Responses to Comments from Reviewer #2

General comments:

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.

Response: We highly appreciate the critical comments on this work. After carefully studying the comments from the reviewer, we have substantially revised the manuscript. We have tried our best to revise the manuscript to show the novelty and scientific value of the TraceME platform. We hope the revised manuscript is not only useful for people who are interested in the traceability framework but also helpful for a wider modeling community. Please find our point-by-point replies below.

Below are some major comments that lead me the above conclusions:

Major Comment:

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.*

Response: Thanks for pointing this issue out. We have made a substantial revision to make the traceability clearer before the introduction of TraceME. First, a few sentences have been added in the introduction to define the method of traceability analysis and the reasons for developing it into a platform. Then, a description of the scientific workflow in TraceME is provided in the supplementary materials (end of this text), and more technical descriptions about TraceME have been incorporated. Furthermore, we have demonstrated the necessity and usefulness of TraceME for the modeling community. We showed that TraceME is useful not only for MIPs but also for specific modeling groups. For example, the TraceME has been applied to CLIM5.0 to evaluate the effects of different climate forcings (i.e., CRUNCEP and GSWP) on the simulated land C storage dynamics. As shown by the following figure, there is a 2-fold difference in global C storage capacity in CLM5.0 between the forcings of CRUNCEP and

GSWP. Such difference is jointly contributed from net primary productivity and C residence time.

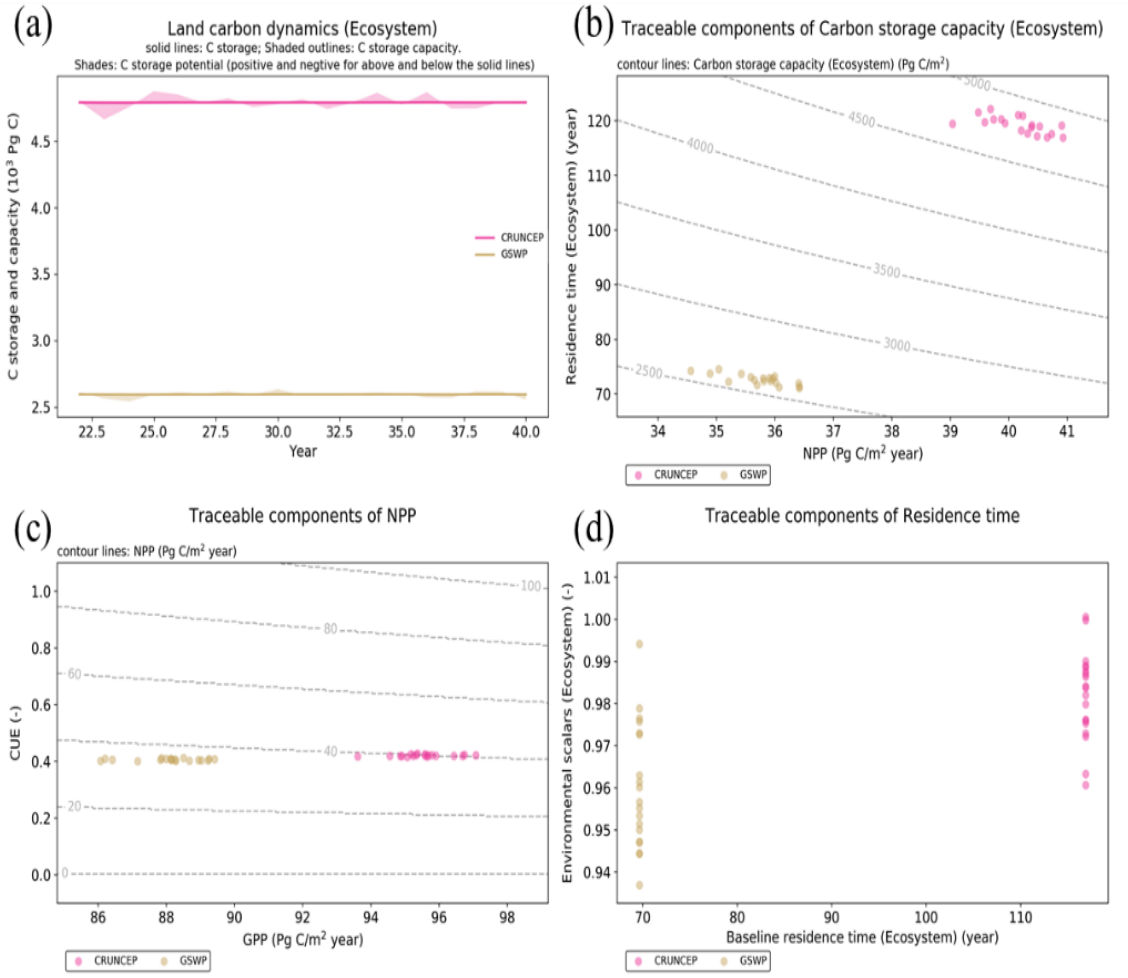

Figure R1. The results of CLM5.0 with two different forcings (CRUNCEP and GSWP) come from TraceME. (a) Land carbon dynamics, carbon storage is decomposed into carbon storage capacity and potential. (b) Carbon storage capacity is decomposed into NPP and residence time. (c) NPP is decomposed into CUE and GPP. (d) Residence time is decomposed into environmental scalars and baseline residence time. All simulated data is the normal simulation from 1921 to 1940 after the spin-up.

Comment 2: *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.*

Response: Based on the suggestions from Reviewer #1, we have removed the highlights of "automatic" and "sharable" in the revised version. The main focus of the revised manuscript is

traceability analysis. We added one more case of CLM5.0 to show that the TraceME is applicable to a specific model to help the impact of different climate forcings. We believe the publication of the TraceME platform is helpful for the readers due to four reasons. First, during our collaboration with different CMIP model teams, we realized that many modelers know their models well on some specific processes, such as GPP or total carbon storage, but usually cannot explain why those processes are different from other models. Second, it still hard to understand how different versions of a specific model simulate different land carbon cycles from CMIP5 to CMIP6. Third, for many readers who use the CMIP results but not run the models, the TraceME platform is very useful for them to identify the key uncertainty components in different regions. Lastly, the large uncertainty issue if also emerged in other components in Earth system models, such as the hydrological cycle and nutrient cycle. The publication of the TraceME platform might be helpful for them to develop traceable tools to evaluate their models.

In a recent virtual training course organized by Northern Arizona University, we used the TraceME to teach the uncertainty analysis of CMIP models. Based on their feedback, we found this online version of the manuscript is very helpful to guide the trainees to understand the scientific importance of the model uncertainty and its traceability analysis.

Comment 3: *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 factors 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.*

Response: Thanks for the suggestions. We have removed the statement in lines 77-78 in the revised version. The reviewer has raised three important concerns in this comment, including why only focus on the land carbon cycle, how to consider disturbances and land use change in the traceability framework, and how to apply the traceability analysis to other components of ESMs. We do appreciate the reviewer for these important questions, which have been deeply discussed in the revised version. Below please find our brief replies:

First, the current version of TraceME focuses on the land carbon cycle mainly due to two reasons. One reason is the large uncertainty of the land carbon cycle in the recent CMIPs, and the other reason is that the traceability analysis has its theoretical foundation on the land carbon cycle. We are testing similar traceability analyses on other components of the ESMs, but some theoretical developments are still needed. In the revised version, we have revised the introduction to narrow the scope to the land carbon cycle. We emphasized that traceability is also an important need for evaluating other highly uncertain components in the current generation of ESMs.

Second, the theoretical basis of the traceability analysis could be traced back to Luo & Weng (2011), which has demonstrated that the land carbon cycle is a dynamic disequilibrium system. The dynamic disequilibrium of the land carbon cycle is jointly driven by the internal properties and external forcings of the ecosystem carbon cycle. The traceability analysis is mainly developed on the internal properties of the land carbon cycle, which can be described as a matrix equation. The external forcings, such as disturbances and land use change, can influence the different components in this equation. The following matrix equation describes the land carbon cycle in the CLM5.0, and it shows that the external forcings can affect carbon dynamics through different components or processes. The TraceME is developed based on such matrix equations, and it can be used to evaluate the impacts of disturbances and land use changes if the model provides the simulations with forcings with and without disturbances or land use changes. We have added one paragraph to discuss this topic in the revised version.

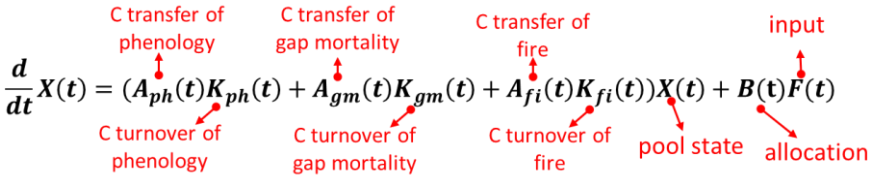

Figure R2. The matrix equation for the carbon dynamics of CLM5.0.

Third, TraceME (v1.0) has not considered all terrestrial processes, such as hydrology, radiation, and land-atmosphere interactions. We are working hard to incorporate other processes such as hydrological and nutrient processes. We have added more discussion about the limitations of this version of TraceME and highlight other processes in future developments.

Comment 4: *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'.*

Response: This is our mistake to ignore the 'Code Availability' section. We have ported the TraceME system from the local server (which is not easily accessible from the external network) to an external server so that reader can access it. The address of TraceME is http://traceme.org.cn. We also have provided the complete code on GitHub (https://github.com/ECNU-RCGCEF/TraceME).

Comment 5: *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.*

Response: Thanks for the suggestions. We have provided a more detailed technical description of TraceME in the revised version. The major revisions include the following aspects:

First, we introduced more details about CAFE infrastructure and the specific improvement in TraceME in *Section 2.1*. TraceME is developed based on CAFE. CAFE is a collaborative analysis environment where multiple data servers could work together to fulfill users' requests automatically. During the data analysis process, users only need to find data of interest, select analysis functionality to use, define analysis parameters, submit analysis tasks, and finally get the results. The logical structure of the CAFE system consists of one central node, several working nodes (data servers), and several web portals. The central node maintains descriptive information about all the data archived in each working node to support the data query. The working node is responsible for the analysis of the model data, according to users' requests. Web portals provide a browser-based GUI for researchers to interact with the whole CAFE federation.

Since it is hard to meet the requirement of systematic multivariate analysis in CAFE, we further develop the TraceME system to perform dedicated Traceability analysis for CMIP6 land models. The major enhancements that TraceME has enabled are the following:

1. Multi-variable interaction processing to satisfy systematic traceability analysis. It involves the task submission module (merging multi-variable into one task), adding a multi-variable preprocessing module, database module (multi-result systematic query), and structured results storage.
2. Deployment of the traceability analysis module and a systematic evaluation module based on python and R languages into CAFE
3. Fine-tuning of the connection to the Python language in CAFE.

Second, a description of the scientific workflow in TraceME is provided. We have incorporated it into the manuscript during the revision.

Third, we have setup a publicly accessible website of TraceME at this URL http://traceme.org.cn. Traceability analysis code is on the GitHub: https://github.com/ECNU-RCGCEF/TraceME.

Comment 6: *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.*

Response: Thanks for pointing this issue out. We have removed the discussions on 'automatic' and 'shareable' in the revision. Instead, we have added discussions on some new technical issues, such as how to federate multiple institutional clusters of CIMP data. These issues could be more important for developing model evaluation tools like TraceME. Then, a description of the scientific workflow in TraceME is provided in the supplementary materials (end of this text). We have incorporated it into the manuscript in the revision.

Minor comments:

Comment 7: *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'?*

Response: Thanks for pointing out this issue. We have revised the content in *Line 45-47* to:

"For example, the traditional methods used in model evaluation, are mainly using statistical approaches to measure the performance of models and generally treat all model components and their different metrics equally, but this ignores the relative contributions of them on model performance (Schwalm et al., 2010; Xia et al., 2013)."

Comment 8: *Line 47-48: 'it is not independence among models' => unclear.*

Response: We have removed it and revised 'it is not independence among models' to 'models share their components and are not independent of each other'.

Comment 9: *Line 49: '80% of the variance' => the variance of what ?*

Response: We have revised 'variance' to 'uncertainty' in the revised manuscript.

Comment 10: *Line 55: dramatically => dramatic*

Response: Done as suggested in the revised manuscript.

Comment 11: *Line 74: land information system => unclear what does this mean.*

Response: We have removed the 'land information system' in the revised manuscript.

Comment 12: *Line 109: it needs a new platform => a new platform is needed : : :*

Response: Done as suggested in the revised manuscript.

Comment 13: *Line 113: automatic and shareable platform => "an" automatic and shareable platform*

Response: Done as suggested in the revised manuscript.

Comment 14: *Line 189: the externally forces => external forcings ?*

Response: Done as suggested in the revised manuscript.

Comment 15: *Line 190: is always deviate from = > please check the grammar here.*

Response: We have revised 'so the X_C is always deviate from' to 'the X_C is always deviated from' in the revised manuscript.

Comment 16: *Line 251: that had been submitted results => 'been' should be removed.*

Response: Done as suggested in the revised manuscript.

Comment 17: *Line 685: positive above the soil lines => 'soil' should be 'solid'*

Response: Done as suggested in the revised manuscript.

Comment 18: *Line 594: composed into => decomposed into?*

Response: Is it 'composed into' in *Line 694*? We have revised 'composed into' in *Line 694* to 'decomposed into'.

Comment 19: *Line 360: needs to some new characteristics => check grammar*

Response: We have removed 'to'.

Comment 20: *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.

Response: We have removed the citation of *Song et al. 2019* in *Line 401-403*.

Comment 21: *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.*

Response: Thanks for the suggestion. We have removed the content in Line 104-105.

Comment 22: *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.*

Response: We will remove the discussion about 'computational efficiency' in the revised manuscript.

References

Luo, Y., Weng, E.: Dynamic disequilibrium of the terrestrial carbon cycle under global change, Trends. Ecol. Evol., 26, 96-104, 2011.

Schwalm et al.: A model-data intercomparison of CO_2 exchange across North America: Results from the North American Carbon Program site synthesis, J. Geophys. Res., 115, 2010.

Xia et al.: Traceable components of terrestrial carbon storage capacity in biogeochemical models, Global. Change. Biol., 19, 2104-2116, 2013.

Supplementary materials: the scientific workflow of TraceME

Within the workflow of TraceME, user can filter data of interest from the entire system, and the selected data is then packaged into a task and delivered to the assigned work node for data processing, which includes data pre-processing, traceability analysis, and evaluation, and finally, the evaluation results are output and visualized for the users (Fig. S1). The scientific workflow is essential for TraceME to realize online automated model evaluation. The detail of the workflow will be described below.

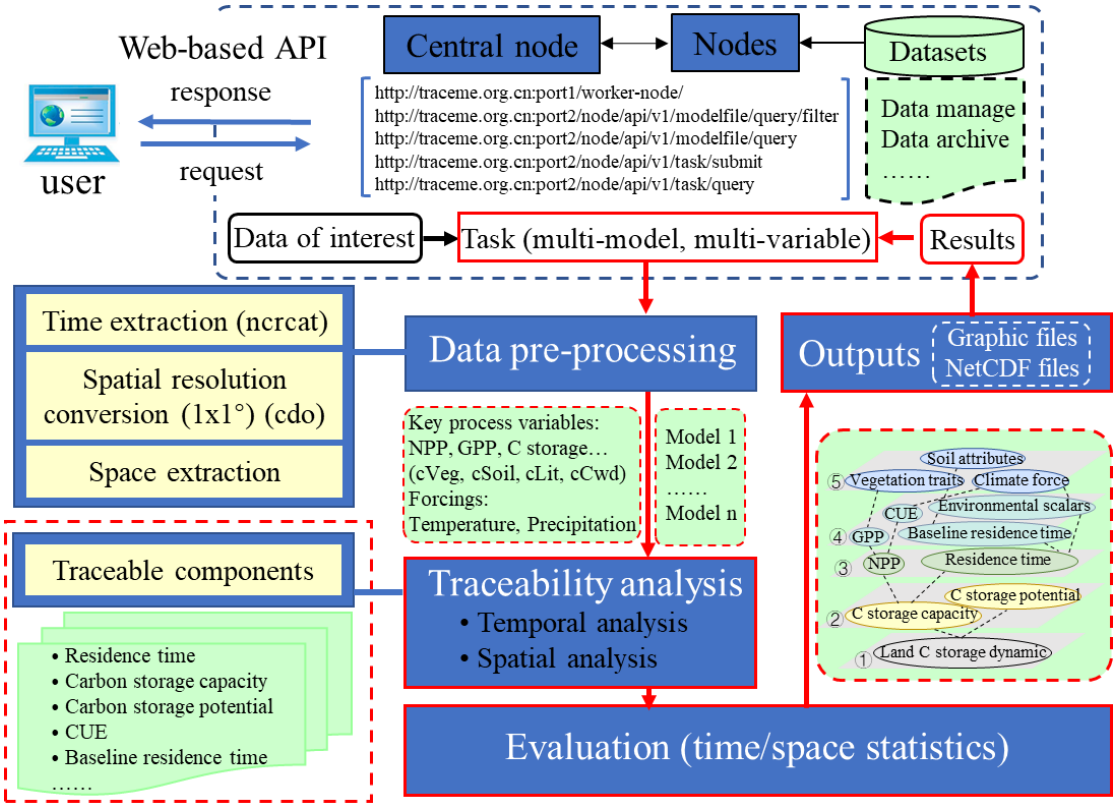

Figure. S1 The workflow of TraceME.

The function of TraceME providing for users to filter data mainly comes from the collaborative framework of CAFE and its various web application-programming interfaces (API). This function includes data source collection, data query and filtering, and submitting the task of selected data. Data is stored on individual work nodes, which can automatically parse data information to the database of work node according to the root directory of the data source by a specific API (http://{host}: {port}/{work node name}/web/parser). Then the central node collects all data information from all work nodes and provides it to the "Search" page for

users to query and filter data by APIs (Fig. S1). After users submit their selected data, the system packages all the information of data into a task for subsequent processing. TraceME focuses on evaluating models systematically with multiple variables, while the framework of CAFE is that one variable is a task. Thus, we have added new modules to support the task with multiple variables and multiple models.

Data processing in TraceME mainly includes three steps: data preprocessing, traceability analysis, and evaluation (Fig. S1). Among them, data preprocessing is mainly inherited from the CAFE framework, and we modify it to accommodate the multi-variable processing and archiving. When a task is submitted, the central node will arrange a work node for data processing. For the system to process all kinds of data uniformly, the data needs to be preprocessed first, which includes time and space extraction based on user selection and spatial resolution conversion (1x1°) by calling the tools of NetCDF Operators (NCO) and Climate Data Operator (CDO). In this version of TraceME, according to the needs of systematic traceability model evaluation, the variables from models include key process variables (NPP, GPP, and carbon storage) and forcing factors (temperature and precipitation). These preprocessed data are then submitted to the traceability analysis module written by python. With the framework of traceability analysis, land carbon storage can be decomposed into various traceable components, such as carbon storage capacity, carbon storage potential, residence time, carbon use efficiency (CUE), and baseline residence time along a temporal or spatial axis. These components are the primary objects for evaluation, and in the current version, the model evaluation includes the standard deviation of these components among models and the variance contribution of these components to the uncertainty of land carbon storage based on a hierarchical partitioning method that is written by R language.

After traceability analysis and evaluation, TraceME (v1.0) provides systematic results about the evaluation, including the figures and nc-format files of each traceable components and their variance contribution to the uncertainty of simulated land carbon storage by the models (Fig. S1). This involves task management, structured results storage, and visualization. Each task in TraceME (v1.0) has a unique task ID and is recoded the ownership of the task, the

information about data and work node, the status of processing, and the results through the database (MySQL). The “My tasks” page of TraceME displays the status of the task and the structured results, and it also provides the available links to download them for users via various API (Fig. S1).