Reviewer 1 suggests "the paper could be elaborated a bit more in the introduction in order to better discuss related papers and embed the present study in the context of model evaluation activities"

There have not been many studies on benchmarking vegetation models, and we cite those that have been carried out in the introduction. However, systematic benchmarking of climate models has been undertaken for much longer (Randall et al. 2007) and provides a more general context for our work, which should be acknowledged. We will include text describing the benchmarking in climate models as a new paragraph in the introduction and expand our description of those papers that do attempt to benchmark vegetation models or advocate approaches to such benchmarking.

Reviewer 1 suggests the addition of a new criterion in the list of principles: "Any data set used in a transparent benchmarking should be free to the scientific community. I would suggest to only use benchmarking data set that are usable by other modelling groups - otherwise, different benchmarking exercises cannot be compared on the same grounds."

We agree wholeheartedly with this point. All the datasets used in this study are freely available. We also make all of them available for download from the website given in the Acknowledgments so that other modeling groups can conduct benchmarking exercises without having to assemble the data sets from scratch. We will add the requirement that all target data sets must be freely and completely available as a new point in the list of principles, and emphasise that any future extension or improvement of the proposed benchmarking system should only be based on datasets freely available to the scientific community.

Reviewer 1. The comparisons of the seasonality (2.3.3) looks a bit complicated to me. Kobayashi & Salam (2000); van Oijen et al. (2011) have shown that the MSE (the squared part of the RMSE) can be decomposed to three elements:

$$MSE = (\langle X_i \rangle - \langle X_{i'} \rangle)^2 + (\sigma_i - \sigma_{i'})^2 + 2(\sigma_i \sigma_{i'})(1 - r)$$

Bias^2 + Variance diff^2 + Phase error^2

The meaning of the squared data bias is obvious, the second term indicates differences in the fast variability, and the lack of correlation *r* between X_i and $X_{i'}$ is a very simple estimator for phase errors. Wouldn't this last term do the job in this benchmarking exercise?

This is an interesting breakdown of MSE, which is actually an alternative way of expressing our three stage comparisons of mean variables rather than the seasonality measures we use. In our comparisons, stage 1 is the same as a combination of the squared bias, variance difference & phase error; stage 2 is squared variance difference and phase error; and stage 3 is the squared phase error by itself. Our approach can be used with normalized metrics, whereas MSE is not normalized – normalized values have the advantage that

they can be compared across models and processes, which is why we adopt them in this study.

Reviewer 3 suggests “adding more comprehensive review for available metrics for benchmarks (other than the metric used here, e.g. NME, MPD, MM), and its pros and cons will improve the manuscript for the development of comprehensive benchmarking system”.

A comprehensive review of potential benchmark metrics would considerably increase the length of the current paper, and furthermore a more comprehensive evaluation of different sorts of metrics is currently being prepared by members of the GREENCYCLES project. However, we agree with the reviewer that it would be useful to explain more clearly why we have chosen the specific metrics we use out of the range of metrics available. Metrics generally fall into three categories: non-normalised metrics (e.g. RMSE), metrics normalized by observational uncertainty, and metrics normalized by observational variance (which are the metrics used in our paper). We plan to explain these differences in the section describing the metrics and to make it clear that our selection of metrics normalized by observational variance reflects the need to compare performance between processes (which rules out non-normalized metrics) and the lack of information about observational uncertainty for many global data sets (which rules out metrics normalized by observational uncertainty.

Reviewer 3 suggests “it will be helpful if authors discussed the priority of required benchmark dataset to improve future DGVMs development at the end of the manuscript in relation to the Luo et al. (2012) although it may requires the subjective decisions”.

In our discussion, we emphasise (a) that it would be useful to include either time-varying or additional snapshots for vegetation cover and height. (pg 31 line 22) and (b) that the release of the primary data from a more comprehensive set of the FLUXNET sites (pg 32 line 16) would both be highly desirable and considerably improve vegetation-model benchmarking. These would certainly be the priorities that we recommend. Beyond this, we have little to add beyond what has already been discussed by Luo et al. (2012), except the requirement (also raised as an issue by **Reviewer 1**) that all benchmarking data sets should be freely available to the research community. We will add this criterion to our section about the principles of dataset selection and we will emphasize this important point in our final discussions.

Reviewer 3 has made a number of minor comments that help to clarify points in the text, and we are grateful for these and will deal with them in the revised text as follows:

1) page 15729, line 21: Could you add SeaWifs dataset website?
The dataset is available to download from http://bio.mq.edu.au/bcd/benchmarks/ as stated in the Acknowledgments, but we will add the SeaWifs dataset webpage (http://oceancolor.gsfc.nasa.gov/SeaWiFS/) reference in the text.

2) page 15731, line 1: What is the definition of the ’disturbed’ sites? The data has ’Managed’ category in Management variable, but there is no single ’disturbed’ category in that.
The original classification indicates sites that are managed, recently burnt, recently cut clear, fertilized and irrigated. We used the shorthand “managed or disturbed to cover all of these sites, which were excluded from our analyses. We will clarify this in the text.

3) page 15735, line 7; page 15735, line 16; page 15737, line 2: Is it ambiguous to refer just Table 2 for each application. It will help if you could add reference numbers or categories in the column 4 of the Table 2.

Table 2 provides the equation of each of the metrics, the value yielded for perfect agreement/disagreement, and lists the comparisons for which we use the metric in this paper. It is perhaps unclear from the text why we are referring to the table when each metric is introduced, and the column headings are also a little opaque. We will clarify the purpose of this table in the text, and change the table headings so that these are clearer.

4) page 15737, line 11: Table 2 show 1 for perfect disagreement. But in the text, it shows 2 for complete disagreement. Could you explain the consistency about this?

Thank you for spotting this inconsistency. The correct value was given in the text. The value in the table was incorrect and will be changed.

5) page 15739, line 5: Please provide LPJ version number used in this study

We used LPJ v1.2, but with a common driver (for I/O) for LPX and LPJ. We will clarify this in the text.

6) page 15742, line 24: Is underestimate of NPP for the variance? If so, please clarify this.

Thanks for spotting this confusing statement. The NPP is over-estimated, but the NPP variance is underestimated. We will clarify this in the text.

7) page 15742, line 26: I couldn’t understand where the values 1.26-0.56 are come from.

We inadvertently gave the scores obtained when comparing to only the Luyssaerts et al NPP sites. The correct values should be 0.86 – 0.50, equivalent to a 42% improvement in performance. We will put the correct values in the text, although the point that the improvement in SDVM performance is much greater than the improvement in DGVM performance is still valid.

8) page 15772, table 5: It seems first appearance of "Annual average 1998-2005" should be "Annual average 1950-2005".

Thanks for spotting this mistake. We will correct the table.

9) page 15747, line 3-4: Table 6 shows LPJ has marginally better performance with Luyssaert et al. (2007) dataset at the site locations, however, the text says all three models performs better against the Beer et al. (2010).

LPJ performs very marginally better against the Luyssart et al. (2007) data set than the Beer et al. (2010) data set, but the change is not significant. Both LPX and SDBM perform much better against Beer et al (2010) than at the Luyssaert sites. Thus, we think that our point is valid, and that access to the wider set of site data used for Beer et al. (2010) would improve benchmarking. We will however, correct our assertion to make it clear that only two of the models show a pronounced improvement and that the results with LPJ are similar for both data sets.

10) page 15752, line 18-19: Is this related to the 'at sites' score comparison between Luyssaert et al. (2007) and Beer et al. 2010 in Table3? If so, please add reference to that in the text for clarification.

It is in relation to Luyssaert et al (2007) "at sites" score comparions, and we will clarify this in the text.