Response to reviewer comments on “Assessment of model estimates of land–atmosphere CO2 exchange across Northern Eurasia” by M. A. Rawlins et al.

Anonymous Referee #1

Authors present a multi-model assessment of the carbon fluxes across the North Eurasia in last 50 years. Models are driven by observation-based climate data. Authors conclude that the soil carbon storage increases in last decade as compared to first decade of the analysis period, which happens despite decline of the soil carbon residence time due to faster decomposition rates owing to higher temperatures and a longer warm season. The test of the models against the GPP and NEP observations are made too. The findings are interesting for climate impact assessments and recommendations are also made for future model improvements. Manuscript is well written and has sufficient scientific value to be accepted for publications.

However it is recommended to give authors a chance to make minor corrections, additions to the discussion part.

There are several factors not covered by the model analysis that need to be reflected in the discussion, concluding remarks:

> We thank the reviewer for the detailed comments which helped to improve the manuscript. Our responses to all comments are included below.

(**A1**) Fire regime change. Carbon harvesting by fires left out of scope, for convenience and fairness of the model inter-comparison. On the other hand accounting for fire fluxes would greatly complement the assessment of the carbon sink made in this study.

> Only five out of the nine models analyzed account for any form of disturbance (e.g. Table 2 model summary in revised manuscript). Of these, four treat fire and three land-use change. Given the limits imposed by only four models accounting for fire, we have chosen to not analyze or discuss this disturbance in any great detail. We do however mention the need for accounting for disturbances in models. Lines 444-446 and 534-536.

(**A2**) Mismatch between modeled and satellite driven (MOD17) GPP pattern was mentioned without a hint at underlying cause. It could be under-representation of the edaphic variability across the landscapes. Do soil and drainage efficiency maps used in modeling reflect it properly? It may also be a cause of the problems with matching the observed GPP and NEP at flux tower sites.

> We now provide more information relevant to the mismatch. We have expanded Table 2 to include more differences among the models, including climate forcing data (Table 1) and certain key parameterizations (Table 2). We have added a statement that the simulation protocol allowed for the choice of a model's driving datasets (lines 109-110). Given the wide range in model forcings, initial conditions, and representations of dynamic vs static vegetation, we concluded that it is not possible attribute mismatches to more specific factors.

(A3) Role of the nitrogen cycle feedback in increasing net carbon uptake in warming climate has been discussed only briefly. While the role of nitrogen cycle is varying between participating models, all models predict similar sign of sensitivity to climate change, which show some improvement since discussion by Sokolov et al, (2008). Table 2 states the nitrogen limitation is not included. That gives impression that detailed nitrogen cycle is not that needed. There are many processes that do need more explicit treatment of nitrogen cycle in northern high latitudes, like increased soil nitrogen availability due to decomposition of the stored organic matter in the thawed permafrost.

> We have examined more closely how nitrogen limitation affects GPP, and also have included additional information about the role of the nitrogen cycle (as in nitrogen limitation). Nitrogen limitation is especially important as high latitude soils are predominantly nitrogen limited. Thawing permafrost, however, may result in the exposure of more nitrogen which could potentially increase above ground productivity (Shaver et al., 1992). Few vegetation models have implemented nitrogen cycling, and thus do not represent nitrogen limitation. Generally speaking, models which do not include nitrogen limitation would tend to overestimate terrestrial carbon uptake in the presence of enhanced CO2 fertilization (Thornton et al., 2007). Only one of the models examined in this study, the CLM4.5, has the effects of nitrogen limitation on photosynthesis implemented in its simulation. We mentioned the and added a reference at lines 366-369. We point to the critical importance of improvements in modeling N cycle processes at line 532-534 in the Conclusions of the revised manuscript.

Shaver, G. R., Billings, W. D., Chapin, F. S., Giblin, A. E., Nadelhoffer, K. J., Oechel, W. C. and Rastetter, E. B.: Global Change and the Carbon Balance of Arctic Ecosystems Carbon/nutrient interactions should act as major constraints on changes in global terrestrial carbon cycling, BioScience, 42(6), 433 441, doi:10.2307/1311862, 1992.

Thornton, P. E., Lamarque, J.-F., Rosenbloom, N. A. and Mahowald, N. M.: Influence of carbon-nitrogen cycle coupling on land model response to CO2 fertilization and climate variability, Global Biogeochem. Cycles, 21(4), GB4018, doi:10.1029/2006GB002868, 2007.

(**A4**) Technical comment:
Fig 12 is not easy to read.

> We have rewritten the caption with a stepwise description of the trend estimation and configuration of the graphic. We also use two examples in the caption.

References:
A. P. Sokolov, D. W. Kicklighter, J. M. Melillo, B. S. Felzer, C. A. Schlosser, and T. W. Cronin, 2008: Consequences of Considering Carbon–Nitrogen Interactions on the Feedbacks between Climate and the Terrestrial Carbon Cycle. J. Climate, 21, 3776–3796. doi: http://dx.doi.org/10.1175/2008JCLI2038.1

A. J. Dolman (Referee)

This is a useful and timely paper, in general well written that compares the output of 9 different models for GPP, Respiration and NEP over a large regions in Northern Eurasia. The model output is benchmarked against eddy covariance data from 4 sites and against a satellite remote sensing product of GPP (MOD17). While the general line is good, the paper misses several opportunities to present a clearer analysis and hence at times the paper ends in a somewhat lacklustre description of numbers, rather than that it tries to identify which process description in models produces what behaviour. As an example in the discussion the impact of nitrogen is mentioned, but nowhere in the paper is an attempt made to use the fact that two of the 9 models incorporate a nitrogen cycle. The authors also conclude that the model's treatment of respiration needs to be improved, similar to previous authors (e.g. Dolman et al., 2012), but again do not use differences between the models to shed more light on how they see this improvement. I would therefore suggest that the authors consider the differences between the models (Table 2) more in their explanation of the results.

> We thank the reviewer for the insightful critique of our manuscript. Quegan et al. concluded that "...the representation of heterotrophic respiration and disturbance appears to be inadequate in the DGVMs..." Our findings build on the results of that study and others which point to issues with model depictions of carbon cycle dynamics across parts of northern Eurasia. Detailed responses to all comments are found below.

> Above we described, in response to referee #1, the additional analysis and discussion around the issue of nitrogen limitation on photosynthesis We indicate that four models account for fire (Line 446). Regarding differences in respiration, we have been unable to identify any specific structural representation among the models that would clearly explain the model differences. We speculate that the amount of exposed soil carbon, and soil temperature, are dominant controls. We have included a sentence pointing to high soil carbon amounts as a potential contributing factor for the elevated model simulated ER at several of the tower sites, relative to the respective observations. This at line 372-374. At line 374-376 we reference a study linking soil carbon pool size to heterotrophic respiration rates. A more detailed analysis of below-ground processes is beyond the scope of this study. We have also added additional information (supported by a study published in BGD, Koven et al., 2015) on the reduction in soil carbon residence time. The overarching conclusion is that increased productivity is the major driver for residence time deceases, as opposed to warming-related respiration increases. Lines 479-487.

> Regarding uncertainties in the model GPP and NEP estimates, we include in the revised manuscript two maps of the coefficient of variation (Figures 5, Figure 8b, Figure 14b). These maps show that model agreement is better in the forest/tiaga biome than the tundra and steppe areas. CV patterns are mentioned at lines 246-248 (GPP) and 264-266 (NEP). The CV estimates for residence time are higher, as described at lines 319-321. We also point to these uncertainties in the Conclusions at line 519-521.

> Detailed replies to specific comments are made below.

Koven, C. D., Chambers, J. Q., Georgiou, K., Knox, R., Negron-Juarez, R., Riley, W. J., Arora, V. K., Brovkin, V., Friedlingstein, P., and Jones, C. D.: Controls on terrestrial carbon feedbacks by productivity vs. turnover in the CMIP5 Earth System Models, Biogeosciences Discussions, 12, doi:10.5194/bgd-12-5757-2015, http://www.biogeosciences-discuss.net/12/5757/2015/, 2015.

(B1) p 2260 line 6-12. This is an complicated sentence to read, certainly for the introduction. Please reformulate. The reference to Cox et al., 2000, is a bit strange. This paper, while pointing to the importance of a carbon-climate feedback, mainly identifies the tropics as the key region.

>We have reworded the sentence as suggested and removed the reference to Cox et al., 2000. Lines 30-34.

(B2) p 2260 line 15-17. It is important to stress that, yes there may be increase in GPP at the beginning of the season, but also an extended period for respiration at the autumn (e.g. Parmentier et al., 2011 show that this may lead to no change in the net flux).

Parmentier, F. J. W., van der Molen, M. K., van Huissteden, J., Karsanaev, S. J.,
Kononov, A. A., Suzdalov, D. A., Maximov, T. C., and Dolman, A. J.: Longer grow-

ing seasons do not increase net carbon uptake in the northeastern Siberian tundra, J. Geo- phys. Res., 116, G04013, doi:10.1029/2011JG001653, 2011.

> We have added a sentence that cites Parmentier et al. 2011 and emphasizes that warming may also lead to increases in respiration which offset productivity increases, resulting in low net carbon uptake. Line 39-42.

(B3) page 2060. It is important to name the different period for which the numbers quoted in this section are obtained. This is relevant to the discussion of trends later on in the paper.

> There is no page 2060. It is not clear whether the comment refers to page 2260 or 2261. There are no numbers in 2260 and the 2nd paragraph of 2260 mentions the relevant time periods.

(B4) page 2262 line 6 efficacy means according to Wikipedia the capacity for beneficial change (or therapeutic effect) of a given intervention. I am sure the authors mean efficiency. That being said, the error analysis is a bit poor, basically the error is defined as modelled minus means. With these comparisons I would suggest that more advanced metrics, such as Nash-Sutcliffe efficiencies or other allow a clear interpretation of the results, give more depth to the analysis and might also help with meeting my earlier comment on the character of the analysis).

> From Merriam Webster dictionary: 'efficacy' is "the power to produce a desired result or effect" (see http://www.merriam-webster.com/dictionary/efficacy). We retain the word. However, we now make more clear that we are assessing how well the models reproduce measured fluxes as captured by the towers and remote sensing measurements. We have included the Nash-Sutcliffe coefficient of efficiency (E) metric in the interpretation of the results. Lines 211-229.

(B5) Page 2263. It would help if the authors could say something about the modelling protocol that was followed: what was the spinup procedure, which forcing data (or model) sets were used etc. In fact, the big question here is do all model start from the same initial condition in 1960, or are there already big differences after spinup. A big question is also why some models are apparently run as DGVM, while it is also possible to fix the land use. Some of the differences may thus be due to different land use types, rather than process description.

> Each model group was free to choose how the simulations, including vegetation specifications and dynamics, were constructed. We have added additional information on the modeling protocol in the revised manuscript. The revised text now describes (lines 109-111) "how each model was free to choose appropriate driving data sets for atmospheric CO_2, N deposition, climate, disturbance, soil texture, and other forcing data." The driving data sets are listed in Table 1. Spin up was also specific to each model, but it was conducted to support the delivery of simulation results starting in 1960. Given our results, it is very likely that there are considerable differences in flux and storage magnitudes among the models in the years prior to 1960. We have added information in the discussion relevant to spinup for the ISBA model and its trend in accumulated NEP. Line 448-454. A tightly controlled model intercomparison would likely shed more light on differences in GPP, ER, and NEP across the models. With the wide range in initial landcover, spinup details, and dynamics through time for some of the models, we are unable to say more on these issues in this study. We point to several of these differences in the discussion section of uncertainties, lines 360-369.

(B6) Page 2264. Is the Zotino site used, the fir, or pine forest. The fir is rather a strong sink (see Dolman et al., 2012, figure 4. I am also rather surprised that there are only four sites used. There is more data available. Why was that not used?

> We have added details on the sites. We are using the data for the RU-Zot tower contained in FLUXNET. The data for other research sites near Zotino are not publicly available. The FLUXNET web site at ORNL (http://fluxnet.ornl.gov/site/721) identifies the site's Plant Functional Type as "Evergreen Needleleaf Trees" and the IGBP land cover as Woody Savannas. The metadata do not indicate "fir" or "pine forest" type. In addition, we have now listed the landcover for three research sites near Hakasija at line 148-149. We have also modified the sentence on line 151 to point out that Hakasija and Zotino are in higher productivity area compared to the other two sites. Regarding additional sites, we have used all of the data for the study region that is publicly available in the Le Thuile dataset These data were processed using a consistent process (e.g. screening and gap filling, estimation of component GPP, R fluxes from NEE). The lead of the FLUXNET Science Management Committee has confirmed that we are using all tower records available. For a number of reasons, limitations on data availability are likely a particular problem for studies across Eurasia, as opposed to North America. Given the limited data available we have added a note of caution (line 346-347) regarding study findings drawn from evaluations from such a small number of sites. In the Conclusions we emphasize the need for additional flux-tower measurements.

(B7) Page 2065. It would help to cite the references against which the MOD17 products appears to have been "extensively evaluated"

> We have referenced several studies in which the MOD17 product was evaluated over northern, boreal-Arctic land areas. We cite four papers: Heinsch et al., 2006; Turner et al., 2006; Zhang et al., 2008; Zhao et al., 2010. Lines 167-169.

(B8) Page 2266. Line 4. assessment, not assessments.
Change made to singular.

(B9) Page 2066. While accepting whether this is to some extent subjective, I do not quite agree with the description of the model at match. Please be as objective as possible. A model like ISBA is clearly way off of properly simulation both GPP and R, and only comes back at NEP because they cancel by and large. Also the timing is not always consistent, some models are a month off. That is in a growing season of 2-2.5 month a substantial mistake, even if the use of monthly means would exaggerate such a mismatch. See also my earlier remark about using more objective metrics.

> We have reworded several statements and removed others, and feel we are being objective with the evaluations. To this end we have computed and presented additional model evaluation statistics in a newly added Table 5. For GPP and ER we show the Nash-Sutcliffe coefficient of efficiency (E) and the refined index of agreement (Willmott et al., 2012). The values of these additional error metrics reflect the salient conclusions that we have drawn from the mean errors shown in Table 5. As stated at lines 223 to 229, the overarching results are that (i) ER errors exceed GPP errors and (ii) errors in g C m-2 are greater at the more southern, higher productivity sites. We have pointed out that "across model standard deviations in areas of small positive and negative NEP are often a factor of ten or more greater than the across model mean, highlighting the large uncertainty in NEP at local scales." Lines 442-444. Lastly, we have added a statement in the Conclusions, stating that... "The models exhibit a wide range in spatial patterns and regional mean magnitudes." Line 511.

(B10) Page 2267. line 7-10. I do understand what you are trying to say here. Please reformulate.

> As originally drafted, our statement pointed to the fact that errors between NEP estimated by the models and flux-tower measurements, just before and after winter, are much lower for Chersky and Chokurdakh, relative to Hakasija and Zotino. Here the observations show NEP at between 10 and 30 g C m^{-2} $month^{-1}$ just before and just after winter. No large outliers appear among the models. We have reworded the sentence (line 207-209) to simply state that "Errors in the magnitude and timing of NEP prior to and following the dormant season are much smaller at Chersky, and to some extent Chokurdakh." In the next sentence we conclude the paragraph with "However, a lack of available tower-based data during the colder months limits the robustness of our assessments during that time of year." Lines 209-210.

Page 2267 line 25-29. A score of less than 50% would suggest you might as well use climate (T,P) and empirical relation between these two variables and GPP or even a random number generator. Please realise these results are really bad, and question the use of these models. So try to reformulate this and emphasise that the models are doing an extremely poor job here.

> We agree that several of the models show poor performance in reproducing the spatial pattern in MOD17 GPP. We have also added a statement in our recommendations that the use of individual models should be avoided in the absence of a rigorous validation against observations across the area of interest. Line 542-544.

(B11) Page 2268 line 10-12. Most models have hardwired a relation between GPP and R, so this high correlation does not say much.

> We do not use the correlation to emphasize model performance. Rather, the correlation helps illustrate that relationship with each model. Using the figure we also point to the range across the mean GPP and ER.

(B12) 2270. While I appreciate the discussion on residence times, I do find it questionable to calculate this, given the discrepancies in GPP, R, let alone soil carbon stocks (see also remark on spin up). I wonder whether this should not be tuned down a little and mentioned that as a results of wrong fluxes, a very large variability in residence time is obtained, with opposing trends, and signs.

> We agree that there are large uncertainties in estimated residence time given the wide range in GPP and ER. However, given that the models all simulate a decline in residence time, we retain the text, but also refer to the relatively high coefficient of variation (CV) values produced from the across-model mean and standard deviation. This statement is found at lines 264-266. These high CVs are indeed related to the large range and uncertainty in soil carbon amounts. In the results section we have added in the revised manuscript that soil carbon varies by an order of magnitude across the models. Line 302.

(B13) Discussion. 1) I really miss a remark on thermokarst, changing hydrology or other cryogenic processes. This is crucially important in this part of Eurasia. 2) A discussion where use is made of the differences in model processes and parametrizations would bring more depth to the analysis, than the current line which basically states that respiration, soil dynamics and productivity need to be improved. These three processes are of course the core purpose of the models, so either one concludes that they fail for that purpose, or one should make an effort to go a little bit more in depth.

> Table 2 includes additional details on model representations of cryogenic processes, including snow insulation, talik formation, soil hydrology and thermal dynamics. We have also added a paragraph (lines 497 to 507) which describes important model elements and environmental processes most critical to effects of warming and landscape change on carbon fluxes, including land to atmosphere fluxes of both CO2 and CH4. Given the lack of any consistency in model setup among these processes, we see no potential for the "use" of these differences among the models. Regarding failure, we are not performing an evaluation of these processes, so model effectiveness can not be determined. While we appreciate the suggestion, we see no benefit to the reader in any great in-depth explanations of the range of processes across the model suite which deal with snow insulation, talik formation, soil hydrology and thermal dynamics.