We would like to thank referee #2 for the helpful and constructive comments. Following are the specific referee comments (in blue) together with the replies of the authors. Additions and changes to the paper text are written *italic*.

General comments

1) Referee comment:
The article is a bit lengthy and could be shortened a bit. For example I would suggest to shorten section 2.4 by 50%. Breaking up long sentences and checking the text with a native speaker could improve the clarity.

Author reply:
We think it is important to spend some effort on describing the differences between PBL top and MH and the different diagnostics. However, we removed some of the details and shortened the section from ~850 words to ~550 words:

Estimating the simulated mixing height (MH) is a non-trivial task. The height diagnosed from the model state (here referred to as PBL top, z_{PBL}) usually differs from the height up to which the CO_2 profile is constant (z_i). The PBL top determines the amount of entrainment zone air that is integrated into the PBL mean CO_2 mole fraction C_m (schematically shown in Fig. 5). We can regard this difference of mixing ratios as an error ΔCm:
(Eq. 4)
Since both PBL schemes use their own method to diagnose the PBL top, (see Appendix A for details) ΔC_m may even be inconsistent between both simulations, which would affect the results of our study. This could be circumvented by using the CO_2 gradients between ML and free troposphere to determine the MH, but in reality this information is usually not available, except for stable situations where tall tower data (up to 330 m) is available. Accordingly, we restrict ourselves to a method that more generally diagnoses the PBL top rather than the MH itself, knowing that this approach will cause some error ($\Delta Cm \neq 0$ ppm). If ΔCm is consistently arising for both simulations, the CO2 mismatch will be unaffected.
We tested independent methods to diagnose the MH from the simulated meteorology (offline, i.e. after simulation finished). We applied different formulations of the Bulk-Richardson number method (Vogelezang and Holtslag (1996)).
Our offline MHs differed only slightly from the YSU online diagnosed ones, which is not surprising since the YSU scheme is based on the Bulk Richardson method (Appendix A). In contrast the offline diagnosed heights of the MYJ simulation were in general much higher than online calculated ones. For the nighttime the differences between MYJ and YSU MHs even changed sign when using different Richardson formulations.

We compared ΔC_m to see which MH is most suitable. For the calculation of the second term on the RHS of Eq. (4) we used half the diagnosed PBL top z_{PBL} for z_i in cases with a well developed ML. This seems to be a reasonable choice because the entrainment zone

can be 0.4 z_i thick (Stull, 1988) and in well mixed conditions C_m is equal C(z) for any height z in the ML. We used the criterion $z_{PBL} > 600$ m for the choice of well developed ML cases. Fig. 6 shows an example of monthly averaged ΔC_m at 12:00 UTC. For YSU ΔC_m is similar for online and offline MHs whereas for MYJ the online diagnosed MH seems more appropriate. A possible explanation might be the combination of how the threshold of the TKE (turbulent kinetic energy) that defines the PBL top in the MYJ scheme was chosen and how the transport of scalars in the PBL is related to this height (Appendix A). Similar results were reported by Hu et al. (2010).
Therefore we decided to proceed using the WRF online diagnosed PBL heights for our study since it seems to relate better to the effective MH and ΔCm is consistent between the two schemes. For the remainder of this paper we will refer to the online diagnosed height as the mixing height (MH). The MH for the YSU and the MYJ simulations will be denoted with zi,ysu (corresponding to zi,truth in the pseudo-data experiment) and zi,myj (zi,model) respectively.

We edited the paper again and tried to simplify sentence structure. A native speaker read the text and we improved language in general.

2) Referee comment:
A general comment is that in this paper the reference 'truth' scheme is YSU to which the MYJ is compared. It should be stated somewhere that it is not clear from this synthetic experiment which scheme actually is better, when looking at the differences one could expect there should be a clear winner. It might well be that the YSU scheme is too diffusive, especially over ocean surfaces. That both schemes, though delivering quite different MLH, work pretty well for PBL meteorology might be caused by compensating effects in for example wind speeds and surface heat fluxes. Timing of ML growth is an important parameter that can help us to discriminate the better PBL scheme when observations op MLH are available.

Author reply:
The aim of the paper is not to validate the PBL schemes and to make qualitative statements about each scheme's performance. We selected two PBL schemes which are commonly used in current model setups for regional tracer transport simulation studies and are known not only to represent PBL meteorology realistically (Borge et al. 2008, Hu et al. 2010), but give also reasonable good results in conjunction with CO2 transport e.g. Sarrat et al. (2007). The choice which of the simulations represents the "truth" was arbitrary and does not affect the main statements of the paper (significant differences in simulated CO_2 concentrations, performance of the local correction with known mixing height). However, we added two sentences to the introduction to state the objectives of the paper more clearly (p 28174 l11):

The choice of the YSU simulation to represent the "truth" is arbitrary. Any of the two schemes might be closer to reality in some or all instances which doesn't affect the mentioned goals of the paper.

We also added a sentence in the conclusion addressing the last point of possible PBL scheme validation (p28192 l9):

These subtle timing differences might be used in comparisons to observed MHs to validate the PBL schemes in the future.

3) Referee comment:
For some of the stations used in this study actual observations of MLH (and CO2 observations (vertical gradient along the tower, which allows to check for the well mixed condition and approximate the true CO2 MH when this is below the tower top level))) should be available in the study period...

We agree with the referee that such datasets are very important for validating the simulations and the reshuffling method against observations. However, we believe this has to be studied carefully with simulations comprising other tracers like Radon and CO and including comparisons to flux measurements in order to isolate model-data mismatch in tracer concentrations due to transport deficiencies from those due to flux errors. The additional complexities and amount of work involved would be clearly beyond the scope of the current paper. Therefore, we decided to concentrate our study on the theoretical groundwork by using a synthetic data experiment in which we are able to control CO_2 fluxes.
At time of writing a follow up study including the before mentioned validation work is conducted and the results will be presented in the near future.

Textual Comments

Technical corrections of the referee have been addressed. Below are the answers to specific comments.

4) Referee comment:
p28172 l1-6: This is too general, in winter conditions in mid-latitude temperate climate and in (sub)arctic climate (even more generally during conditions with subsidence at large synoptic scale), also during daytime not always a CBL develops. Entrainment occurs at every upward change of the CBL top, not only by thermal overshoots. The mixed in air is not always free troposheric air but might be from (stable) residual layers where emission signals remain from previous day or night. The effects of entrainment can be much larger then several ppm.

l14: after sunset no incoming radiation heats the surface any more, the earth loses heat because of long wave transmission and the surface and surface air cools down leading to a stable stratification of the PBL. Turbulence does not cease but is reduced. There is still vertical diffusion though it is strongly reduced. Strong winds due to synoptic conditions or orography may cause sustained mechanic turbulence preventing the development of a stable PBL.

We agree the description is not valid for all possible weather conditions. However, the entrainment of residual layer air is explained in the paragraph lines 14-26 and mechanic turbulence is mentioned in lines 20-22. Nevertheless, we reformulated the two paragraphs (l 1-26) to make them more differentiated and improve clarity:

On the diurnal scale, several vertical mixing processes can affect the concentration of CO_2 in the PBL. During daytime in situations when a convective boundary layer (CBL) develops (summer, clear sky) photosynthetic uptake is diluted up to the height of a turbulent mixed layer (ML) within CBL on hourly time scales. Entrainment of air situated above the ML (free troposphere or residual layer) is caused by vertical advection and overshooting thermals. Such processes affect time-mean CO_2 concentrations in the mixed layer on the order of several ppm, but also alter other properties of the ML like moisture, temperature, and the mixing height itself (McGrath-Spangler and Denning, 2010). Due to the vigorous mixing in the CBL on small timescales variables like potential temperature, water vapor and CO_2 are approximating constant vertical profiles (Stull, 1988). In fact, various studies used this simplification for column mass budgeting approaches to directly determine CO_2 exchange fluxes (e.g. Wofsy et al., 1988; Chou et al., 2002; Laubach and Fritsch, 2002; Bakwin et al., 2004; Helliker et al., 2004; Aubinet et al., 2005b).
When turbulence is reduced as less radiation heats the surface after sunset, the colder layer of air near the surface is decoupled from the warmer well-mixed part of the ML. During these times the latter part, called residual layer (RL), is not directly affected by surface forcings, because air parcels are confined to the lower part of the PBL by a capping temperature inversion. Hence, tracer profiles in the RL stay relatively constant with time (Yi et al., 2001b). In the night, when respiration fluxes dominate and there is only weak mixing in a Stable Boundary Layer (SBL) CO_2 can accumulate near the surface. In the SBL mixing can still occur due to wind shear and surface friction up to several hundred meter (Stull, 1988). When sun rises again the capping inversion becomes weaker due to increased heat fluxes into the SBL. During the growth of the new mixed layer RL air is entrained, which causes a rapid dilution of CO_2 molecules (Gibert et al., 2007).

5) Referee comment:
p28173 l 25: reference and correct version number for EDGAR emission data and resolution deployed is missing

Author reply:
This is mentioned in the method section (2.1). Now we also added the information to sentence:
In addition, we diagnose biospheric fluxes from satellite reflectance data at 500 m horizontal resolution updated every 8 days and use the recent 2005 EDGAR emission (0.1 x 0.1 degree grid) to consider anthopogenic flux contributions (Source: EC-JRC/PBL. EDGAR version 4.1. http://edgar.jrc.ec.europa.eu/, 2010).

6) Referee comment:
p28175 l7: you mean a 10 cells (i.e. 100km) thick border zone excluded on all sides of the inner domain?

Author reply:
Yes, this amounts to about 12 % of the simulated volume. We modified the sentence to improve clarity:

In addition, 10 grid cells (~ 100 km) at each domain border were excluded to minimize direct influence of the lateral boundary conditions.

7) Referee comment:
p28176 l1-4: In principle there is a feedback between vertical mixing and the surface temperature, latent heat and following cloud development (influencing radiation) and thus on assimilation rates, this would complicate the comparison as the focus is on influence of the MLH on concentrations, but this still should be mentioned

Author reply:
We agree and added a sentence to the paragraph:

To reduce the complexity of the study we neglect feedback effects between vertical mixing, surface temperature, soil moisture, latent head and cloud cover (differences in radiation) which may influence NEE.

8) Referee comment:
p28178 l1: I would prefer not to do this simplification, as application of the hydrostatic equation is very simple to derive the height dependent molar air densities. This simplification introduces artefacts up to 20% in the correction factors which is simply too large

Author reply:
The simplification was motivated by the fact that MH data alone is insufficient to apply the hydrostatic equation. Nevertheless, to be more accurate we recalculate the analysis with the ratios of molar air density. We made the following modifications:

We changed section 2.2 to state dependence on the density profile consistently:

$$C_{\mathrm{m,truth}} = (C_{\mathrm{m,model}} - C_{+}) \frac{z_{\mathrm{i,model}} \rho_{\mathrm{m,model}}}{z_{\mathrm{i,truth}} \rho_{\mathrm{m,truth}}} + C_{+} \qquad (1)$$

with

$$\rho_{\mathrm{m,j}} = \frac{1}{z_{\mathrm{i,j}}} \int_0^{z\mathrm{i,j}} \rho(z) dz$$

and

$$C_{\mathrm{m,j}} = \frac{1}{z_{\mathrm{i,j}} \rho_{\mathrm{m,j}}} \int_0^{z\mathrm{i,j}} \rho(z) C(z) dz, \text{ for j = truth, model}$$

We deleted the paragraph (p28178 l1) including equation (2).

We modified the last sentence in the section:
We tested the correction method with a conceptual 1-D model of the PBL which will be discussed in the following section

In this context we also updated equation (4) in section 2.4 for consistency in notation:

$$\Delta C_{\mathrm{m}} = \frac{1}{z_{\mathrm{PBL}}\rho_{\mathrm{m,PBL}}}\int_0^{z_{\mathrm{PBL}}} \rho(z)C(z)dz - \frac{1}{z_{\mathrm{i}}\rho_{\mathrm{m,i}}}\int_0^{z_{\mathrm{i}}} \rho(z)C(z)dz$$

We recalculated the reshuffling now using Eq. (1) and updated section 3.2 which shows those results. Please also see our reply to referee comment 10). The following changes were made:

We updated Figures 9, 10, 11 and 12 to show the results of the reshuffling using Eq. (1).
We changed sentence p28184 l18:
Here we show the bias in simulated CO_2 and the bias reduction after applying the reshuffling method Eq. (1) to each land pixel and the full simulation period of the MYJ fields.
We changed sentence p28187 l5:
Like the bias, the random error (standard deviation of bias) is reduced, albeit at lower rates of 10–20 % during daytime (Fig. 10b and 12a), and 30–50 % during nighttime (Fig. 12b) when the CO_2 mismatch is spatially more uniformly distributed (cf. Fig. 11a and c) and thus conceptual deficiencies (i.e. 1-D correction of the 3-D simulation) of the correction seem to play a minor role compared to daytime (more on this in the discussion section).

9) Referee comment:
p28181 l14: there are many tall towers that measure profiles up higher than 100m, even up to >330m agl

Author reply:
Yes, such data might help in stable condition, e.g. in winter or nighttime if mechanical turbulence is not too strong. A note has been added. See also reply to referee comment 1).

10) Referee comment:
p28183 l14: more common is to use 50m as lower boundary for MLH, half of grid cell height in a particular setup does not seem like a rational choice

Author reply:
The choice to make the minimum MH depending on the model grid was based on the fact that in such cases CO_2 accumulates in the first model layer whose height varies with space. This variation in effective mixing height would not have been accounted for if we'd fixed the minimum height to 50 m. We agree that choosing half of the first model layer is not a reasonable choice either. So we recalculated the analysis fixing the minimum height to the top of the first model layer, which is usually 35-45 m. With this change we noticed a better performance of the correction up to 10 % during nighttime. The minor effect during daytime is expected, because during daytime the MHs are mostly

well above the first model level and the bias amplitude is smaller. We had to update section 3.2 accordingly. Please see also our answer to referee comment 8) for a list of all the changes we made.

11) Referee comment:
p28135 l25-28: this effect might also be explained by advection of day time depleted air masses reaching the site, during the well mixed conditions during the day the uptake is accumulated in the ML. Local NEE does not need to play a big role.

Author reply:
We assume the referee is referring to page 28185. We did not say that local NEE caused this drawdown. We agree that there is horizontal advection of air depleted in CO_2, but also the advected air is depleted in CO_2 due to shallow mixing along the path. We added a sentence to prevent misunderstandings:

Sudden bias increases in the late afternoon may be caused by fast decaying turbulence in the MYJ simulation while the photosynthesis is still dominating fluxes for 1–2 h (cf. PDM 19:00 LT, Fig. 9). Because in the MYJ simulation the mixing is weaker, the air masses advected in the PBL to the station are depleted in CO_2.

12) Referee comment:
p28186 l1: It is known for Cabauw that observed MH are below 200m for 60% of the time, in August nighttime MH is usually 50m or lower. So a night time offset of 300-400m allows to falsify the 'truth' values. Antropogenic emission influence in the Cabauw model grid cell depends strongly on exact grid cell configuration of WRF and the EDGAR grid as the city of Utrecht is at relatively small distance.

Author reply:
We modified the sentence mentioning the likely overestimation of the YSU simulation and mention *EDGAR* to highlight the dependence on our setup of the anthropogenic fluxes:

The comparably large night bias at Cabauw is caused by firstly a larger bias in MH which is in the early morning (01:00–05:00 LT) between 300 and 400 m , likely caused by too vigorous mixing in the YSU simulation, whereas the bias at all the other sites in the range of 200 to 300 m (cf. Fig. 7). Second and more importantly the EDGAR anthropogenic emissions in our model set up contribute strongly to the bias (contribution to surface CO2 at CBW between 01:00–07:00 LT above 25 ppm in MYJ, all other sites below 5 ppm).

See also the changes made in p28192 l15 as reply to referee comment 15).

13) Referee comment:
p28189 l17: ongoing research and refer to a 12 year old publication?

Author reply:

We added a more recent publication as well.

Seidel, D. J., C. O. Ao, and K. Li (2010), Estimating climatological planetary boundary layer heights from radiosonde observations: Comparison of methods and uncertainty analysis, J. Geophys. Res., 115, D16113, doi:10.1029/2009JD013680.

14) Referee comment:
p28190 l8-l10: lagrangian methods are as good as quality of the underlying meteorological fields allows, the vertical movements in frontal systems not resolved in e.g. WRF are not generated by the lagrangian transport models themselves. The lagrangian method might be able to resolve better the local flow field. But after improvement by assimilation of MLH observations by the mentioned methods also lagrangian models will improve further

Author reply:
The paragraph is misleading. We meant to use lagrangian transport simulations to identify and possibly account for situations in which contributions from horizontal advection play a major role in downstream budgets. We reformulated the paragraph:

To identify synoptic events in which horizontal advection dominates the column mass budget one might use particle trajectories from a lagrangian transport model together with a term for horizontal advection in the budget formulation (e.g. Aubinet et al., 2005a). But these trajectories can only be as accurate as the underlying meteorological drivers allow and the required computational efforts make such method unfeasible for inversions. Thus, assimilation approaches like 3D/4DVAR or Extended Kalman Filters that use observation based MH to optimize the model-state for tracer transport seem preferable.

15) Referee comment:
p28192 l15: would be better to express the bias relative to the source signal

Author reply:
The relative bias has been added:
However, large peaks in bias were found in point comparisons (grid cells including observational sites) related to timing differences of turbulence (−6 ppm during daytime, -2 % relative bias) as well as influence of nearby anthropogenic emission sources in the model setup (Cabauw, 30 ppm at nighttime, 8 % relative bias).

We also added the relative bias and its definition to section 3.2 (p28185 l5):
The relative bias is calculated from $<C_m-C_t>/<C_t>*100$.

And additionally to the bias peaks (p28185 l13 and p28186 l1):
For instance the high altitude station Pic Du Midi has a negative peak (−6 ppm, -2 % relative bias) at 10:00 local time (LT) …

The comparably large night bias at Cabauw (8 % relative bias) is caused by…

16) Referee comment:
p28193 l16: known -> presumably better constrained (???)

Author reply:
We agree with the referee as there might be considerable uncertainties in these fluxes. We now use the words "*better constrained*".

17) Referee comment:
Fig 3 : mu moles-> mu mol. To avoid confusion I would suggest to pick other line colours for the right graph

Author reply:
The legend text and figure have been edited accordingly.

Changes to acknowledgements

We added a sentence to the acknowledgements:
We would also like to thank the anonymous referees for their helpful comments and Julia Marshal for proof reading of the manuscript.

References:

Borge, R., Alexandrov, V., del Vas, J. J., Lumbreras, J., and Rodriguez, E.: A comprehensive sensitivity analysis of the WRF model for air quality applications over the Iberian Peninsula, Atmos. Environ., 42, 8560–8574, doi:10.1016/j.atmosenv.2008.08.032, 2008.

Hu, X.-M., Nielsen-Gammon, J. W., and Zhang, F.: Evaluation of Three Planetary Boundary Layer Schemes in the WRF Model, J. Appl. Meteorol. Clim., 49, 1831–1844, doi:10.1175/2010JAMC2432.1, 2010.

Sarrat, C., Noilhan, J., Dolman, A. J., Gerbig, C., Ahmadov, R., Tolk, L. F., Meesters, A. G. C.A., Hutjes, R. W. A., Ter Maat, H. W., Perez-Landa, G., and Donier, S.: Atmospheric CO2 modeling at the regional scale: an intercomparison of 5 meso-scale atmospheric models, Biogeosciences, 4, 1115–1126, doi:10.5194/bg-4-1115-2007, 2007.