Anonymous Referee #1
Received and published: 14 March 2012

This paper is the result of a discussion at a workshop organised by the International Land Model Benchmarking project. It presents the general concept of benchmarking and discusses how the community might go about benchmarking land models. Somewhat disappointingly, it does not go any further and actually make any recommendations about specific benchmarks to be used. For someone already working in this area, the paper therefore does not present anything new. Nonetheless, the paper still is useful as an introduction to benchmarking land models. I gave it to several of my PhD students to read and they found it helpful. However, we had a number of comments that would help to improve the paper.

Thanks the reviewer for her/his recognition that this paper is a useful introduction to benchmark analysis. We also agree with the reviewer that our original manuscript did not make specific recommendations about benchmarks. In this revision, we are more critical to evaluate knowledge gaps and highlight the future needs to define benchmarks. We particularly added "Section 4.1 Criteria of benchmarks" to propose what would make benchmarks.

Since benchmark analysis is still at the infant stage of research, we do not have exact answers for many of the issues. This revision has tried to address many of them as much as we can. We hope the revised manuscript will be useful not only for students and post-docs to learn about benchmark analysis but also practitioners to understand the knowledge gaps and research challenges.

The chief comment was that the paper is too vague. It promises to define what is meant by benchmarking, but gives only a very general definition of the concept, not a specific definition in the context of land models. In particular, the paper does not explain the differences between benchmarking, model evaluation and model validation. This distinction needs to be much more clearly drawn. Many of the things that are presented as benchmarks in this paper are not, in my view, actually benchmarks, but rather examples of model comparison against data, or data sets. A benchmark is not just a data set against which a model is compared; it is a specific measure of comparison against that data set, and preferably one that takes into account the information content of the data set (Abramowitz 2005). Tables 2 – 4 claim to give "sample benchmarks" but the things listed in these tables are not benchmarks – they are data sets which could be used to derive benchmarks.

The reviewer comment is very useful for us. We added "Section 4.1 Criteria of benchmarks" to propose qualifications of available data to become benchmarks. The first two sentences in the section read, "What would be qualified to be benchmarks have not been carefully discussed in the research community although several studies have evaluated performances of land models against available data. In general a set of benchmarks has to be objective, effective, reliable, and relatively simple." The paragraph explains those criteria.

We agree that some of the published studies on benchmark analysis are more similar to model evaluation and hope this paper will stimulate further discussion on what constitute a benchmark and benchmark analysis.

In addition, Sentences in lines 80-88 briefly describes model evaluation and model validation. With an added section "Section 4.1 Criteria of benchmarks", we hope to distinguish benchmark analysis from model evaluation, model validation and model comparison against data or data sets.

Given that the target audience is non-practitioners, some more detailed examples of how to carry out a benchmarking analysis would also be appropriate. Examples are mentioned in passing but not explained in detail. For example, Figure 2 is presented as an example of a benchmark analysis but it is unclear what makes this a benchmarking analysis rather than a model evaluation, nor are the steps involved in this benchmarking exercise explained. One of the key points missing in this example is how the models should be evaluated against the benchmark – the caption says "A well functioning model has to match the observations" – but neither model appears to do so. What are we to make of this? How are these discrepancies handled in the benchmarking framework?

As we explained above, we have revised the manuscript to identify major challenges in future benchmark analysis. Hopefully, our paper will be useful not only for students and post-docs to learn about benchmark analysis but also practitioners to understand the knowledge gaps and research challenges. The main objective of this paper is to propose a framework for benchmarking land models. Detailed examples of how to carry out a benchmark analysis are beyond the scope of this paper.

We agree with the reviewer that some of the published studies on benchmark analysis are more similar to model evaluation. Benchmark analysis is supposed to identify strengths and weaknesses of models. When models do not match observations (or benchmarks), the benchmark analysis hopefully provides clues to improve the model(s).

It would also be appropriate to include some material about how we would move towards a standardised set of benchmarks. This paper suggests a number of things that would be good to have in benchmarks, but leaves rather open the question of how we will actually move towards developing a community-wide benchmarking system. Is there a roadmap for achieving this?

We added "Section 4.1 Criteria of benchmarks" toward standardised set of benchmarks. It appears that the research community should develop widely acceptable benchmarks first for a community-wide benchmarking system.

I disagree with the discussion about benchmarking against FACE data. The paper says "The LPJ model matched the NPP response to elevated CO2 observed in four FACE experiment in temperate forests, which provided more confidence in predictions of response in other biomes". This study did not provide more confidence in predictions of response in other biomes. Hickler et al themselves commented that more data are needed to test the modelled responses in boreal and tropical biomes, which were predicted to differ from responses in temperate biomes. An important point that needs to be stressed, is that benchmarking is a way of highlighting where models are going wrong, but it does not show that they are right.

The point is well taken. We modified the paragraph to acknowledge the point.

Finally, the title should be "A framework for benchmarking land models" not a framework of.

Changed as suggested