Final authors response to comments to the manuscript BG-2012-433

Dalmonech and Zaehle Constraints from atmospheric CO2 and satellite-based vegetation activity observations on current land carbon cycle trends.

REVIEWER #1

Reply to major comments:

I first thank the reviewer for his useful comments. It follows the reply to the list of reviewer's comments (in bold).

(1)The title is a bit misleading, and reads more like a 'data-model assimilation scheme'. I suggest changing it such that is more consistent with the key objectives of the paper (e.g. Towards a more objective evaluation: : :.).

We agree and propose a revised title: Towards a more objective evaluation of modelled land carbon trends using atmospheric CO2 and satellite-based vegetation activity observations

(2) In the abstract it states that ' : :The selection of observational characteristics (traits) specifically considers the robustness of information given the uncertainties in both data and evaluation scheme: : :'. When I first red this I became very excited, because this would indeed be a significant improvement. But after reviewing, I didn't really see how uncertainties were explicitly treated. Just not using the observations when they are not as robust (e.g. optical satellite data over the tropics) does not have much merit in my opinion. One could rank the observations in regards to robustness. For example, satellite vegetation data are in general most robust at seasonal time scales, and least robust at quasidecadal time scales (Tucker et al. 2005; ref. given in ms). There is a related discussion about this point in the Introduction on Page 16090 (lines 1-8) as well, but the stated 'philosophy' based on a null-model and how this deals (or does not deal) with uncertainties in observations and evaluation scheme are not clear to me. Since this is a key point of the paper, the specific approach and thinking behind it should be made as transparent as possible to increase the impact of the study.
(3) Since this study uses only a small subset of available observations (e.g. see Randerson et al. 2009), it appears that the key contribution to the existing state of the art in model data comparisons is the extension towards a more objective evaluation scheme including quantitative model performance (ranking) measures. I, therefore, suggest emphasizing this portion of the study much more through restructuring the paper. The following comment (main comment (4)) is also related to this.

We agree with the reviewer that the methodology is the key innovation of this manuscript.

We first apologies to have been unclear about the scope of the method regarding the treatment of uncertainties. Common to most of the evaluation schemes, data and model errors are not considered explicitly. This is the major limitation of the study and will be explicitly stated in the revised version of the manuscript. Errors are not considered because unknown in most of the case, and because, at exception of the cost-function metric, metrics do not take in to account biases and random errors in their mathematical formulations. However, the cost-function metric, requires the knowledge of the structure of model and data error (see i.e. review of Raupach etal.2005 GCB).

What we meant to stateis that given i) uncertainties in data and observation, operators to link model and data exist, and ii)data error and structural errors are often not known or provided quantitatively, our metrics and traits have been focussed in characteristics that are robust against these unquantifiable uncertainties. In addition, we provide the lower benchmark (see below) which allows us to provide an objective evaluation of the numerical value of the statistical comparison against a null-model, accounting for the uncertainties in the observation operators.

We follow the suggestion of the reviewer emphasizing aspects as the 'objectivity and robustness of the method' and in the revised introduction we make clearer key points of the approach following what is reported thereinafter:

1) We decided to use direct observations and dataset that were not strongly model-mediated (this explains why a restricted number of reference dataset was used, and we did not use upscaled products), and with a global coverage (this explains why we used atmospheric CO2 and satellite based vegetation activity)

2) To bypass part of the problem of uncertainties of the data sources (that are unknown), we aimed to provide traits and metrics that can be robust against the use of different satellite datasets (and not i.e. using several dataset and providing then weights to the final scores as in Randerson et al.2009, since anyhow the definition of a weight represents a personal author choice).To do this we used traits based mainly on sign of changes and specific traits in time. With regards to the decadal time scale, GIMMS-NDVI has been shown to be suitable for long term trend study (see Beker et al.2011, in the ms).

3) The use of a lower benchmark (the null-model) helps to partially circumvent the limits due to the uncertainties of the evaluation approach: A part of the uncertainties in the evaluation scheme emerges because the interpretation of the numerical value of the metric is not straightforward. While this is simple for the perfect data-model match (i.e. root mean square error equal to zero, or correlation equal to 1), this is ill-defined for the case of a mismatch. The idea of our work is that we provide a lower bound to the metric by calculating the statistic for a reference case (neutral model). Any numerical value of the metric in question better than this reference value implies that the land surface model under investigation adds information to the signal, and the linear distance between the numerical values of the metric for the reference and the 'perfect fit' case provides an objective measure of how good the model is relative to the null model. To ease comparison amongst traits, the final metric is always scaled to the lower benchmark and in the range 0-1.

This third point is particular important when considering the uncertainties in the transport model, since the lower benchmark is obtained using the same transport model. Scaling the metric to the lower benchmark (the CO2 signal obtained from fossil fuel emissions and net ocean net carbon fluxes only, implying a neutral land), highlights the contribution of modeled land fluxes to match the observation. Hence, the final score number is cleaned by the contribution of other CO2 source/sink other than the modeled land fluxes. This statement is also linked to the reply of comment #5.

In the revised version, we will make changes to clarify the motivation, scope and limitations of the approach following the points reported below

After line 12 of page 16089 we will include:

- Several global model evaluation analyses have been published in the last decades with respect to land model performances of the carbon cycle (Anav, et al.2013,Cadule et al.2009, Blyth et

al.2009,Randerson et al.2009, Heimann et al.1998), however they differ with respect to reference dataset used, selection of the observational traits and their computation and mathematical formulations used to quantify the data-model mismatch. This is cause of uncertainties when it comes to rank several land surface models Recent model benchmarking initiatives (Randerson et al., 2009; Luo et al.,2012) have underlined the need for the development of a standard set of tests and metrics applicable to any land surface model at different spatial and temporal scale.

- In addition to a lack of standards, a key challenge in evaluating global biosphere models comes from the uncertainties in observations. From a perspective of data-model mismatch quantification, given uncertainties in data and observation, operators to link model and data exist, however data error and structural errors are often not known or provided quantitatively (e.g. Raupach et al.2005).

- This study adds hence a new component to the attempt to move toward a more standard protocol by defining novel tests and quantitative model performance measures that are robust against the mentioned unquantifiable uncertainties. In order to provide a robust and a more objective evaluation framework we first selected only a parsimonious number of reference datasets that are as much as possible direct observations . Atmospheric CO2 and remote sensing data of vegetation activity were thus selected taking advantage of the complementary information contained in atmospheric CO_2 observations and remote sensing data of vegetation activity.

- We selected ecologically relevant information from 'observations' which helps constraining model projections, but robust with respect to the choice of satellite based dataset or transport model. The traits and metric selection was carried in the direction to provide model test sensitive to differences in signs of the changes and difference of phase detected between data and model

- Compared to previous studies, the analyses performed here attempt to provide a more objective evaluation approach by imposing a lower acceptable model performance measure (baseline benchmark) based on the assumption of a Null-model, i.e. a model that does not show any trend in the quantity under investigation. This information are used to set a lower bound for the metric aim to quantify how much information the land surface model adds to the signal and to quantify hence how good the model is relative to the null model.
- Despite no uncertainties in data or model were quantitatively included in the benchmarking framework,we show how the informations extracted in the data are able to disentangle model biases.

(4) Figure 2 captures a big portion of the key results. But after reviewing the paper I was unable to understand how the global scores were really derived. I suggest to make this point much more transparent.

In the new version of the ms, the section 2.4 will be re-nominated as 2.4 The baseline benchmark and final scores. In the revised section, we will detail how the aggregation to the global scores were made, as follows:

First the scores are computed for the model output and are considered as raw-metrics (Mor in eq.1). The same statistics is applied to return the metric for the lower benchmark case (Mref in eq.1). The latter result is used to scale linearly the metric of the model to a new 0-1 metrics according to eq.1. where 1 indicate perfect data-model match and 0 indicates that the model is not able to perform better than a system without land.

For the satellite-based scores, the global score is the average of the score computed for each transcom3 region, whereas for the CO2-station based scores first the scores for each station were averaged by latitudinal band. The global score was then derived as the average of the score computed by latitudinal band. In the polar plot, 0 score.

(5) Section 3.5.3. I wonder how meaningful any model data comparisons are that involve the atmospheric CO2 growth rate (e.g. Fig. 10). This metric integrates numerous carbon sink processes at various spatial and temporal scales. This also points to a general problem in model-data intercomparisons in regards to using observations that are difficult to interpret. I would refrain from using such 'traits' as it is not clear what understanding could be gained in such comparisons.

It is clear that using atmospheric CO2 as target, we are considering a mixed signal coming from different regions and processes (but this holds also for the mean seasonal cycle of CO_2, or most other observables at larger spatial scales).There is a lot of evidence to support an important role of the land's response to climate variability at interannual-decadal time scale in the interannual variability of the atmospheric growth rate of CO2 (Cadule et al.2009 in the ms, Peylan et al.2005 in the ms,Keeling et al.1995 in the ms, , Cox et al.2013 on Nature), as even suggested by the reviewer (see his comment 7). Also, the long-term trend in atmospheric CO_2, given fairly well-known emissions and net ocean fluxes (le Quere et al. 2009) provides information on the land C uptake that we expect a land model to simulate.

The fact that a trait is not linked to a specific single process, does not mean that can not be used to indicate that the model is not able to return what we capture in the observations, i.e. we can still diagnose the performance of a model. The precise point of this paper then is to link this information to independent information (from satellites in this case) to see whether we can isolate the reasons for this mismatch with alternative constraints. Admittedly, the manuscript does not completely achieve this task, as sufficient information to constrain all aspects of the land carbon cycle are not available. The merit of our work is that we try to move behind metrics that only considers the mean state of the system/climatology of the variables of interest (i.e. the mean seasonal cycle of atmospheric CO2). In the perspective of evaluating models that are used to return future carbon projection, also the climate-induces variability at interannual time scales and longer time scale has to be of interest in the evaluation.
In the revised manuscript, we will strengthen the discussion to underline the point made above.

Reply to minor comments:

(1) Page 16096, lines (13-21). Using observations in the most meaningful way. The CO2 record derived from observing stations is a complicated signal as it contains a long-term trend (from ff burning emissions) and a superimposed seasonal cycle due to plant activity. Simply doing a trend analysis on monthly CO2 data, as stated is not a meaningful metric, and one has to extract the seasonal cycle in a prior step (see Keeling et al. 1996; ref. given in ms). Extraction of the seasonal cycle, however, relies on knowledge about the underlying trend (which is not known), and for that reason the amplitude of

the seasonal cycle is considered the most robust signal (and hence used by several authors as a carbon cycle metric) followed by the downward and upward shifts. Unless the authors can convincingly show that their seasonal cycle xtraction method does account for the complexity in these data, I would not have any trust in the bserved monthly CO2 trends (MT) and thus also not in the corresponding data model comparison.

The monthly trend was performed on the seasonal signal of the atmospheric CO2 defined as the signal that retains only the frequencies higher and equal to the annual frequency, according to Thoning et al.89. We have shown (in Figure 5a) that the resulting trends are not affected by ocean and fossil signals. We conducted further experiments that were not reported in the manuscript to ensure that the observed monthly trends are not due to biases in the wind fields or seasonality of the fossil fuel emissions.

The increased terrestrial carbon uptake diagnosed in summer (fig.5a) is consistent with the occurrence of positive trends of vegetation activity in the northern hemisphere (high latitudes) and likely to be linked to the negative correlation that we detect between time of onset and mean annual temperatures, implying earlier springs. It is also consistent with analyses of the long-term trends in seasonality in long-term monitoring stations (Ralph Keeling, pers. comm.)

Since we applied the same filtering technique on the modelled and observed concentration data sampled with the same time resolution, any biases inferred during the extraction process of the signal would affect observed and modelled signals in the same manner, such that the metric is robust against these potential biases. However, the observational trait extracted is also returned, and hence it is robust, when a different type of filtering that the one used in the ms, is performed.

In the manuscript at the section 2.3 first lines, we make clearer that the outcome of the filtering is a seasonal component up to the annual frequency and an interannual signal, including all the frequency lower than the annual.

(2) Page 16096, line (9). ‘: : :.in both standard and modeled fluxes: : :’. What are the standard fluxes?

We refer to the standard fluxes as terrestrial net carbon exchanges inferred from atmospheric concentration data using the TM3 transport model (Roedenbeck et al.2005)

(3) Page 16097, lines (4-6). ‘A direct comparison of : : :’. I would highly disagree with this sentence. We have now observations, including satellite-derived products (e.g. MODIS GPP products, or GPP from upscaled FluxNet which could be used to evaluate magnitudes of simulated GPP. In fact, the absolute magnitude is as important as the phase to get the carbon sinks right (which is also discussed later in the paper when modeled and observed atmospheric CO2 are compared). I suggest changing this sentence.

In the manuscript we are referring specifically to remote sensing data, hence satellite-based vegetation indexes (i.e. NDVI,EVI), FAPAR or LAI retrieved by satellite data.

The beginning of the section 2.3.2 of the revised ms will be rewritten according to what hereinafter reported:

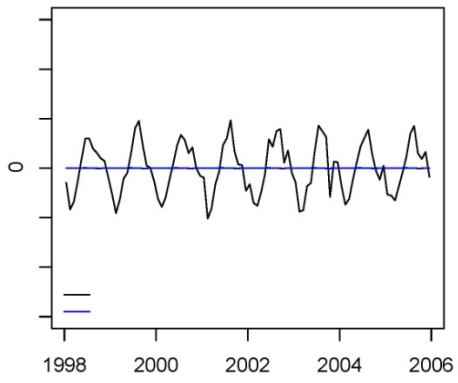
Comparing the FAPAR data or NDVI in absolute terms is not a suitable strategy due to a series of general issues including different retrieval and post-processing algorithms used to compute the final estimated FAPAR/NDVI in different satellites products, to remove i.e. cloud contamination and atmospheric corruption etc. In addition the radiances recorded by satellites differ from the way to compute radiation extinction at the land surface in land surface models. This difference does *a priori* not allow for a perfect match between observations and model.

In a recent paper published in Journal of climate, Dahlke, C., Loew, A., & Reick, C. H. (2013). Robust identification of global greening phase patterns from remote sensing vegetation products. Journal of Climate, 25, 8289-8307. doi:10.1175/JCLI-D-11-00319.1, they clearly show the emerging absolute differences. And this further supports the author's statement in the manuscript. All these issues influence hence the selection of the traits and how information can be extracted from the 'observed' temporal series in the most robust way.

With regards to the GPP-upscaled dataset, this is a model-mediate. In our work we used, where possible, direct observations or dataset that were as much as possible not mediated by models, as support of a robust-approach to the model evaluation. In general, the outcome of an upscaled product is dependent on the statistical algorithm or model used to upscale the site level information, and its representativeness depends also by the representativeness of the site level information (i.e. observation could mainly cover a specific biome but lack in others biomes as well important). Despite the used of the GPP-upscaled dataset has been shown to be usefull for the improvement of a land surface model (Bonan et al.), we did not use this dataset for a quantitative comparison.

(4) Page 16097, lines (10-26). How did the authors calculate the 't-onset' dates from the fAPAR records? The explanation given here is entirely non-transparent.

The sentence will be rewritten in the new version of the manuscript in section 2.3.2. The proxy of the time of onset of vegetation activity (fAPAR in our study) is calculated on the seasonal signal, that is zero centered after the low frequency informations have been removed from the original signal (according to the filtering method).The time of onset of vegetation activity is then the point in time of the upward zero-crossing point of the seasonal signal. See figure below as exemplar. A similar figure will be included in the Appendix.



(5) Page 16099, lines (17-18). Why didn't the authors also exclude tropical Africa? Are there viewer clouds over this portion of the tropics?

Thanks for pointing this inconsistency in the text out: grid cells with dominance of tropical forests in Africa were also excluded by the analysis. We clarify this in the new manuscript.

(6) Page 16103, lines (12-13). The text here states that data for Pt. Barrow are shown in the corresponding Fig. 5, but in Fig. 5 caption it seems it is the Alert observingstation. For this comparison shown in Fig. 5, see also my minor comment (1).

We correct the name in the manuscript. It is the station of Alert, not Barrow.

(7) Page 16104, lines (1-2). Not sure about the logic here. Since the mean seasonal cycle comparison showed good performance of JSBACH, the mismatch in Fig. 5b may not be as much due to an 'asynchrony of photosynthesis and respiration' but more related to divergence in observed and modeled climate sensitivities of photosynthesis and respiration. Here, I may also add that reproducing interannual or longer-term variability is a much stringer test than corresponding comparisons at seasonal time scales.

We agree with the reviewer that the mismatch in fig.5b might be due to divergence in observed and modeled climate sensitivities of photosynthesis and respiration, this is actually what we implied. We will re-write the sentence.

(8) Page 16106, lines (11-14). In relations between spring phenology and land sur-face arming, why did the authors use annual temperatures? I would expect spring temperatures, or at least cold season temperatures, would be more suitable here.

The annual mean temperature was used because, at that spatial scale of analysis, was the variable providing the most robust information in terms of detection of significant correlation (but we tested several explaining variables, data not shown). The problems starts with defining "spring" temperatures consistently for boreal and temperate ecosystems in the northern and southern hemisphere. A further justification of the simplification of using annual mean temperatures is the fact that at coarse resolution there are several vegetation types within the same grid cell, with likely different temperature sensitivity. Hence any detailed analysis considering i.e. winter mean temperature, growing degree days (GDD) etc, are more suitable for high spatial/temporal resolution studies. Obviously, this correlations is only an emerging empirical relationship, and we do not imply causality.

REVIEWER #2

I thank the reviewer for his critical and useful comments and for checking cautiously manuscript typing/figures and tables. It follows the reply to the list of reviewer's comments (in bold).

The only major concern I have is that I miss a discussion of the approach's limitations to some extent. Often, climate and biosphere traits are only weakly correlated, which has implications for the interpretation of results.

We agree that the weak correlation requires careful interpretation, and this is what we have provided in the discussion section of the manuscript. Importantly, weak correlations do not mean insignificant correlations. We tested the significance of statistical relationship and took in to account number of independent informations. When comparing trends, correlations, and covariations (with or without link to climate variability), we underline the tendency of the system to respond in a specific way to external forcing/climate, or to respond instantaneously or with some lag. The metric selection works in this direction: they are sensitive to difference in sign and difference of phase detected between data and model, but do not rely on the slope of the regression, since this is less likely to be robust and well-determined by the data at hand. We will reemphasize in the discussion, section 3.4 and 3.5,that the qualitative trend information is the information that we are looking, and that this is more robust that any direct numerical comparison of the observed trends.

Similarly, uncertainties from other sources are often listed but could be better addressed in the interpretation of results; the uncertainties from the transport model are explicitly discussed at appropriate places, but ocean uptake, land-use emissions, satellite data interpretation could be extended.

Land-use emissions: is a model output, not an external dataset. We added a reference to the manuscript, in which the method and its implications is described (Reick et al. in review: Reick, C., Raddatz, T., Brovkin, V., & Gayler, V. (submitted). The representation of natural and anthropogenic land cover change in MPI-ESM, *Journal of Advances of Modelling Earth Systems*, *4*).

Satellite data: In the revised manuscript, we will add a keyreference (Dahlke, C., Loew, A., & Reick, C. H. (2013). Robust identification of global greening phase patterns from remote sensing vegetation products. Journal of Climate, 25, 8289-8307. doi:10.1175/JCLI-D-11-00319.1) in the methodological section, in which we already discuss how we derived information from satellite data in forms of traits that were robust toward the use of different satellite dataset. In particular traits that were not dependent on the absolute value of the satellite-based variable were used, but based on information on time, correlations, and sign of the changes.

Ocean uptake: the uncertainties of the estimate ocean fluxes(and other observational datasets) were not included in the metrics as previously stated. However we used one of the best available ocean C-fluxes product (based on Takahashi dataset as priori information and involving several biogeophysical ocean models). They showed (Mikaloff et al.2007 in the ms) how the inverted fluxes of natural CO2 sink/sources are robust (limited sensitivity to the ocean model used) and not limited by the few observational coverage (that is not true instead for the inverted fluxes on land). And they provide an upper estimate of 0.25 Pg C y-1 as uncertainties on the global anthropogenic CO2 uptake by ocean (Mikaloff et al.2006 in the ms). We have quantified the contribution of the ocean signal to the metrices and found that most metrices are not strongly influenced by the ocean signal, hence our conclusion appear robust against uncertainties in these fluxes.

The identification of suggested mechanisms in JSBACH responsible for analyzed discrepancies between observations and model is not always clear and could be explained a bit more.
The main focus of the paper is the development of a benchmarking system, not the evaluation of JSBACH, therefore - and to keep the manuscript at a reasonable length - we did not provide detailed explanations of the model deficiencies. However, in the revised version, we will rework the text to explain the most pertinent model-data mismatches better.

The presented evaluation scheme should be discussed more in the context of existing evaluation schemes (site-level dynamics for water and carbon, global pools and dynamics, space for time etc.).

A list of points in the manuscript has already underlined how this study differs/provides a novel contribution to what has been already done

I)the complementary use of both vegetation activity related dataset and atmospheric CO2
II)the use of bounded metrics, using the lower benchmark to scale and that avoid misleading scores and it makes the quantitative assessment less dependent by the choice of the mathematical formulation of the metric;
III) extensive use of variability in addition to multi-annual mean properties, which makes the metrics more related to the system dynamics.

In the revised manuscript, we will strengthen this novelty by adding a comparison to the existing global benchmarking schemes. To avoid an unnecessarily long manuscript, we will not discuss site-level evaluation schemes in detail, as this is not the focus of our study.

Almost all figures could be improved with respect to explanations in text, legends and captions.

This will be taken in to account in the new version of the manuscript.

SPECIFIC COMMENTS TO THE MANUSCRIPT

*** title: the paper is not on constraints on current trends but on evaluation of current projections, I'd suggest to title the paper "Atmospheric CO2 and satellite-based vegetation activity observations in evaluations of carbon cycle projections"**

We agree and propose a revised title: Towards a more objective evaluation of modelled land carbon trends using atmospheric CO2 and satellite-based vegetation activity observations

*** 16088/20 (and elsewhere): model failure is a strong word: I'd suggest "model deficiencies"**
Thanks, this will be changed in the new version.

*** 16091/23ff: I don't know too much about the details of TMs but I would imagine that uncertainties of reanalysis wind fields is higher in remote areas?**

There surely is a trade-off between the accuracy of wind-fields and surface emissions in remote stations. An assessment of the impact of modeled concentrations using TM3 suggests that the uncertainty in remote stations is lower than in continental areas, supporting our choice of remote stations as more robust (MPI Biogeochemistry , technical reports 5-6: http://www.bgc-jena.mpg.de/bgc-systems/pmwiki2/pmwiki.php/Publications/TechnicalReports).

*** 16091/26-16092/8: This needs a bit more explanation. After reading it several times, some understanding dawns but it remains unclear how the robustness is assessed here.**
The sentence has been rewritten in the new version of the manuscript in order to be less misleading.

*** 16092/13: what is the "collection g"? Can't find it in Tucker or Beck**
"g" is part of an internal versioning system that we are acquainted for (from documentation) and which may be useful to mention for the GIMMS user community. We have added the documentation as additional reference for this data set.

*** 16093/21: how was the aggregation performed?**

We used conservative regridding. This information will be added to the revised manuscript.

*** 160595/3-4: It would help my flow of reading if you'd specify the actual years used: 80s (1982-1991), 90s (1992-1997), 2000s (1998-2006)**
This will be changed in the new version.

*** 16095/6-7: confusing: "by mean of normalized"; rephrase sentence.**
'The systematic quantitative assessment of the correspondence of anomalies and trends in simulated vegetation activity and net C exchange is performed using normalized metrics". The sentence has been rewritten in the new version of the ms.

*** 16096/10: This trait checks for large regional inconsistencies between: : :?**
The sentence has been rewritten in the new version of the manuscript. Specifically the trait as reported in 16096/10 checks for the inconsistency between regional seasonal carbon fluxes returned from inverted fluxes and from the model.

*** 16096/22: I'd suspect that the original atmospheric CO2 time series has multiple zero-crossing dates: how was the data aggregated (running mean, trend?) to identify a single zero-crossing date? BTW: Section 2.1.1 does not specify the temporal resolution of the CO2 data.**

The temporal resolution of CO2 data is the original resolution as recorded at the monitoring station (hourly to daily/weekly). Data have been filtered according to Thoning et al.89: the seasonal signal is the signal that retains high frequency information only (>= than the annual frequency) and cut very high frequency information (daily-10days).The signal is hence smooth and multiple zero-crossing dates do not occur. For each year the algorithm looks for the downward zero-crossing point of the seasonal time-series. This zero crossing point leads the minimum of the signal for the specific year. This information are included in the new version of the manuscript.

*** 16097/21: what is the zero-crossing point in fAPAR data which I would expect to be always positive?**
The proxy of the time of onset of vegetation activity (fAPAR in our study) is calculated on the seasonal signal, that is zero centered after the low frequency informations have been removed from the original signal (according to the filtering method).The time of onset of vegetation activity is then the point in time of the upward zero-crossing point of the seasonal signal. (see figure reported above).

*** 16097/22: I'm very much in favor of evaluating correlations between climate and vegetation signals, as climate in models can be substantially skewed as you write earlier. Why do you here use linear differences between observed and modeled months of t-onset/t-max, even though these should directly reflect biases in driving climate data seasonalities?**

We apply the principle of ocam's razor in defining the metric, and there seems to be no justification for a non-linear metric. We did not find a good predictor of t-onset/t-max that could be used as a climate-bias independent metric..The differences between the most frequent month of t-onset/t-max are an indication of the differences in the mean field of the variable of interest.In our study, in which the benchmarking has been applied to the model run in offline mode (but see line 16110/11), the detected differences might reflect i.e. different response of leaves development to temperature/soil hydrology processes. Identifying, the additional model-data difference attributable to climate biases in coupled climate-land studies is nonetheless possible by comparing the scores or an uncoupled and coupled simulation.

*** 16102/18: wouldn't it rather be "the east and south of the North America Temperate region"?**
Yes,the sentence has been rewritten.

*** 16102/9-14: I feel that the "data not shown" would be better suited to support the claimed overestimation of the amplitude than the latitudinal gradient shown in Fig 3b. If not, this requires more explanation.**
The sentence has been rewritten in the new version of the manuscript. Fig 3b represents the amplitude of the mean seasonal cycle with the latitude; hence the figure 3b suits the statement about the 'overestimation of the amplitude'

*** 16102/26 "owing to" instead of "responsible to" or do you mean "responsible for"?**
"Responsible for"

*** 16103/10: how does this conclusion refer to selecting only areas with one vegetative season (16097/10)?**
The sentence has been rewritten in the new version of the manuscript. The statement refers principally not to tropical evergreen forests, but to area dominated by deciduous broad leaved forests in the tropical latitudinal band. The occurrence of one growing seasons as expression of leaves phenology is driven by seasonality in climate.

*** 16106/1: instead of "results not shown" the previous studies could be cited I guess?**

The paper did not pass peer-review yet. The sentence will be removed in the revised version.

*** 16107/4: NDVI in regions with: : : is mainly driven: : :**
The sentence will be rewritten according to suggestion.

<u>COMMENTS TO FIGURES AND TABLES</u>

*** Figure 2: LTT in right-hand panel should be names V-LTT as in Table 3 to avoid Confusion**

This will be changed in the new version.

*** Figure 2: what is the meaning of blue/red/black traits?**

In the new version the traits label have the same colour.

*** Figure 3a: what is the meaning of the color coding?**

The colour coding indicates the latitudinal band. In the new version this will be written explicitly.

*** Figure A2: what is the horizontal line?**

Is the mean of the scores reported in the same figure. In the new version this will be written explicitly.

*** Figure 8 seems to use at least 2 different tones of blue – does it have any meaning?**

In the new version of the manuscript only one tone of blue/red will be reported.

*** Figure 7/8: it seems the areas masked out differ (Greenland) – and why would northern Greenland not be masked out?**

In the new version of the manuscript we correct the masks. In Greenland are already missing satellite data of FAPAR in most of the region. The grid cells that we additional masked out, are grid cells where, despite some vegetation is growing, the area is dominated by ice.

Technical corrections as suggested by reviewer will be included.