Biogeosciences Discuss., 10, C5738–C5744, 2013
www.biogeosciences-discuss.net/10/C5738/2013/

Biogeosciences
Discussions
Open Access

Interactive comment on "Evaluation of biospheric components in Earth system models using modern and palaeo observations: the state-of-the-art" *by* A. M. Foley et al.

A. M. Foley et al.

amf62@cam.ac.uk

Received and published: 6 October 2013

We would like to thank the reviewers for their constructive comments on this paper. Our point-by-point responses to the specific comments of the Anonymous Referees 1 and 2 are given here. Both referees also made several minor editorial comments, which have been remedied.

Referee comments are in italics. Our responses are in plain text. Text that has been added to the manuscript to address a referee comment is in bold.

General comments of Referee 1

Section 2 and 3 are well written and will serve as a comprehensive review for model benchmarking. One problem I noticed in the section of palaeo data use is that Harrison et al. submitted is cited many times there, but details of their work is written ambiguously, so please update the description about their study if it is published.

Several references that can be updated were brought to our attention, and have been updated. In particular, the Harrison et al. paper is now available and so can be properly cited as Harrison et al. (2013): Harrison, S.P., Bartlein, P.J., Brewer, S., Prentice, I.C., Boyd, M., Hessler, I., Holmgren, K., Izumi, K., and Willis, K. (2013) Model benchmarking with glacial and mid-Holocene climates. Climate Dynamics DOI 1007/s00382-013-1922-6. The description of this work has also been updated as follows:

Page 10949, line 27. (see Harrison et al. (2013) in which mid-Holocene and LGM simulations from the CMIP5 archive, and from the second phase of the Palaeoclimate Modelling Intercomparison Project (PMIP2), were evaluated against observational benchmarks, using goodness-of-fit and bias metrics).

I felt section 4 like a book chapter, and the structure is not very consistent (i.e. subsections doesn't seem to itemize the equal level topics). I would remove section heading 4.1 and just start the paragraph after the heading of section 4; and shift 4.2, 4.3, and 4.4 to 4.1 and 4.2, 4.3. Also "Emergent constraints" subsubsection could be a separate subsection from "Recommendations for model evaluation methodologies" because that topic seems rather independent from a narrow sensed "model evaluation methodologies" written in the subsection (like Fig 4.). This may also apply to the Table 5. Making them consistent with the structure drawn in Fig 4. would help readers to understand the section easier.

The heading structure of section 4 has been revised to indicate more clearly how each point relates to model evaluation. The opening of the section on emergent constraints has also been revised to integrate it into the overall section and clarify the place of emergent constraints within a model evaluation framework, as follows:

Page 10967, line 10.

4.4 Utilise emergent constraints

Emergent constraints can also provide valuable information for model evaluation, as they convert the extensive short-timescale information available for the contemporary period into longer-timescale constraints on the Earth system sensitivities that are most important for the 21st and 22nd centuries (e.g. climate sensitivity to CO_2, or carbon cycle sensitivity to climate). Observational data on short timescales do not relate directly to these sensitivities, and analogue approaches, which evaluate ESM sensitivity against known changes in the past, are also limited by observational data, as the analogue events in Earth's past are not as well characterised as those in the contemporary period.

Page 10944, line 22: Could you add references of benchmarking ocean model evaluation? All of the literatures cited here seem for the terrestrial model.

The following ocean model evaluation references have been added:

- Najjar, R. G., Jin, X., Louanchi, F., Aumont, O., Caldeira, K., Doney, S. C., Dutay, J.-C., Follows, M., Gruber, N., Joos, F., Lindsay, K., Maier-Reimer, E., Matear, R. J., Matsumoto, K., Monfray, P., Mouchet, A., Orr, J. C., Plattner, G.-K., Sarmiento, J. L., Schlitzer, R., Slater, R. D., Weirig, M.-F., Yamanaka, Y. and Yool, A.: Impact of circulation on export production, dissolved organic matter, and dissolved oxygen in the ocean: Results from Phase II of the Ocean Carbon-cycle Model Intercomparison Project (OCMIP-2), Global Biogeochemical Cycles, 21(3), n/a–n/a, doi:10.1029/2006GB002857, 2007.

- Friedrichs, M. a. M., Carr, M.-E., Barber, R. T., Scardi, M., Antoine, D., Armstrong, R. a., Asanuma, I., Behrenfeld, M. J., Buitenhuis, E. T., Chai, F., Christian, J. R., Ciotti, A. M., Doney, S. C., Dowell, M., Dunne, J., Gentili, B., Gregg, W., Hoepffner, N., Ishizaka, J., Kameda, T., Lima, I., Marra,

J., Mélin, F., Moore, J. K., Morel, A., O'Malley, R. T., O'Reilly, J., Saba, V. S., Schmeltz, M., Smyth, T. J., Tjiputra, J., Waters, K., Westberry, T. K. and Winguth, A.: Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean, Journal of Marine Systems, 76(1-2), 113–133, doi:10.1016/j.jmarsys.2008.05.010, 2009.

- Stow, C. a., Jolliff, J., McGillicuddy, D. J., Doney, S. C., Allen, J. I., Friedrichs, M. a. M., Rose, K. a. and Wallhead, P.: Skill assessment for coupled biological/physical models of marine systems, Journal of Marine Systems, 76(1-2), 4–15, doi:10.1016/j.jmarsys.2008.03.011, 2009.

- Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., Halloran, P., Heinze, C., Ilyina, T., Séférian, R., Tjiputra, J. and Vichi, M.: Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models, Biogeosciences Discuss., 10(2), 3627–3676, doi:10.5194/bgd-10-3627-2013, 2013.

The following ocean data reference has also been added:

- Oliver, K. I. C., Hoogakker, B. A. A., Crowhurst, S., Henderson, G. M., Rickaby, R. E. M., Edwards, N. R., and Elderfield, H.: A synthesis of marine sediment core δ^{13}C data over the last 150 000 years, Clim. Past, 6, 645-673, doi:10.5194/cp-6-645-2010, 2010.

Oliver et al. (2010) is cited in the text as follows:

Page 10948, line 6. Marine sediment cores have been used extensively to generate sea surface temperature reconstructions (e.g. Marcott et al., 2013), and to reconstruct different past climate and carbon cycle variables (see review in Henderson, 2002) related to ocean conditions. For example, δ^{13}C is used in reconstructions of ocean circulation, marine productivity, and biosphere carbon

storage (Oliver et al., 2010). However, the interpretation of these data is often not straightforward, since the measured indicators are frequently influenced by more than one climatic variable (e.g., the benthic δ^{18}O measured in foraminiferal shells contains information on both global sea level and deep water temperature).

Page 10945, line 9-14: It is not very clear for me why the use of opposite directional use of the same function could verify the comparison.

We agree that the statement about the opposite directional use of radiative transfer functions was unclear, and did not communicate the key point we wish to convey. This point is that while overlap between model and data, such as similarities between functions used in the retrieval of satellite data and those used in a climate model, complicates validation, the additional information may still be useful in validating some part of the model. We have revised the text as follows:

Page 10945, line 9. However, while the functional overlap between model and diagnostic complicates the validation exercise, the additional information provides an opportunity for partial validation. For example, a retrieved product can help to evaluate whether the initial assumptions of the atmospheric or surface state of a model are correct.

Page 10957, line 6: It seems Jakob and Tselioudis 2003 GRL is more appropriate to be cited here rather than Jakob 2003.

At the reviewer's suggestion, we have included a reference to Jakob and Tselioudis, 2003.

Page 10988, Table 2: Could you explain "mixed variables" in (1)? If you mean comparison of continuous and categorical, writing the definition in the parentheses, like "mixed(continuous and categorical) variables" would help readers understand.

By "mixed variables" we mean different variables. An example would be variables with

different units. We have amended the table to communicate this:

Page 10988, Table 2. Implicitly supposes that x and y are comparable, so is not suited to mixed variables (e.g. variables with different units)

General comments of Referee 2

Section 4 appears to contain a collection of disparate subtopics in subsections of various lengths joined together as recommendations. This section could be significantly improved by reorganization.

The heading structure of section 4 has been revised to indicate more clearly how each point relates to model evaluation. The opening of the section on emergent constraints has also been revised to integrate it into the overall section and clarify the place of emergent constraints within a model evaluation framework.

Page 10941, lines 9–10: The multi-model mean should always perform averagely since it is an average. You may wish to more clearly explain what you wish to say about model performance for regional climate.

Schaller et al. (2011) found that the multi-model mean outperforms the ensemble members on which it is based when assessing skill based on climatological means. When skill is assessed based on the ability to reproduce global fields of climate variables, the multi-model mean does not score as highly. They state that "the more grid cells, metrics or variables are aggregated, the better the performance of the MMM becomes". We have revised the text to communicate this concept more clearly as follows:

Page 10941, lines 9–10. However, Schaller et al. (2011) have demonstrated that the multi-model mean outperforms individual models when the ability to reproduce global fields of climate variables is evaluated, but does not consistently outperform the individual models when the ability to simulate regional climatic features is evaluated.

Page 10968, lines 17–21: This sentence suggests that a larger number of model tests

may reduce the ability of an investigator to understand or interpret model results. This makes little sense, but the authors may have a point there that is not conveyed clearly or correctly.

A large number of metrics could be applied to a model and the individual metrics might be each easily interpreted. Yet, a combination of many different metrics would be a challenge to interpret. As discussed in the paper, there is also great subjectivity associated with combining many metrics into a single number. Given that the results of model evaluation studies are often referred to by others within, and beyond, the field of Earth system modelling to gain an insight into model skill, this high level of subjectivity is a challenge. We have modified the section to better communicate this as follows:

Page 10969, lines 15. Benchmarking models against a set of well-chosen observations (Sect. 2), and using appropriate metrics (Sect. 3), should be considered a vital step in model evaluation. While individual metrics might each be easily interpreted, a combination of many different metrics could be a challenge to interpret, particularly when very different scores in metrics that measure different aspects of model performance need to be reconciled. Therefore, while it may be tempting to simply evaluate the performance of the model against every dataset that can be found (and indeed a "perfect" model should be able to withstand such a test), if this comes at the expense of being able to interpret the results then it may be more beneficial to focus on a smaller set of tests which target key model outputs.

Interactive comment on Biogeosciences Discuss., 10, 10937, 2013.