Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-76-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "TraceME (v1.0) – An online Traceability analysis system for Model Evaluation on land carbon dynamics" *by* Jian Zhou et al.

Anonymous Referee #2

Received and published: 26 June 2020

In this work, the authors developed the TraceME system, in order to address what they argue are the three core challenges of ESM evaluation: the untraceable of model outputs, the lack of automatic algorithms and the high computational cost. They therefore built a cloud-based evaluation system, which, according to the authors, is traceable, automatic and sharable. The system was built on a previously established collaborative analysis framework of CAFE. I do believe that the traceability framework, which has been continuously developed by a few authors in this study since 2012, is a very useful one to expose model structure differences and errors in simulating land carbon cycle processes. But I am not convinced that substantial advances in terms of scientific model development have been made in this specific work to warrant its publication in

Geoscientific Model Development.

There is large room for improvement toward being more rigorous in writing and better logical flow in the present work. Very often, the authors either laid a too much wide background and then end up with a much narrower implementation, or used a lot vague expressions to justify the added value of their work. Throughout the whole text, a better and more rigorous justification for the novelty and usefulness of TraceME is needed, especially in a sense to the wider modeling community in contrast to those who are interested in traceability framework. Below are some major comments that lead me the above conclusions:

Major Comment #1: Line 23: 'the untraceable model outputs' pre-assumes the readers' knowledge on traceability framework and assumes traceability is foremost important in evaluating ESMs. I am not convinced on this. I believe every modeling group, when looking at their model performance in development cycles, would try to 'trace' the error into its underlying processes and understand the causes. In this sense, there is no model output that is 'untraceable'. The justification for the necessity of TraceME for the wider modeling community, and its usefulness in day-to-day model development has not been demonstrated in the paper.

One core argument for 'automatic' and 'sharable' evaluation platform would be to help identify model errors and improvement directions. If this is only for some key MIPs like CMIP5 or CMIP6, then it seems that analyzing the output on this platform by the authors and making the webpage available for different modeling groups would be sufficient. This would further raise doubts on whether there is value for this work to be published and for the tool to be available for the whole modeling community. There is a lack of evidence in the paper that modeling groups would indeed be interested to visit the platform and use it in their work. In the contrary, the figures contained inside make it more like a normal science paper. If by reading the paper figures, modelers would already have the information needed, I doubt they would visit the platform. Then the 'sharable' key feature would be not that useful either.

Major comment #2: The authors discussed in several places of the Introduction section the mounting challenges of evaluation of ESMs and cited the large volume of data from CMIP projects but ultimately nailed down only to its land component, or more specifically, the land carbon cycle component. In this case, the advantage of traceability seems only valid in evaluation of the land carbon cycle models. This point weakens the importance of their work and leaves the introduction scope of evaluation of ESMs (especially the 1st paragraph there) unmatched to what the authors actually delivered finally. Even for evaluating land carbon cycle models, I think the traceability framework oversimplifies the complexity of the land carbon cycle process. Disturbances, land use change and land management become increasingly important in carbon cycle models, can the traceability framework accommodate the differences in these factor among models? The conclusion in lines 77-78 seem unfair for other evaluation tools because the traceability framework is based completely on the idea of pool size and residence time, and finds its best application in carbon cycle models but not in others. The ESMs evaluation also includes those on hydrology, radiation and land-atmosphere interactions. The authors seemed ignoring these in their traceability framework.

Major comment #3: I downloaded the code provided at the end of the paper. There seems only a few python and R scripts with several hundred lines. There are not any user guides or documentation. No weblink for TraceME was provided in the paper either (I hope I did not miss it though). The modeling community is left only reading the paper and wonder how they can use this tool. This is at odds with what the authors claim that TraceME is 'sharable'.

Major comment #4: For a paper focusing on model development, descriptions on the technical aspects of the development, e.g., on the technical roadmap selection, implementation details, code structure and platform architecture, description of the key but new processes in contrast to previous model versions, usually take an important part in the paper. But the technical description on the TraceME development is rather weak in this paper. The only section on this topic might be Section 2.1. But the description

is vague and general. It is unclear what is the novelty in TraceME compared to CAFE, and which part of work has been done by engineering support and which by the authors, and what is the technological novelty. I cannot believe with the several hundred lines of python and R scripts provided by the authors in the 'Code Availability' section would make such a complex platform as described in the paper.

Major comment #5: Key arguments for TraceME by authors include automatic algorithms, sharable and saving the need to download data. The concept of 'automatic' is vague. For the results presented in the paper, I agree the authors make these figures automatically because the scripts must be extensively tested. But the authors do not show that beside what they have presented, if modeling groups want to use the platform practically, how much flexible and automatic could it be? If indeed it's useful, the data uploading and downloading would be unavoidable.

Minor comments:

There are many places the writing is causal and ambiguous and needs to be improved, to name a few examples:

Line 45-47: some articulations are needed here. Current statements are a little too general. Does 'their' in 46 refer to 'metrics', how can these metrics have 'indirect effects'? What are these 'indirect effects'?

Line 47-48: 'it is not independence among models' => unclear.

Line 49: '80% of the variance' => the variance of what ?

Line 55: dramatically => dramatic

Line 74: land information system => unclear what does this mean.

Line 109: it needs a new platform => a new platform is needed . . .

Line 113: automatic and shareable platform => "an" automatic and shareable platform

Line 189: the externally forces => external forcings ?

Line 190: is always deviate from = > please check the grammar here.

Line 251: that had been submitted results => 'been' should be removed.

Line 685: positive above the soil lines => 'soil' should be 'solid'

Line 594: composed into => decomposed into?

Line 360: needs to some new characteristics => check grammar

Line401-403: I don't see how the citation of Song 2019 fit here. Song et al. is based on site level which is at a completely different scale of what has been presented in the paper.

Line 104-105: the citation of data volume for CMIP5 and CMIP6 has not direct relevance. I guess nobody would download and analyze all the data for all variables. Focusing on several variables would not lead to download more data in CMIP6 than CMIP5 unless spatial resolution dramatically increases.

Line385-388: I understand 'computational efficiency' as how many tasks are done given a unit of computation resource. The author argued that automated computation increase efficiency, but this was not proved in the paper.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-76, 2020.