Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-138-RC1, 2016

Geoscientific Model Development Discussions

Interactive comment on "Data-mining analysis of factors affecting the global distribution of soil carbon in observational databases and Earth system models" *by* Shoji Hashimoto et al.

Anonymous Referee #1

Received and published: 3 August 2016

General comments

==========

This manuscript describes an analysis, using machine-learning algorithms, of what factors affect the distribution of soil organic carbon (SOC) in both observational databases (e.g., the Harmonized World Soil Database) and earth system models (ESMs, in particular CMIP5 data). This is an important and interesting topic, as our understanding of the factors governing the spatial distribution of SOC is poor. Data-driven algorithms such as those used here offer the possibility of novel, quantifiable insights into both natural processes and model weaknesses.

This ms is thus promising, but ultimately significantly weakened by a series of problems. First, poor presentation: many parts of the text are unclear; some of the figures need re-thinking; results are at times presented confusingly.

Second, there's no reproducibility, which is shocking (see #6 below). In particular, no code or data availability is specified, no software details given, nor are the methods fully complete or understandable.

Finally, and related to the previous point, there's a lack of insight and applicability. Most of the discussion is a rote recitation of formulaic points; the authors need to expand on the genuinely interesting parts, and offer more interesting, novel insights on how their results apply, and will be useful, to future work. The lack of reproducibility (above) means that it's also not clear how any of this would inform or be useful for modelers seeking to improve their software and science.

In summary, there are interesting points here, but the current ms needs substantial revisions in almost every area for clarity and presentation, reproducibility, and insight.

Specific comments

==========

1. Page 1, line 1: I'd suggest either "Data-mining analysis of the global. . ." or "Factors affecting the global. . ."

2. P. 1, lines 9, 20, and 26-27: these three short sentences could be deleted with no real loss

3. P. 1, l. 25: "elucidate the nature" of the databases? Confusing

4. P. 2, l. 4: what recent study?

5. P. 3, l. 8-9: divided over what spatial scale? Some more detail in this entire paragraph would be useful

6. Methods: need to give version numbers CDO, R, and all packages used. Also, I'm shocked at the complete lack of any mention of data or code availability (no, that one sentence on p. 7 doesn't count). It's 2016, and I expect all code and data (at least that backing the main results) to be included as supplementary info, or posted in a repository. It's not acceptable to produce results from a black box; see also http://www.geoscientific-model-development.net/about/code_and_data_policy.html

7. P. 4, l. 17: "Relationships with a mean annual temperature were relatively close to each other" – what does this mean? Clarify

8. P. 4, l. 34: "The contribution of each variable varied between ESMs" ?

9. P. 4, l. 36: "large inconsistencies. . .demonstrated low contributions" – what?

10. P. 5, l. 23-24: this is an interesting point, and should be expanded upon. What are the implications, if the seemingly wide variety of CMIP5 models in fact uses a much smaller number of fundamental assumptions or modeling approaches? I'm pretty sure that Kathe Todd-Brown made this point in one of her papers; see also Alexander et al. (2015), 10.5194/gmd-8-1221-2015

11. P. 6, l. 13-14: "The use of temperature sensitivity. . ." - ?

12. P. 7, l. 2-3: would such model-data fusion ever be possible, given the extremely long running time of modern ESMs?

13. Table A1: an URL or reference for each model would be useful

14. Table A2: this classification was applied to. . .? Where is it from?

15. Figures 1 and 2: these are so tiny I'm not sure they convey any information, really

16. Figure 6 should be the central, most important figure of the entire paper–showing how variable importance compares between observational databases and ESMs–but it's very difficult to see what's going on. I'd suggest re-thinking this, and carefully considering the most effective way to show this

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016.