We would like to thank the reviewer for his valuable and helpful comments. The author response to these comments is given below.

Referee 2:

One of my two major concerns is in the comparison of simulated CO2 concentrations between the models, and also between the models and the actual observations, in Section 3.0. The authors state that "Inter-model differences [. . .] are about a factor of two smaller than the standard deviation of model-observation differences" (p.1277 lines 20-22). This means, in layman's terms, that the models are CLOSER to one another than they are to the actual observations (which could be caused either by errors in the underlying fluxes, or transport errors that are consistent between the models). The very next sentence, however, states "[. . .] the evaluation of these models against observations indicates the model bias in opposite signs, which explains the larger intermodel bias compared to the model (model-measurement) bias" (p.1277 lines 22-24). In layman's terms, this means that the observations fall in between the two models, and by extension that the models are FURTHER APART from one another than they are from the actual observations. This seems contradictory at face value, and difficult to assess visually from Figure 3. In thinking through this more, however, I realized that this seemingly contradictory result can be explained by two potential factors. (1) If the model-predicted concentrations are smoother (at the 3-hourly time scale) than the actual observations, but similarly smooth to each other, then the variance of the difference between either model and the observations is simply representative of this difference in smoothness rather than a measure of the similarity between the two model predictions in terms of overall variability over the examined period. (2) The standard deviations in lines 20-22 are calculated at the 3-hourly scale, whereas the bias in lines 22-24 is calculated (presumably) over the entire 29-day simulation; the fact that the models are more similar to each other in the first case, but more similar to the observations in the second case, may indicate that there is a time scale issue here, i.e. that the relative performance of the models relative to each other, and relative to the observations, is a function of the time resolution. Both of these factors could easily be examined using the existing results (e.g. by looking at the variance of the three time series, by comparing models to one another and to observations at for example, 12-hourly, daily, synoptic time scales), and I think that such an analysis is necessary to substantiate the presented argument.

We use standard deviation as the measure of uncertainty between models as well as model and observations. Since standard deviations are found to be about a factor of two smaller in inter-model comparisons than in the model-observation comparisons, we could argue that models are closer to each other than they are to the observations. The mean of the difference ("bias") does not give this information, as it is averaged over the whole period.

Following to your argument on the time scale issue, we examined data from the 163 m level for different time-scales and the results are given in the table below (Table 3). The mode-to-model standard deviation is about a factor of two smaller than the observations-to-model standard deviation, irrespective of the time-scale. It confirms that there is no significant dependence on time-scale of analysis on model discrepancies.

The text is modified as follows:

p.1277, line 22
included the following sentence:
"In addition, different time scales are used for this comparison in order to confirm that there is no significant dependence of time scale of analysis on model performance relative to each other, and relative to the observations (see Table 3)."

p.1292
included Table 3:
Table 3. Summary statistics of inter-model and measurement-model comparisons for different time scales of analysis. A tower level of 163 m at Ochsenkopf tall tower observatory for August 2006 is used for this analysis. See Table 2 caption for abbreviations.

Time scale	*System*	*sd (ppm)*	*Ratio of sd* of Obs - Model to std of WRF - STILT*
3 hourly	*WRF-STILT*	1.8	1.0
	Obs-STILT	3.0	1.7
	Obs-WRF	2.9	1.6
12 hourly	*WRF-STILT*	1.4	1.0
	Obs-STILT	2.3	1.6
	Obs-WRF	2.2	1.6
24 hourly	*WRF-STILT*	1.1	1.0
	Obs-STILT	1.9	1.7
	Obs-WRF	1.6	1.5
36 hourly	*WRF-STILT*	0.9	1.0
	Obs-STILT	1.8	2.0
	Obs-WRF	1.6	1.8

My second major concern is with the last sentence of the abstract (p. 1268 lines 23-24), and the last paragraph of the manuscript (p. 1286 lines 10-14). Here, the authors suddenly jump from a comparison of WRF and WRF/STILT to a very broad conclusion about the use of WRF/STILT as an adjoint for WRF. Although I understand that this may have been the ultimate goal of the analysis that was performed, the paper is not at all *aimed at testing the feasibility of using WRF/STILT as an adjoint to WRF, and the results of the analysis do not support this conclusion. For example, if this were indeed one of the main goals of the analysis, then much more quantitative and specific criteria would have needed to be defined that would characterize the types of errors /discrepancies that would, or would not, be acceptable if WRF/STILT is to be used as an adjoint for WRF. The authors make no attempt to do so. Instead, they present discrepancies (which they find to be rather large), discuss their possible causes, and run sensitivity tests to evaluate these possible causes and/or reduce their impacts. This is a great analysis, but does not support the final conclusion listed above. The only other step the authors take in the direction of this last conclusion is to compare the relative discrepancy between the models to the overall model-data mismatch. However, no objective criteria are presented in doing this comparison, and the comparison itself needs further investigation (see my other major concern above). I think that the simplest solution is to remove the last sentence of the abstract and the last paragraph of the manuscript. The manuscript will make a fine paper without them, and a substantial additional (and potentially substantially different) analysis would need to be conducted to support this conclusion.*

The last sentence from the abstract will be removed from the text.
p. 1286 lines 10-14 is modified as follows:

"Nevertheless, the similarity of the results provided by WRF and WRF/STILT at high resolution as well as the fact that the inter-model differences are a factor of two smaller than the model-observation differences and about a factor of three smaller than the mismatch between the current global model simulations and the observations, suggests the usefulness of STILT as an adjoint model of WRF. To achieve the definitive proof to justify the use of STILT as an adjoint of WRF, one would further need to carry out quantitative analysis of error characteristics between the models and to perform successful inversion using this model framework.

Other Comments

From the introduction, it was not entirely clear whether the manuscript is simply comparing CO2 fields from a prescribed set of fluxes, or optimizing these fluxes as part of the analysis. Figure 1 further contributes to this confusion because of the dual-direction arrows connecting WRF to VPRM, and STILT to VPRM. In fact, the manuscript focuses on comparing atmospheric distributions from a given set of fluxes, and therefore does not attempt to use WRF and WRF/STILT to independently optimize fluxes in an inverse modeling framework. The fluxes that are prescribed in the two models, however, are not identical (as described in Sections 2.1, 2.2, and 3.1.2), because they are driven by each model's (somewhat distinct) temperature and radiation fields. This somewhat confounds the analysis, and I am not sure whether it adds to the overall argument / discussion, which is to focus on the atmospheric transport as simulated by the two models. At a minimum, I think that the authors need to clarify the use of prescribes vs. optimized fluxes in the introduction. However, they may also consider simplifying the manuscript by using identical fluxes in both models.

We clearly state in the introduction that the Eulerian and the Lagrangian models use the same surface fluxes (p. 1271 line 8). STILT uses temperature and radiation fields as given by WRF as well as the same VPRM parameters to compute biospheric flux fields. As described in Section 3.1.2, the only reason for discrepancies in the biosphere fluxes is due to the temporal interpolation of these fields within STILT. However the associated error due to this flux discrepancy is found to be negligible.

p. 1271 line 2 will be modified as follows:

"Resulting footprints (sensitivities to upstream surface-atmosphere fluxes) are then mapped to high resolution surface fluxes that are prescribed from a diagnostic biosphere model, anthropogenic fluxes, as well as initial/lateral boundary conditions from a global model."

While reading the introduction, I immediately wondered about the impact of model resolution, the use of online vs. offline meteorology, etc. between the two models. The subsections of Section 3.1 then go on to a thorough analysis of the impact of these factors, but it would have been nice to let the reader know that these factors will be examined in the subsequent sections

p. 1271 line 18 will be modified as follows:
"… transport and mixing. Consistency between model parameters and its impact on tracer simulations is examined in detail."

The authors do a nice job of examining some of the possible causes of the discrepancies between the two models in Section 3.1, but some other factors are not discussed. Two of these are the choice of 100 particles per receptor/time and the use of a maximum of 3 days for the backward trajectories (p. 1275, lines 17-18). In addition, the impact of the use of the dynamically-adjusted "horizontal size of the grid cells for resolving the footprint" (p. 1275, lines 9-13), the need for which is related to the choice of using only 100 particles per release, is also not examined as a possible cause for discrepancies. These choices should at the very least be discussed. Preferably, the sensitivity of results to these choices should be examined.

We repeated the experiment with 1000 particles where found no significant changes in the results. The mean differences between two model simulations are negligible (in the range of 0.02 to 0.04 ppm), which confirms that our results show no sensitivity to number of particles used in STILT.
This information is added in the text as follows:

p.1278, line 9:
"….coordinate transformations during data processing procedures, (3) sensitivity to number of particles used in the Lagrangian model, (4) differences …..
p.1278, line 15:
"The results show no sensitivity to the number of particles used in STILT, giving rise to negligible bias (0.02 to 0.04 ppm) between STILT simulations with 1000 particles instead of 100. This confirms that the discrepancy is not caused by the choice of number of particles in the STILT. "

the use of a maximum of 3 days for the backward trajectories (p. 1275, lines 17-18).

Since we use relatively small domain (600 km x 600 km), a maximum of 3 days is sufficient for all particles to leave the domain. This information is added in the text as follows:

p.1276, line 4:

included the sentence:

"A maximum of 3 days is sufficient for all particles to leave the outer domain."

I encourage the authors to consider moving Section 2.3, which describes the model domain and period of simulations, to the first subsection in Section 2, as it would be helpful to have this information prior to reading the current Sections 2.1 and 2.2. In addition, I may have missed this, but the full extent/size of the model domain, illustrated in Figure 2, is not described clearly in the current manuscript.

We still think that it would be better to provide information on the model domain and the period of simulations after briefly describing the model framework. The details on the size of model domain are now included in the text as follows:

p.1276, line 26:

"…a single day in 2008 (20 October 2008) for a domain centered over Ochsenkopf in Germany (see Fig. 2). The outer and inner domains have a total area of about 600 km x 600 km and 250 km x 250 km respectively.