RESPONSE TO REVIEWER #2 FOR GEOSCIENTIFIC MODEL DEVELOPMENT: MANUSCRIPT EGUSPHERE-2025-1380

BY Zhongwang Wei, Qingchen Xu, Fan Bai, Xionghui Xu, Zixin Wei, Wenzong Dong, Hongbin Liang, Nan Wei, Xingjie Lu, Lu Li, Shupeng Zhang, Hua Yuan, Laibao Liu, and Yongjiu Dai

We thank Reviewer #2 for thoughtful and constructive feedback. This Response to the Reviewer file provides complete documentation of the changes that have been made in response to each individual comment. Reviewer's comments are shown in plain text. Authors' responses are shown in purple. Quotations from the revised manuscript are shown in blue.

1. This paper describes new cross-platform software system for evaluation and comparison of land surface models using a broad suite of metrics, statistics and comparison methods. Authors clearly demonstrate OpenBench's capabilities with various examples. Figures are comprehensive and clear. The manuscript is written very clearly, with few grammatical errors, and therefore I have few comments in this regard.

➔ Thank you very much for this positive assessment of our manuscript. We greatly appreciate your recognition. I will address each of your comments and propose revisions to improve our manuscript.

2. Regarding the software itself, I appreciate authors efforts to provide an easily accessible and runnable code base along with sample data for testing. However, I note if users follow "usage" instructions from the github repository README, there is no file provided for "nml/main.nml", so the program fails. I was able to run the more complex example with sample data using the file "main-Debug.nml", but I recommend authors update the codebase to provide a highly simplified "main.nml" for initial user testing, and clearer instructions on how to adapt the codebase for custom models/dataset analysis.

➔ Thank you for this valuable feedback. We have thoroughly revised the README of the GitHub repository to provide an accurate overview and clear step-by-step instructions for users. Additionally, we have developed and included a comprehensive user manual (located in the doc folder) that contains complete installation instructions, step-by-step tutorials from basic to advanced usage, clear examples demonstrating how to adapt OpenBench for custom models and datasets, and troubleshooting guidance for common issues. We have also ensured that all referenced configuration files, including a test namelist for initial testing, are properly included in the repository with clear documentation of their purpose and usage.

3. An internet connection is required for some plotting functions (e.g. to download Cartopy coastline), while some HPC environments may not have internet connectivity. Without internet connectivity, the program fails. A programmed exception to exclude downloading coastlines etc would improve functionality.

➔ Thank you very much. This is indeed a common constraint in many institutional computing systems. Since Cartopy is the main package for our plotting functions and Basemap is no longer actively maintained, we cannot exclude Cartopy without significantly compromising the visualization capabilities that are central to OpenBench's functionality. To address this issue, we have incorporated comprehensive troubleshooting guidance into readme and our user manual, which offers detailed instructions on how to manually download and install Cartopy map files in offline environments. This includes step-by-step procedures for pre-downloading the requisite coastline and boundary data on internet-connected systems and subsequently transferring them to HPC environments, as well as providing configuration instructions for directing Cartopy to utilize these local data files.

4. Regarding the manuscript, authors may wish to comment in the paper on the name “OpenBench”, and reduce reference to this being a “benchmarking system”, as readers may have a different interpretation of “benchmarking”. To my understanding, the broad meaning of benchmarking is comparison with a well-

defined standard, or an a-priori performance expectation (e.g. see introduction and explanatory figures in your reference Best et al., 2015). This software undertakes evaluation and comparison without explicitly benchmarking (using the definitions in Best et al.,). However, I recognise that others in the community use "benchmarking" differently (e.g. in ILAMB). This could be commented on in the paper.

➔ Thank you for the excellent point regarding the conceptual distinction between "benchmarking" and "evaluation". We agree that the term "benchmarking" can have different interpretations within the modeling community. In the strictest sense, benchmarking implies comparison against well-defined performance standards or a priori expectations, as described by Best et al. (2015). However, following the precedent established by community tools such as ILAMB (International Land Model Benchmarking), we use "benchmarking" in the broader sense of systematic model evaluation and comparison against observational datasets, without necessarily establishing predetermined performance thresholds. We have added the following content to address this issue (P16L280-P17L284):

"It is worth noting that, although we refer to OpenBench as a "benchmarking system" in accordance with community convention, the tool primarily functions as an evaluation and comparison framework rather than adhering to strict benchmarking with predetermined performance standards. This design choice affords users the flexibility to establish their own performance criteria while benefiting from standardized evaluation methodologies."

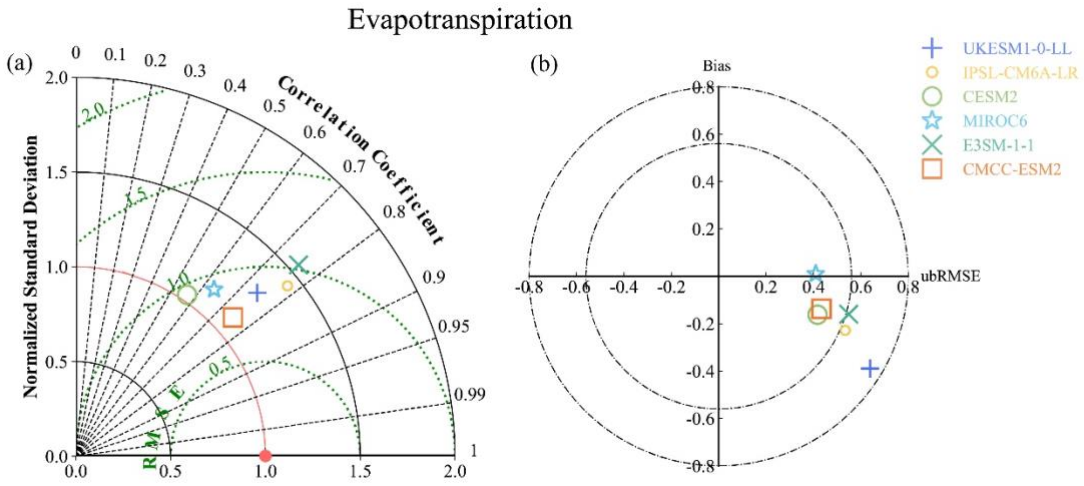
5. Some referenced models, datasets or studies are not properly referenced. For example: CLASS, CABLE, PLUMBER2. Please include relevant references.

➔ Thank you very much. We have included relevant references.

6. Also ensure all acronyms are defined. For example, I cannot find a definition for uRMSD used in Figure 10. Overall, figure captions could be improved by reducing or explaining acronyms.

➔ We sincerely appreciate the reviewer's careful attention to terminology accuracy. In the revised manuscript, we have made the correction as below: The originally

labeled "uRMSD" has been corrected to "ubRMSE", and "RMSD" has been updated to "RMSE". All acronyms throughout the manuscript have been verified and properly defined. Figure captions have been reviewed to ensure technical terms are either spelled out or properly referenced to their definitions in the text. These changes improve accessibility for readers and maintain consistency across the manuscript. Thank you for highlighting this issue.



7. Please ensure software in Table 2 is properly named. For example ESMVal should be ESMValTool, and PALS has changed their name to modelevaluation.org.

➔ Thank you very much. Corrected. We have also reviewed all software names in Table 2 to ensure they are properly named.

8. Overall, I see great potential to this work, and congratulate authors for this contribution. I look forward to integrating OpenBench into my evaluation workflow.

➔ Many thanks. We are delighted to hear that you plan to integrate OpenBench into your evaluation workflow. Please feel free to contact us via our GitHub repository or reach out directly if you encounter any issues or have suggestions for improvements.