Anonymous Referee #3
Received and published: 4 April 2012

The review paper by Luo et al. attempts to describe a procedure for benchmarking land models, as a summary of discussions at a workshop of the "International Land Model Benchmarking project". Unfortunately, the manuscript seems to be merely a rather unbounded collection of ideas and examples of what could be part of such a benchmarking effort. It describes four major steps, but it fails to define clearly what these steps should contain for setting up such a benchmarking procedure. A review paper on benchmarking procedures for land surface schemes (or more generally land models) would be a valuable contribution to the literature, but in its current form, I cannot recommend this manuscript for publication in Biogeosciences.

Thanks for the comment. This revision took a more critical approach to benchmarking land models. We have made revision on many parts of the manuscript to critically evaluate benchmarking procedure. In particular, we added one section in Section 4 Benchmarks as defined references (i.e., Section 4.1 Criteria of benchmarks) to define how benchmarks should be set up. Hopefully our substantial revision is satisfactory to the reviewer.

Major remarks

- The term "land models" is very general, and it should be pointed out much earlier that "land models" can have greatly different functions (as stand-alone models or as part of a larger model framework). A short comment on this appears only in section 5 (p. 1914, l. 14).

The first introduction paragraph has described that land models can be embedded in earth system models.

- One of the main shortcomings of the examples and list of sample data sets that can be used (Tables 2-4) is that the selection seems to be based mainly on data availability, and not on data suitability. Not all processes in land surface-atmosphere exchange are equally straight-forward to compare with observations, and the lists given are mixtures of state variables and rate variables, and of prognostic and diagnostic variables. If this paper wants to contribute to the setup of a benchmarking procedure, it should describe more clearly on what basis benchmarking data sets are to be selected (or on which basis the example data sets were selected).

The reviewer made an important point. Not all available data would become good benchmarks. To distinguish data availability from suitability for benchmarking land models, we first change the table captions for Tables 2-4 to “Candidate Benchmarks …”. We also added Section 4.1 Criteria of benchmarks to discuss the issue on what basis benchmarking data sets are to be selected.

- Section 5 discusses the weighting of individual data comparisons in the benchmarking data set, but it fails to define how these weights should be established. A number of different weighting measures are discussed (p. 1915), but there is no discussion of their suitability for a possible benchmarking procedure, and it remains open which considerations should lead to a selection of these.

We agree with the reviewer that it remains an open question which weights should be used in benchmarking land models. We hope the knowledge learned by the data assimilation community would be useful to for benchmark analysis.

- Because of the exploratory character of the study, the examples provided in the text and figures are of great importance. It is, however, not clear whether these should be treated as a careful selection of measures and comparisons, or whether they are largely illustrative. The objective with these should be phrased more clearly. In any case, the examples should be either generalized, or explained in greater detail. E.g., in Fig. 2 it is unclear what the 6 figures (and their abbreviations MBC, ICE, etc.) are illustrating, and in Fig. 5 it is unclear which observations (parameter, data set, etc.) is used to derive the model skill from.

Figures are used to illustrate some points in benchmark analysis. We have explained them in more detail.

- A few issues that are not highlighted in this study, or not enough:(1) How to treat uncertainty in observations/observation-based estimates when using these data for benchmarking? (2) The problem with feedbacks is touched upon a few times, but not discussed in great detail. The outcome of land models within a modelling framework (e.g. as used for the CMIP5 simulations mentioned in the text) depend not only on land model parameterizations and descriptions, but also on (biases in) the input data to these models. (3) Are correlations between the parameters that are part of a larger benchmarking data set an issue? How should these correlations be accounted for? Minor remarks

We agree with the reviewer that some of the issues may not be fully explained, largely due to limited space in this paper. Nevertheless, lines 426-427 did discuss uncertainties in data and their influences on model performance measures. Feedback was explored in one paragraph in Section 3 (lines 236-244). All of them are great topics, which should be extensively explored in the future.

- p. 1907, l. 12: "radiation fluxes" should be replaced by "energy fluxes".

Changed as suggested.

- I was not familiar with the term "nitrogen capital" used (p. 1908, l. 19, 24). The authors may want to consider replacing it with "nitrogen availability".

Nitrogen capital is the total ecosystem nitrogen content. We replaced the term to avoid confusion.