Referee 1:
We would like to thank T. Nehrkorn for his valuable and helpful comments. The author response to these comments is given below.

General Comments:

One problem with the way the material is presented is that the material dealing with the month-long tower simulation results is somewhat disjointed from the aircraft case

As mentioned in section 2.3, the aircraft observations were analyzed to provide further qualitative insights on models' performance/discrepancies in simulating the vertical structure of the tracer transport. The analysis was important to pin-point where, why and how models treated tracer advection differently which was found to be the reason for the models' discrepancy and this cannot be done by analyzing only tower data at fixed heights. We agree that the aircraft observations on August 2006 would be ideal to remain the consistency of tower simulation results (regarding time period). However, the more importance was given here to figure out the causes for possible mismatch between models.

Specific Comments:
p. 1276, line 3: 100 particles: Did you test whether results were significantly different for larger numbers of particles?

Yes, we repeated the experiment with 1000 particles where found no significant changes in the result. 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 now 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. "

p. 1276, line 5-7: Which version of STILT was used for these runs? Please clarify whether this version switches all particles to the large domain (irrevocably) at once, or allows separate particles to use d01/d02 winds as needed?

We used the STILT version that uses d01 winds once a particle leaves the inner domain (d02).
This information is now added in the text as follows:

p.1274, line 1:

"A brief description of STILT is given as follows. We used the STILT repository version 608, checked out on 3 July 2009."

p.1277, lines 11-25: The model-model agreement seems to be better at upper than at lower tower levels. Do you have an explanation for this?

On comparing statistics of model-model agreement given in Table 2, we would argue that it is only "slightly" better at upper levels compared to lower levels during August 2006. If it were significant, it could have been due to larger influence of advected tracer transport at lower levels.

p. 1278, line 1 and Figure 2: Please mark the location on figure 2. It appears to be right near the d02 domain boundary. Can you comment on if and how the treatment of transport across nested domains in WRF and STILT differs and how this might affect the model-model mismatch?

We'll update figure 2 by marking the location. We would not expect any significant inter-model difference in transport across nested domains in WRF and STILT. However, we agree to verify this by comparing models' performance for a flat area well within the d02 domain. This information will be updated in the manuscript.

p. 1282, line 15-20: Could you explain what you mean by "interpreted loosely"? I understand the problem you were facing (STILT provides footprints, but they were obviously not quite right since the modeled transport of the fossil fuel sources was not correct), but it's not clear how you solved this (finding a source location with a WRF plume reaching the aircraft location). Was it trial and error around the general STILT footprint vicinity, or based on the location of actual sources? If the latter, it would be helpful to show maps of the sources and/or the STILT footprints.

We located the "likely source region by looking at STILT footprints as well as relatively strong fossil fuel emission source as given in the emission map. You are right that we cannot assess the source region by following the footprints alone (the ideal case) since STILT had the issues with fossil fuel transport. This was solved with the aid of emission maps.
This information is now added in the text as follows:
p. 1282, line 19
"… the footprint was interpreted loosely with the aid of emission map"

p. 1284, line 20 - p.1285, line 9: This paragraph does not really belong in this section.

Perhaps make it a separate section (something like "Sensitivity of results to model resolution"), before section 3.1.

We agree, and made it a separate section "3.3 Sensitivity to model resolution".

p. 1284, line 25: I'm not sure I would agree that the results exhibit a strong sensitivity to resolution: the curves are essentially flat past 12 km, and the two data points at 2 and 6 km also don't give an unambiguous result (with a dip for the model-model r^2 at 6km).

It is true that the curves for R^2 are flat after 12 km, however the standard deviation of the inter-model differences are higher at lower resolutions (please see error bar in Figure 7b).

We have modified the text accordingly
p. 1284, line 26:
removed the word "strong".

p. 1285, lines 1-9: I'm not sure I completely follow the argument here. Extrapolation of the curves from 2km to 0 is not necessarily representative of what one would see if one actually did those computations at those very high resolutions, since the assumptions made in the formulation of the turbulence parameterization (and others) start to break down, and one would have to use different (e.g., LES) approaches

Here our argument is that both Lagrangian and Eulerian models will converge to identical results when increasing the spatial resolution, assuming no artifacts in turbulence parameterizations on Eulerian models on increasing the spatial resolution (hypothetical assumption). Also the speculation is that if one could achieve this high resolution with sufficient turbulence parameterization, the models results will not be necessarily influenced by inter-model difference in parameterizations since small scale features will be better resolved. Following to your argument and to make the interpretation clearly, we modified the text as follows:

p. 1284, line 29
removed the sentence: "The inter-model differences for the ideal case can be obtained by qualitatively extrapolating the resulted curve (2–24 km) to a horizontal resolution of 0 km"
p. 1284, line 29
included the sentence: "Hence it is expected that the inter-model differences become smaller at higher resolution (<< 2 km) and in this case the remaining differences between …."

p. 1285, line 8
removed the sentence:" That is, both models are expected to give identical results at a spatial resolution of 0 km.

p. 1286, line 14: I have a minor quibble with the conclusion stated here. Reasonably close agreement between concentrations simulated by the two models demonstrated here

is certainly a necessary condition for using STILT as the WRF pseudo-adjoint, but not sufficient. The definitive proof of its usefulness in this role would be a successful inversion with WRF as a forward model and STILT for the adjoint.

We agree with your interpretation. We will make this point clearly in the text 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."

Table 1: Since you used the K-F cumulus scheme (I assume only in the 6km grid?), convective fluxes are not used in STILT. How, if at all, does parameterized convection affect the tracer transport in WRF? Was there enough convection during August 2006 for this to have a significant effect on the results?
Kain-Fritsch cumulus scheme in WRF-Chem (used for both domains) does not transport the tracer convectively, so not using the convective fluxes in STILT makes no differences between two models. Hence it does not affect model-to-model mismatch. However it should be noted that in reality there is likely an impact of CO_2 from convective transport by clouds.

p. 1272, line 16

included the sentence: "Also note that convective fluxes are not used for tracer transport, but used cumulus parameterization scheme for meteorological parameters."

p. 1277, line 25
included the sentence: "Not using convective fluxes for tracer transport in both models can likely be the reason for the large model-observation differences as compared to the inter-model differences, if there is an impact on observed CO_2 from convective transport by clouds. However, the comparison of model performance for non-convective periods (time series excluding the data where convective rainfall is greater than 0.5 mm) has showed no reduction in the standard deviation of model-observation differences (not shown) and this confirms that there is no impact of deep convection on these mismatches."

Table 1 is modified as follows:
included the term "for outer domain":
"Cumulus - Kain-Fritsch (new Eta) scheme (for both domains)"

Technical corrections

Updated as it is suggested