Response to anonymous referee #2:

We thank the referee for their review of our manuscript and provide responses to the questions and suggestions below:

A much more rigorous description of the two model setups and their differences is required. There is currently no mention of the vertical resolution although this is later reported as being important for e.g. STE fluxes. In its current format the model description in Section 2 is lacking information and is written in a way that makes it unclear which modifications are applied to the new TM5 model version and which to the higher resolution.

We now add more details regarding the vertical resolution employed which is identical to that described in Huijnen et al. (2010). In the interests of brevity we did not include these specific details but will address this point by adding the following sentence:

Although TM5-MP can adopt all 60 vertical levels provided by the ECMWF ERA-Interim analysis, we employ 34 vertical levels for this study with higher resolution in the troposphere and the upper-troposphere-lower stratosphere (UTLS).

An identical model version is compared in our study, with the only change between simulations being the horizontal resolution that is employed. This allows us to attribute the changes shown to the use of increased resolution. We now add the following sentence to the end of the first paragraph in Sect. 2.

The following model description pertains to both 3° x 2° and 1° x 1° simulations discussed in this manuscript.

A paragraph should be added in Section 2 to clearly describe all details for the two model integrations used for this work. In particular, the authors should specify:

- start date of integrations and run length

- chemical initial conditions and spin up periods

- details of the analysis used (horizontal, vertical and temporal resolution).

Are these the same for both model resolutions or are they different? Are the simulated model years the same used for the emissions and observational datasets?

We now provide extra details regarding the simulations as requested by the referee. Again the only model parameter changed between the simulations is the horizontal resolution, increasing from 3° x 2° to 1° x 1°. We only use 2006 observational data and clarify this in Sect. 2.3:

We choose a range of ground-based and airborne measurements taken at diverse locations during the year 2006 representing different chemical regimes.

The use of EMEP observations in it's current state is confusing. In line 242-243 the authors state that sites in "Norway, Finland, The Netherlands, Belgium, Poland, Germany, Spain, Italy and Portugal" are used for comparison. Why just use EMEP stations from the above and not the ones in other countries? Or is this a mistake (see later)?

This was an oversight and has now been corrected in the text, thus:

… where we exploit measurements taken at various background sites in Norway, Finland, The Netherlands, Belgium, Poland, the Czech republic, Germany, Great Britain, Spain, Slovakia, Italy and Portugal. The number of sites used for comparisons of trace species other than O_3 is smaller due to data availability.

In Fig 3, EMEP sites from " Finland, The Netherlands, Belgium, Poland, Slovakia and Italy" are used, aggregated by nation. The authors should explain why just these six countries? Why aggregate the sites? Poland and Slovakia should be mentioned in the list of sites in Section 2!

For the sake of brevity, we choose not to show individual stations but rather aggregates as has been presented in other studies (e.g. Williams et al., 2013), since we feel a station by station decomposition is not the ideal presentational form. The aggregates shown cover a significant range of latitudes throughout Europe and we wish to show comparisons for the entire European domain to provide confidence in the model performance. We now change the text to:

Figure 3 shows comparisons of simulated and observed mass mixing ratios of surface O_3 at EMEP sites across Europe (www.emep.int; Aas et al. 2001), with countries chosen so to cover a range of latitudes.

In Fig 4, four EMEP sites are selected for comparison. Why these 4 sites? Again the sites in the Czech Republic and Great Britain used in this figure are not listed in Section 2.

These sites were chosen to show the diverse changes that can occur for different locations. Again the range in latitudes and longitudes shown for the European domain is broad, where only a limited number of EMEP stations measure NO and NO_2 therefore identical composites as those for O_3 cannot be presented. The biases for all stations are shown in Table 6.

In Fig 5, two sites are selected for comparison and again no explanation as to why those specific sites are used.

For brevity we choose to show a high and low NO_x rather than an extended set. Table 7 does provide the seasonal biases across all stations if the reader is curious as to the behavior at other sites.

Similarly, Table 6 and 7 use yet two different subsets of EMEP stations for comparison without explaining the reason for their choice.

Tables 6 and 7 present seasonal biases for **all** EMEP stations which measure NO and NO_2. Therefore the selection is dictated by data availability rather than by the authors. We modify the table heading accordingly: *Values are shown for both the 3° x 2° and 1° x 1° simulations for all stations with available data.*

In Fig 8, four selected EMEP sites are shown (from Norway, Germany, Austria and Slovakia). Same issues as above.

See explanations given for previous referee comments related to EMEP sites above. The number of stations measuring HNO_3 is a small subset of the total number of stations in the EMEP network.

If the purpose of the comparison with EMEP is to evaluate the model performance in the new configuration, as well as addressing the differences in model resolution, the current analysis is not convincing. Comparison of model data with tropospheric ozone column from satellite would help better evaluate model performance on the global scale. This could also lead to better evaluate the model ozone profiles which currently show significant discrepancies with MOZAIC data. Further comparison with EMEP surface sites (and other campaign data) would then add to the analysis, so far as the comparison is done across all suitable sites and a clear explanation is given if only a subset of sites is selected.

Due to the lack of stratospheric chemistry and microphysics in TM5MP we actually employ tropospheric ozone columns from the Multi-Sensor Re-analysis (van der A et al., 2010) for constraining the overhead ozone column (i.e.) the total column is nudged towards the observed value. Details of this method are given in Sect 2.1. This means an independent comparison against satellite data is not feasible as a large fraction of O_3 exists above the tropopause, which is the threshold where the nudging constraint is applied. The tropospheric component of any total column value is notoriously difficult to retrieve (e.g. de Laat et al., ACP, 2009) with a high uncertainty in the value, further compounded by sampling (totally cloudy skies). EMEP comparisons are regularly used to assess the accuracy of air quality models and have high temporal coverage throughout the day for all seasons therefore act as an excellent dataset for evaluating near surface O_3 as long as interpolation is done well. We also provide comparisons against two independent campaigns which cover multiple days and locations, allowing an assessment of the vertical profile in the tropopause which the referee fails to mention. We feel that our conclusions are robust as to the effect of higher horizontal resolution of tropospheric O_3, with similar behavior seen across independent comparisons.

In Section 5 the authors provide an analysis of budget terms for tropospheric ozone and compare these at the two different model resolutions. They state: "the chemical tropopause calculated for 3x2 is applied for the analysis of 1x1 budget terms to ensure that a valid comparison is performed". However, if convection and convective transport is significantly different in the two model resolutions (as the authors suggest) the position of the chemical tropopause at 1x1 should be at a higher altitude in the tropics compared to 3x2. Using the 3x2 chemical tropopause to analyse 1x1 budget terms is in my view inconsistent and the reduction in STE term with increasing resolution is likely to change If the 1x1 chemical tropopause is used.

Initially we did not impose the 3° x 2° tropopause definition onto the 1° x 1° budget analysis using the 150ppb gradient to diagnose the chemical tropopause for both simulations. The resulting burdens are entirely different as a different total mass of air is compared, making the STE component unrealistically large and the analysis incompatible with the profile comparisons shown throughout the manuscript. Many CTM studies adopt a climatological tropopause such as that provided by e.g. Lawrence et al., ACP, 2001 for their analysis in order to address this total mass issue.