<u>Response to interactive comment on "Assessing stratospheric transport in the CMAM30 simulations using ACE-FTS measurements" by Felicia Kolonjari et al. from Referee #2</u>

RC = Reviewer Comment
AR = Author Response

RC: This manuscript describes a detailed set of comparisons between specified dynamics simulations with the CMAM chemistry climate model and satellite-based observations of stratospheric long-lived tracers, with the goal of identifying discrepancies and therefore errors in the simulated stratospheric dynamics. The topic is suitable for ACP, and the conclusions reached by the authors are generally reasonably well supported by the analysis presented.

AR: The authors thank Reviewer #2 for their helpful comments on the manuscript. We have addressed both the general and specific comments below.

General comments:
RC1. I have no doubt that the "advanced" comparison technique–which samples the model data along the actual line of sight of the satellite instrument–is the best way of minimizing sampling errors in the CMAM vs. ACE-FTS comparison. But, the explanations given in Sec. 3.3 don't quite make sense to me. If the difference between the advanced and intermediate methods is so small, as shown in Fig. 4c, this implies that the "line of sight" sampling is actually making very little difference to the sample means. The difference between advanced and basic (fig 4b) is much larger, which means that the most important source of sampling error has to do with a bias in the distribution of samples within each 5deg latitude bin (which is fixed by performing the 2D horizontal interpolation rather than using the closest neighbor gridpoint). It's possible these issues would be easier to sort out if Fig 4 showed differences between the 3 methods and the full model sampling, rather than differences between the advanced sampling and the other sampling methods. Or perhaps BASIC-FULL, INTERMEDIATE-BASIC, and ADVANCED-INTERMEDIATE. In any case, I believe the logic of the explanations here could be sharpened.

AR: We chose to represent this comparison as the difference between the Advanced method and the others to illustrate the differences between what would be "observed" by ACE-FTS compared to the full model and common sampling techniques. For a model at T47 resolution, it is true that the intermediate and advanced sampling are very similar. However, for models with finer spatial resolution, the differences would be more apparent. Therefore, we have used the advanced method for ACE-FTS/model comparisons. The text has been clarified to minimize the expectation for a comparison of each method to the full model output.

Original sentence: "To determine the impact of sampling the model output at varying levels of detail, three methods were tested and compared to the full model output (the CHAM30HR output at all latitudes and longitudes for each 5 ° latitude bin) between June 2004 and May 2010."

New sentence: "To determine the impact of sampling the model output at varying levels of detail, three methods were tested by sampling the full model output (the CMAM30HR output at all latitudes and longitudes for each 5° latitude bin) between June 2004 and May 2010."

RC2. No doubt there is much more information given in the Ray et al. (2016) paper, but Sec. 6 requires a little more guidance on the set up of the TLP simulations. At some point, in passing, we learn that there were 480 TPL simulations, but it is not said how these simulations differ; presumably certain input assumptions are varied in the different simulations, but not, it seems, w and epsilon directly. Also, the terms w* and epsilon should be better defined. At some point, w* is introduced as a mean tropical upwelling, but it is later used to quantify vertical motion in the extratropics. It's also not really clear if epsilon is a prognostic or diagnostic variable, and how it depends on height, latitude, time, etc.*

AR: To maximize clarity and minimize additional text, the second paragraph of section 2.3 has been moved to the beginning of section 6 and the first sentence of section 6 has been moved to the end of section 2.3.

The sentence moved to section 2.3 from the first sentence of section 6 has been changed to: "The TLP model is used here to identify the changes to the CMAM30HR tropical upwelling and effective mixing that may improve the comparisons between ACE-FTS and CMAM30HR. However, this tool does not identify a specific mechanism but it can isolate seasons and regions in which changes are required."

The following sentences were added to the end of the paragraph moved (now the first paragraph of section 6): "There were 480 simulations initialized with different combinations of w^* (velocity of tropical upwelling) and ε (the mixing efficiency) settings. The fraction of the CMAM30 w^* used to initialize the TLP model ranged from 0.20 to 1.24 and the ε ranged from 0.18 to 1.50. In each TLP simulation run, the relationship between mean circulation and mixing is constrained by the vertically-averaged mixing efficiency [Ray et al., 2016]. The mixing efficiency in the TLP model is defined as $\varepsilon = \alpha / \lambda\tau$, where α is the ratio of tropical to extratropical mass, λ is the rate of the mean circulation influence, specifically the mass flux out of the tropics due to mean circulation, and τ is the mixing time or time scale for mass flux between the tropics and extratopics [Ray et al., 2014; Garny et al., 2014]."

RC3. I have trouble following the logic from Figure 15 to 16. Figure 15 seems to show that best agreement with the ACE-FTS measurements is achieved running the TLP model with parameter settings which produce the smallest w and largest epsilon values (at least in Fig 15a,b,c). In fact, it seems that the range of TLP simulations is not large enough to find the actual best agreement with ACE-FTS–a point which could be discussed. But then, in Figure 16, it is implied that e.g., best agreement with ACE-FTS is produced with no significant change in w* values in the tropics. Something seems inconsistent here.*

AR: To clarify the link between changes in CMAM that may improve the comparisons to ACE-FTS, we have updated Fig. 16 to include the individual contributions of both CFC-11 and CFC-12. The thresholds used for all latitude ranges in the new Fig. 16 have also be reduced to 0.2 (as was used for the tropics) to make the regional comparisons more consistent. Deciding to use a more

restricted but consistent threshold leads to no result in some seasons for some geographic regions, perhaps suggesting that the range of *w and *ε* values should be extended.**

RC4. For many of the difference contour plots, it would be helpful if the colorbars were chosen such that positive differences could be more easily differentiated from negative differences. An example is Fig 6, where it is very difficult to know whether the CMAMACE differences in the UTLS are positive (like the middle stratosphere differences) or negative.

AR: The colorbars of the difference contour plots have all been changed to a red/white/blue scheme where white is zero.

Specific comments

RC P1, l11: "The model consistently: : :" could be taken out of context–this conclusion is specific to the trace gases examined in this study (and probably wouldn't apply to ozone, for example).

AR: We have clarified this statement.

Original sentence: "The model consistently overpredicts tracer concentrations in the lower stratosphere, particularly in the Northern Hemisphere winter and spring seasons."

New sentence: "The model consistently overpredicts tracer concentrations of CFC-11, CFC-12, and N_2O in the lower stratosphere …"

RC P1, l14: the "too little isentropic mixing" should probably be connected if possible to a height or range of heights.

AR: We have clarified this statement.

Original sentence: "In particular, the CMAM30 simulation exhibits too little isentropic mixing in the June-July-August season."

New sentence: "In particular, the CMAM30 simulation exhibits too little isentropic mixing in the tropical lower stratosphere during the June-July-August season."

RC P2, l1: I'm not sure this is the only reason for the increase in interest in stratospheric transport–and it's a bit of a chicken and egg problem.

AR: We have included a reference to Butchart (2014) to support this statement.

Original sentence: "Interest in stratospheric transport has increased over the last 20 years as a result of significant developments in stratosphere-resolving general circulation models (GCMs) and chemistry-climate models (CCMs) (e.g., Pawson et al., 2000; Eyring et al., 2005; SPARC-CCMVal, 2010; Gerber, 2012)."

New sentence: "As highlighted by Butchart (2014), interest in stratospheric transport has increased over the last 20 years as a result of significant developments in stratosphere-resolving general circulation models GCMs (e.g. Pawson et al., 2000; Gerber, 2012) and chemistry-climate models (CCMs) (e.g. Eyring et al., 2005; SPARC-CCMVal, 2010)."

*RC P2, l4: *Accurate* projections rely on good models.*

AR: Accurate has been added.

Original sentence: "Projections of stratospheric ozone and climate rely on the ability of these models to simulate stratospheric transport and chemistry."

New sentence: "Accurate projections of stratospheric ozone and climate rely on the ability of these models to simulate stratospheric transport and chemistry."

RC P2, l5: Definitely the distribution of long-lived trace gases depends on the BDC…for short-lived species it may not have that much influence.

AR: The term "long-lived" has been added to the sentence.

Original sentence: "It is clear that the transport of chemical tracers will be impacted by changes in the BDC, which will in turn influence ozone recovery projections, lifetimes of ozone depleting gases, and mass exchange between the troposphere and stratosphere (Butchart, 2014)."

New sentence: "It is clear that the transport of long-lived chemical tracers will be impacted by changes in the BDC, which will in turn influence ozone recovery projections, lifetimes of ozone depleting gases, and mass exchange between the troposphere and stratosphere (Butchart, 2014)."

RC P3, l14: "has" or "have"? The word choice depends on whether the models project changes in the BDC (plural) or evidence of those changes (singular).

AR: Based upon the feedback of Reviewer #3, this sentence has been changed.

Original sentence: "Fundamental questions remain as to the mechanisms driving the stratospheric circulation because there has been evidence of changes in the BDC that has not been projected by models (e.g., Butchart, 2014; Mahieu et al., 2014)."

New sentence : "Discrepancies between observations and model projects may be due to the short time scales of observation systems (Butchart, 2014; Hardiman et al., 2017)."

RC P4, l31: ACE-FTS measurements have high vertical resolution, not the instrument itself.

AR: That is correct, the ACE-FTS retrieved profiles have a high vertical resolution. The sentence has been clarified.

Original sentence: "ACE-FTS is ideal for studying 30 the vertical structure of constituent gases from cloud tops to 100 km; it is particularly useful in the upper troposphere and lower stratosphere because of its high vertical resolution (Hegglin et al., 2008)."

New sentence : "ACE-FTS is ideal for studying the vertical structure of constituent gases from cloud tops to 100 km; the retrieved profiles are particularly useful in the upper troposphere and lower stratosphere where the vertical resolution is approximately 3 km (Hegglin et al., 2008)."

RC P8, l2: For comparisons of model to measurements, it's probably more intuitive to treat the measurements as the truth, and show relative differences of the model with respect to the measurements, rather than the mean of the measurements and model. It's of course not a big deal as long as it is clear how the calculation is being done, but I feel the simpler the calculation, the easier it is to interpret.

AR: The comparisons between measurements and model output of tracer concentrations throughout the paper used the same method of calculating comparisons with respect to the mean of the measurement and model. The authors agree that the comparison would be more intuitive with respect to the measurements. Therefore, all relative difference calculations have been updated throughout the paper and are now relative to the measurement.

RC P8, l8: this statement of significance applies to the 1 sigma confidence level. For 2 sigma, I guess all differences would be not significantly different from zero.

AR: True, we have clarified the sentence to reflect this detail.

Original sentence: "Over the time period compared, CMAM30HR appears to overpredict CFC-11 at all HATS sites while the CFC-12 comparisons are not significantly different from zero for all but two sites in the Southern Hemisphere."

New sentence : "Over the time period compared, CMAM30HR appears to overpredict CFC-11 at all HATS sites while the CFC-12 comparisons are not significantly different from zero within one standard deviation for all but two sites in the Southern Hemisphere"

RC P9, l24-26: Apparent contradiction between "increased isentropic mixing" and "slower shallow branch".

AR: You are correct. There is a problem with this description. Relatively old mean age in the lowermost stratosphere is due to either a relatively strong upper branch of the BDC compared to the lower branch, and/or less isentropic mixing since that process brings young tropospheric air

into the lowermost stratosphere. In the LMS, increased mixing brings in young air because it's primarily tropospheric.

Original sentences: "... CMAM30 clearly has older air in the extra-tropical lowermost stratosphere. It is potentially caused by either stronger downwelling of the older air from above, consistent with a stronger BDC, or increased isentropic mixing of tropospheric air from lower latitudes (e.g. Hegglin and Shepherd, 2007). Therefore, the differences in age appear to suggest a slower shallow branch or a faster deep branch of the BDC."

New sentences: "... CMAM30 clearly has older air in the extra-tropical lowermost stratosphere. It is potentially caused by either stronger downwelling of the older air from above, consistent with a stronger BDC, or reduced isentropic mixing of tropospheric air from lower latitudes (e.g., Hegglin and Shepherd, 2007). Therefore, the differences in age appear to suggest a slower shallow branch or a faster deep branch of the BDC."

RC P12, l19: the end of this sentence could be misconstrued, as of course there has been vertical interpolation applied in the translation to the vertical levels of the ACEFTS retrievals. I would suggest to remove this last part, and write ": : : location of the tangent points with altitude".

AR: We have made the change suggested.

Original sentence: "... no consideration of the variation in geographical location of the tangent points ..."

New sentence: "… no consideration of the variation in geographical location of the tangent points above or below 30 km."

RC P13, l25: But the INTERMEDIATE sampling technique is also limited to the 30 km tangent altitude, and the differences to the advanced method are much smaller. Indeed, by construction, differences between a method using the variable tangent heights and one using only the 30 km tangent height should be zero at 30 km (which appears to be the case in the advanced-intermediate comparison). Therefore, the differences between basic and advanced, which are strongest around 30 km, cannot be due to the line of sight sampling.

AR: The basic sampling uses a 30 km tangent height but the nearest model column and the intermediate sampling produces a profile by computed the bilinear interpolation of the concentrations of the nearest 4 grid boxes at each point along the column. It is the intermediate sampling that we should expect to have a zero difference with the advanced sampling at 30 km. This is the case in Fig. 4c.

RC P15, l26: how are large disagreements in the north polar region so confidently connected to problems with tropical upwelling? Could this not be an issue with mixing across the polar vortex?

AR: The sentence has been changed.

Original sentence: “The large disagreements in the north polar region during winter and spring indicate that the upwelling portion of the BDC across the different seasons is not well characterized by CMAM30HR.”

New sentence : “The large disagreements in the north polar region during winter and spring indicate that the downwelling portion of the BDC across the different seasons is not well characterized by CMAM30HR.”

RC P22, l17: I would avoid the term “mixing levels” as it could be taken as meaning isentropic surfaces.

AR: This has been changed to mixing efficiency throughout the manuscript, which is more accurate.

RC P22, l19: Is this result based on looking at the best agreement between CMAM and ACE-FTS under the natural variability of the CMAM simulations? Located here within the discussion of the TPL, it comes across a little as one has tuned CMAM, which I think is not the case. Also, “a reduction from the fitting estimate” is unclear to me, is this the best fit of the TPL parameters to the CMAM climatology?

AR: The TLP model was tuned to the CMAM30 output before it was used to test a range of w^* and ε values. Line 16 of page 22 states “... the TLP model was tuned to be representative of CMAM30HR by fitting estimates of the mean tropical upwelling (w^*) and mean mixing levels (ε) …”. The reduction from the fitting estimate refers to the change required in the CMAM30HR run based on the tuned TLP model.

RC P22, l21,22: First sentence implies best agreement when epsilon is increased–which, based on previous description I take to mean a mixing rate should be increased. But the following sentence says mixing “times” should be increased, which actually means rates should decrease. Some clarification here would be useful.

AR: We have decided to revise the paragraph to clarify our meaning.

Original paragraph 2 of section 6:

“A reduced mean circulation would likely correspond to less mixing and longer mixing times. Mixing levels are defined in the TLP by the ratio of horizontal mixing mass flux to horizontal mean mass flux scaled by the width of the tropical pipe region (Garny et al., 2014). Ray et al. (2016) found that the CMAM30HR simulations best match the ACE-FTS measurements when the w^* is between 0.27 mm/s and 0.32 mm/s (a reduction from the fitting estimate of 0.4 mm/s) and ε ranges from 0.7 to 1.2 (an increase from the fitting estimate of 0.55). Based on the in-mixing time profiles shown by Ray et al. (2016), it is apparent that the CMAM30HR mixing times need to increase to slow down the mixing at all levels, although the differences are only significant in the lower part of the stratosphere, below 20 km, and above 24 km. These are physically consistent changes since the mean circulation is driven by wave breaking, which also causes mixing between the tropics and extratropics.”

New paragraph 2 of section 6:

"Ray et al. (2016) found that the CMAM30HR simulations best match the ACE-FTS measurements when the *w is between 0.27 mm/s and 0.32 mm/s (a reduction from the fitting estimate of 0.4 mm/s) and ε ranges from 0.7 to 1.2 (an increase from the fitting estimate of 0.55). Ray et al (2016) found that since ε is inversely proportional to both λ and τ, there is a compensating effect with changes in *w** (λ) or τ. For the CMAM30HR changes derived, *w** needed to be slowed down significantly below 20 km, and above 24 km. For constant ε that would result in larger τ (less mixing). However, Ray et al (2016) found that ε also needed to be increased so there needed to be more mixing than would result from slower *w** and constant τ, but not enough of an increase in ε so that the mixing times were less (more mixing) than CMAM30HR has currently. With the increase in ε, mixing times are reduced but still longer than the current CMAM30HR mixing times."**

RC P23, l5: Figure 15 is quite dense, and really could use better description in the main text and figure caption. The "level of agreement" between CMAM and ACE-FTS needs to be explained fully, what kind of quantity is this? It took me some time to determine that the white-to-black shading and the white isolines were describing the same quantity. Also, there are bluish boxes which are barely detectable in the plots, are these the "agreement matrices"? As mentioned in the general comments, these agreement boxes don't seem consistent with the ACE-FTS "agreement" scale.

AR: We have clarified the text to make it clear that the white isolines are the same quantity as the grey color contours but we are not sure what the blueish boxes the reviewer is referring to. We do recall that on printing the paper sometimes, weird blueish boxes appeared on these plots. Please refer to the electronic version of the paper and check to see if the blue boxes are there as well.

Original sentence: "The white contours illustrate the agreement isolines."

New sentence: "The white-to-black shading is reinforced by white contours of the same quantity to illustrate the comparison more clearly."

RC P23, 19: How are these thresholds chosen? 0.65 seems like a rather lenient agreement threshold, since the agreement values shown in Fig 15 go as low as 0.1.

AR: As noted above, the thresholds have been changed to be 0.2 for both CFC-11 and CFC-12 in all regions (tropics and extratropics). This is reflected in the updated Figure 16.

RC P23, l20: what chemical species is Fig 16 based on?

AR: CFC-11 and CFC-12 were both used in Fig 16. Figure 16 has been changed, as previously noted.

RC P24, l32: If mixing and the meridional residual circulation are both driven by Rossby wave breaking, it is hard to see how insufficient mixing could be the cause of a too-rapid BDC.

AR: Our results show that the CMAM30 circulation needs to be slower but the mixing efficiency, *ε*, needs to increase. As explained in response to previous comments, increasing ε does not necessarily mean increasing mixing.