S6), and tables with the fit coefficients for the models (Tables S2 and S3).

2) The authors have carefully investigated the factors affecting CO in the southern hemisphere. They have developed insightful analyses supporting their concept of these factors and tested the fidelity with which the four models replicate these factors. This work has put them in the position of having the best insight into model modifications required to improve that fidelity. At present the final section is more of a summary of the analysis than a true conclusion section. I suggest that the Conclusions Section be rewritten not as a summary of the analysis, but rather with the goal of clearly presenting the authors' recommendation of how these four models in particular, and all similar atmospheric models in general, need to be improved to accurately simulate the vertical profiles and seasonal cycles of CO in the southern hemisphere.

We have renamed the final section Summary and Conclusions and have added the following paragraph presenting recommendations for priority improvements:

"**The results presented here, along with the companion analysis of the SHMIP models presented in Zeng et al. (2015), point to biogenic NMVOC emissions and chemistry as clear priorities for improving atmospheric chemistry models in the remote SH. Isoprene and monoterpene emissions from tropical and SH sources remain highly uncertain even in state-of-the-science emission models like MEGAN and LPJ-GUESS (Holm et al., 2014; Stavrakou et al., 2014). In many data-poor parts of the world where**

biogenic sources are expected to be dominant, constraints on emissions are limited by fundamental uncertainties in the factors that cause plants to emit isoprene and other NMVOCs (Pacifico et al., 2009). Improving the process-based NMVOC emission models used to drive atmospheric chemistry models will be key to improving model ability to simulate the background atmosphere. Despite many recent advances, fundamental uncertainties also remain concerning the chemistry of NMVOC oxidation (Naik et al., 2013; Achakulwisut et al., 2015), with large impacts on CO as shown here. Ongoing work to advance our understanding of isoprene oxidation pathways, particularly in the low-NOX environments characteristic of much of the SH (e.g., Bates et al., 2014; Peeters et al., 2014; Liu et al., 2013b), should significantly improve simulation of SH CO production."

Relatively minor issues to address:

1) The bottom panel of Figure 1 needs more discussion. It seems to indicate that there are primary emissions of CO from biogenic sources, but I assume these must be the secondary source of CO produced from the biogenic emissions of isoprene (and monoterpenes?). Please clarify.

We include both a small primary CO source from biogenic emissions and a much larger secondary source (with seasonality given in the figure using isoprene emissions). We have now clarified this in the text:

"Figure 1 shows the mean seasonal cycle of primary CO emissions **from biomass burning, fossil fuel, and biogenic sources as well as** biogenic isoprene emissions (a proxy for secondary CO production) in the SH tropics and extra-tropics used in the standard SHMIP simulations."

We have also clarified in the figure legend:

"**The bottom panel shows biogenic emissions of both primary CO (black, left axis) and isoprene (gray, right axis), the latter used as a proxy for secondary CO production.**"

2) A particularly interesting feature that is only briefly touched upon in the paper deserves further discussion. Figure 3 shows that the peak of the seasonal cycle in the marine boundary layer is delayed by about a month from that observed in the free troposphere. What is the cause of this delay? Is it simply very slow transport between the free troposphere and the boundary layer? None of the models appear to reproduce this delay. Does this indicate that the models fail to strongly isolate the boundary layer from the free troposphere? Is this also indicated in Figure 5 where the observed uniform vertical profile below 2 km is not well reproduced by the models, at least in some seasons?

We have expanded the discussion of Figure 3 in Section 2.3 as follows:

"Figure 3 shows the median observed seasonal cycle of CO at Cape Grim averaged over 0–2, 2–5, and 5–8 km altitude bins (black line). The observations show increasing CO mixing ratios with altitude in all months, as previously reported by Francey et al. (1999) in an analysis of five years of the same dataset. Peak mixing ratios were observed in austral spring during the tropical BB season. **At altitudes below 2 km, the seasonal maximum occurred in October, as seen also in flask samples collected in surface air. This October peak in the boundary layer appears to represent a 1 month offset from higher altitudes, where peak CO was observed in September. However, the September maximum above 2 km is not statistically significant and is skewed by a large number of samples from September 2000 collected as part of the SAFARI aircraft campaign (Pak et al., 2003). As no measurements were made in other months in 2000, the SAFARI data cannot be considered indicative for the purposes of evaluating the annual cycle. Indeed, when**

these data are removed, the CGOP observations show peak CO mixing ratios in October at all altitudes, as shown in gray in Fig. 3 (note however that September and October remain statistically indistinguishable above 2 km). There does still appear to be a small delay between the boundary layer and the free troposphere, which may be indicative of slow mixing of transported BB plumes into the boundary layer.

The colored lines in Fig. 3 show the simulated seasonal cycles in the Cape Grim background region for the four individual SHMIP models. Despite large differences in absolute mixing ratios (discussed below), the models are generally able to reproduce the shape of the observed seasonal cycle especially above 2 km, as expected from previous studies (e.g., Shindell et al., 2006). **In the 5-year mean, the models show peak mixing ratios in September rather than October at all altitudes, but this timing varies from year to year. A particularly strong September peak is simulated by all models for 2005, reflecting significantly enhanced BB emissions in South America and southern Africa in the GFEDv3 inventory for this year (and leading to an outsize influence on the 5-year mean). None of the models capture the delay in peak mixing ratios in the boundary layer, suggesting errors in model representation of vertical mixing and/or boundary layer heights, both known issues in atmospheric transport models (e.g., Gerbig et al. 2008; Locatelli et al, 2013).** Model ability to match **other aspects of** the seasonal changes in the relationship between different altitudes is varied and is the subject of further discussion in Sect. 3. ”

3) Section 4.1 - The authors conclude "The differences in CO2.5 between models are much smaller than differences in total CO, especially in summer-autumn, suggesting inter-model differences in meteorology and transport are small relative to other drivers of variability." This conclusion requires some further discussion. In summer (DJF) and autumn (MAMJ) all models exhibit very small vertical gradients, which could indicate that all models have vertical mixing that is rapid with respect to the 25 day CO2.5 lifetime. In that case, large inter-model differences in meteorology and transport would not be apparent, but coupled with other altitude dependent inter-model differences, perhaps could still be significant drivers of variability. Later discussion in the paper (e.g. Figure 9) does indeed identify important differences in transport between models that drive inter-model variability. The discussion of this issue should be clarified.

This is an important point. We have moved this sentence to the start of a new paragraph, which now also draws on results from Zeng et al. (2015):

“The differences in CO_{25} between models are much smaller than differences in total CO, especially in summer-autumn. **This is consistent with results from Zeng et al. (2015), who examine CO_{25} columns over the entire SH and find both the magnitude and distribution to be similar across the four models. They also show that the small inter-model differences in CO_{25} columns are not reflected in the distributions of total CO columns, indicating a limited role for horizontal transport differences as a source of inter-model variability, at least over the 25 day lifetime of the tracer. As seen in Fig. 6b, all four models show similarly shallow CO_{25} vertical gradients in DJF and MAMJ. These similarities could reflect similar transport of primary emissions to the Cape Grim region and/or similarly rapid vertical mixing relative to the 25 day tracer lifetime, which would obscure the role of transport differences in driving inter-model variability. Given the lack of primary CO sources near the Cape Grim region, the latter is unlikely to have a major impact on primary CO gradients but may be important for inter-model differences in secondary CO. We explore this effect further in Sect. 5.**”

4) On pg. 27540, the authors write "A major component of the mean difference be- tween models is the difference in the simulated mean OH background, shown in Table 1." This statement does not really appear to be correct. The two models with the largest difference in mean CO mixing ratios (Figures 2 and 3) are TM5 and CAM-chem; yet these two models have similar global mean tropospheric OH concentrations (Table 1). The two other models that have intermediate mean CO have similar, but significantly larger, mean OH. This discussion should be clarified.

We have removed the sentence and replaced it with a more accurate summary of the results from Zeng et al. (2015):

"These differences in background CO are influenced by a number of factors including grid resolution, meteorological drivers, and chemical mechanisms as discussed in detail by Zeng et al. (2015). **In particular, they find that consistent inter-model differences in the SH CO background are largely driven by differences in CO production efficiency, with an additional contribution from differences in oxidizing capacity (especially for TM5, which has the lowest OH of the four models as shown in Table 1).**"

5) The observational data sets were collected over the period of 1991 to 2011 and the modeling is done for 2004-2008. There should be an expanded discussion of any long-term changes in the ambient CO concentrations over this two-decade period, with a focus on justifying the neglect of any such changes in the analysis presented.

We have expanded the discussion of our analysis of long-term changes in CO concentrations at Cape Grim. We have added the following paragraph to Section 2.3:

"Because of the temporal offset between CGOP (1990s), HIPPO (2009–2011), and the SHMIP simulations (2004–2008), we do not compare individual flights or profiles but instead focus on average behavior seen across multiple years in the observations and models. **Multiple studies have shown that trends in SH CO over similar time periods are either small (Zeng et al., 2012; Worden et al., 2013) or insignificant (Warner et al., 2013; Yoon and Pozzer, 2014), depending on the period and region analyzed, especially when El Niño years are neglected. We evaluated long-term CO trends specific to the Cape Grim region over the 1991–2008, 2004–2008, and 1991–2011 time periods relevant to this work using CSIRO flask samples collected in surface air at the Cape Grim Baseline Air Pollution Station. Results of this analysis, shown in Table S1 in the Supplement, indicate that CO trends at Cape Grim over these periods were not statistically significant on either an annual basis or for any individual season, justifying our use of long-term temporal averages.**"

We now also refer to this analysis in Section 3.2:

"This is consistent with a number of recent studies (and with our own analysis of Cape Grim surface flask data, **Sect. 2.3**) showing observed trends in SH CO are much smaller than interannual variability (Zeng et al., 2012; Wai et al., 2014; Warner et al., 2013; Worden et al., 2013; Yoon and Pozzer, 2014). "

6) The HIPPO investigators are acknowledged in the paper, but are not included as coauthors. Were they invited to join as coauthors? If not, such an invitation should be considered.

The lead HIPPO PI and all HIPPO investigators involved in the CO measurements used here were invited to join as co-authors. They collectively opted for acknowledgement rather than co-authorship given the limited use of HIPPO data in the final manuscript.

Additional changes:

In response to external comments on the manuscript, we have made a few additional modifications as described below.

In Section 3.1, we now discuss likely causes of the differences between CGOP & HIPPO:
"The figure shows that although the relative variability in CO (thin lines) differs somewhat between HIPPO and CGOP, there is generally overlap in the observed ΔCO from each dataset (thick lines). **Small differences between the two are likely driven by (1) BB plumes from Africa and South America experiencing more dilution during transport to the Pacific than to Cape Grim, and (2) sampling of Australian BB outflow during HIPPO but not CGOP. Both of these factors should be most influential in austral winter-spring, when SH burning is at its peak (also the period when the datasets show the most variability).**"

In Section 2.3, we now clarify our use of the Cape Grim background region rather than a single grid square containing the observation site for comparison to observations:
"For comparison with observations from CGOP, which measured only clean background air, we sample each model over the Southern Ocean southwest of Tasmania. We reduce the influence of model spatial variability on the comparisons by averaging each model over four representative grid squares in this **region (referred to hereafter as the Cape Grim background region). These grid squares, shown in Fig. 2, were chosen to minimize the influence of outflow from the Australian continent (which we cannot filter directly as only monthly mean model output was archived and radon was not simulated as part of SHMIP). We tested the influence of our choice of sampling region by also performing our analyses using either the grid square containing the CGOP profiles or the nearest ocean-only grid square (as done for TRANSCOM, e.g. Law et al., 2002; Loh et al., 2015). We found that changing the sampling region did not significantly impact the shape of the model profiles or the relative differences between the models, suggesting our results are robust to this choice.** Coordinates of the grid squares **in each model that define the Cape Grim background region** are given in Table 1, **with minor differences stemming from model resolution and grid spacing as shown in Fig. 2**."

In Section 3.2, we have expanded discussion of the biomass burning contributions in SON, and specifically the model-observation difference above 7 km:
"Fire emissions have been shown to influence Australia and the Cape Grim region via long-range transport in the mid-upper troposphere (Bowman, 2006; Gloudemans et al., 2006; Pak et al., 2003), driving the enhanced gradient above the surface in these months. **Simulated tracers of regional influence (CO_{25}, described in Section 4.1) show peak contributions from southern African BB at 4-7 km and from South American BB at 6-10 km.**

The ability of the models to capture the observed BB enhancement indicates that the models (all using GFEDv3 emissions) are successfully capturing the long-range transport of BB sources. **The main exception is the positive gradient simulated from 7-8 km in SON (versus the observed decrease over this altitude range in CGOP). The cause of the discrepancy is unclear. In the models the increase above 7 km reflects a larger contribution from South American than African BB at these altitudes, primarily in October. BB plumes are likely very dispersed at these altitudes following long-range transport, and this dispersion complicates simulation of the gradient. The otherwise good agreement between observed and simulated JA-SON gradients suggest**s that there has not been significant change in the major SH burning source regions that contribute to

background CO in the **Cape Grim** region since the 1990s (when the observations were collected)."

In the first paragraph of Section 4, we have clarified that the years used for sensitivity testing are representative of the full simulation:
"**As seen in the figure, simulated profiles during the 2004–2005 test period are generally similar to those for the full SHMIP period (Fig. 5).**"

In Section 4.2, we have expanded the discussion of differences between Figs. 6b and 6c:
"… indicative of the impacts of **OH-driven chemical loss. The CO_{OH} lifetime varies spatially and seasonally (due to OH variability), and in winter-spring can be significantly longer than 25 days. As described by Zeng et al. (2015), the CO_{OH} mixing ratios therefore provide a more realistic metric than CO_{25} for evaluating the combined impacts of transport and loss of primary CO.** "

In Section 4.3 (paragraph 2) we have added discussion of the methane contribution:
"**Chemical production of CO originates from oxidation of both methane and NMVOCs, and inter-model variability in the vertical gradients may reflect contributions from both. In remote regions, the methane source dominates the CO burden while the NMVOC source dominates the variability (Pfister et al., 2008). Differences in the methane mixing ratios in the four models (Table 1) are thus more likely to affect overall concentration differences (e.g., Fig. 3) than differences in the vertical gradient. However, the methane contribution cannot be quantified from the archived SHMIP output. Instead, we perform a final sensitivity test to** evaluate the role of the NMVOC source in driving the simulated CO vertical gradients. Figure 6d shows the result of replacing MEGAN-CLM biogenic emissions with LPJ-GUESS for isoprene and monoterpenes. **Methane, OH, and other emissions remain unchanged from the standard simulation.** Since emissions are the same across models, they cannot explain inter-model variability; however, they can help attribute sources of model-observation bias as well as provide insight into the dependence of the simulated vertical gradients on biogenic **NMVOC** sources. The figure shows that **relative to the standard simulation**, the LPJ-GUESS emissions reduce the simulated CO vertical gradient in summer-autumn in all models. In winter-spring, the differences are negligible. **The small increases in gradient from Fig. 6c to 6d reflect both methane and NMVOC contributions (which are smaller but still significant in LPJ-GUESS).** These results present a picture consistent with the previous sensitivity tests; namely, that observed vertical gradients are driven in winter-spring by primary BB emissions and in summer-autumn by secondary CO, **largely of** biogenic **NMVOC** origin."

We have corrected an averaging error that resulted in strange behavior for NIWA-UKCA at 7-8 km in August 2006. This led to updated versions of Fig. 5 (now more consistent in JA with Fig. 6) and Table 1.