The authors would like to thank all three referees for their thorough review of the manuscript. We appreciate their helpful comments and suggestions, and we have included them into our revised manuscript as good as possible. In the following, we address the comments and suggestions in detail. As several important points, mainly regarding the creation of our dataset and how that was described in the paper, have been addressed by more than one referee, we have decided to answer first to these common points. Afterwards, we will address the remaining comments from each referee separately.

Suggestions concerning improvement in wording and grammar, mislabeling of figures as well missing or badly formatted citations are not mentioned here, but we have taken them into account for the revised version of the manuscript. We appreciate the careful reading of the manuscript by the reviewers.

Answer to comments raised by several referees

1. Observed deviations of OR from global average (page 6189/ 6190)

In this part of the introduction, we mention observations from Hateruma Island that indicate variation in OR_{ff}, correlated with the origin of air. Referees 2 and 3 have stated that the sentence comparing observations with calculated OR_{ff} is not clear.

We have now rephrased the sentence to clarify which part refers to observed and which part to calculated OR_{ff}, and where the calculated ratios come from.

2. Connection between UN and EDGAR data (P 6191/6192)

All referees have expressed the need for a clearer description how the EDGAR and the UN data are connected. It has been asked to clarify why and how these data were merged. Besides, the resolution of the resulting OR_{ff} seemed to be not clear.

For clarification, we have now added a diagram (Figure 1 in the revised manuscript) that illustrates the different steps in the creation of the COFFEE dataset, and also makes clear which data is available in which resolution (gridcell, national level, usage type level). In addition, we have modified the description of the methodology in the text according to the referees' suggestions. Following the suggestion of Referee 2, we have replaced the rather general term “categories” by “usage types” throughout the paper. In answer to the question by Referee 1, we have added an additional sentence how the usage types in the EDGAR and UN dataset were merged. In addition, a table with all EDGAR usage types and their corresponding UN categories has been added as a supplement to the manuscript.

3. UN data prior to 1996 (P6192 L18-19)

Referees 1 and 2 have asked about the UN data in the years 1990-1995 that had not been included into our dataset. There has been the question how many data were missing and whether we had any quantitative criteria for defining "unrealistically large variations".

The problem with the missing data was not its high percentage, but which part of the data was missing: If there was no data for a few years for a certain usage type in a country, and this usage type was aggregated with another one into one EDGAR category, this leads to discontinuities in the oxidative ratio if both categories have a different fuel mix. To give an extreme example, let's say an EDGAR usage type includes two UN usage types, one only using oil and the other one only gas. If, in a certain country, no data exist in the first two years for the first type and in the next two years no data exist for the second type, the OR of this category would change from 1.95 in the first two years to 1.44 in the next two years.

In addition to missing data, sometimes the consumption for certain countries and usage types was just unrealistically low in certain years or showed other strange variations. To decide whether a change was "unrealistic", we have checked the deviation of data from single years from the linearly fitted fuel consumption for their usage type and country. We also compared the time series on the usage type/country level with the CO_2 emissions from EDGAR as well as country/fuel type level consumption with the BP dataset.

Our data check also showed some missing data/unrealistic variations in the years after 1995, but by far not as many. Therefore, we were able to extrapolate the missing/bad years.

4. Extrapolation of EDGAR data

Referees 2 and 3 have asked about the extrapolation of the EDGAR data from 1996 onward. There have been worries that this was a quite heavy extrapolation, in addition it has been asked for a better or more current citation.

Unfortunately, there is no better citation for the extrapolation performed by Sander Houweling. However, this extrapolation has been made in the same way as the extrapolations performed by the EDGAR team for the FastTrack2000 dataset, and the process-based factors have been provided by them as well. (This was maybe not written clear enough in our paper. We have clarified it more in the revised version). This extrapolation is not only used by us, but also for other purposes.

We are currently in the progress of upgrading our dataset to the most recent version of EDGAR (4.1), to have the advantage of the higher spatial resolution of the emissions (0.1° x 0.1° instead of 1° x 1°). This update will still take some time, but once finished we will make the results available on the same website as the current COFFEE dataset (we have added this announcement also in the conclusion part of our paper). Although we think that the higher resolution and the more current information on fossil fuel use will improve the resulting dataset, we do not think (based on our comparison between EDGAR 3.2 and EDGAR 4.1 so far) that this information changes any of the messages of our paper.

5. Temporal factors in COFFEE (P6193/Section 3.2)

The referees have raised several questions and concerns about the temporal structure in COFFEE. They have asked to include some more information how the original EDGAR time structure was created and which modifications we have made to it. Referees 2 and 3 have expressed their concern about using the – slightly modified – Dutch time structure globally. Given the limitations of this temporal structure, they think our discussion of the sub-annual variations in the COFFEE dataset is too intensive.

For better description of the temporal structure, we have included a link to the EDGAR website that describes the original timefactors, explaining how they are created and what their limitations are. In addition, we have clarified for which categories we have reversed/suppressed the seasonal cycles in the Southern Hemisphere respectively the Tropics.

As to the general limitation of the temporal structure – both the original EDGAR factors and our modification – we agree to the referees that this is far from being perfect. We also would have preferred to use better factors; however, to our knowledge, there is no dataset with global information about the sub-annual variations in fuel use available. As mentioned by Referee 3, high-resolution inventories with better temporal structures exist locally, but are limited to Europe and the US. Merging all of the available information on fossil-fuel-related time structures into the COFFEE dataset would of course lead to improvements locally, but would also require considerable effort. It probably makes more sense to use the COFFEE dataset as a global overview to get an idea whether the effects of variations in OR_{ff} are locally important – and then to combine OR_{ff} with a local high-resolution fossil fuel inventory that also includes a more correct temporal structure.

But we agree to the referees that the limitations of our temporal structure should be pointed out more clearly, thus we have done this in the revised manuscript – at the introduction of the temporal factors in section 2 as well as in the discussion section 3.2.

As for the length of the discussion in section 3.2, we do not think that it is too extensive. Even if the resulting temporal structure only reflects what we put into the dataset, we do not think that this is completely evident from our description, maybe not even from the look at the EDGAR timefactors: Of course, it is clear from the text that the seasonal cycles are reversed in the SH and lower in the tropics, but this does not say anything how large the seasonal variations are in the different latitudinal bands, or compared to the size of the short-term fluctuations (or about the shape of these short term fluctuations). Thus section 3.2 is planned to give an illustration to the reader – and the potential COFFEE-user – how the complete time structure looks like in COFFEE (and now also to point out the caveats for interpretation of this structure)

6. Comparison of COFFEE to CDIAC and BP data (Section 3.2/Figure 3a)

Referees 2 and 3 have critized that our comparison of the three dataset is too long, taking into account the fact that they are not completely independent.

We have followed the referees' suggestion and removed the detailed comparison from the manuscript, as well as the CDIAC and BP OR_{ff} from Figure 3a. Following the suggestion by Referee 2, we have just written one sentence stating the general agreement between the global average OR_{ff} from these datasets.

7. Comparison of REMO and TM3 simulations to observations at Ochsenkopf

As we have shown that there a significant sensitivity differences between the global and the regional model simulations, Referees 1 and 2 have suggested to compare simulations from both models to observations - at a station that is part of the REMO domain, for example the Ochsenkopf station.

We actually have done such a comparison at the Ochsenkopf station, and had originally planned to include it in the paper. However, the biospheric influence at

OXK is much stronger than the fossil fuel influence, thus it also dominates the OR derived from the observations (This is also the case for our other European station, Bialystok). A comparison of the observations with the fossil fuel part of the model, as done for Hateruma, does therefore not make sense here. When adding model simulations for the biosphere, one needs to keep in mind that also the biospheric OR exhibits deviations from its global average that are not accounted for in the models. So, the comparison at OXK required an additional plot (biospheric simulation), a longer explanation (including at least some discussion of the variations in the biospheric OR), but does not really give more information concerning the performance of the two models for OR_{ffp}. Given the length of our manuscript, we thus have decided against this comparison.

However, as two of the referees have asked for such a comparison, and some readers might be interested as well, we have decided to include a comparison plot at OXK as a supplement to the paper.

In addition, we have decided to use the observations at OXK as an illustration for the discussion in section 4.3. ("Whether the contribution of OR_{ffp} to the total atmospheric oxidative ratio is significant, depends on the relative strengths of fossil fuel signals compared to the influence of other processes, e.g. related to biospheric activity"). Following a suggestion by Referee 1, we have included the observed OR for OXK and HAT in Figure 6.

Answer to remaining comments from Referee 1

1. Could the authors indicate in the paper whether the lower sensitivity of CO2 and O2 variations of TM3 with regard to synoptic events is more due to the lower spatial rather than the lower temporal resolution of TM3 compared to REMO?

   The lower spatial resolution is the dominating factor, but the low temporal resolution reduces the variability even further.

2. It would be worthwhile information to state the percentage of CO2 emissions in NH, TR and SH to the total CO2 emissions. This information can either be included into Figure 2a or mentioned in the text.

   That is indeed useful information. We have added this information in the caption of Figure 2a.

3. The daily variations of ORff are expected to vary with seasons (heating vs. mobility issues). I miss information regarding this additional effect.

   Indeed the relative contribution of the different usage types changes throughout the year, influencing the diurnal cycle. However, these changes are rather small and do not change the general structure of the daily variation. We have added a sentence about this in section 3.2.

4. Would it be feasible to include the in-situ oxidation ratios (observed O_2 vs. observed CO_2) in Figure 6a to compare the actually observed O2/CO2 ratios with the fossil fuel influence?

We have adapted this suggestion partly, by adding the observed OR for OXK and HAT (see the discussion on the common points above). We have only done it for these two stations to prevent the plot becoming to crowded. Doing it for all stations, would not be possible anyway, as in-situ measurements are not performed everywhere, most stations have only flask sampling (with much lower sampling frequencies, e.g. daily or weekly).

5. P6189.l22-24: Sturm et al., ACP 2006 also mentioned variable oxidation ratios.

   This is correct, and there are also other papers (e.g. Stephens et al, J. Atmos. Oceanic Technol, 2007) discussing this issue. However, these papers has not been cited here, since the observed variations there have been attributed to deviations of the biospheric oxidative ratio from 1.1 rather than being caused by variations in fossil fuel combustion.

6. P6193, l13: Is there a specific reason to select the year 2006?

   The reason for choosing this year was just data availability: The UN data was available until 2006 when we have started creating COFFEE, and has just recently been updated (what has been included in the dataset). But as also pre-processed in-situ data for Hateruma was available for 2006 only, we decided to use 2006 for all plots to have the same year everywhere.

7. P6193, l19-20: What is the percentage of those omitted gridcells (dominated by cement production) to respect to the total and where are those gridcells located?

   OR_{ff} is zero (or close to zero) for 0.2% of all gridcells with a defined OR_{ff}. There are no larger areas dominated by cement production, but rather single gridcells distributed all over the world. A high number of those gridcells is found e.g. in China, India, and the US.

8. P6194, l4: You list Russia, Argentina and Canada, why not Mexico?

   We just listed three examples for countries with higher gas consumption; Mexico could be added to the list as well as several other countries.

9. P6194, l12: …for the year 2006, results for the other years are comparable). Does this statement also hold for the spatial distribution? If yes, it should be added under 3.1.

   This statement holds as well for the spatial distribution, but there is a slight difference between spatial and temporal changes: Spatial patterns in emissions and thus also in oxidative ratios do not change completely over the years, but there are changes, e.g. growing coal emissions from the Asian countries. The magnitude of these changes is quantified in Figure 3 (will be even more clear when adding the local changes in CO_2 emissions, as you have suggested) and discussed on page 6195. On the other hand, seasonal and short-term variations for the CO2 emissions are mainly dominated by the temporal factors that do not change on a yearly basis. The only difference in the temporal variations is thus caused by the distribution of the emissions among the categories.

10. P6195, l1-6: Figure 2 b and d. are the x-axis scales correct? To me, it would seem more realistic when the lower CO2 emissions occur on Saturday and Sunday?

This is indeed the case – apparently something went wrong with labeling of the axis, there seems to be a one-day shift. The labels have been adjusted now.

11. P6195, l17ff: In Figure 3, it would be good to include the CO2 emission changes (geographically) besides the ORff.

This is a good idea - we have added such a plot to Figure 3.

12. P6200, l3: What means an error larger than 0.05? Can you express this with a criteria on the linear correlation (r, r2).

We have decided to use the error of the linear fit rather than the linear correlation, as the error could be included into the plot graphically and is just more illustrative than the linear correlation. Our significance criteria (error > 0.05), corresponds roughly to the R^2 being larger than 0.8.

13. P6200, l5…that a significant part….From Figure 5c, about 30% seems to be statistically reliable (see previous point)

This is true. However, being statistical reliable refers mainly to the absolute value of the oxidative ratio, another point is the variability of the OR, which is capture quite well by our model.

14. P6200, l11-12: The statement about using the ORff information to test and improve the transport models is quite strong. Is this more a wish or do you really believe that this is important data for that particular purpose.

Our formulation is indeed a bit strong here. We have changed the sentence to "the information on OR_{ff} …contributes to the evaluation of transport models."

Answer to remaining comments from Referee 2

1. P6198 L17 Please be a bit more specific about how the 5-day timescale was chosen, and how sensitive the results are to this choice.

As stated in the paper, we wanted to get an oxidative ratio that is representative for the synoptic timescale, rather than getting seasonal variations. As the timescale of the simulated synoptic fossil fuel events at different stations seemed to be mostly somewhere between one day and one week, we focused on this range and performed linear regressions for different time steps (1,3,5,7 days), but also tried one with a longer time (30 days). The shorter the time series, the more fluctuations are seen in the retrieved OR_{ff}. Another point was the comparison with observations, where gaps in the data occur from time to time. With fewer values per individual fit also the quality of the fit is decreasing, which lead to exclusion of more values (that did not fit our quality criteria). The 5-day timescale seemed a good compromise to show enough variability while having enough significant data left.

2. P6198 L19-20 When I look at the plots (OR_{ff} at OXK), it actually does look to me like the black trace (TM3) is consistently a little higher than the blue REMO). Please address this.

This indeed seems to be the case for most of the time. However, we rather think that this is a coincidence. The difference does not seem significant, and we can not think of a reason why TM3 should give higher OR_{ff} values at this station. We have compared TM3 and REMO simulations for other stations within the REMO domain, and we do not see systematic differences – sometimes the TM3 values are slightly higher, sometimes the REMO values, and in other cases none of models is consistently higher.

3. P6199 L11 Strictly speaking, you don't know that OR_{ffp} is changing at Hateruma; you only know that OR_p is changing, and you use other lines of evidence (such as backtrajectory analysis) to attribute this to fossil fuels.

This is correct. To clarify that, we have changed the sentence to: "The reason for choosing this station is the above mentioned **fossil fuel related variation in OR_p** that has been observed at this station."

4. P6200 L1-12: I would like to see a more focused discussion of the months of November and December. This is the time with well constrained observations and (in early November), a noticeable disagreement between the model and data. What do the authors think is going on here? It would also be good if the authors could spell out how much better things would be at Hateruma if they were to use a constant, but locally correct value for OR_{ff} (instead of the globally appropriate 1.4).

Considering your first point, we are not completely sure what the reason for the disagreement in early November is, but we think it is probably due to ocean influence. One of us (Minejima) has run model simulations for Hateruma using the Flexpart model, and the OR_p derived from model simulation including ocean + fossil fuel agrees better to the observations in this period then the fossil fuel signal (OR_{ff} from the fossil fuel part of Flexpart does not differ strongly from that of TM3 for this period). In general, one needs to keep in mind that our method with the 5-day-regression is good to get an overview of the variability and the absolute of OR_p and also to compare the distribution of OR at different stations – but it is probably not the best way to really understand single synoptic events. For that purpose, it makes more sense to pick synoptic events from the observations/simulations and calculate the OR_p per event. Such a detailed investigation of synoptic events is done for Hateruma in another paper, written by some of us, that has just been accepted to ACPD (Minejima et al, Analysis of $\Delta O_2/\Delta CO_2$ ratios for pollution events observed at Hateruma Island, Japan. ACPD 11, 15631-15657, 2011)

As to your second point, as OR_{ff} =1.4 is general an overestimation at Hateruma, the use of a locally correct value would indeed be helpful. Minejima et al have observed that between Oct 2006 and Dec 2008, more than half of the assigned fossil fuel events are caused by air masses coming from China, thus the use of OR_{ff}=1.1 already gives better results than OR_{ff} =1.4. This is discussed in [Minejima et al, ACPD, 2011] and we have added a sentence concerning this to our manuscript as well. However, especially for this station, the variability in OR is an additional interesting factor, as several countries with different fuel mixes are so close together.

5. P6201 L15: What about Kumukahi? This shows variations in OR comparable to ALT and BRW but is not discussed at all.

ALT and BRW were just picked as examples for remote stations with strong variations in OR_{ff}, KUM could be mentioned there as well. As seen in Figure 6b, the difference $\Delta O2_{ff}$ resulting from these variations is negligible in the case of Kumukahi as well.

6. P6204 L6-8 I am probably simply failing to understand something here, but I am not entirely convinced by figure 7b and 7c. The authors use Figure 7b to claim that there is an increase in the interannual variability of APO fluxes when a variable OR_{ff} is used. What puzzles me is that the black and red curves in Figure 7c show very similar amounts of variability (at least to my eye). If 7b really does show the variability in APO fluxes that result from introducing variable OR_{ff}, and values around Hateruma are positive, shouldn't Figure 7c reflect this increased variability?

   Indeed the increased variability is not easily seen in Figure 7c, since it is really small compared to the variability in the APO fluxes itself. To make this small difference visible, we had first thought of showing the difference ΔF_{APOff} instead of the total flux in Figure 7c. However, we also wanted to give an impression how large the differences are compared to the total APO fluxes and their uncertainties. Therefore we have decided to split the information – Figure 7 a-b and d-e give the information where and in which direction changes occur, whereas Figure 7c gives the information how large it is compared to the total flux – and on this scale it becomes clear that the difference in variability is rather negligible.

7. P6189 L2 and throughout the paper: The authors move freely back and forth between OR and $\alpha_{F.}$ I know that both are used in the literature, but within a single paper, there should be a single choice. Please pick one and stick with it.

   There is a reason for using both OR and α_F: As stated on page 6189, line 3, we use the terms α_F and α_B for the global average oxidative ratios, whereas OR is the general term, e.g. when talking about the oxidative ratio at a certain location. This difference is important in 3.2, when we are writing about the different global averages – introducing the term α avoids writing "the global average OR_{ff}" all the time. As this apparently was not clear in our manuscript, we have added an additional sentence in section 3.2 to make the distinction clear.

8. P6193 L8 The acronym "COFFEE" is a charming one, but it's actually a little bit deceptive, since the dataset includes biomass burning (not a fossil fuel). While I don't expect you to abandon your COFFEE acronym, you should at least point out this bit of artistic license.

   Yes, strictly speaking this is correct. (There has also been some discussion between the authors whether to include biofuels in the dataset or not). However, the percentage of biofuels included in COFFEE is less than five percent – therefore COFFEE users and reader might forgive us to not explicitly including the part in the acronym.

9. P6207 L20: In what sense is this "average"? Is it a spatial average, a temporal one, or both? Please be explicit.

   It is the spatial (global) average, and the variations are temporal variations of this spatial average. To clarify that, we have reformulated the sentence as follows: "Spatially averaged across the whole globe, the mean OR_{ff} calculated from COFFEE varies between 1.39 and 1.42 from year to year."

10. Figure 6: It may be clearer if you were to make the map Fig. 6a and then move

the other two to the right side of the panel and relabel as 6b and 6c. Also, line 7 of the caption should read "REMO is used when available..." Finally, why are the whiskers in 6a solid lines, while those in 6b are dashed?

We have followed this suggestion and changed the order of figures and the caption accordingly, as well as used the same type of whisker lines in both plots now.

Answer to remaining comments from Referee 3

[As page and line numbers given by the referee refer to the submitted manuscript instead of the discussion paper itself, the respective numbers for the discussion manuscript are added in brackets]

1. I question the validity of figure 3a. It seems that much of the long-term time variations were based on country-level extension in time. How can a pixel-level trend be determined?

Although the extrapolation of the CO_2 is mainly based on country level data, the information on the distribution of categories goes into the extrapolation as well, and the ORff varies per country **and usage type**. The pixel-level changes in OR_{ff} are also determined by how strong which usage type contributes to the CO_2 emissions/O_2 uptake at this location.

2. Page 2, line 13 (P6189, line 2): should you not have a delta symbol in the denominator of this expression?

Yes. The delta symbol has been added.

3. Page 5, line 30/31 (page 6193, line 8): I would note the download website here as well as where you have it in the conclusions portion of the paper.

This is a good idea. We have added a sentence about the download website at this point.

4. Page 8, line 32 (page 6196, line 25): I think you want to note that this is the most recent sink estimate that incorporates APO. There are many estimate of carbon uptake and many of these vary from the manning and keeling estimates.

We have now clarified in the text that is the most recent sink estimate **based on O_2 measurements** – and only for those it makes sense to discuss the influence of the uncertainty in $\alpha_{F.}$

5. Page 9, 1st para: this clearly relates to my comment above: : :. But, this puzzles me further. It seems to me that you are using the IAV as a form of uncertainty? You seem to agree that the uncertainty that manning and keeling identify is the dominant but go on to state that this cannot be quantified for coffee. But, it seems that the 0.04 value could be used and worked into figure 3. You appear to have done that but I cannot make out any other lines on the plot as the color is very weak.

The purpose of our discussion is to assess whether neglecting the IAV of OR_{ff} leads to a larger error in the calculation of the sinks, not using the IAV as a form of uncertainty. However, as it can be seen from Figure 3 (or will be seen now, after making the ±0.04 value clearly visible), the range of year-to-year variation is much smaller then the value of 0.04. We hope this will be clearer now, after removing the discussion on the different datasets and just using one measure for the uncertainty (which now is also more visible in the plot).

6. Page 10, lines 1-13 (page 6198, line 12-19): you use "perceived" in 3 places in this para (and in figure 4, figure 5). I am not sure what is intended by this word. Do you mean "simulated" perhaps? Or maybe "calculated"??

The term "perceived oxidative ratio" generally denotes the ratio of atmospheric ΔO_2 to atmospheric ΔCO_2 – this can refer to both simulation and observations. This terminology is consistent to that of an earlier publication (van der Laan-Luijkx et al, ACP 2010).

The term "perceived" comes from the fact that it describes the atmospheric OR as it is perceived at a given location, as opposed to the ratio of O_2 to CO_2 sources/sinks. The "total perceived oxidative ratio" OR_p would correspond to the atmospheric oxidative ratio as it can be measured at this location, where as the "perceived fossil fuel oxidative ratio" $OR_{\mathbf{ffp}}$ denotes the part of the atmospheric OR caused by fossil fuel emissions (in contrast to $OR_{\mathbf{ff}}$, the ratio of the O_2 uptake to the CO_2 emissions.)

For the model results, the meaning of perceived OR_{ffp} is indeed equivalent to "simulated/calculated" - however, strictly speaking, the term "simulated OR" is not correct here, as CO_2 and O_2 are simulated, and not the ratio itself. As stated in page 10, line 8-17 (page 6198, 13-18), the perceived oxidative ratio as shown here is determined from a running regression of O_2 vs. CO_2, and thus also depends on the timescale for the regression. These sentences were also meant as a definition of the term "perceived oxidative ratio", but maybe this was not clearly stated. We have now put the term "perceived oxidative ratio" in quotation marks on the first use, to clarify that the following sentences are a definition of this term.

7. Page 10, lines 6-8 (page 6198, line 11-13): I cannot understand what regression is being performed. It seems like perhaps a regression over time is being performed on the oxidative ratio as derived from the simulation? And this time regression is of 5 days duration? So, 5 points? The purpose of this is to capture the OR coming from a surface gridcell rather than a mixture of sources as might be expected from an ambient measurement? This section needs a lot more explanation as it appears critical to the subsequent discussion and conclusions.

Maybe part of the question is already answered by our answer to the last comment, but to clarify it a bit more: As the oxidative ratio is defined as the ratio of O_2 and CO_2 **changes**, the usual way of determining this ratio is to plot O_2 vs. CO_2 and look at the slope. Thus the linear regression here is not a regression of OR_{ff} over time, but a regression of O_{2ff} vs. CO_{2ff} to determine the ratio. In the case of Figures 5 and 6, this is done for the simulated CO_2 and O_2 signals, in Figure 7 the same regression is performed for the O_2 and CO_2 observation. The result gives an impression of the mixture of sources influencing the atmospheric OR at the measurement location in the given timeframe – here 5 days. This is typical timeframe for synoptic events.

8. Page 19, line 7 (page 6209, line3): recommend citing the European high res inventory and/or the US vulcan inventory here.

Both inventories are now cited at this point.