Responses to Comments from Reviewer #1

General comments:

This paper discusses the so-called traceability analysis that the authors have published and applied since 2013. The motivation for the paper is the development of a cloud-based software built on the CAFE framework which implements their method and is launched via web interface. The authors position their analysis as the answer to what they consider are the deficiencies in the current state of model evaluation, describe how their software and algorithms work, and then show analysis results from a selection of CMIP6 models.

Thank you very much for your careful reading and valuable suggestions.

Comment 1: *The first is that the novel portion of the paper, the TraceME software, is not a substantial advance in modeling science. TraceME is a web interface to an analysis script which runs on their server. There is some sophistication in that source data can exist on multiple nodes, but this is the CAFE framework and not TraceME itself. The authors neither provide a link to test out the operability of the software they describe or even a screenshot of the user interface. The reader is left to trust the authors that this software exists and is functional as they say it is.*

Response: Thank you for raising this important issue. TraceME has two significances: performance and traceability. Performance means the model evaluation process could be facilitated by just accessing to the online system, submitting analysis tasks, and getting analysis results back. Researchers do not need to download any original model data. Traceability means the model evaluation could be done based on the Traceability analysis capability.

Although CAFE framework does provide an online process environment where multiple model data servers can automatically collaborate with each other to fulfill users' analysis tasks, there are three specific enhancements that are just enabled in TraceME:

1. Multi-variable interaction processing to satisfy systematic traceability analysis. It consists of the task submission module (merging multi-variable into one task), the multi-variable preprocessing module, the data query module (multi-result systematic query), and the result maintenance module.
2. Providing a traceability analysis module and a systematic evaluation module based on python and R languages, and having them integrated with the CAFE framework.
3. Fine-tuning of the connection to the Python language in CAFE.

The TraceME (v1.0) system was deployed in a local intranet environment for internal usage only. It is now ready for public access. The website is http://traceme.org.cn. The screenshots below show the "data Search" page, task list page, and the analysis results page. Details about the workflow of TraceME are provided in the supplementary material (end of this text), which is expected to be incorporated into the manuscript during the revision.

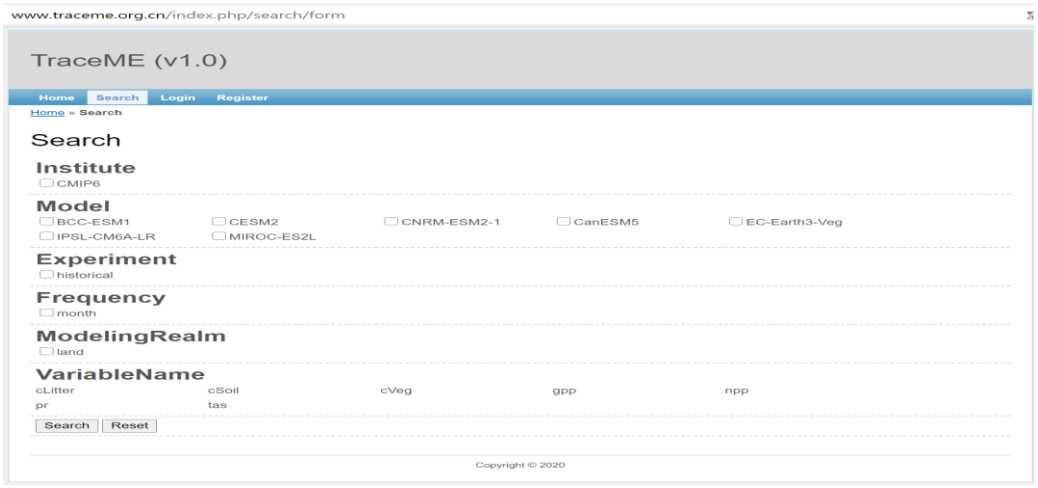

Figure. R1. The screenshot of the Search page.

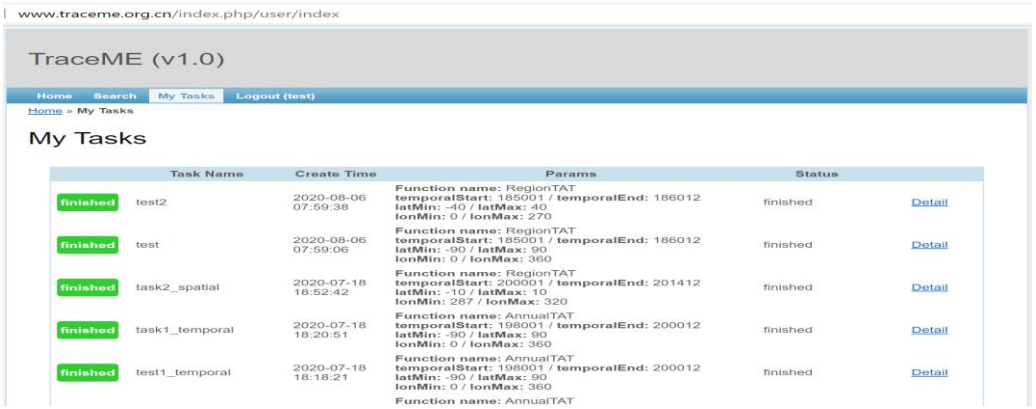

Figure. R2. The screenshot of the My Tasks page.

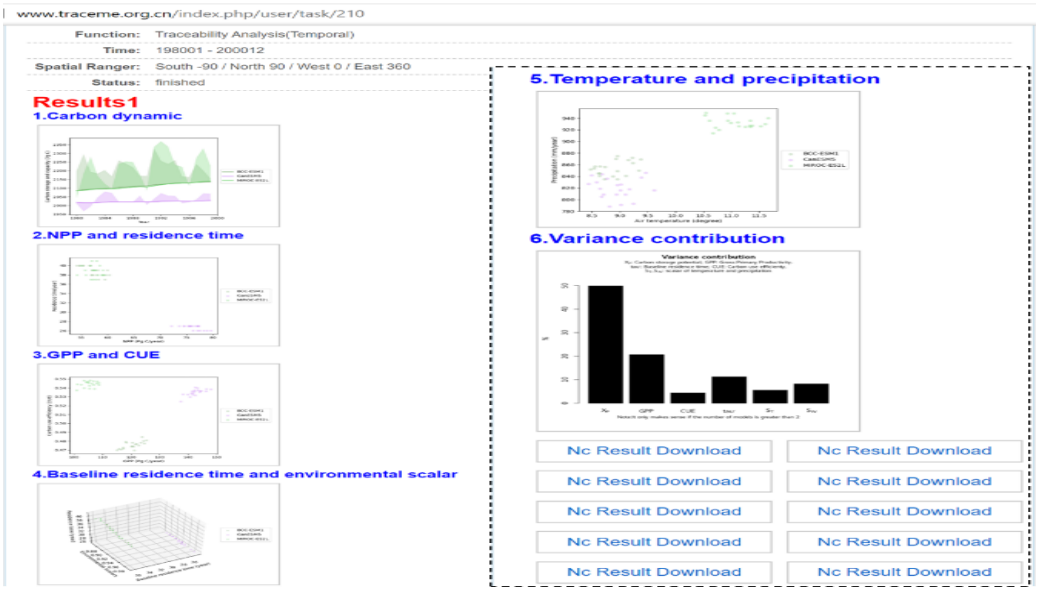

Figure R3. The screenshot of the results page.

Comment 2: *My second reservation is that a significant portion of the paper puts forth a viewpoint of the state and needs of model evaluation that is poorly supported. The authors use ambiguous terms such as 'traceable', 'shareable', 'indirect effects','high computational cost', and 'automatic'. They use these terms to describe their viewpoint on the deficiencies of current model evaluation, but do not say in detail what these terms mean.*

Response: Thanks for providing this comment. Since this question is highly related to your comment 4 and 12 below, we provided explanations for 'indirect effects' in the response to comment 4, and 'traceable', 'shareable', and 'automatic' in the response to comment 12.

We added more explanations on these terms during the revision in the *Introduction* and *Discussion* sections.

The term 'high computational cost' is not accurate. We have removed it during the revision.

Specific Comments:

Comment 3: *lines 22-24: You assert that the main challenge of using observations to evaluate ESMs is the untraceability of model outputs. It is not clear to me what this means precisely or why it is true. Why is it the 'main' challenge among many others?*

Responses: Thanks for raising this issue. The model evaluation process usually consists of three steps: downloading the model output data and archiving them locally, pre-processing the data to be suitable to be analyzed, utilizing a dedicated program to finish the evaluation.

Among many challenges in these three aspects that researchers need to address to fulfill model evaluation, data volume and traceability are major issues that we would like to highlight in this study. Taking the Coupled Model Intercomparison Project (CMIP) as an example, the total volume of CMIP Phase 5 (CMIP5) data is about 1.5PB, while the ongoing CMIP Phase 6 (CMIP6) data is expected to be 40-60 PB. Such a rapid explosion in data volume largely resulted from the increase in model complexity. As one featured component in CMIP models, land carbon cycle has been becoming highly complex during the past two decades. However, it's modeling uncertainty has been very high in CMIP5, which will still exist in the CMIP6 (Please see the below figure). Many recent studies have recognized that it is urgent and important to understand why these models perform so differently among the CMIP models (e.g., Bonan, et al., 2019). Indeed, a tool which can trace the model-to-model difference based on the big CMIP data is urgently needed. From the model evaluation perspective, it is highly needed to quantify the structural sources of the uncertainty within the complex models.

We agree with the reviewer that *Lines 22-24* are not clear. We have revised them to make the above point clearer.

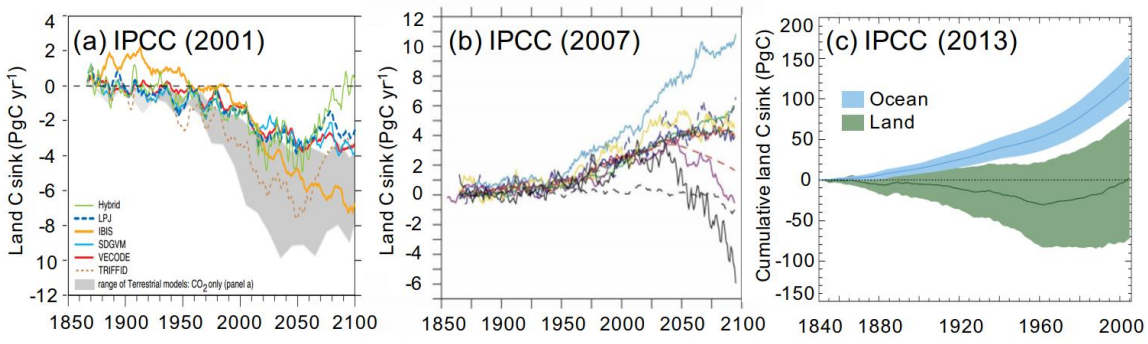

Figure R4. The uncertainties of land carbon sink simulated by models from IPCC reports.

Comment 4: *lines 44-45: What does 'generally treat all metrics equally' mean? The referenced ILAMB package, for example, does not treat every statistical measure equally. Neither does it treat measures from all data sources equally. Also, what do you mean by 'indirect effects'? ILAMB also considers variable-to-variable relationships including metrics such as Koven's inferred carbon turnover time [1]. I am having trouble envisioning what the authors mean by this statement.*

Responses: Thanks for pointing out this issue. We have removed that statement in the revised version. To the best of our knowledge, only the ILAMB has used a weighted approach to score different models based on the data. The ILAMB is the current paradigmatic package leading the international model evaluation activities, so we introduced it as the novel effort rather than a traditional method in the second paragraph of the Introduction section. It is great that the Koven's carbon turnover time has been used as one scoring matric in the ILAMB. We have made it clear in this revised version.

Comment 5: *lines 56-59: It does not follow that 'an automated computation and shareable platform' is essential because of a increase in the amount of data. Environmental computing has utilized computing centers for many decades in response to computational and data costs. Furthermore, these computing centers are becoming more user-friendly. For example, NERSC now supports Jupyter notebooks [2] which allow you to script analysis on your browser without needing to move data around. There are also cloud resources which give compute and storage capacity to anyone at low or no cost. This is a trend across many disciplines and even in the private sector.*

Responses: We appreciate the reviewer for introducing the computing centers and related progress. We agree with the reviewer that data centers are becoming more user-friendly. In the revised version, we have removed the emphases on "automated and shareable" as the key features in the TraceME. The major novel characteristic of TraceME is traceability, and we have made it clearer in the new version. The examples of NERSC and cloud-based resources have been cited as one example in our revised manuscript.

Comment 6: *lines 99-102: This point is misleading. You also are downloading large volumes of data, you are just automating it for the user. They would still need to wait while it downloads or they would benefit from you having pre-downloaded it for them. This is how the community does analysis already. Users can download data once into a project group directory on an institutional cluster where many scientist can perform their analysis. In the case of the CMIP6 archive, much of it has been copied onto NERSC drives where it is directly available to the community via a Jupyter notebook interface. There was even a multi-institutional hackathon [3] to collectively work to push results out faster. The point is that there are many ways around the need of downloading large amounts of data. If access to these institutional clusters is an issue, this is a need that the community should address.*

Responses: To avoid possible misleading, these lines have been removed during the revision. We greatly appreciate the reviewer for rising the new important issue on accessing to multiple institutional clusters of CMIP data. We have added a few sentences in the *Discussion* section to discuss this issue. In brief, we agree with the reviewer that institutional clusters are continually enriched. Taking CMIP6 for example, there are 30 data nodes available right now (https://esgf-node.llnl.gov/search/cmip6/). According to the WGCM Infrastructure Panel (https://www.wcrp-climate.org/wgcm-cmip/wip), no single institution has committed 40-60PB storage to host all the CMIP6 data in one place. As highlighted by the reviewer, federating multiple data nodes/ institutional clusters to enable a new collaborative analysis environment is needed. We also have discussed how the TraceME system can bring the traceability analysis to

benefit the CMIP6 evaluations across different institutional clusters.

Comment 7: *line 118: Jupyter notebooks [4] are another widely used solution which you should reference.*

Responses: Thanks for the suggestion. Jupyter notebook has been referenced in the revised version.

Comment 8: *lines 141ff: It is interesting that CAFE can deal with data sitting in different locations. However, I wonder how scalable this idea is. If the required data is large, then the runtime of TraceME will be dominated by download times. This may be acceptable for a relatively small analysis (few variables for a few models at monthly resolution), but could be on the order of days/months if higher temporal frequency is to be analyzed.*

Responses: Thank you for raising this issue. TraceME follows the design principles of CAFE: 1) analysis functions are available on the data server and are accessible through a website; 2) the analyses are done remotely on the data server and users do not need to download the data and archive them locally; 3) users only need to download the analytical results as pictures, graphs, or text files. The files in native formats, e.g. NetCDF files, are also available upon request; 4) multiple data nodes could collaborate with each other so that data archived on different nodes could be analyzed automatically and collaboratively if they are specified in one request. The download time is dramatically minimized.

The reason that we selected the CAFE framework is that it is capable of linking multiple data nodes together to fulfill users' data analysis requests. For example, there are 30 data nodes in CMIP6. Except for some supernodes that could replicate some selected data archived in other nodes, data hosted in one node won't be available in other nodes. Since researchers always need to intercompare multiple models, they need to download CMIP6 data of interest from individual data nodes individually.

CAFE enables a new way to access to CMIP6 data. It is capable of automatically linking these nodes together to build a federation. It provides a data search panel for researchers to search for CMIP6 data of interest, presents a list of built-in data analysis functions to select. Researchers only need to submit a model intercomparison request. CAFE can automatically locate the data nodes where the CMIP6 data need to be analyzed reside, forward the request to those data nodes. Each data node can then finish the request automatically, and then forward the result back to the node where the request was forwarded. All the results will be finally presented to researchers in the end. During this whole process, researchers neither download the original data nor know where the data actually reside. More details about the data downloading can be found in Xu et al. (2019).

One further novel development in TraceME is that the CMIP6 model data has been pre-processed to be "yearly" data, so that the volume of data needs to be analyzed in TraceME has been dramatically reduced. We have made it clearer in the *Discussion* section of the revised version.

Comment 9: *lines 152ff: Where is the web-based UI? It feels strange to see you advertise this 'shareable' technology and then not have access to explore just what it is. What are the limitations of what I can trace? Does it depend on what you have previously downloaded? Can I upload my own model output? Can I edit the analysis script that is run? Or does it rather run on the limited models you have predownloaded and only the analysis you have setup? If this is the case, a web UI seems superfluous. You could simply upload all possible results to a website for community perusal. In fact, this is what ILAMB does and how the service is most used.*

Responses: We are truly sorry for the inconvenience that this has caused. We now have made the TraceME available at this URL: http://traceme.org.cn. Details about this website have provided in the response to your first comment, and the workflow of TraceME has been described in the supplementary material (end of this text).

To facilitate the traceability analysis for the CMIP6 data in TraceME, we have pre-downloaded land model variables for all the CMIP6 models. Yes, the data that can be analyzed are those available in the TraceME system. At its current form, TraceME just pre-downloaded the selected CMIP6 model data to enable this new traceability analysis capability for the community. We are building a large CAFE federation in China by federating two CMIP6 data nodes and setting two more ones. Once it is ready, the TraceME system will be able to provide analysis capability for all the BCC and CIESM land process model output. Although the concept of CAFE is general-purpose, it focuses on the CMIP5 and CMIP6 data at its current stage. Therefore, it relies on the file naming and organization conventions that CMIP5 and CMIP6 follows. Uploading users' own model output cannot be supported at the moment, as the main focus of CAFE is enabling easy access to data servers. As the internal data management module can only recognize CMIP5 or CMIP6 datasets, uploading CMIP5 or CMIP6 compliant datasets is doable.

We really appreciate the reviewer for sharing with us the experience of ILAMB. We have recently learned the ILAMB package and discovered many novels and great features that can facilitate the application of TraceME.

Comment 10: *lines 234-235: So the data must be moved to the central node, doesn't this mean download times will dominate your analysis? How is this computationally efficient?*

Responses: Sorry for the confusion. There is no need to transfer data from data nodes to the central node. In this version, we have removed the statements about 'computationally efficient', and provide more explanations on the role of the central node during the revision.

Comment 11: *lines 244-245: Does this mean that we are restricted to using models as they were uploaded a year ago? A lot of model data has been uploaded and updated. Or will the web execution of TraceME automatically query a search of ESGF and redownload these model outputs?*

Responses: Thank you for raising this important issue. In the current version of TraceME (v1.0), we pre-download CMIP6 model data and manually keep the modeling data updated based on the ESGF server. We are currently adding more CMIP models to the TraceME website. In the future version, we are actually working on two data downloading functions. One is building a large CAFE federation in China by federating two CMIP6 data nodes and setting two more portals. Once it is ready, the TraceME system will be able to provide analysis capability for all the BCC and CIESM model data without pre-downloading any of them. The other is having a close collaboration with WGGM Infrastructure Panel and ESGF to enable a new collaborative analysis environment for all the CMIP6 nodes. If this is the case, CMIP6 data inter-comparison work would be dramatically facilitated, and the need to download a large volume of original CMIP6 data would be maximally minimized. Further introduction to the workflow of TraceME is ready in supplementary materials (end of this text), which is expected to be incorporated into the manuscript during the revision.

Comment 12: *line 360: You should expand on what you mean by each of these terms.*

** What does 'traceable' mean beyond the execution of your analysis? Why is this aspect of model evaluation so critical?*

** Does 'automatic' mean executable from a web form? If so, ILAMB has had this for 2 years in the work done by Mark Piper [5]. Also, on each commit to the master branch, ILAMB deploys automatically on Azurepipelines [6], downloads observational data, and runs a test. Also as you mention, both ILAMB and ESMValTool have workflow that make the (parallel) computation of a huge suite of model evaluations automatic. If this is not what you consider 'automatic', then what is 'automatic' and why is what the community is doing insufficient?*
** What does 'shareable' mean? The ILAMB package generates a hierarchy of evaluation results that are browseable in a web page that you can distribute to the world by simply uploading it to a web-accessible location. If that is not 'shareable', what is and why is it so important to model evaluation? Furthermore, not every group wants a shareable solution, say for quick verification tests they do not want accessible.*

Responses: Thanks for the suggestions. We have made substantial modifications to this part according to the great suggestions.

First, in our study, 'traceable' refers to the traceability analysis method, which can trace the structural sources of the uncertainties of key model components in land models. For example, carbon storage dynamics can be decomposed into carbon storage capacity and potential, and NPP can be decomposed into GPP and CUE. This method can systematically quantify the structural sources of uncertainty in land models. The increasingly complex model and the requirement for model development also drive model evaluation to provide more instructive information to better understand the sources of uncertainty. Traceable methods would be a great development and supplement to the traditional methods of statistical comparison. For example, the results of our traceability analysis can be used as a supplement to the scoring criteria of ILAMB. Besides, for example, ESMValTool (v1.0) uses a log file to record all information on a task, which is also considered as the traceability of the results (Eyring et al., 2016), while TraceME (v1.0) adopts the database (MySQL) to record it, which belongs to the content of technical description and will not be mentioned in this discussion.

Second, 'automatic' in our study refers to the web-based workflow that can automatically run the processes, such as searching, downloading, managing, preprocessing, and analyzing data. As the reviewer has mentioned, many other model evaluation tools or web-based systems, such as ILAMB and ESMValTool, also can run automatically execute a workflow for model evaluations. 'automatic' is not a novel property for model evaluation and just a technical function of TraceME. Thus, we plan to remove 'automatic' in the revision.

Third, 'shareable' in our study mainly means that model evaluation requires a platform for sharing data, which can reduce the time for users spending on repeating searching, downloading, managing, and preprocessing data, to improve the efficiency of model evaluation community. In the revised version, we have removed the emphasis on the 'shareable' feature, and only discussed the benefits of the TraceME platform for the evaluation community. We appreciate the reviewer for the introduction about the characteristics of ILAMB and have added it in the *Discussion* section of the revised manuscript.

Comment 13: *lines 385f: I disagree with you that model evaluation needs to be more efficient. ILAMB may take a long time in serial execution, but this is why it was written to launch in parallel on several institutional clusters or even a laptop/workstation. I am aware that the entire ILAMB CMIP5v6 comparison runs in a few hours. Given the decadal span between MIPs, I contend that the speed of our analysis is not the bottleneck. Beyond this, there are scripting tools and packages specifically designed to handle parallel and fast evaluation (see dask [7] and xarray [8] among others).*

Responses: We agree with you on this comment, especially what the ILAMB team has

promoted in recent years. References to these packages have been added. We also removed this statement.

Comment 14: *line 387: You argue that there is an 'absence' of automation and then explain how ILAMB and ESMValTool both implement it?*

Response: We have removed the statements on automation in the revised version.

Comment 15: *line 396: Unfortunately there is no substitution for technical training. You can setup a system like TraceME which automatically runs analysis. Yet someone has to setup and maintain that system. As software stacks change, it will break. Models will need to be added and updated. The analysis script will need to change. Others will want to upload their own scripts. How will they do this? There is a great amount of technical work that is needed to keep such a setup running and useful. What you have done is made running a relatively narrow task simple, which is by far the easiest part of the work.*

Responses: Thanks for this very important comment. We have removed the statements on the aspects of automatic and sharable in this version, so these sentences have been removed in this version. We agree with the reviewer that there many challenges to keep model evaluation systems like TraceME. The great package of ILABM has provided us many ideas to improve the TraceME. Also, one real challenge approaching to us is that how researchers worldwide could effectively finish their CMIP6 data analysis work when the total volume of CMIP6 data could be 40-60PB, while it is just about 1.5 PB for CMIP5. We also agree with the reviewer that what TraceME has pre-downloaded is just a very small part of this huge data archive. We have setup a dedicated CMIP6 data download speed test website (http://www.cmip6speedtest.cn/) to know how fast downloading CMIP6 data in a different part of the world could be. Browser-based data transfer speed testing against 28 data nodes from 44 testing cities has been finished. The mean download speed is just 6.12 MB/s.

One basic idea is that CMIP6 data nodes are expected to take more responsibility other than just disseminating the model data. If researchers have to stick with the original scenario that they need to download, archive and process the CMIP6 model data by themselves, then finding a win-win solution is impossible. Dealing with this issue is a long-term goal of the CAFE system, and the TraceME will gradually adopt some applicable features in its future versions.

Comment 16: *lines 404ff: You have not solved the issue of data transfer, you have hidden it. And it is not really hidden either. When the user clicks on your web interface and then has to wait, perhaps days, while the data is downloaded to your central node, it will not feel terribly automatic.*

Responses: We are sorry for not making this part unclear to you. TraceME does pre-downloaded several important variables-related CMIP6 land model data. Therefore, during the traceability analysis process, there is no need to further download CMIP6 data. The workflow of TraceME is provided in supplementary materials (end of this text). This part is going to be appended to the manuscript during the revision. We have also seriously discussed the ideas from the reviewer's suggestions based on ILAMB. We have realized the advantages of ILAMB for evaluating CMIP models, and we will try to collaborate with the ILAMB team for facilitating the model evaluations on Earth system models.

References

Collier et al. The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, J. Adv. Model. Earth. Sy., 10, 2731-2754, 2018.

Eyring et al. ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP, Geosci. Model. Dev., 9, 1747-1802, 2016.

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.

Xu et al. A collaborative analysis framework for distributed gridded environmental data, Environ. Model. Softw., 111, 324-339, 2019.

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.

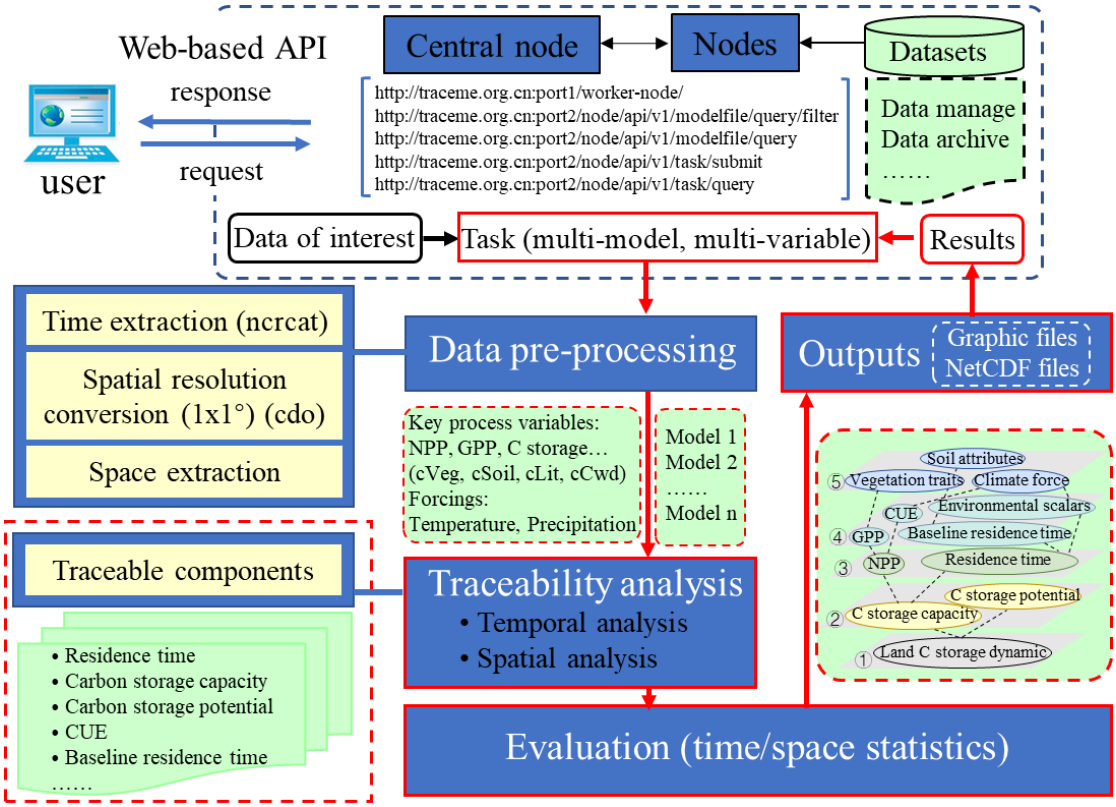

Figure. S1 The workflow of TraceME.

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).