We thank the reviewers for their insightful comments which will improve the manuscript considerably. Below we answer to their comments, and provide further information and data how we will consider their suggestions in our manuscript. The reviewer comments are in black, our answers are in light blue.

Reviewer #1

General comments

The paper presents the results from four simulations with the LPJ-GUESS dynamic global vegetation model (DGVM) driven with climate data for the Tortonian obtained from two AOGCM simulations using 280 and 450 ppm CO2. The resulting global vegetation distributions are compared with proxy data from about 170 sites (mostly located in temperate regions), with results from similar simulation studies, and with additional evidences on Tortonian vegetation e.g. from fossil mammals or phytoliths. Methodologically, the authors distinguish between an analysis at global scale (section 4.2) and an analysis at regional scale (section 4.3). While for the global analysis they introduce an "agreement index" to compare the site data with simulation data, the analysis at regional scale is almost completely qualitative. At both scales the authors conclude that paleo evidence is in better agreement with a lower CO2 value. By their particular simulation setup, they also conclude that its mostly the climate effect of CO2 that determines the resulting vegetation distribution and not the physiological effect of CO2 fertilization.

There are only few studies of Tortonian climate taking advantage of the knowledge on vegetation-climate interactions encrypted in DGVMs. Insofar, the study provides a timely contribution to the research on pre-Quaternary climates. But methodologically the paper could be improved in three aspects:

First, the statistics behind the comparison between fossil data and model results is not really convincing. Partly this may be because the authors tried to keep the presentation short, but more fundamentally, important aspects of a robustness analysis of their statistical approach are missing (details follow below).

We have actually performed multiple robustness tests for the analysis. Like the reviewer mentions, most of these were left out of the manuscript because we wanted to keep the presentation short. We will provide these in the supplementary material as per requested and answer the more detailed points below.

Second, the regional analysis (section 4.3) is rather unrelated to the global analysis (section 4.2), although it would be easy to repeat the statistical analysis performed globally also regionally. Surely, the data base is quite small for some continents, but by adding such an analysis one would get a clear impression why at a regional scale the study must stay qualitative.

We think that the regional analyses and discussion of these are important and particularly interesting for researchers with a regional focus. We fully agree that applying the statistics at the regional scale might not be very meaningful, not only because of the small sample size, but also because we cannot expect a global vegetation model driven by a global climate model to be very accurate at the regional scale. Furthermore, for the discussion of regional scale aspects, we also rely on other independent evidence, such as fossil mammals, phytoliths and isotopes that indicate open conditions for North America. These have been outlined in the existing text, and we will modify it further in the revised version. Therefore, we don't agree that including these statistics (if this is what the reviewer suggests) to show that they are not meaningful at this scale would be helpful. However, if required, we could present them.

Third, in the regional discussion a clear concept is missing for judging whether the differences seen in PFT distribution, biome distribution, tree fraction, and grass fraction between the 280 ppm and the 450 ppm simulation results are large enough to allow an interpretation towards a higher or lower atmospheric CO2 concentration. Therefore, I do not see that this qualitative discussion is appropriate to vote for or against a high or low CO2. Instead, I would suggest to consider this qualitative regional analysis to be a check for the consistency of the continental vegetation patterns seen in their simulations with results from simulations of other groups and with evidences from aditional fossil data.

We thank the reviewer for raising this point. We agree that we might have stretched the regional interpretation in the manuscript. We will correct this, focusing more on evaluation compared to other studies and only mentioning an indication of lower or higher CO2 concentrations when the pattern is very clear, such as in North America, where the more open vegetation under low CO2 clearly corresponds better with the paleobotanical data and other independent sources of evidence. We will thus focus more on how well our model produces the regional and continental vegetation patterns during the Miocene (as compared to paleobotanical evidence and other modelling studies). We will also add additional proxy evidence to the qualitative discussion. These include well-known samples from fossil mammals, isotope data and sedimentary records from Europe and North America. For North America and Southern Europe we will also use the existing phytolith data.

More detailed comments

1. Visual inspection suggests that the difference in biome distribution between simulated and reconstructed potential vegetation for today (Figs. S1A and S1B in the Supplement) is larger then the simulated Tortonian differences between low and high CO2 (Figs. 1A and 1B). If this were true, the authors should explain why they can derive the main result of their paper from simulations that are within the range of model errors. I suggest that the

authors apply a rigorous similarity/dissimilarity statistics to their biome distributions to quantify the model errors and compare them with the size of the signal they intend to interprete.

We agree that in its original form the manuscript does not present sufficient analysis of the model uncertainties and signal size. We were reluctant to use the statistical similarity/dissimilarity metrics to analyse biomes for our main comparison for reasons that we outline below. However, we now agree with the reviewer that a statistical comparison can provide useful insights. Therefore, we have now evaluated our biome simulations with the Kappa statistic, which is a standard for comparing modelled and a reconstructed biome distributions (e.g. Hickler et al. 2006) and the results show acceptable agreement between our present-day simulation and the PNV reconstruction, in particular if biomes are aggregated to a more general level. However, the pure numbers should not be over-interpreted for the reasons we outline below.

The first drawback of comparing Kappa scores for biomes is that Kappa does not include any "degree of difference" mechanism which can be important when considering more than two categories. For example, there is a much smaller conceptual difference between a "tropical grassland" and a "tropical savanna" than there is between a "tropical grassland" and a "boreal parkland", but that difference is treated identically when calculating Cohen's Kappa. This can be ameliorated to some extent by aggregating to megabiomes as done by Pound et al. (2011) (which we will present in a revised manuscript), but is inevitably present to some extent. A weighting can also be attempted, but this introduces subjective decisions.

The second argument against comparing potential natural vegetation (PNV) biome distributions using Kappa is that PNV biome classifications themselves introduce uncertainty. Potential natural vegetation cannot be measured directly (it no longer exists due to human influence) and so must be reconstructed. There is uncertainty in such reconstructions as evidenced by the differences between PNV biome maps: for example, the horn of Africa is predominantly covered by "tropical deciduous forest" in Haxeltine and Prentice (1996), but is dominated by "dense shrublands" in Ramankutty and Foley (1999). Similarly, the extent of the "tropical deciduous forest" biome in Southern Africa varies considerably between the two maps. Even the biomes categories themselves vary between the maps as different authors make different distinctions. Our experience is that kappa statistics applied to compare different PNV maps can indicate as bad agreement as the one between a model and a PNV reconstruction, when biomes are not aggregated to coarser classes. There are also subjective choices when classifying model output which introduces uncertainty. For example, how much tree LAI or tree cover constitutes a forest? How much for a savanna? The choices for these numbers are not well-motivated and can change the biome boundaries considerably. Concerning the paleobotanical data, we deliberately did not derive biomes because classifying fossil sites into biomes introduces large

uncertainty arising from interpreting the fossil record in terms of vegetation cover.

Quantifying Model Uncertainty using Kappa

To better present model uncertainty we will present the comparison of reconstructed and modelled potential natural biomes (Fig S1a in the original manuscript) at 0.5 degree resolution (native resolution of the biome map and the PGF forcing data). We will aggregate to fewer biomes, similar to the megabiomes of Pound et al. (2011), simplifying the forest types to two per climate zone (evergreen and deciduous) and combining some arid biome types. The biome classification will be presented explicitly in a table as an appendix. Initial investigations give a Kappa value of 0.44 between data and model (similar to the agreement of Hickler et al. (2006) and constituting "fair" agreement by Monserud (1990)). The per-biome scores show that the model does poorest in the most arid biome types, which are not important for the main results derived in this paper, and the analysis gives a kappa of around 0.6 for the main forest types. Given the "degree of difference" effect and sources of uncertainty discussed above, we consider this level of similarity to be sufficient for our interpretation.

For consistency we will use the same biome classification for the Tortonian biomes which will simplify the discussion a little. This change will make no difference to the Agreement Index results as it is purposefully constructed without involving any biome classification. We will also present the dominant PFT map for the present day (Figure S1B in the manuscript) at 0.5 degree resolution.

Quantifying Effect Size using Kappa

The kappa between the biomes from 280ppm and 450ppm Tortonian runs is 0.62. Given that these biome maps are produced with identical methodologies (they use the same model structure differing only by the effect of CO_2 concentration on vegetation and climate, they utilise the same biome classification and hence have the same subjective choices, and they involve no data-originating uncertainty), we argue that we do see a sufficiently large signal for our interpretations.

Furthermore, the Kappa between the Tortonian 280ppm biomes and the PGF control run biomes is 0.56. Considering again that these maps are produced with identical methodologies, this indicates that we can distinguish Tortonian vegetation with 280ppm CO_2 and present day vegetation (in answer to reviewer 2's second point). Comparing the Tortonian 450ppm biomes and the PGF control run biome gives a Kappa of 0.38. These scores will be included in the updated manuscript.

In summary, we believe that our vegetation model uncertainties are reasonable (given the uncertainty in the method of quantification) and our effect sizes are large enough to support our interpretation. We will include this information in the revised version of the manuscript by reporting the kappa values and showing the present day maps at higher resolution. Note also that we used a DGVM that has been generally benchmarked and used for climate impact studies in a very large number of studies (see http://iis4.nateko.lu.se/lpj-guess/LPJ-GUESS_bibliography.pdf for a list of LPJ-GUESS publications)

2. The concept of the "Agreement Index" needs further explanation. I failed to understand how the "fractions" that characterize PFT status are obtained from LPJ-GUESS. It is said that they are derived from the LAI (p. 2249, line 19), but the authors did not explain this relation.

We will include further elaboration of the method in the manuscript. To answer the reviewer briefly here: the "fraction" (or "relative abundance") of a PFT in a gridcell is the LAI of the PFT in the gridcell divided by the total LAI in the gridcell. The LAI values are the growing season maximum values and they are averaged over a 30 simulation year period.

3. In view of the various problems with paleo-botanical data, there is indeed no ideal way to compare them with model results. And surely the Agreement index (AI) introduced by the authors could be one way to quantify agreement. Nevertheless, this index is based on a number of arbitrary decisions: (i) the choice of fractional ranges for the different PFT 'statuses', (ii) the choice of numbers for the quantification of the different types of agreement (table 1); and (iii) the choice of the null hypothesis. To explain the latter a bit more: Instead of assuming that all possible values for the agreement (values -2 to 2) have equal probability, one could also assume that all fractional values for the "data" and the "model" have equal probability which would give a different random distribution ("null" distribution) of AI values. In my opinion there is no good argument for either of the choices (i) to (iii). Therefore it is not clear whether the results based on the particular choices for the AI are robust. The authors claim to have addressed robustness with respect to (i), but did not present these results. Robustness with respect to all aspects should be demonstrated in the paper (or in appendices) by varying the particular assumptions (i) to (iii).

Yes, we agree with the reviewer that we should have provided more information about the robustness of the method. As we mentioned above, we have actually performed some of these, and just left them out due to length limitations. We will include them in the supplementary material and will also include some more description in the manuscript main text. We are still working on point (iii) but can immediately offer answers to points (i) and (ii)

(i) Choice of fractional ranges

As mentioned briefly in the manuscript, a factorial study was carried out with the following values for the fraction ranges.

Min for trace: 0.025, 0.05, 0.075 (original was 0.05)
Min for sub- dominant: 0.075, 0.15, 0.3 (original was 0.15)
Min for dominant: 0.5, 0.75 (original was 0.5, doesn't makes sense to have "dominant PFT" with a fraction less than 0.5)

The results are shown for the 450 ppm run versus the 280ppm in Figure 1. Our default boundaries are marked with the red star. Overall conclusions:

1) It is clear that the 280ppm gives better agreement than the 450 ppm in almost all cases. The exception (big black square) has a huge sub- dominant range from 0.075 to 0.75 which will include a lot of PFTs, and therefore has very little differentiating power.

2) The boundaries control the absolute value of the agreement index much more than they control the difference between the 280/450 runs, which suggests that our result is robust against changes in the boundaries. We could have chosen different boundaries to get either better differentiating power or higher values (in terms of absolute numbers) or even both, but we wanted to check robustness, not tune our method, so we retained our initial choices.

(ii) Choice of numbers for the quantification of the different types of agreement

Table 1 shows the AI scores and ranges when different numbers are used to quantify agreement/disagreement between statuses. In all cases the score is higher for the 280 ppm run than the 450 ppm run.

4. The arguments for introducing the new AI measure of data-model agreement (p. 2249, lines 13-17) are not convincing: The authors simply state a personal preference ("We prefer a metric that . . . ") but do not explain why the other metrics (Salzmann et al. 2008; Pound et al. 2011; François et al. 2011) should be discarded. In fact, it would be good to know whether those other approaches would reveal similar results when applied to the data used by the authors. I personally feel, that in particular the method by Francois et al. (2011) is the most objective because it generally distrusts a comparison of data diversity with model abundances (in the terminology of the authors, p. 2248 bottom) by comparing only presence/absence. Moreover, if dispite all warnings such a diversityabundance comparison is attempted (as done by the

authors with their Agreement Index), why not using the classical rank correlation which is known to be statistically robust?

We thank the reviewer for pointing this out and agree we should be more exact in our reasons for developing the AI rather than using the other methods. We include a more detailed discussion of the reasoning for not using existing methods or classical statistics below and will include these reasons in a revised draft of the manuscript. We will also provide additional statistical analyses to prove the robustness of our results.

We have calculated both Pearson's product moment correlation coefficients and Spearman's rank correlation coefficients for the 280ppm and 450 ppm scenarios per PFT and for the entire dataset and present them here in Fig 2. As mentioned in the original text, these do not prove to be particularly illuminating. The per-PFT coefficients do not show a consistent trend favouring a particular CO_2 scenario. Furthermore, the Spearman's rank for the full dataset is virtually identical for both CO_2 scenarios, but the Pearson's coefficient indicates better correlation for the 280 ppm CO_2 scenario than for 450 ppm CO_2 (0.53 vs. 0.42). This could be interpreted as weak evidence that the 280 ppm CO_2 scenario agrees better with the paleo-botanical data. We will include these additional analyses in the manuscript, and as indeed not all applied statistics clearly favor the low CO2 scenario, we will emphasize the uncertainties more. Note that we already formulated the title quite carefully, as: "Climate–vegetation modelling and fossil plant data suggest low atmospheric CO2 in the late Miocene." The wording "suggest" should indicate that we cannot be sure, as often the case in paleoclimate research. However, one should keep in mind that our qualitative regional discussion (where supported by sufficient data) also tends to favor the low CO2 scenario.

Regarding the other comparison methods: Salzmann et al. (2008) present a map of the inconsistency between model and data. Whilst a visual comparison is useful, we wanted to add a quantitative method to discriminate between the two CO_2 concentrations. The later study of Pound et al. (2011) uses Cohen's Kappa to determine biome agreement, both the 27 'native' biomes from BIOME4 and a 7 "megabiome" classification. This does offer a single statistic which could be used for hypothesis testing. However, (as discussed extensively in point 1.) there are drawbacks with using Kappa to compare biome classifications and with biome classifications themselves. So whilst comparisons of biomes are clearly useful visual aids and can be a useful cross-check (see our response to point 1), we decided to use only information on PFT fractions for our main analysis and therewith minimize subjective choices and classifications.

As the reviewer points out, the work of François et al. (2011) offers a method for determining agreement between paleobotanical data and simulated vegetation which percentage agreement per PFT based on presence/absence. These per-PFT scores could conceivably be combined to produce overall agreement scores, taking care that PFTs which are mostly absent from the fossil record do not unduly affect the final result. However, our study is

different in nature to that of François et al. The study of François et al. was a regional study with a relatively high degree of taxonomic precision (ie. a more detailed PFT set), whereas our study is global with appropriately coarser taxonomic resolution (ie. a relatively simpler global PFT set). By means of example, there are 8 purely temperate PFTs in the CARAIB version used in François et al. 2011 compared to only 2 in the default LPJ-GUESS configuration and 4 in the configuration used in our study. Thus by exploiting a high degree of taxonomic precision, presence/absence data were used effectively in the regional study of François et al. In our global study, each PFT spans a much larger geographical extent and there are fewer PFTs at each site for which to make presence/absence comparison. Thus we expect the effective differentiating power of such presence/absence to be lesser. So rather than using detailed taxonomic resolution and presence/absence information, we sought to exploit the abundance/diversity fractions which we believe has useful information and so is worth attempting despite our previous warnings. For this reason we developed the Agreement Index and introduced statuses beyond presence/absence.

The Agreement Index also allows easy assignment of a zero-weighting when PFTs are absent from a site in both the fossil record and model (contribution in this case is zero). It also allows an (admittedly subjective) method to tackle the "degree of difference" effect which causes problems for Kappa analyses which involve more than two classifications with differing conceptual degrees of similarity, as mentioned in point 1. This is done by assigning the value -2 for very strong disagreement and the value +2 for correctly matching dominant PFTs, as this must necessarily include at least 50% of the PFT and defines predominant biome functioning. A similar effect could be achieved by weighting the Kappa scores depending on the degree of difference, but this would also require subjective choices. The subjective choices involved in this method are motivated in an obvious and transparent way and can be (and were) tested relatively easily (see point 3).

We will modify the text in the manuscript to explain the above arguments in more detail.

5. With Fig. 2 the authors want to demonstrate that their results differ from the null hypothesis of random agreement. And indeed, the AI values for the 280 ppm and the 450 ppm simulation are well off their "null model". But they did not demonstrate that the difference between the AI values obtained from their two simulations with different CO2 is significant. If naively one would add the spread of the null model to the AI values from the two simulations, they would be statistically indistinguishable. Therefore the authors must plot into Fig. 2 also the full distribution of their results for the two experiments to allow judgement of signifance concerning their difference – maybe the authors added those Z-scores exactly for that purpose, but it's nor how they were computed. But plotting the individual distributions would in any case be more informative.

We agree that we could have provided more information on the difference between the AI values from different models. It also appears that the text which explains the distribution in Fig 2 is unclear and we will remedy this. To clarify here, each of the 25,000 frequency counts in Fig. 2 is the mean AI score from matching all 167 fossil sites to 167 random gridcells (not of the AI per site or per PFT). Thus there is no meaningful "full distribution" to plot on Fig. 2 for the two experiments because each experiment only yields a single frequency count of the type plotted in Fig 2 (ie. the mean of all the 167 fossil sites compared to simulated vegetation). It may be that the "full distribution" to which the reviewer is referring is the 'per site' or 'per PFT' AI values (or 'per site per PFT' AI values) but that quantifies a different variability from that in Fig 2. The variability in AI between sites is not inconsiderable (see Figure 1 in the original manuscript for an idea of the variability between sites) but we don't believe this sheds any light on the issue of distinguishing the mean AI values of the two CO_2 scenarios. Similar arguments apply for the distribution of AI per PFT.

In the first instance, the distribution in Figure 2. aims to show the mean value of chance agreement. This seems to be clear enough (although we will test other means of assessing chance agreement as discussed in point 3.(iii)). One can then look at the AI values for each Tortonian scenario and conclude that both scenarios do indeed offer better agreement than chance. In the second instance, the standard deviation of the same distribution aims to quantify the natural variability in chance agreement and so how much better the Tortonian scenarios are than random chance, and how much better one scenario is than the other. The traditional *p*-value interpretation would be the probability of getting a random combination of gridcells giving better agreement than the Tortonian scenario. These are $p < 10^{-8}$ and $p < 10^{-13}$ for the 450 ppm scenario and the 280 ppm scenario respectively. We can conclude, reassuringly but not surprisingly, that both our reconstructions are very much better than chance. Furthermore, the 280 ppm scenario is clearly better than the 450 ppm but differences in such very small *p*-values are not helpful, so instead we report the difference in units of standard deviation (*Z* scores), in this case 1.7. We believe this difference sufficiently supports our conclusion that the 280 ppm run agrees better with the fossil record than the 450 ppm run.

We realise that this logic relies on the assumption that matching random model gridcells to the fossil record gives an adequate representation of chance agreement. We chose this method because it will give ecologically consistent PFT compositions (no unrealistic combinations of boreal and tropical PFTs for example) and so is a more stringent test than some random numbers (which could give such unrealistic combinations). As mentioned above, we will examine other methods of estimating random agreement and discuss them in the revised manuscript.

Minor comments

p. 2246, line 25: The authors note that they transfered the soil parameters of the AOGCM to LPJ-GUESS. This provokes the general question to what extent the water cycles in the AOGCM and LPJ-GUESS are consistent, and whether inconsistencies in evapotranspiration fluxes might affect the results for the vegetation distribution.

To clarify, the soil parameters transferred from the AOGCM to LPJ-GUESS refer to static soil parameters (for example texture), not state variables such as soil water content. This means that each model, ECHAM5/MPIOM and LPJ-GUESS, has a fully independent hydrological cycle with different process representations, but with the latter model being driven (in terms of input precipitation and temperature) by the former. The hydrological cycle of LPJ-GUESS is therefore still fully internally consistent. The different representations in each model could be termed an "inconsistency" but is an inevitable consequence of the method of forcing a vegetation model with climate model output (the "asymmetric, iterative offline coupling" we mention on page 2247, line 5) and is commonly done for such studies (eg. François et al. 2011; Pound et al. 2011). We do not feel it appropriate to compare in detail the hydrological cycles of ECHAM5 and LPJ-GUESS here, but would like to point out that hydrological cycle of LPJ-GUESS (as implemented in the related model LPJ-DGVM) has been benchmarked in Gerten et al. 2004.

p. 2247, lines 18-28: The authors describe a number of modifications they introduced to LPJ-GUESS, but not why these modifications were necessary for their study. For the modified bioclimatic limits they claim improvements for present day biome distribution (lines 18-20) but do not demonstrate the improvements. It is only claimed (p. 2248, lines 11-12) that the modern biomes are reproduced "reasonable well". For such a claim one needs a measure, but this is not provided. Moreover, the main issue of the study depends on the model's reaction to changing climate and CO2. Therefore, some comments why the authors trust the model's response to such changes would be helpful.

With regard to the bioclimatic limits, the main effect was to remove treeless areas in South China, Argentina and Florida (see Smith et al. 2014, Figure 2(C) for the model version which does not include nitrogen limitation). This was an artifact whereby in these areas it was too warm for temperate trees to establish, but too cold for tropical trees, which resulted in treeless belts. In other words, there was a mistake in the model, which we corrected, with the main result that the model correctly simulates forests in south-eastern Asia. The other changes to bioclimactic limits were made for consistency with Sitch et al. (2003) and make very little difference. The introduction of Temperate Needleaved Evergreen (TeNE) trees, and the splitting of shade-Intolerant boreal/temperate Broadleaved Summergreen trees (IBS) into Temperate shade-Intolerant Broadleaved Summergreen trees (TeIBS) and Boreal shade-Intolerant Broadleaved Summergreen (BIBS) was intended to better compare the model results to the fossil record and because we believe that, with these changes, functional characterisitcs of the global vegetation are represented more appropriately. However, we will describe the reasoning for these

changes in more detail.With regards to the model's ability to capture present day biomes, we refer to our answer to point 1 which includes a Kappa measure and higher resolution maps for a more detailed visual comparison. We will also mention in the text that the biomes produced by LPJ-GUESS without our modifications can be seen in Smith et al. (2014) (their Figure 2(C)) and we will discuss our improvements relative to that. Furthermore, we will also include text to mention that LPJ-GUESS (and the closely related LPJ-DGVM model) has been benchmarked against various observations including, for example, NPP (e.g. Zaehle et al., 2005; Hickler et al., 2006), modelled PNV (Hickler et al. 2006; Smith et al. 2014), stand-scale and continental-scale evapotranspiration (AET) and runoff (Gerten et al., 2004), vegetation greening trends in high northern latitudes (Lucht et al., 2002) and the African Sahel (Hickler et al., 2005), stand-scale leaf area index (LAI) and gross primary productivity (GPP; Arneth et al., 2007), forest stand structure and development (Smith et al., 2001, 2014; Hickler et al., 2004), global net ecosystem exchange (NEE) variability (Ahlström et al. 2012, 2015) and CO_2 fertilisation experiments (e.g. Hickler et al. 2008; Zaehle et al. 2014; Medlyn et al. 2015). Many of these benchmarks are constantly repeated by the LPJ-GUESS consortium (of which Hickler is a member, unpublished). Regarding the CO2 response, the model without nitrogen limitation most likely overestimates CO2 fertilisation (see e.g. Hickler et al. 2015), which implies that our conclusion that the climate forcing is more important than the physiological CO2 effects for distinguishing the low and high CO2 scenario for the late Miocene is robust, which we will discuss in the manuscript.

p. 2251, lines 10-11: Here the authors announce a table in the supplement relating fossil plant taxa and PFTs. But such a table is missing. Please add that table since a large part of the study is based on this classification. Instead there is an un-numbered table in the supplement listing the study sites.

We were referring to the LPJ-GUESS PFTs listed in Table S1. However we will add a table relating fossil plant taxa to the PFTs.

p. 2252, line 16 and Figs. 1a an d 1b: It would be good to refer to Appendix B for references to the biome classification. Even better in my opinion would be to serve the readers by providing a table with the rules for the biome classification.

Yes, this will be done.

p. 2255, line 7: What are the "two reasons"? I cannot identify them in the following text.

The two reasons are increased seasonality in Central Europe, and increased openness in the Iberian Peninsula and in modern Turkey. However, we agree that this should be rephrased and will be reworded in the revised version of the manuscript.

Table 1: I guess the row headers should be shifted.

Thank you for pointing this out, we will ensure this is correct in the final proofs.

Supplement Fig. S2: This figure should in my opinion be shifted to the main part of the study, because it shows that in certain regions (e.g. the Iberian peininsula) the proxy-data are not informative about the value of atmospheric CO2.

Yes, this is a good idea and we will do so.

Reviewer #2

This paper presents a reconstruction of late Miocene vegetation using a dynamic vegetation model driven by the climatic outputs of climate model runs for two different partial pressures of CO2 in the atmosphere, 280 and 450 ppmv. These partial pressures reflect the range of atmospheric CO2 pressures that have been reconstructed from proxy data for the late Miocene. The authors compare the vegetation reconstructed with palaeovegetation data available for this time period. They also compare in detail their results with late Miocene vegetation model reconstructions published in the literature. For the comparison with the data, they build an agreement index (AI) which is an interesting and relatively novel aspect of their work. Since the AI is significantly higher for the low CO2 (280 ppmv) case, they conclude that climate and vegetation modeling suggest low CO2 in the late Miocene and so would favour the lower values in the range exhibited by the proxies.

The paper is generally well written, scientifically sound and with some clearly novel aspects with respect to previous work on the subject. I am thus in favour of its publication in Climate of the Past. I just have a few comments or suggestions that the authors might want to address.

(1) Section 3.4 : your comparison at the PFT level and associated statistics is presented as a new method for model-data comparison. However, as mentioned by the authors, François et al. (2011) have also performed a similar comparison at the PFT level, and contrary to what is said here, they also used the PFT diversity from the data (see for instance their table 7 and the comparison with model NPPs in their figure 6), although only presence-absence is used in their kappa calculation. What is the advantage of your AI index compared to the more traditional kappa method ? Kappa can also be averaged over sites or over PFTs. The statistical study on kappa presented here for AI (which is really interesting and the most novel contribution of this paper) is also possible for kappa. You just define more classes (abundance classes) that may also be involved in the kappa method, but actually have not been involved because of the large uncertainties on model PFT abundances. Models are certainly more robust in evaluating presence/absence than

abundance. Morever, as mentioned in your section 3.4, it is not obvious that PFT diversity from the data can directly be compared to model abundances. Even presence/absence in the data may be uncertain due to the PFT assignment scheme in the data (see again François et al., 2011). This may also critically depend on the number of PFTs in the classification used. This might be discussed somewhat more, because the associated uncertainty might have some impacts on the conclusions reached.

We thank the reviewer for his insightful and positive comments. We apologise for mis-representing the work in François et al. (2011), we meant to state that PFT diversity was not used to provide a quantitative measure of agreement, and will amend the text accordingly.

Our reasons for not using Kappa and for using abundance data beyond presence/absence are detailed in our answer to reviewer 1's comment 1. We would also argue that the coarser taxonomic resolution of our global PFT gives sufficient robustness in terms of presence/absence and abundance to use abundance fractions. Furthermore, we agree that whilst it could be possible to use Kappa on model abundances classes (neatly avoiding the uncertainties of biome classification whilst still utilising abundance/diversity data); such a method would still suffer from the "degree of difference problem" where a mismatch between the absent category and trace category would be treated as severely as a mismatch between absent and dominant categories. It also offers no obvious way to remove or zero-weight the contribution from PFTs which are absent in both the data and model at a given site. We will discuss these points in the revised text.

(2) Section 4.1, figure 2 : it might be interesting to add on figure 2 the AI that would be obtained with present-day (control run) model vegetation (when comparing to palaeodata). Is it significantly different from the AI for the 450 and 280 ppmv late Miocene configurations ? If it is close to the 280 ppmv late Miocene case, it might mean that your model is not fine enough to discriminate between the present-day vegetation and the late Miocene one.

As described in our answer to reviewer 1's point 1, we will provide statistics to quantify the differences in modelled vegetation between today and the Tortonian. The Kappa between the present day control run and the Tortonian 280 ppm run is 0.56 and the Kappa between the present day control run and the Tortonian 450 ppm run is 0.38. Given that identical methodologies were used to derived these biomes (ie. using the same model), we argue that we our model is indeed fine enough to discriminate. However, we don't think that presenting the AI for the present-day vegetation is meaningful for addressing the research questions addressed here.

(3) Section 4.3.1 : the characteristics of Miocene vegetation in Europe is indeed as discussed here the widespread presence of temperate deciduous trees, with some temperate evergreens in the south. Evergreens are however

different from present-day Mediterranean (drought-tolerant) evergreen trees, since data show the presence (not dominance) of temperate evergreen perhumid trees. This is a very important climatic constraint from the point of view of the data, while your model does not separate between drought-tolerant and perhumid temperate evergreen trees. The impact of this simplification on the results should be discussed, or at least it should be mentioned. Also, your figure S2 indicates that the SI index strongly varies from one site to the next. This is an important result that shows that there are still some features that are not well captured by the model (or possibly it might be a problem in the interpretation of the data). It would be interesting to discuss figure S2 in the main text.

Yes, we agree with the reviewer on both these points and will address them in a revised manuscript, with more emphasis on the shortcomings of the model in this regard. For Europe, it would indeed have been better to discriminate sclerophyllous (drought-adapted) evergreens, as in a regional version of LPJ-GUESS (Hickler et al. 2012) or a global version with hydraulic architecture (Hickler et al. 2006), but this type is not included in the current global standard version of LPJ-GUESS.

(4) Section 5 (Summary and conclusions): In view of the large uncertainties on climate models (including other boundary conditions than CO2), vegetation models and PFT classification, I am not sure that models can really provide a strong constraint on palaeo-CO2. It is interesting to learn that you model is more consistent with low CO2 in the late Miocene, but this is a very indirect constraint. I would suggest that you reformulate the last sentence of your conclusion to make the statement less direct (there are uncertainties and it may be model-dependent, so we may need to study the same problem with other climate/vegetation models).

We fully agree with the reviewer that there are still large uncertainties in climate models, the applied vegetation model and the applied analyses. We have been aware of these uncertainties, but apparently some of the formulations indicated too much certainty. Thus, we will reformulate the last sentence of the conclusions and other key sentences throughout the manuscript. We nevertheless believe that our indirect evaluation of two plausible CO2 concentrations for the Tortonian and other aspects of the manuscript (e.g. state-of-art climate modelling and DGVM applied to simulate Tortonian vegetation, novel approach for comparison with paleobotanical data, separating direct climatic and physiological CO2 forcing) represent an interesting contribution to the science on Tortonian climate and ecosystem dynamics.

(5) Some small typos:
P 2254, line 10: 'possibly because' P 2262, line 25: 'Fig 1a and b'does not correspond to the present-day biome map, it should be figure S1 P 2263, line 7: 'It ´ also shows a band P 2263, line 12: 'particularly shrubs'

Thanks for pointing these out, we will correct them.

References

Ahlström, A., Miller, P.A. & Smith, B. 2012. Too early to infer a global NPP decline since 2000. Geophysical Research Letters 39, L15403.

Ahlström, A., Raupach, M.R., Schurgers, G., Smith, B., Arneth, A., Jung, M., Reichstein, M., Canadell, J.P., Friedlingstein, P., Jain, A.K., Kato, E., Poulter, B., Sitch, S., Stocker, B.D., Viovy, N., Wang, Y.-P., Wiltshire, A., Zaehle, S. & Zeng, N. 2015. The dominant role of semi-arid ecosystems in the trend and variability of the land CO2 sink. Science 348: 895-899.

Arneth, A., Miller, P.A., Scholze, M., Hickler, T., Schurgers, G., Smith, B. & Prentice, I.C. 2007. CO2 inhibition of global terrestrial isoprene emissions: Potential implications for atmospheric chemistry. Geophysical Research Letters 34: L18813.

François L, Utescher T, Favre E, Henrot AJ, Warnant P, Micheels, A., Erdei, B., Suc, J.P, Cheddadi, R. and Mosbrugger, V.: Modelling Late Miocene vegetation in Europe: Results of the CARAIB model and comparison with palaeovegetation data. Palaeogeogr., Palaeoclim., Palaeoecol., 304, 359–378, 2011.

Gerten, D., Schaphoff, S., Haberlandt, U., Lucht, W. and Sitch, S.: Terrestrial vegetation and water balance – hydrological evaluation of a dynamic global vegetation model. Journal of Hydrology, 286, 249–270, 2004

Haxeltine, Alex, and I. Colin Prentice. "BIOME3: An equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability, and competition among plant functional types." Global Biogeochemical Cycles 10.4 (1996): 693-709.

Hickler, T., Smith, B., Sykes, M. T., Davis, M. B., Sugita, S. and Walker, K.: Using a generalized vegetation model to simulate vegetation dynamics in northeastern USA. Ecology, 85, 519-530, 2004.

Hickler, T., Eklundh, L., Seaquist, J., Smith, B., Ardö, J., Olsson, L., Sykes, M.T. & Sjöström, M. 2005. Precipitation controls Sahel greening trend. Geophysical Research Letters 32: L21415.

Hickler, T., Prentice, I. C., Smith, B., Sykes, M. T. and Zaehle, S.: Implementing plant hydraulic architecture within the LPJ Dynamic Global Vegetation Model. Global Ecology and Biogeography, 15, 567-577, 2006.

Hickler, T., Smith, B., Prentice I.C., Mjöfors, K., Miller, P., Arneth, A. & Sykes, M.T. 2008. CO2 fertilization in temperate forest FACE experiments not representative of boreal and tropical forests. Global Change Biology 14: 1.12.

Hickler, T., Vohland, K., Feehan, J., Miller, P. A., Smith, B., Costa, L., Giesecke, T., Fronzek, S., Carter, T.R., Cramer, W., Kühn, I., and Sykes, M. T.: Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model. Global Ecology and Biogeography, 21, 50-63, 2012.

Hickler, T., Rammig, A. & Werner, C. 2015. Modelling CO2 impacts on forest productivity. Current Forestry Reports 1: 69-80.

Lucht, Wolfgang, et al. "Climatic control of the high-latitude vegetation greening trend and Pinatubo effect." Science 296.5573 (2002): 1687-1689.

Medlyn, B.E., Zaehle, S., De Kauwe, M.G., Walker, A.P., Dietze, M.C., Hanson, P.J., Hickler, T., Jain, A.K., Luo, Y., Parton, W., Prentice, I.C., Thornton, P.E., Wang, S., Wang, Y.-P., Weng, E., Iversen, C.M., McCarthy, H.R., Warren, J.M., Oren, R. & Norby, R.J. 2015. Using ecosystem experiments to improve vegetation models. Nature Climate Change 5: 528-534.

Monserud, R. A., & Leemans, R. (1992). Comparing global vegetation maps with the Kappa statistic. Ecological modelling, 62(4), 275-293.

Pound, M.J., Haywood, A.M., Salzmann, U., Riding, J.B., Lunt, D.J., Hunter, S. A: Tortonian (Late Miocene, 11.61–7.25 Ma) global vegetation reconstruction, Palaeogeogr., Palaeoclim., Palaeoecol., 300, 29-45, 2011.

Ramankutty, N., & Foley, J. A. (1999). Estimating historical changes in global land cover: Croplands from 1700 to 1992. Global biogeochemical cycles, 13(4), 997-1027.

Smith, B., Prentice, I.C. & Sykes, M.T. 2001. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. Global Ecology & Biogeography 10: 621-637.

Smith, B., Wårlind, D., Arneth, A., Hickler, T., Leadley, P., Siltberg, J., and Zaehle, S.: Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. Biogeosciences, 11, 2027-2054, 2014.

Zaehle, S., Sitch, S., Smith, B. & Hatterman, F. 2005. Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics. Global Biogeochemical Cycles 19: 3020.

Zaehle, S., Medlyn, B.E., De Kauwe, M.G., Walker, A.P., Dietze, M.C., Hickler, T., Luo, Y., Wang, Y.-P., El-Masri, B., Thornton, P., Jain, A., Wang, S., Warlind, D., Weng, E., Parton, W., Iversen, C.M., Gallet-Budynek, A., McCarthy, H., Finzi, A., Hanson, P.J., Prentice, I.C., Oren, R. & Norby, R.J. 2014. Evaluation of 11 terrestrial carbon–nitrogen cycle models against observations from two temperate Free-Air CO2 Enrichment studies. New Phytologist 202: 803–822.

	AI 280 ppm	AI 450 ppm	Max	Min
Standard	-0.67	-0.96	4.7	-11.5
Absent-Absent = 1 (default = 0)	4.43	4.06	10.5	-11.5
Dominant-Dominant = 1 (default =2)	-0.91	-1.13	4.2	-11.5
Both of the above	4.19	3.9	10	-11.5
Minor disagreement = -1, disagreement = -2, major disagreement = -3 (default = 0,-1,-2)	-4.9	-5.23	4.7	-21.5

Table 1. Overall Agreement Index (AI) scores for the 280 ppm and 450 ppm Tortonian runs, as well as the minimum and maximum values calculated with different scores assigned for levels of agreement.

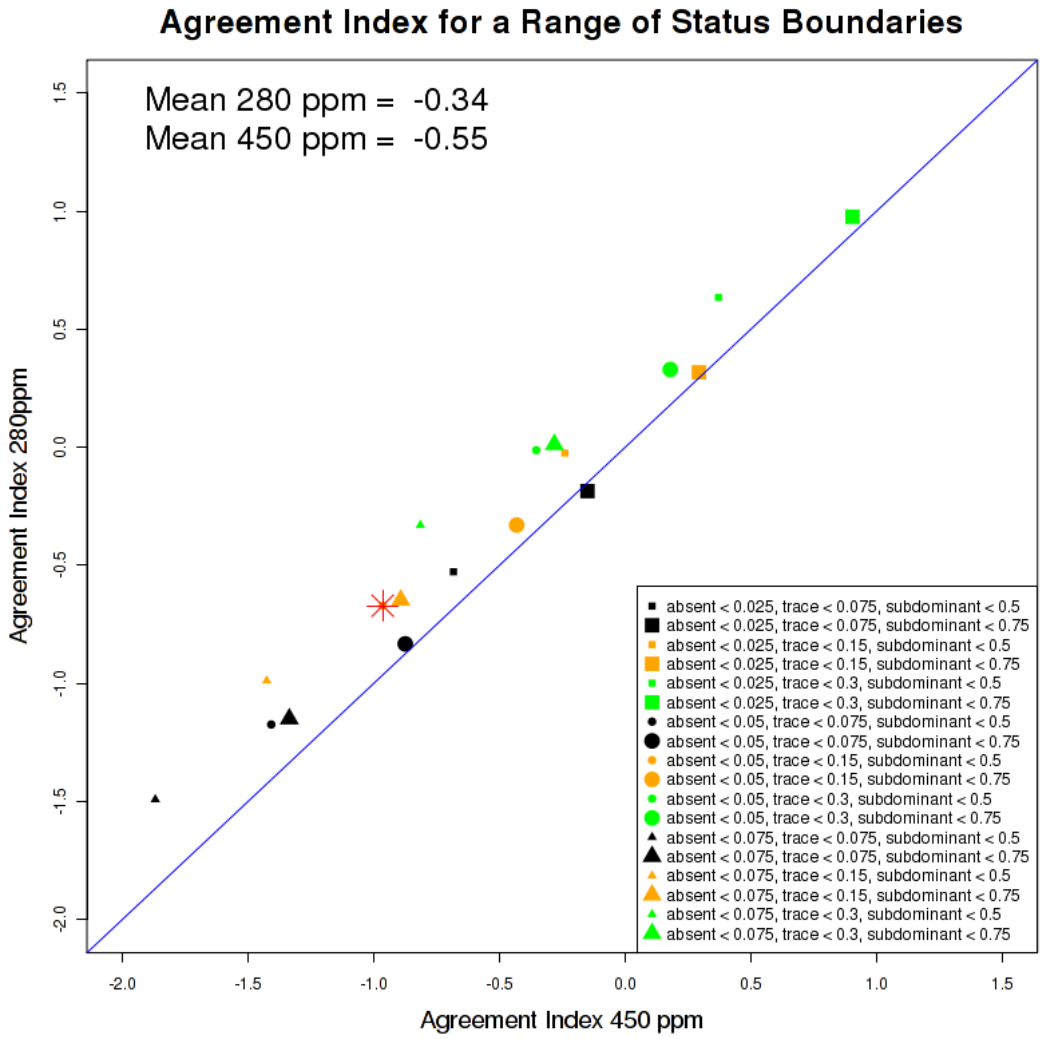

Figure 1. Agreement Index (AI) values for the 280 ppm and 450 ppm runs for different fractional boundaries of the AI statuses.

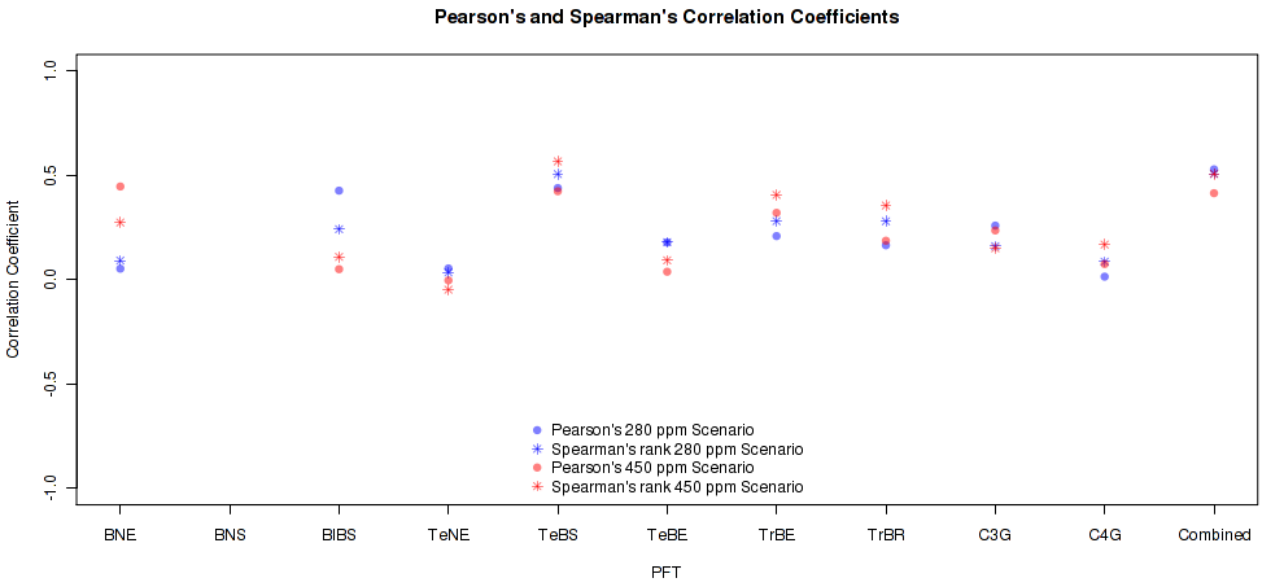

Figure 2. Pearson's product moment correlation coefficient and Spearman's rank correlation coefficients between the paleobotanical data diversity fractions and the simulated LAI fractions for the 280 ppm and 450 ppm CO_2 Tortonian scenarios.