Review of "Including vegetation dynamics in an atmospheric chemistry-enabled GCM: Linking LPJ-GUESS (v4.0) with EMAC modelling system (v2.53)"

The paper describes the atmosphere-to-land coupling of EMAC to the LPJ-GUESS dynamic vegetation model. Resultant vegetation properties that will become important during the two-way coupling step are benchmarked against observations for offline and two resolutions of the newly-coupled LPJ-GUESS in order to determine if discrepancies are due to the underlying vegetation model or climate biases from EMAC. I find it quite refreshing that the authors have focussed on the one-way coupling of part of the land surface as an initial step of ESM development, rather than including the full coupling in one paper. The paper includes a substantial description of future development priorities, which helps frame this analysis in the wider ESM development. It will be interesting to see if this step-by-step approach helps aid these future developments and evaluation of the final ESM. The models' performance seems pretty impressive in the comparisons in this study, which is definitely a good sign for thing to come.

Whilst individual parts of the paper are well explained and easy to understand, the structure of the paper as a whole is quite confusing and could do with some work. I also share some of reviewers 2 concerns around attribution of model discrepancies to the vegetation model, simulated climate or model resolution, as well as the modelling protocol description. However, after going through the manuscript a number of times, I think most of these concerns could also be down to the paper structure. I also have additional questions about the benchmarking methodology which might just need clarification, but could potentially require new analysis. Due to the suggested restructuring and number of general questions, I'm afraid the review is quite long.

I notice that I seem to be replacing reviewer 2, so I start by briefly addressing responses to reviewer 2 (none of which actually need a fresh response from the authors) before some suggestions regarding paper structure, followed by general methodological comments and ending with specific suggestions to the text.

Responses to reviewer 2.

One of the main concerns raised by reviewer 2 was that the papers focus on natural vegetation at the exclusion of land use coupling. The reviewer included a rather odd comment, suggesting that the authors should consider land use because they are from Europe - almost as if the development of **global** models should be based on our own circumstances or immediate surroundings?! It is perfectly fine to focus on natural dynamic vegetation in this paper, whilst noting (as the authors do) that land use will need including in future developments towards a completed ESM. Dynamics in natural vegetation behaves very differently from agricultural systems, and often require separate consideration (e.g. (Burton et al., 2019)). The paper also accounts for the discrepancy between the simulation of natural vegetation and observations including land use during benchmarking in what seems an appropriate way that is consistent with previous benchmarking studies (Kelley et al., 2013). I, therefore, feel the authors have addressed this point adequately.

I do, however, share the reviewer's concern about attribution of simulated discrepancies to LPJ-GUESS model structure, EMAC climate and model resolution; as well as the description of the model protocol, including the prescription of nitrogen, even in the revised manuscript. I have incorporated these into the remainder of my review.

Paper structure

The paper mixes future developments into the introduce and model description and has combined benchmarking methods with results, which at times makes the manuscript hard to follow. On the first read, the paper seems very short on results and discussion, and it's only on subsequent reads that I realised the authors do actually present enough results and make some very interesting discussion points. But these are lost throughout other sections of the paper. It would be much easier to follow the paper's narrative if it was restructured into a more traditional format (i.e an introduction; methods split between model description, modelling protocol, benchmark data and benchmark metric; results and discussion including future work, and then the conclusion). Below, I've highlighted some point that could be moved to more appropriate sections of the manuscript. But I have probably missed some, and I hope the authors can identify some more in the next iteration.

Abstract
The parts of LPJ-GUESS that are and are not being used in this study should be explicitly separated in the abstract. Lines 5-8 list a lot of processes that, despite being important in the Earth System, are not being considered for evaluation of this study. This list of processes could be moved into the introduction. Only some processes/couplings in the description of LPJ-GUESS on lines 10-14 are switched on, and not all of those are evaluated.

The abstract is very short on benchmarking methods and results. It should describe what variables are being benchmarked, and give a brief description of attributions of model performance between LPJ-GUESS itself, climate model biases and resolution.

Introduction
The introduction should mention the importance of attributing biases in simulated vegetation to the vegetation model deficiencies, GCM biases or resolution effects - especially as this probably the most important piece of analysis during benchmarking.

A lot of the model description should be moved to later sections. For example:

- How climate is aggregated and passed from EMAC to LPJ-GUESS (i.e page 2, line 32 “In both modelling systems…” to end of next paragraph on page 3 line 21) is more relevant to the methods (ie current section 2.3?) and most could be moved there, although it will be worth briefly mentioning that LPJ-GUESS has been used in ESMs before. Also, I’m not entirely sure it’s that relevant for this study to describe in detail how LPJ-GUESS is coupled to other GCMs (although maybe the authors disagree on this point…?), and I would suggest making only **brief** comments on past couplings when they use the same technique as the one described in this study.
- Page 3, line 31 “In addition…” to the end of the introduction could be summarized with the details moved to the discussion/future work.

Model descriptions and protocol (section 2 and 3)
Much of the model descriptions include a description of processes/coupling that have either not been implemented or aren’t switched on and therefore not relevant for the results that follow. These are useful to give a wider context to the coupling presented here but should be moved to other parts of the text to help make it much clearer what the authors are actually using in this study. Also, some of the modelling description, coupling implementation and simulation protocol seem misplaced within the methods. For example:

- Unless the authors use fire in their results, all of the paragraph at the bottom of page 5 could be moved to a “future work” section in the discussion. Maybe the authors do use (and during revision, go on to evaluate/discuss) GlobFIRM. In which case the first sentence can be kept.
- The last bit of the 1st paragraph on page 6 (line 7 onwards) could be moved to Appendix A, as it has more to do with modification and how to run the model and isn’t necessary information to evaluate the quality of the coupling or benchmarking (which the rest of the paper is dedicated too).
- The first half of the paragraph starting on line 24 on page 6 could be merged with the line 3-5 on the same page to avoid repetition.
- The “next steps” to the end of the following paragraph on page 6, line 26 to page 7, line 7 could be moved to the discussion.
- The 1st sentence of the paragraph starting line 16, page 8 should be moved to the model description (section 2.2)

Model evaluation.
The second half of the 2nd paragraph on page 9 feels like a discussion on poor model performance... before the results are actually presented! The NME scores actually turn out to be pretty reasonable so this slightly negative statement doesn't just seem misplaced but also takes away from the results that follow. This should, therefore, be moved to the discussion, and phrased in a slightly happier way.

Dataset descriptions and comparisons are combined for each variable. I can see the logic here, but it makes the m/s feel very jumpy, especially given the interactions between e.g. carbon to tree cover, tree cover/height to biome and biomass etc. The authors should consider putting descriptions of comparison datasets and benchmarking metrics into an earlier, more traditional methods section. That way, you can also describe how benchmark datasets are processed in one place (I'm sure I saw "conservative remapping" more than once.)

This should be followed by the comparisons, presented in an order which helps link how model errors in different variables affect one another and are affected by resolution and climate.

Discussion and Future work and development plans

There's no discussion section yet, but much of the text identified above could be moved into one. When rewriting, it should be made clear in the text which coupling/model developments are implemented but not yet assessed and what is still to be implemented (this shouldn't require too much effort as the authors have already demonstrated this quite nicely in Figure 1). Once this is done, development and evaluation priorities could be better linked to model deficiencies identified in the results/discussion (as has been done with things like disturbance rates etc).

General comments

Model descriptions

More description of the land surface scheme (outside of LPJ-GUESS) would be helpful. Specifically, as the paper only deals with one-way coupling, what vegetation cover/distributions does EMAC actually see in these simulations? And where are they obtained from? Are there any other, none-dynamic land surface properties that are relevant?

For the stochastic processes described, are these processes truly random? Or do they use semi-stochastic seeded random number generators? I.e if you performed the exact same simulation twice, would you get the same answer?

Modelling protocol

On line 18 of page 6, what is meant by "LPJ-GUESS provides fractional vegetation cover, leaf area index, daily net primary productivity and average height of each PFT to EMAC"? From Figure 1, I'm pretty sure this means that the coupling is technically implemented but not turned on for this study. But the text sounds a bit like EMAC is using information in LPJ-GUESS. The authors should make it clear what is and isn't turned on.

I am a little lost as to what the simulation actually represents? The solar forcing and CO_2 concentration of 367ppm suggest present day (and the authors should state which years this concentration is from). However, nitrogen deposition is from the 1850s - suggesting pre-industrial/early historic. What is the reason for the mismatch? I know the authors have said why in the responses, but a better definition of what these runs represent might help explain the mismatch in the paper. And how does Figure C1 show that there is no impact of nitrogen limitation?

What time period are the sea surface temperatures from?

On the whole, the run sounds like an equilibrium run. Was CRU-NCEP detrended to match (both for the spin-up and the final 113-year run)? And overall, what do the runs represent? An equilibrium version of the present day? Or a pragmatic spin up that could be used for further transient runs? Pragmatic is fine - we're all climate modellers and we know computer resources are too limited to run all the perfect runs we might want. But it would help when interpreting the results to better define the runs.

What resolution was the CRU-NCEP run? If it was different than the T42 and T63 runs, might this have a difference?

Was the 500-year spin up for the coupled EMAC-LPJ-GUESS, or was EMAC spun up using a separate protocol before being coupled to LPJ-GUESS? Either is a valid protocol to follow given EMAC doesn't actually see simulated vegetation properties, and it is not entirely evident from the text which is used. Either way, the spin-up protocol for the EMAC part of the model should be described. Did the 100 year period without N limitation follow an initial 500-year spin up? Or does the 100 years with N limitation + 400 years with N limitation constitute the full 500 years spin up?

How was the trend in PFT extension and height tested at the end of the spin-up? And does the inter-annual variability in vegetation refer to extension and height, or other vegetation properties as well? Was the trend in carbon pools tested?

Null models and metric interpretation

I can't see any reference to null models described in (Kelley et al., 2013; Kloster and Lasslop, 2017). I *think* a potential reason these have not been included is because the benchmarking is used to compare performance across models, rather than quantifying model performance itself. If this is this case, then it should be clearly stated. The 2nd paragraph in section 4, page 9 is probably a good spot to add this. However, I would point out that scores taking into account land use (in Table 2) look pretty good, and using null models may help highlight this.

As NME is basically absolute mean error, changes in scores are directly proportional to the distance away from the observations, so can be interpreted as % improvement/degradation in model performance. Maybe this could be explained when introducing NME and used when describing the scores? i.e for ree cover with LUC, T63 represents a degradation in performance of (0.69 - 0.62)/0.69 10.14% compared to CRUNCEP runs.

Choice of comparisons

Before introducing the datasets, the authors should justify their choices of variables for comparison, particularly why they are important to assess when coupling vegetation to a GCM.

There is no comparison of important biogeochemical earth system fluxes (i.e NPP/GPP, respiration, ET, methane, aerosols etc). The authors should justify why fluxes aren't considered or else consider adding these comparisons to their analysis. Especially as the title promises coupling to a "chemistry-enabled GCM" where fluxes will be important.

As the model should be in equilibrium given the modelling protocol (?) then the authors could also consider demonstrating the model's carbon is in equilibrium as well, i.e net ecosystem exchange is zero - a good basic test for an ESM in equilibrium runs.

From the description of biome reconstructions in section 4.1, it seems like biome comparisons are being used partially as a way of benchmarking several vegetation properties, including LAI. There are LAI products which could be used for benchmarking if LAI is an important coupling variable (Myneni et al., 2002)?

Vegetation cover comparisons

I know little about biome maps, but (Haxeltine and Prentice, 1996) seems a little bit old. (Oak Ridge National Laboratory, n.d.; Olson et al., 2001) maybe a bit newer? The use of (Haxeltine and Prentice, 1996) over other products should probably be justified somehow.

(Kelley et al., 2013) used the Manhattan Metrics (MM) for vegetation cover comparisons. Why was NME used instead? For two item comparisons (ie tree cover vs none-tree cover, as in this paper), I *think* scores obtained using MM and NME would be proportional to one another. If the authors can confirm this is the case (probably with a bit of maths in the response to this review), then using NME will be okay, and could be a better choice as NME is normalised by the variance around the mean of the observations, providing a more intuitive score. But the change in metric should be explained.

MODIS MODD44B Collection 6 measures woody cover of a height > 5m. Was LPJ-GUESS tree cover of less than 5m removed? If so, how? Does LPJ-GUESS use shrub PFTS? If so, were they included in tree cover comparisons?

Previous cover-based benchmark comparisons also compare herbaceous/total vegetation cover and details on leaf type and phenology (Burton et al., 2019; Kelley et al., 2013; Rabin et al., 2017). Some of these might not all be so relevant for ESM benchmarking, but the authors should either perform them or briefly justify the omission of at least total vegetation cover.

Fire

The manuscript describes what I *think* is the current fire model (GlobFIRM), as well as plans for future fire model development, at several points in the introduction and model description section. Yet there is no mention of fire in the results or discussion, except a brief comment about possible underestimation of fires effects on canopy height. If fire is important for any of the variables tested (e.g for veg distribution, as it is for (Bond et al., 2005; Burton et al., 2019; Hantson et al., 2016)) or for further development of ESM coupling, then surely it should be benchmarked as well? Simple burnt area and emission benchmarking is described in (Kelley et al., 2013; Rabin et al., 2017) and could be applied.

Impact of spatial resolution

The authors suggest at several points that degradation of performance of vegetation properties between T63 and T42 runs demonstrate poorer performance in EMAC at coarser resolutions. It's well established that changes in GCM resolution has an impact on model performance (e.g. (Kuhlbrodt et al., 2018)). However, I'm not sure the comparisons presented in this study lend any evidence to support this. It could be LPJ-GUESS performs worse at coarse resolution, and/or is sensitive to the aggregation of inputted climate information. Unless the authors can justify this statement another way (i.e driving LPJ-GUESS with CRU-NCEP gridded to the two scales of grid - which I'm sure would be much more effort than it's worth? Or making more use of the climate maps in the appendix), I would rephrase this argument to suggest the model as a whole (i.e EMAC+LPJ-GUESS) performs better at the increased spatial resolution, particularly in the first paragraph on page 17 and lines 4 and 5 on page 18.

To explore resolution effects further, I do actually think it would be interesting to see T42 in the other maps. There are clearly some big differences in biome cover in some parts of the world between model resolutions, including some interesting changes in the biomes in carbon-rich Amazon, Indonesian and South Asian forests (changes which I don't think are mentioned in the text?). It would be useful to see where these differences are coming from (including a T42 tree cover, height, and biomass maps and linking these to Figure B1-B3 climate where T42 is already plotted out).

Figures

Why is the background in different maps in figures 3-5 blue? Shouldn't they be white to match the part (a) figures?

There are streaks of missing data running across the Sahara and Arabian Peninsula and southern Australia what I think is in the (Avitabile et al., 2016) region of the biomass observations map. What's causing this?

Specific comments

Given the timing of the submission, I wonder if this model will be contributing to CMIP6. If so, the authors may wish to point this out somewhere?

Remove most of the instances of "state of the art". The authors either clearly demonstrate that the models they are coupling are the latest version, in which case the phrase is redundant, or use it when there is no extra justification as to why the model is "state of the art", in which case we are left wondering why it warrants the description.

Page 2, line 17 is the phosphorus cycle considered in EMAC or LPJ-GUESS? If so, add to the model description.

Page 2, line 26 add “being” between “actively” and “developed”

Page 4, line 23/24: “comparatively detailed”. Compared to what?

Page 4, line 29: replace “phenology” with “phenological response”

Page 4, line 29: C4 is just for grasses right? If so, state.

Page 4, line 30: Are any of the woody PFTs shrubs? If so, say so. If not, don’t worry.

Page 5, paragraph starting line 5: remove the information of version 3 of LPJ-GUESS. Unless version 3 is going to be used somehow (?), it is not relevant to this study. Apart from that, this is actually a very clear and concise overview of how LPJ-GUESS works!

Page 5, line 20-22 “However, monthly climate data… distinct rain events”. Is this climate interpolation actually used in coupling to EMAC? Or is it used when driving with the CRU-NCEP in this study? If not, remove.

Page 7, line 6-7: remove “Whilst it's not within the scope … dependence on spatial resolution” and if necessary, merge the remainder of the sentence with the next. The paper does later test biases/resolution that affects dynamic vegetation. And it’s self-evident that your aim is to test the coupling and veg dynamics, not the wider GCM.

Page 8 line 1: state what resolution or T42 and T63 are. (ie no. lat x lon. Deg at the equator or something like that). I know spectral resolutions are a little confusing in these regards, but it would be useful to give an idea of how much coarser T42 is relative to T63. And maybe discuss somewhere what applications each resolution is likely to be used for.

Page 8, line 8 remove “at no point were external climate datasets used”. I think the authors mean that no external climate datasets were used in the T42 and T63 runs. But that's pretty self-evident, so stating it here sounds like at no point in this analysis was climate data used, right before describing how climate data was used.

Page 8, lines 19-22: Explain how Fig C1 shows that PI N deposition does not result in additional N limitation. And additional when compared to what?

Page 9, line 4: This paragraph starts off with a very long list of previous benchmarking of LPJ-GUESS. It should be reduced to just state that LPJ-GUESS has a long history of development and evaluation, and then pick a couple of key references. And recent ones preferably using the version of LPJ-GUESS used here (if I remember correctly, (Gerten et al., 2004) for example, was pre-GUESS?)

Page 9, line 10-11: Remove “it is beyond the scope…. dynamic vegetation model”. I’m not sure it is beyond the scope to suggest changes to the dynamic vegetation model that might affect/be affected by the coupling. Especially because the authors do later describe some changes (i.e to disturbance rates, fire modules etc) that will change the dynamic vegetation model.

Page 9, line 31 replace “unity” with “1”

Page 10, line 6: The sentence “Whilst we can’t expect … forced using EMAC climate” seems to be making the same point as the start of the paragraph, and should be moved/merged into somewhere around lines 3 and 4.

Page 10, line 15: The paragraph starts off a little negative. Maybe just remove everything before the first comma and start the sentence by saying “Knowledge of EMAC biases…”.

Page 10, line 32: Add “Fig” before “2”

Page 11, line 24: Briefly explain conservative remapping.

Page 11, line 25: removed “as would be expected by a state-of-the-art DGVM”. I’m sure I would expect some DGVMs to do a rubbish job. The fact that you have a combined model which does kind of okay is pretty impressive and shouldn’t be understated.

Page 12 Please add dataset reference to figure 2 caption.

Page 15, line 22: remove “(lower is better)”. It’s already been described and the reader is reminded in the table caption.

Page 15, line 26, 27, sentence stating “For biomass…” states that biomass performance gets worse when accounting for LUC. I might be reading this wrong, but from Table 1, it seems the biomass results are actually getting better with the LUC modification, not worse…?

Page 15, line 31, sentence starting “In particular”. The authors state that the background disturbance rate could be changed to increase vegetation carbon in the T42 run. However, as far as I can tell, this disturbance rate is also used in the CRU-NCEP run as well (?), where biomass is generally too high. This suggests the problem is actually climate biases, and that EMAC should be improved to get a better representation of biomass. The authors may be putting this forward as a pragmatic way of getting the right carbon balance within an atmosphere model which is much harder to fix. If this is the case, then please say so in the text.

Also, how might the simplistic turn over rates be developed?

Avitabile, V., Herold, M. and Heuvelink, G. B. M.: An integrated pan-tropical biomass map using multiple reference datasets, Glob. Chang. Biol. [online] Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.13139, 2016.

Bond, W. J., Woodward, F. I. and Midgley, G. F.: The global distribution of ecosystems in a world without fire, New Phytol., 165(2), 525–537, 2005.

Burton, C., Betts, R., Cardoso, M., Feldpausch, R. T., Harper, A., Jones, C. D., Kelley, D. I., Robertson, E. and Wiltshire, A.: Representation of fire, land-use change and vegetation dynamics in the Joint UK Land Environment Simulator vn4.9 (JULES), , doi:10.5194/gmd-12-179-2019, 2019.

Gerten, D., Schaphoff, S., Haberlandt, U., Lucht, W. and Sitch, S.: Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model, J. Hydrol., 286(1), 249–270, 2004.

Hantson, S., Arneth, A., Harrison, S. P., Kelley, D. I., Prentice, I. C., Rabin, S. S., Archibald, S., Mouillot, F., Arnold, S. R., Artaxo, P., Bachelet, D., Ciais, P., Forrest, M., Friedlingstein, P., Hickler, T., Kaplan, J. O., Kloster, S., Knorr, W., Laslop, G., Li, F., Melton, J. R., Meyn, A., Sitch, S., Spessa, A., van der Werf, G. R., Voulgarakis, A. and Yue, C.: The status and challenge of global fire modelling, Biogeosciences, 13(11), 3359–3375, 2016.

Haxeltine, A. and Prentice, I. C.: BIOME3: An equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability, and competition among plant functional types, Global Biogeochem. Cycles, 10(4), 693–709, 1996.

Kelley, D. I., Prentice, I. C., Harrison, S. P., Wang, H., Simard, M., Fisher, J. B., Willis, K. O. and Others: A comprehensive benchmarking system for evaluating global vegetation models, Biogeosciences, 10(5), 3313–3340, 2013.

Kloster, S. and Lasslop, G.: Historical and future fire occurrence (1850 to 2100) simulated in CMIP5 Earth System Models, Glob. Planet. Change, 150, 58–69, 2017.

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., Hill, R., Graham, T., Ridley, J., Blaker, A., Calvert, D., Copsey, D., Ellis, R., Hewitt, H., Hyder, P., Ineson, S., Mulcahy, J., Siahaan, A. and Walton, J.: The Low-Resolution Version of HadGEM3 GC3.1: Development and Evaluation for Global Climate, J. Adv. Model. Earth Syst., 10(11), 2865–2888, 2018.

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G. R., Lotsch, A., Friedl, M., Morisette, J. T., Votava, P., Nemani, R. R. and Running, S. W.: Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sens. Environ., 83(1), 214–231, 2002.

Oak Ridge National Laboratory: Olson's Major World Ecosystem Complexes Ranked by Carbon in Live Vegetation: An Updated Database Using the GLC2000 Land Cover Product, [online] Available from: https://cdiac.ess-dive.lbl.gov/epubs/ndp/ndp017/ndp017b.html (Accessed 26 April 2019), n.d.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P. and and Kassem, K. R.: Terrestrial Ecoregions of the World: A New Map of Life on Earth A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity, Bioscience, 51, 933–938, 2001.

Rabin, S. S., Melton, J. R., Lasslop, G., Bachelet, D., Forrest, M., Hantson, S., Li, F., Mangeon, S., Yue, C., Arora, V. K. and Others: The Fire Modeling Intercomparison Project (FireMIP), phase 1: Experimental and analytical protocols, Geoscientific Model Development, 20, 1175–1197, 2017.